A Multi-Disciplinary Co-Design Approach to Social Media Sensemaking with Text Mining

David McKendrick Rogers

Thesis submitted in candidature for the degree of

Doctor of Philosophy (PhD)

at Cardiff University

August 2021

DEDICATED TO

JAMES MCKENDRICK ROGERS

ACKNOWLEDGEMENTS

I would like to express my deep gratitude to Prof. Irena Spasić, Prof. Alun Preece, and Prof. Martin Innes for their continuous support, encouragement, and guidance throughout the course of this work, both as PhD supervisors and as colleagues. I would also like to extend this thanks to all of my friends and colleagues at both the Crime and Security Research Institute, and the School of Computer Science and Informatics.

I would like to thank Prof. Ian Taylor and Dr. Andrew Jones for their friendship and encouragement throughout my time in academia, and their belief in me when starting out as a research assistant. I would like to thank Andrew and Ian along with Kieran Evans, Ian Harvey, Andrew Harrison, and Francisco Quevedo-Fernandez for all the late nights and early starts across Europe.

I take great pride in my involvement with the British Universities Ice Hockey Association, and would like to thank Joe, Andy, Rambo, Rob, and Nick for all the fantastic experiences and fun I have had with you all over the years. We have built some really wonderful things and hope to continue doing so for a long time.

Thank you to all of the teammates I have had, and all the friends I have made on the ice at the Cardiff Redhawks over the last 15 years. It has been a pleasure to play alongside all of you and represent Cardiff University across the country. Thank you also to the Shamilton Academicals for all the Tuesday night and Thursday night football that has kept me going over the years.

I would not be where I am today without a wonderfully supportive, kind, and loving family growing up. Thank you, Mum, Dad, Simon, Jonathan, Julien, Grandma and Grandad, Granny and Grandad, Poppy, and all my fantastic Aunties, Uncles, and Cousins. I love and cherish you all.

Finally, and most importantly, I want to thank my Wife and best friend Hannah. You are an inspiration to me and have kept me sane throughout my studies, and then some! I could not have done this without you. I love you.

ABSTRACT

This thesis presents the development of a bespoke social media analytics platform called Sentinel using an event driven co-design approach. The performance and outputs of this system, along with its integration into the routine research methodology of its users, were used to evaluate how the application of an event driven co-design approach to system design improves the degree to which Social Web data can be converted into actionable intelligence, with respect to robustness, agility, and usability.

The thesis includes a systematic review into the state-of-the-art technology that can support real-time text analysis of social media data, used to position the text analysis elements of the Sentinel Pipeline. This is followed by research chapters that focus on combinations of *robustness*, *agility*, and *usability* as themes, covering the iterative developments of the system through the event driven co-design lifecycle. *Robustness* and *agility* are covered during initial infrastructure design and early prototyping of *bottom-up* and *top-down* semantic enrichment. *Robustness* and *usability* are then considered during the development of the Semantic Search component of the Sentinel Platform, which exploits the semantic enrichment developed in the prototype, alpha, and beta systems. Finally, *agility* and *usability* are used whilst building upon the Semantic Search functionality to produce a data download functionality for rapidly collecting corpora for further qualitative research.

These iterations are evaluated using a number of case studies that were undertaken in conjunction with a wider research programme, within the field of crime and security, that the Sentinel platform was designed to support. The findings from these case studies are used in the co-design process to inform how developments should evolve. As part of this research programme the Sentinel platform has supported the production of a number of research papers authored by stakeholders, highlighting the impact the system has had in the field of crime and security research.

TABLE OF CONTENTS

TABLE OF FIGURES

TABLE OF TABLES

CHAPTER 1: INTRODUCTION

1.1 BACKGROUND

Narratives are defined in the Oxford English Dictionary as "A spoken or written account of connected events: a story". Riessman (1993) states that "as nations and governments construct preferred narratives about history, so do social movements, organisations, scientists, other professionals, ethnic/racial groups, and individuals in stories of experience".

When looking at discourse online, a plethora of personalised narratives and viewpoints can be observed, that prove highly informative as to a person's or groups' viewpoint of cultural and political events. It can also be seen that this phenomenal growth of public narrative distributed on social media is having a significant impact on institutions, identities, politics, and behaviours via digital influencing (Dobreva et al., 2019). The relative anonymity provided by the web produces a conducive environment for users to express their opinions with a depth of feeling rarely found in formal communication (Robinson, 2001).

This lack of formality can quickly polarise opinion among an online audience, with readers tending to gravitate towards content that reflects their political views, and the majority disinclined to visit sources presenting opposing viewpoints (Lawrence et al., 2010). Extremist groups can leverage digital influencing to promote and spread narratives and opinions complimentary to their cause (Chen et al., 2008). These narratives are often centred on the subject of conflict; using the emotional nature of war as a mechanism to elevate the meaning and nature of their tasks, providing them with justification (Furlow and Goodall, 2011).

With more and more of the global population engaged with social media, communities and authorities have an additional means of spontaneous and organised response to matters such as social unrest (Procter et al., 2013), natural disaster (Carley et al., 2016), and terrorist atrocities (Cassa et al., 2013, Innes et al., 2018). Social media also acts as a vehicle for political activism by supporting the mobilisation of social movements (LeFebvre and Armstrong, 2018) and the education of the general public to a cause through informal learning (Gleason, 2013).

Whilst it can act as a concentration of voice and cause, online political and social discourse is also becoming increasingly clouded due to the emergence of "fake news" as a formidable tool of modern politics that has thrived in the social media ecosystem (Allcott and Gentzkow, 2017), to the point where social media platforms have had to change functionality in order to mitigate its reach (Allcott et al., 2019). The post-atrocity social media environment has also been shown to be conducive to misinformation surrounding the event, with often conflicting narratives interacting to both harden

and soften facts and opinions (Innes, 2020). This is further exasperated by the algorithms that drive the way in which content is served to a user, focusing their attentiveness on particular persons or topics whilst potentially discarding context, creating powerful instruments of perception (Amoore and Piotukh, 2015).

1.1.1 Social Listening

It is clear then that social media provide a multifaceted window into global and local discourse and there is great value in the ability to analyse and interpret the content, interactions, behaviours and workings of social media platforms and their data. The concept of social listening is the active process of attending to, observing, interpreting, and responding to a variety of stimuli through mediated, electronic, and social channels (Stewart and Arnold, 2018). A diverse range of data can be obtained from social media whose analysis can be relevant to a myriad of research domains such as Health Science (Shepherd et al., 2015, Salzmann-Erikson and Hiçdurmaz, 2017), Environmental Science (Peary et al., 2012, Kay et al., 2015), and Social Science (Törnberg and Törnberg, 2016, McCormick et al., 2017).

Both quantitative and qualitative analysis can be used in social listening to inform researchers' findings and conclusions. Due to the volume of social media data available to researchers, the data lends itself well to quantitative analysis as a means of characterising social media datasets, using methods such as network analysis (Williams et al., 2015), time series analysis (Asur and Huberman, 2010), and text mining (Thelwall et al., 2011). These methods can provide powerful insight into a social media discourse but should not be used in isolation to characterise and understand social media data and can be complimented with a number of qualitative analysis techniques as described in the following section.

1.1.2 Qualitative Analysis

Qualitative analysis focuses on non-numeric and less structured data relative to quantitative analysis, which is generally ordinal in nature, with the data collection process itself being more flexible and inductive (Guest et al., 2012). Several common outputs from qualitative research come in the form of thematic analysis, discourse analysis and narrative analysis:

- **Thematic Analysis -** Focuses on identification and description of both implicit and explicit ideas within the data that are typically in the form of transcribed interviews or other long-form texts (Guest et al., 2012), but can also be applied to content of shorter length such at social media content (Lowe-Calverley and Grieve, 2018).

- **Discourse Analysis** - Primarily focused on how social relations, identity, knowledge, and power is created through written and spoken texts via the study of language and conversation within social institutions (Luke, 1997).
- **Narrative Analysis** - A family of analysis approaches to texts that take on a storied form, incorporating thematic analysis to identify "what" is being told, structural analysis to identify "how" something is being told, and interactional and performative analysis which focuses on the relationships between storyteller and audience (Riessman, 1993).

These outputs are driven by a number of overlapping research designs that can be employed singularly or in concert with one another:

- **Ethnography -** An anthropological method that, whilst relatively broad in definition, focuses on qualitative study of people, societies, and cultures through observations and interviews (Fielding, 2001). Naturally, this form of research has begun to be applied to digital communities and social media platforms, with the term Digital Ethnography being coined to cover this specific evolution of ethnographic research (Murthy, 2008, Caliandro, 2018).
- **Case Study** - An 'in-depth' treatment of a single individual, group, or event, allowing for a detailed study of practices and processes in relation to a particular social context (Innes, 2001). Social media provide an ample source of topic and event driven data, with lexical features such as hashtags providing an ad-hoc means of coalescing texts to a subject (Bruns and Burgess, 2011).
- **Grounded Theory -** A means of deriving conceptualisations through the iterative interpretation of the data, unconstrained from any pre-existing theoretical precepts (Corbin and Strauss, 1990). This results in the categorisation of the data being constructed in a *bottom-up* manner; driven by the content of the data.
- **Mixed Methods -** Is the incorporation of quantitative analysis into the qualitative analysis approach, adopting multiple approaches to analysis as a means of triangulation to ensure that any variance is not biased by the method, and is truly reflective of the underlying phenomenon or trait (Johnson et al., 2007).

The mixed method approach can considered a growing methodological approach to the study of social media data, with big data analysis emerging as a common sub-analysis (Snelson, 2016). As discussed earlier, quantitative analysis is well suited to social media analysis. Furthermore, Text Mining of big data is often considered a form of quantitative analysis that is well suited to supporting qualitative analysis techniques, by aggregating data and reducing vast datasets into much more

interpretable information to which qualitative analysis can be applied (Nikolenko et al., 2017). We discuss this further in the following section.

1.1.3 TEXT MINING

Text Mining is a field of research within Computer Science that provides a means of processing big data from social media, applying techniques such as Natural Language Processing (NLP), Data Mining and Machine Learning on large corpora of text documents, to derive both quantitative and qualitative information (Hotho et al., 2005).

Opinion Mining (or Sentiment Analysis) is a form of Text Mining that combines NLP and Machine Learning with conceptual knowledge to identify emotions in text. It has found many applications in the social media space, with Twitter becoming a particular focus of sentiment analysis (Go et al., 2009b, O'Connor et al., 2010, Cheong and Lee, 2011, Thelwall et al., 2011). This can be attributed to the ease by which Twitter provides access to researchers and the public, and the diversity and volume of conversation hosted on this platform (Pak and Paroubek, 2010).

Ontologies are a form of Knowledge Representation which can support interpretation of written content during Text Mining by providing machine-interpretable definitions of domain-specific concepts and the relations between them (Noy and McGuinness, 2001). These artefacts can be defined through manual development and collective agreement (Noy and McGuinness, 2001) or through automatic generation (Maynard et al., 2008).

Whilst these technologies offer a solution to the big data problem, they should be viewed as a supplement to traditional methods of data collection and analysis rather than their replacement (Lazer et al., 2014). Consideration of a number of weaknesses found in social media data needs to be taken, such as the amount of noise present in a dataset, user demographics, the trustworthiness of authors, and self-selection biases present within a platform's userbase (Gayo-Avello, 2013).

This further pushes the argument for a mixed methods approach to ensure triangulation of findings, the importance of which is even more acute if the data analyses needs to be interpreted rapidly by decision makers who may not have the full understanding of the technology (Preece et al., 2016). The embedding of practitioners into the design, modelling, and analysis elements of machine learning tools help to guide and support the derived information by maintaining relevant scope of data, and building trust in the algorithmic approach (Endert et al., 2014, Dietvorst et al., 2018). Additionally, this can allow for mixed methods research to be performed in parallel to algorithmic analysis (Terveen, 1995, Sheth, 2009), and also reduce the lead time needed for qualitative researchers to extend the information beyond that automatically derived. The timeliness of delivering knowledge obtained through the text mining of social media can prove crucial to

businesses, government and researchers looking to build situational awareness around an event or topic (Gao et al., 2011).

1.1.4 Human-Computer Interaction & Co-Design Software Development

It is essential to support cross-discipline and mixed methods analysis when developing tools to perform text analysis on social media. The ultimate goal should be to design systems that reduce the barrier of entry for qualitative researchers involved in social media research. Human-Computer Interaction (HCI) is an area of research that focuses on how humans can interact with computers through hardware and software. HCI covers a broad range of interactions between human and computer; at the physical level focus is on the mechanics of interaction; at the cognitive level attention is paid to the ways users can understand a system; finally, affective level research looks to identify how attitudes towards a system can be altered in order to encourage re-use (Karray et al., 2017).

With the advent of Social Media HCI principles have been applied to both how users of these platforms interact with the systems and each other (Liu et al., 2008, Innes et al., 2018), and to how to design tools that can provide insight into interactions, between citizens and authorities, to research and business analysts (Reuter et al., 2018). An important HCI consideration when looking at big data analytics is to ensure that users have access to capabilities in ways that empower sophisticated users without overwhelming them, and for less experienced users care must be taken to ensure that data is not misrepresented (Fisher et al., 2012).

The co-design approach to software design, where end users are invested in the iterative design process (Muller and Kuhn, 1993), can provide a conductive ecosystem for developing systems by providing a “third space” between developer and user in HCI (Muller and Druin, 2012). This is driven by the understanding that not one expertise is sufficient to develop a well performing, user friendly system for social media analytics, encouraging developer-user co-operation to ensure that standard work practice is reflected in design and development processes (Hartswood et al., 2002). Co-design’s key motivations are to improve the choice of tools and their ease of use; to encourage collaborative development to elicit tacit knowledge and invisible practices that might otherwise be overlooked; and to observe an iterative process of design to allow for reflection, refinement and re-focus (Spinuzzi, 2005).

The co-design approach can cover a range of methods and practices that can connect stakeholders and developers together:

- **Workshops -** Provide a physical space for co-design to occur that can range of formality, whose goal is to facilitate design discourse between stakeholders (Sanders and Westerlund, 2011). These workshops can employ generative tools and applied ethnography (Steen, 2013) as a means for developers to better understand the user experience of existing practices and systems, identify user needs, and explore new solutions together (Sanders, 2000).
- **Rapid Prototyping -** Provides explorative answers to the feasibility of co-designed systems, concretising ideas and providing tangible products for stakeholders to engage with, in order to iteratively gather user input (Roschelle et al., 2006),
- **Layered Elaboration** – Layered Elaboration revisit design ideas iteratively over the course of development, adding upon them in a straightforward manner. Rather than replacing existing functionality with new features, Layered Elaboration enables co-designers to add and modify ideas without permanently destroying the original (Walsh et al., 2010).

Pirinen (2016) discusses factors that influence the success of co-design activities, identifying through review a series of barriers and enablers that should be considered in order to produce more impactful co-design practice. Importance is also placed on ensuring the continuity of development beyond one-off projects. They encourage focus be placed on mutual communication to create trust, facilitation of collaboration, and the identification of meaningful roles and value added for stakeholders.

Zamenopoulos and Alexiou (2018) break down the co-design into four ways that stakeholders can work together in co-design; *collaborating* by working together towards a common interest or project; *co-operating* by finding synergies across essentially different interests or projects whilst maintaining independent objectives; acting as a *collective* to elicit knowledge, values, and ideas from different members; or simply *connecting* actions and resources. Pirinen (2016) states that tangible pilots and rapid prototypes can act as vehicles of transformation, and that any planning should be open to respond to unexpected situations in real-life contexts.

1.2 Research Question

Looking at the field of crime and security research at the commencement of this work, engagement with social media as a source of social sensing was still in its infancy when compared to fields like health science, business studies, or environmental science. Studies into the reaction to the Mumbai roaming terror attack (Murthy, 2011, Oh et al., 2011), the Boston Marathon bombing (Cassa et al., 2013), and the dissemination of Jihadist material on Twitter and Facebook (Weimann, 2010) began to show how social media can act as a source of information in this field. We see the utilisation of a co-design in the development of crisis support tools, such as the development of crisis management

tools for emergency responders (Kristensen et al., 2006) and for social sensing during an emergency (Hughes and Shah, 2016).

A variety of analysis tools and platforms exist in both the commercial and academic sectors that provide analysis and insight into social media content, but many of the commercial tools have a focus on marketing and brand-management applications. These tools tend to take a black box approach, supporting a number of specific analyses but being hard to re-purpose for new applications, or to integrate new functionality. We look to develop an analysis platform tailored towards sensemaking in the field of crime and security, and in doing so look to answer the following research question:

"How does the application of a co-design approach to system design improve the degree to which Social Web data can be converted into actionable intelligence within an analysis platform, with respect to robustness, agility, usability?"

To answer this question, we developed a bespoke social media analytics platform called Sentinel using an event driven co-design approach. Hughes and Shah (2016) discuss the engagement of a co-designed system with real-world events by their stakeholders during evaluation activities, but our novel approach looks to go further through the utilisation of immersive workshops during real-world events (*situation rooms*) to act as both an evaluation and development in the "third space". The performance and outputs of this system, along with its integration into the routine research methodology of its users, were used to evaluate the three crucial aspects of the research question, whose multifaceted descriptions are presented below.

1.2.1 Robustness

Robustness is the ability of a system to continue to operate correctly across a wide range of operational conditions, and to fail gracefully outside of that range (Gribble, 2001). Jen (2005) discusses two interpretations of robustness by likening them to biological concepts of "mutational robustness" and "phenotypical plasticity". The former refers to the ability of a system to withstand disruption to structure without change in function, and the latter describes the effectiveness of a system's ability to switch among multiple strategic options (Jen, 2005).

These definitions are not orthogonal and can be applied to numerous levels of the analysis platform. Recovery from service outage and the resultant persistence of service is an important case of mutational robustness that this platform must address. The evolution and development of the analysis platform can be assessed against both robustness definitions by observing how the system reacts to the co-design approach; maintaining the mutational robustness of the platform's purpose

within the wider research agenda as needs change and exhibiting phenotypical plasticity through the ease by which development of new interfaces, data types, and features can be achieved.

1.2.2 AGILITY

Agility is another term which holds a number of definitions that can fall within the design and assessment scope of this thesis' research domain. The most immediate definitions come from the agile software development process as laid out in the agile manifesto, which consists of 12 principles that focus on early, frequent, and continuous delivery, design simplicity, and regular user involvement and reflection (Beck et al., 2001, Fowler and Highsmith, 2001).

Dingsøyr et al. (2012) provide a comprehensive review of how these principles have driven agile software development methodology, highlighting a number of interpretations within their survey. They includes the proposition that software development agility involves both the adaption to change, and the fine-tuning of the development process as needed (Henderson-Sellers and Serour, 2005). Other interpretations suggest that agility is the software team's capability to respond and incorporate user requirement changes during the project life cycle (Lee and Xia, 2010), or that agility is the continued readiness to rapidly create, embrace, and learn from change, whilst providing value to the user (Conboy, 2009). It is clear that the co-design process couples very closely with the principles of agile software development through the close involvement of end users in the assessment and re-design processes, something that is fundamental to both concepts.

Beyond agility of the software development process, we can also apply the concept to the services provided by the platform developed. We can consider the service to be agile if it is able to react to user requests in a flexible and speedy manner. We can also consider data to be processed and served in an agile manner if it is done so quickly and with relative ease.

1.2.3 USABILITY

Usability generally refers to the HCI concept of Software Usability, which is "the extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency, and satisfaction in a specified context of use" (Bevan, 2001). As with the other concepts covered in this section, there are multiple definitions and interpretations of usability, of which a full discussion can be found in Folmer and Bosch (2004).

Measurable aspects of usability include binary task completion, accuracy as a quantification of error, users ability to recall information from interfaces, completeness of users' solutions to tasks, quality of output, and expert assessment (Hornbæk, 2006). These aspects can be assessed through usability studies conducted in the form of usability surveys or focus groups (Kontio et al., 2004), through the analysis of application logs and the integration of user tracking software (Gray et al., 1996), or by

using formalised processes such as heuristic evaluation, cognitive walkthroughs, feature inspection, and standards inspection performed by a single expert evaluator (Nielsen, 1994).

We can also make an analogous assessment of the entire analysis platform and tools, with respect to the system's fit within a mixed methods research agenda and workflow. In addition to this, Folmer and Bosch (2004) argue that it is necessary to have both a working system and representative users present in order to assess usability, thus reaction to usability assessments can be costly to an inflexibly designed architecture. The co-design approach can presumably alleviate these costs, as any architecture should be designed in anticipation of the need to change due to user need, and a representative user group is naturally established early on in the development lifecycle meaning usability assessment can begin promptly.

1.3 Research Design

1.3.1 Scoping Task and Function

We use two situational awareness models as drivers for functionality design of the Sentinel platform: the 5Ws framework covered in Roberts et al. (2015) and the sensemaking loop defined by Pirolli and Card (2005). These two models act as a means of defining the scope of the system and will be used later in the thesis the assess its effectiveness as a situational awareness and research tool.

1.3.1.1 User Task

The 5Ws framework is focused upon characterising events, asking the questions *who*, *what*, *when*, *where*, and *why* and acts as a framing device for what tasks a user is expected to complete when using Sentinel:

- **Who?** - Identify participants in events.
- **What?** - Characterize events including crimes and social mobilisation.
- **When?** - Maintain a timeline of linked events.
- **Where?** - Determine locations of interest by both geo-tags and place names.
- **Why?** - Uncover causal links between events.

These questions are intentionally very broad in order to allow the co-design process to have a great degree of freedom leaning into the co-realisation thought that "design emerges and evolves as part of the ongoing struggle of making this particular system work for these particular users, in this particular workplace and at this particular time" (Hartswood et al., 2002).

1.3.1.2 *User Workflow*

We look at the sensemaking loop for intelligence analysis (Figure 1.1), an intelligence based model that defines the data flow and process flow surrounding the transformation of information as it flows from raw information to reportable results (Pirolli and Card, 2005).

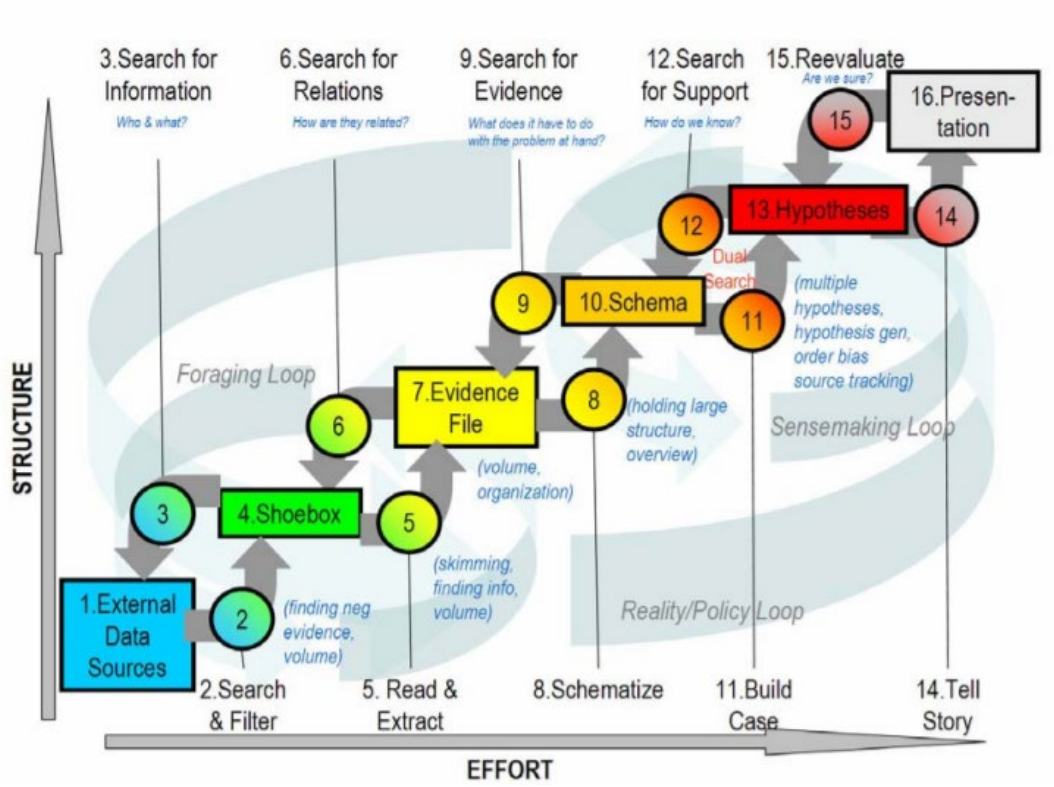


Figure 1.1: The Sensemaking Loop for Intelligence Analysis as Defined by Pirolli and Card (2005).

The processes (circles) and data (rectangles) progress by degree of effort and information, with processes able to shift data both up and down the conceptual axis to form the overall sensemaking loop. A pair of sub-loops that bisect the two processes across the data/theory boundary; with the *foraging loop* focused on information seeking, searching, and filtering, and the *sensemaking loop* which focuses on schematization and presentation of working hypotheses.

In addition to this, Pirolli and Card (2005) describe how information processing is driven by a mix of *bottom-up* (from data to theory) and *top-down* (from theory to data) processes, that are employed opportunistically by an intelligence analyst. This idea shares much with the mixed methods concept of sequential design, which describes the bi-directional analysis between quantitative and qualitative results, with explanatory sequential describing qualitative exploration being followed by quantitative analysis, and exploratory sequential referring to quantitative analysis being explained with qualitative analysis (Creswell et al., 2011).

1.4 THESIS STRUCTURE

The rest of the thesis is structured as follows. Chapter 2 reviews the relevant literature in the form of a systematic review into the state-of-the-art technology that can support real-time text analysis of social media data. This is used to position the text analysis elements of the Sentinel Pipeline within the field of text mining, and to build an understanding of how text mining is being incorporated into other analysis tools and platforms.

Chapter 3 is focused on *robustness* and *agility*. It covers the initial infrastructure design and early prototyping of the Sentinel Pipeline and Sentinel Interfaces, using a number of case studies to act as an early assessment of the development direction and tool performance. We also present some of the early *bottom-up* and *top-down* semantic enrichment incorporated into the Sentinel Pipeline.

Chapter 4 focuses on *robustness* and *usability*, further developing the semantic enrichment employed by the Sentinel Pipeline. We discuss the development of two classification tools focused on emotion and violence, before introducing the Semantic Search component of the Sentinel Platform, which exploits the semantic enrichment developed in Chapter 3 and Chapter 4. Case studies are again presented as part of continuous tool and interface assessment.

Chapter 5 looks at *agility* and *usability*, building upon the Semantic Search functionality to produce a data download functionality for rapidly collecting corpora for further qualitative research. This functionality also allows us to develop a suite of topic modelling tools that allow for rapid sequential analysis. The chapter concludes with a case study into the ongoing COVID-19 pandemic, that further advances the analytical capabilities of the Sentinel Platform and allows us to assess the performance of the Sentinel Platform against the 5Ws and Sensemaking Loop models.

Finally, Chapter 6 discusses the key contributions of this thesis by focusing on the three key aspects described in Section XX, returning to the research question, discussing limitations to the work, before drawing final conclusions and outlining future work.

CHAPTER 2: BACKGROUND SYSTEMATIC REVIEW
REAL-TIME TEXT CLASSIFICATION OF USER-GENERATED CONTENT ON SOCIAL MEDIA

2.1 INTRODUCTION

In order to position our work within the wider social media analysis domain, it was decided that the literature review for this thesis would take the form of a systematic review. The aim of this systematic review is to determine the current state of the art in the real-time classification of user-generated content from social media.

2.1.1 RATIONALE AND RESEARCH QUESTIONS

Chapter 1 introduces several human-centred fields of research that are applicable to this thesis including social listening, qualitative research, and HCI. This review takes on the role of investigating the main technical area of the thesis that the development of the Sentinel pipeline will support. We focus the review on the identification of current and emerging trends in the use of NLP and data mining techniques to extract features that can support text classification, as well as the approaches to text classification itself covering an array of machine learning methods and the data used to train them.

ID	Question
RQ1	What are the main characteristics of data used to train and test real-time text classification applications?
RQ2	What types of text processing and normalisation are required to facilitate classification of user-generated content from social media?
RQ3	Which machine learning methods are used most commonly to implement real-time text classification?
RQ4	How do these methods compare to one another in terms of classification performance?

Table 2.1: Research questions.

This was initiated as a means of informing some of the reassessment process discussed in Chapter 3 (Section 3.3.1.2), the classification tools developed in Chapter 4 (Section 4.4), and the post processing of data downloads in Chapter 5 (Section 5.2.3). We therefore devised a series of research questions presented in Table 2.1, that were used to systematically assess and understand the data, pre-processing, and algorithms utilised by assess the most recent literature in text mining of social

media. The final systematic review was performed between 2018 and 2019 and forms the body of this chapter and a standalone journal submission.

2.2 METHODS

This systematic review follows the Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) guidelines (Moher et al., 2009), which have been adapted from the health science domain.

2.2.1 SEARCH STRATEGY

In order to systematically identify articles relevant to social media related to text classification of user-generated content, we first considered relevant data sources including subscription-based digital libraries curated by reputable journal publishers (Association for Computing Machinery, 2020, Institute of Electrical Electronics Engineers, 2020), free-to-access bibliographies that provide aggregated searching across third party digital libraries (Achilles, 2020, Database Systems Logic Programming - University of Trier, 2020), and self-curated repositories (CiteULike, 2019, Google, 2020) where authors are able to link published works to a third party library. Individual sources are listed in Table 2.2.

ID	Name	Curator	Type
ACM	Association for Computing Machinery	Association for Computing Machinery	Primary
IEEE	IEEE Xplore	Institute of Electrical and Electronics Engineers (IEEE)	Primary
CCSB	Collection of Computer Science Bibliographies	Alf-Christian Achilles	Secondary
DBLP	DBLP Computer Science Bibliography	Trier University	Secondary
CUL	CiteULike	Users	Tertiary
GS	Google Scholar	Automated & users	Tertiary

Table 2.2. Data sources.

A Boolean query was created by combining three major facets of the aims of the systematic review (real-time, classification and social media), whose near-synonyms and hyponyms are presented in Figure 2.1 along with an example of the formatted query string:

title:((real AND Time) OR Realtime OR Live OR Stream) AND*
title:(Classif OR Mining OR Analys* OR Process* OR Monitor*) AND*
((Social AND (Media OR Web)) OR Facebook OR YouTube OR WhatsApp OR Twitter …)

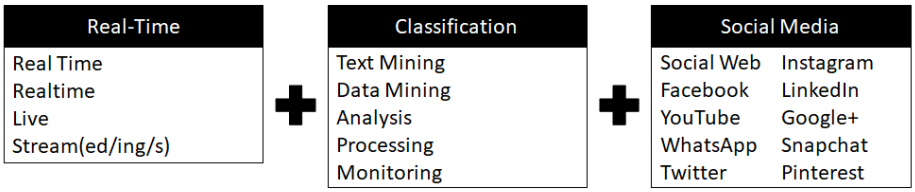


Figure 2.1. Synonym identification for formal query.

The Real-Time and Classification facets were bound specifically to article title to maximize the accuracy of retrieval. The social media facet was matched against the article's abstract and, where possible, the full text. The searches were performed on January 25th, 2019. It should be noted that CiteULike ceased to operate in March 2019 (Wayback Machine, 2020).

2.2.2 Selection Criteria

To further refine the scope of this systematic review, we defined a set of inclusion and exclusion criteria (see Table 2.3 and Table 2.4). To ensure the rigorousness and credibility of selected studies, they were also evaluated against the quality assessment criteria defined in Table 2.5.

ID	Criterion
IC1	The input text represents user-generated content posted on social media.
IC2	The input text is processed and normalized using techniques from NLP.
IC3	The processed text is classified automatically using a machine learning approach.
IC4	There is sufficient evidence that the classification has been or can be used to classify data streams from social media in real time.

Table 2.3. Inclusion criteria.

ID	Criterion
XC1	The article was published before January 1st, 2014.
XC2	The article was not written in English.
XC3	The article was not peer reviewed.
XC4	The article does not describe the implementation of an original application.

Table 2.4**. Exclusion criteria.**

ID	Criterion
QC1	The research aims are clearly defined.
QC2	The study is methodologically sound.
QC3	The method is explained in sufficient detail to reproduce the results including algorithms, their parameters as well as the datasets used for training.
QC4	The results were evaluated systematically in terms of accuracy, precision, recall and/or F1 score.

Table 2.5. Quality assessment criteria.

2.2.3 Study Selection

The search results were downloaded from the given sources (see Table 2.2), converted into BibTex format and then aggregated into a single list. Duplicates were identified and removed through semi-automated title and abstract matching, with source attribution retained based upon the repository hierarchy and then simply on alphabetical order within the highest tier. Abstracts for the remaining citations were then downloaded and added to the corpus manually.

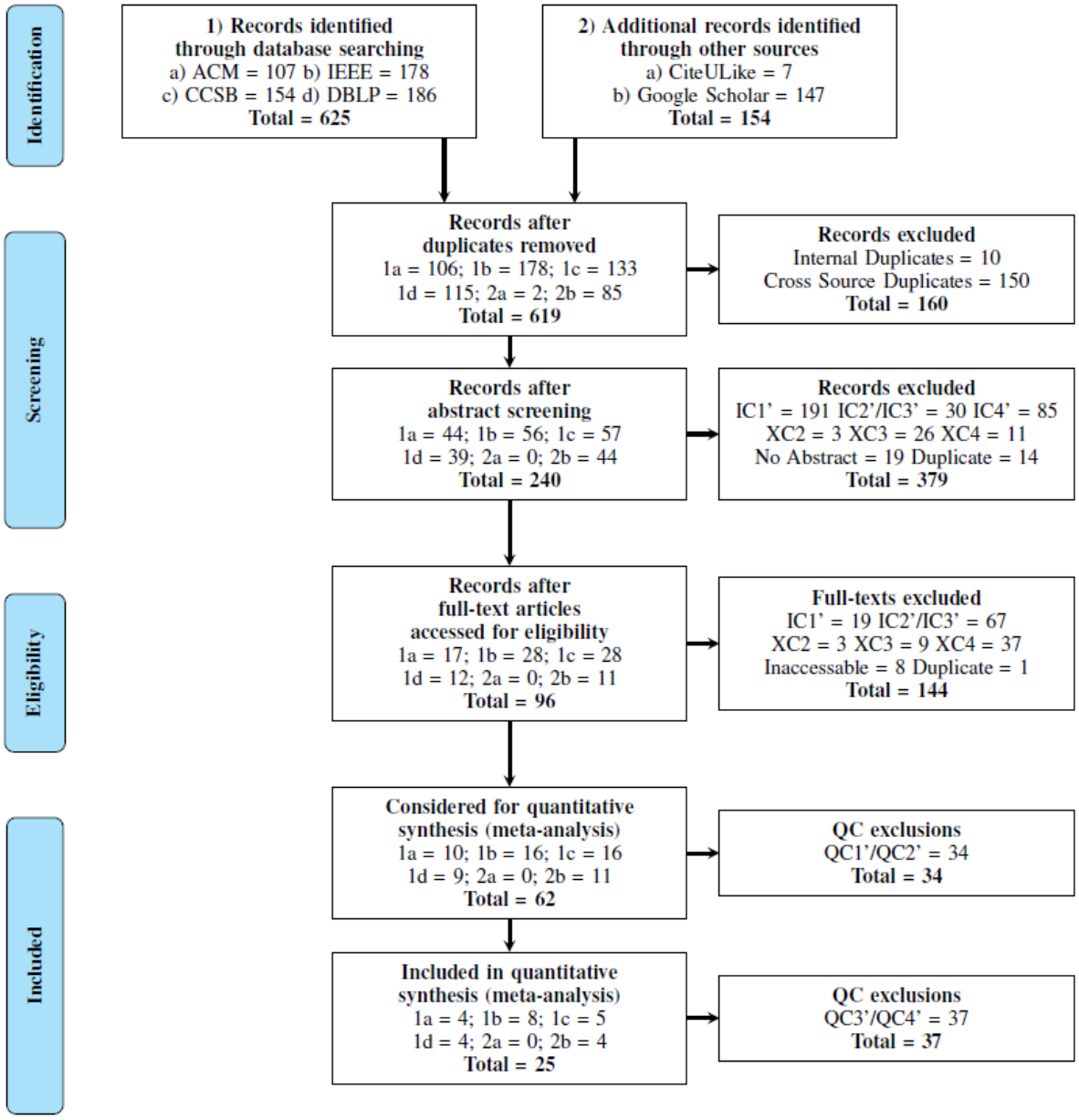


Figure 2.2. PRISMA 2009 flow diagram.

Document screening was performed using the Rayyan QCRI, a web application and a mobile app for systematic reviews (Ouzzani et al., 2016). The abstracts were screened to exclude the articles that

were clearly outside the scope of the review as defined by the selection criteria. Full-text copies of the remaining articles were downloaded automatically using PaperCaddie [1], a bespoke Django application we implemented to source articles using their Digital Object Identifier (DOI). The full-text articles were then assessed against the selection criteria.

Figure 2.2 provides a PRISMA diagram that describes the study selection process. Searches against the four curated sources retrieved a total of 625 documents. A total of 154 documents were retrieved from self-publishing sources. We identified and subsequently removed 160 duplicate documents using PaperCaddie's abstract comparison script.

2.2.4 Data Extraction

Data extraction cards were defined in PaperCaddie to facilitate data synthesis. They allowed for relevant data to be extracted and structured against the following facets: NLP methods, data characteristics, classification methods and evaluation results (see Table 2.6) during the screening process.

2.2.4.1 RQ1: Main characteristics of data sets

Provenance covers the type of social media where the data were published originally. Mainstream social media should be expected here; Twitter and Facebook have been the dominant sources of user-generated content since their emergence in 2006 and 2004 respectively. We extended the sources considered to any platform where social messages could be exchanged publicly, e.g., user comments in news media.

Source refers to the provider of the data set. Some studies choose to benchmark their algorithms using publicly available data sets produced during NLP community challenges such as SemEval (Nakov et al., 2016). Other algorithms will be designed to operate in smaller domains leading to training data being obtained through manual curation or automated collection through an application programming interface (API).

Volume refers to the size of training and test data set used. Where such information was provided, the volume was stratified against the classes.

Scope refers to the search criteria used to collect the data, which includes the search terms, geospatial constraints, the time period when data were published and/or collected together with the motivation behind these choices. Duration was recorded in the order of days, with the exact

[1] http://github.com/rogersdm/papercaddie

number taken when dates are explicitly given, otherwise we estimate to the nearest week or month dependant on the phrasing given by study authors.

Category	Field	Description	Values / Examples
Data	Provenance	Datatype	Tweet, Facebook comment, Article comment, etc.
	Source	Origin of Data	Twitter API, Gold Standard Corpus, Manual Collection, etc.
	Volume	Number of documents	Numerical
	Scope	Coverage of data	Date Range
Pre-processing	Non-linguistic Analysis	SM format correction	Lowercasing, URL Removal
	Morphological Analysis	Abbe et al. (2016)	Stemming, Stop-word Removal
	Syntax Analysis	Abbe et al. (2016)	Tagging, Chunking, Parsing, Lemmatization
	Semantic Analysis	Abbe et al. (2016)	Tagging, Disambiguation, Ontology
	Feature Selection	Abbe et al. (2016)	N-Gram Analysis, Bag-of-words, TF-IDF
Machine Learning	Algorithm	Supervised classification method	Naïve Bayes, RNN, Decision Tree
Results	Precision	$P = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$	Numerical
	Recall	$R = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$	Numerical
	F-Measure	$F_1 = 2 \times \frac{\mathrm{P} \times \mathrm{R}}{\mathrm{P} + \mathrm{R}}$	Numerical
	Accuracy	$A = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}$	Numerical

Table 2.6. Data extraction fields and examples.

2.2.4.2 RQ2: Text Processing and Normalization

Non-linguistic Analysis partially covers the morphological analysis presented in Abbe et al. that covers punctuation and lowercasing (Abbe et al., 2016). On social media, the user-generated content features frequent use of non-linguistic content such as icons and special characters, platform specific prefixes and tokens, and web links. This complicates pre-processing of user-generated content in comparison to traditionally formatted text. To reduce web-based idiosyncrasies in user-generated content, its pre-processing includes, but is not exclusive to, removing characters (via encoding, syntax, without any semantic reasoning), tokenising non-text features, and removing HTML elements such as images and links.

Morphological Analysis covers the decomposition of a stream of text into words phrases, symbols or other meaningful elements, resulting in the extraction of terms from the text which are independent from the information and relationships that is found among them (Abbe et al., 2016). Popular methods that fall into this category are tokenisation and stemming, the former being the segmentation of word-like units from a text (Grefenstette and Tapanainen, 1994) and the latter is the further reduction of these words down to one heading for all variant forms which share a common meaning (Porter, 2001).

Syntactic Analysis is defined by Abbe et al. as methods that are used to determine the structure linking different parts of a sentence (Abbe et al., 2016). They highlight lemmatisation, i.e., reduction of different inflectional word forms to a common base based on morphology and syntax into account, as a common form of syntactic analysis. Part of speech tagging represents a form of syntactic analysis; with the structure and composition of the sentence as a whole or in part being used to determine the grammatical context of its constituent words. Indexing of phrases and grammatical components also fall into this category.

Semantic Analysis refers to the process of interpreting the text usually through the application of domain-specific lexicons, ontologies, and dictionaries. Abbe et al. (2016) state that ontologies based on semantic analysis allow text to be mined for interpretable information about domain concepts, as opposed to simple correlations discovered using statistical information (Abbe et al., 2016). Alongside the use of ontologies and lexicons that focus on domain-specific concepts, word normalisation through slang dictionaries, information extraction through named entity recognition (NER), emoji reference tables and abbreviation expansion are considered within this category.

Feature selection is focused upon reducing the variability present within a corpus through the use of mathematical transformation of texts. Common techniques found in feature selection include term frequency – inverse document frequency (TF-IDF) which is used to identify the most discriminative words in a corpus of documents (Ramos, 2003), the gain ratio which is used in decision trees to calculate the value of a document feature has for classification (Karegowda et al., 2010), and word embeddings which allow for words (and documents) to be represented in a low dimensional vector space (Mikolov et al., 2013).

2.2.4.3 RQ3: Machine Learning Methods

Algorithm is the only field in this section, and broadly covers the algorithms used to perform text classification. Note was also taken on which algorithm outperformed others, with the algorithm presenting the highest accuracy being selected as the best performing, when no preference was stated by the authors. Commonly used supervised classification algorithms include support vector

machines (SVMs) (Cortes and Vapnik, 1995), naive Bayes classifiers (Domingos and Pazzani, 1997), decision trees (Quinlan, 1986), and neural networks (Kim, 2014).

2.2.4.4 RQ4: CLASSIFICATION PERFORMANCE

Precision, Recall, F-Measure and ***Accuracy*** are standard measures used to evaluate classification performance (Goutte and Gaussier, 2005). They are derived from the numbers of true positives, false positives, true negatives, and false negatives obtained when the classification model was applied to the test data. Where presented the measures corresponding to the author's choice for best performing algorithm were taken. A single measure of each was recorded and where these were only presented at a classification category level, overall calculations were made.

2.3 RESULTS

Table 2.7 presents the provenance, scope, source data retrieved from the studies, along with the composition of the data sets used to train and test the algorithms and the reported results. Though individual results cannot be directly compared, we can observe certain trends by ordering the table based on Accuracy and F1 Scores. Table 2.8 presents the full review of the document preparation methodology. The following sections use these tables to discuss the nature of the 25 studies in relation to our four research questions.

2.3.1 RQ1 – PROVENANCE, SCOPE AND SOURCE

It is evident that there is a gap in the literature with regards to diversity of data. Twitter is still the predominant source of data utilised by researchers when working with social media. The ease of access through several robust and established APIs, volume and variety are commonly cited as reasons for its choice. Other sources of data include YouTube comments (Benkhelifa and Laallam, 2018), Sinai Weibo messages (Ma et al., 2018), and a bespoke company-based messaging service developed for IBM (Golestani et al., 2018).

The majority language of the data processed is English; with only 5 out of 25 studies providing non-English data processing of some form. This is most evident in Ma et al. (2018), which worked with exclusively with Sinai Weibo data, a Chinese language platform. Neuenschwander et al. (2014) also worked with non-English data, in this case Brazilian Portuguese tweets focused on the Brazilian stock market. Italian was used alongside English in two of the studies Avvenuti et al. (2015) and D'Andrea et al. (2015). Finally, Vincente et al. (2018) present multilingual text mining, working with data sets from Basque, English, French and Spanish.

APIs are the most popular form of data collection; 14 studies using APIs provided by the social media platforms (Twitter, YouTube, and Sinai Weibo). A total of 8 studies indicated that collections were

bounded by search terms (Middleton and Krivcovs, 2016, Yu et al., 2016, Behzadan et al., 2018, Win and Aung, 2017, Benkhelifa and Laallam, 2018, Ma et al., 2018, Lee et al., 2015, Avvenuti et al., 2015), 2 utilised a geospatial bounding (Nguyen et al., 2016, Șerban et al., 2019), 2 used a combination of search terms and geospatial bounding (D'Andrea et al., 2015, Subramani et al., 2018) and 1 collected using a combination of terms and specific user accounts (Kurniawan et al., 2016).

Publicly available pre-collected data sets were the second most common form of data used. We identified 6 such studies. The majority were gold-standard data sets produced as part of NLP community challenges with Vilares et al. (2014) utilising the RepLab 2014 corpus (Amigó et al., 2014) and Azzouza et al. (2017) and Karanasou et al. (2016) using corpora from the SemEval series (Poursepanj et al., 2013, Nakov et al., 2015). Interestingly, Mane et al. (2014) used data from SemEval 2013.T2 corpus (Poursepanj et al., 2013), but this was obtained through a secondary source whereby on of the classes from the original set was omitted, resulting in a different volume an overall constitution of training data relative to Azzouza et al. (2017) who also used the SemEval 2013.T2 corpus. Both Steed et al. (2015) and Alharthi et al. (2018) source their data sets from previously published articles, with the former sourcing a large 1.6 million datapoint corpus produced by Go et al. (2009a), and the latter citing previous work of their own where they detail the collection methodology (Alharthi et al., 2017). In all cases where the pre-collected data sets were re-used, the details of the original collection method were obtained from the corresponding citation.

Michailidis et al. (2018) resorted to manual collection and curation of their Twitter-based data set through a third-party service FigureEight (then called CrowdFlower), opting to bound their collection criteria through search terms. Golestani et al. (2018) obtained their IBM messaging directly from the company's databases, with no constraints on terms, users or location. Finally, we were unable to identify the origin of data in 3 studies (Cavalin et al., 2014, Vicente et al., 2018, Neuenschwander et al., 2014). The number of days covered by a data set was identified in all but two of the studies (Avvenuti et al., 2015, Rezaei and Jalali, 2017). Studies used data that ranged from less than a day up to just over two years.

Looking across the volumes of data used to train and/or test the text mining approaches, there was a lot of variation between studies with an inter-quartile range spanning from 1,688 to 18,000 documents and a median of 6,126. The data set is positively skewed, with one study (Steed et al., 2015) presenting a training set size that can be considered an outlier falling outside the upper quartile by over three standard deviations. Steed et al. (2015) use 1.6 million documents within their training set which is an existing auto-generated training set produced by Go et al. (2009a).

Study		Source				Data				Algorithms		Results			
Citation	**Task**	**Type**	**API**	**Man**	**PrePub**	**Range (d)**	**Vol**	**Classes**	**Dist**	**Tot**	**Best**	**Acc**	**P**	**R**	**F1**
Kurniawan et al. (2016)	Event	Tweet	Y			7	35184	2	E	3	SVM	99.77%	99.65%	99.89%	9.77%
D'Andrea et al. (2015) *	Event	Tweet	Y			0.17	2660	2	E	5	SVM	95.75%	95.30%	96.50%	95.80%
Nguyen et al. (2016)	Event	Tweet	Y			30	5000	2	U	4	BN		94.20%	96.60%	95.40%
Benkhelifa and Laallam (2018) §	Topic	YouTube	Y			122	10K	2	E	1	SVM	95.30%	95.35%	95.35%	95.35%
Behzadan et al. (2018)	Event	Tweet	Y			4	21K	2	U	1	CNN	94.72%		94.57%	94.62%
Alharthi et al. (2018)	Topic	Tweet			Alharthi et al. (2017)	111	6126	3	B	1	SVM	93.10%			92.66%
Subramani et al. (2018)	Topic	Tweet	Y			108	618	2	U	1	LogReg	92.50%			
Win and Aung (2017) ǂ	Event	Tweet	Y			6	1045	3	Ma	3	Linear	92.02%	91.20%	92.00%	91.30%
Steed et al. (2015)	Emotion	Tweet			Go et al. (2009)	80	1600K	2	E	2	NB	90.00%			
Michailidis et al. (2018)	Sentiment	Tweet		Y		364	17360	2	B	4	SVM	90.00%	86.00%	85.00%	85.50%
Middleton and Krivcovs (2016)	Event	Tweet	Y			1	1045	2	U	5	DT		*69.00%*	*98.00%*	88.00%
Șerban et al. (2018)	Topic	Tweet	Y			256	9353	2	U	4	CNN	85.40%			85.20%

* - Non-English language corpus, ǂ - Favours speed over Accuracy/F-Measure, § - Non Tweet data source

italic - Values estimated from graph(s), underline - Values calculated from presented data

E - Equal distribution, U - Unequal distribution, B - Balanced distribution, Ma - Distinct majority class, Mi - Distinct minority class

Table 2.7. Combined study results (Table continued on next page)

Study		Source				Data				Algorithms		Results			
Citation	**Task**	**Type**	**API**	**Man**	**PrePub**	**Range (d)**	**Vol**	**Classes**	**Dist**	**Tot**	**Best**	**Acc**	**P**	**R**	**F1**
Karanasou et al. (2016)	Sentiment	Tweet			Nakov et al. (2015)	30	12529	3	Ma	4	SVM	85.10%			
Avvenuti et al. (2015) *	Event	Tweet	Y				5069	2	U	1	DT	83.50%			
Rezaei and Jalali (2017) ǂ	Sentiment	Tweet	Y				9903	2		2	DT	82.51%			
Cavalin et al. (2014)	Topic	Tweet				15	1910	3	Ma	1	NB	82.00%			
Lee et al. (2015)	Topic	Tweet	Y			736	2000	2	U	4	NBM		81.10%	81.10%	81.10%
Yu et al. (2016)	Sentiment	Tweet	Y			2	200	2	E	5	SVM	77.00%			
Golestani et al. (2018) §	Topic	IBM		Y		30	130K	2	U	5	NBM		73.00%	74.00%	73.50%
Neuenschwander et al. (2014) *	Sentiment	Tweet				153	922	2	U	3	NBM		73.40%	73.50%	73.40%
Mane et al. (2014)	Sentiment	Tweet			Poursepanj et al. (2013) *reduced*	364	1466	3	Mi	1	NB	72.27%			
Vicente et al. (2018) *	Sentiment	Tweet				15	12273	3	Mi	1	SVM	70.43%			
Vilares et al. (2014)	Topic	Tweet			Amigó et al. (2014)	214	28088	7	Ma	1	Linear		69.85%	72.43%	69.81%
Ma et al. (2018) *§	Topic	Sina Weibo	Y			546	160K	2K		4	RNN		67.30%	66.50%	66.90%
Azzouza et al. (2017)	Sentiment	Tweet			Poursepanj et al. (2013)	364	3813	3	Mi	1	Rules	55.96%			

* - Non-English language corpus, ǂ - Favours speed over Accuracy/F-Measure, § - Non Tweet data source

italic - Values estimated from graph(s), underline - Values calculated from presented data

E - Equal distribution, U - Unequal distribution, B - Balanced distribution, Ma - Distinct majority class, Mi - Distinct minority class

Table 2.7. Combined Study Results - Continued

Citation	Task	Non-linguistic	Morphological	Syntax	Semantic	Reduction	Techniques
Vilares et al. 2014	Topic	- lowercasing - # and @ removal - url tokenising		- dependency tree - lemmatisation - PoS	- emotion dict index	- boolean vectors - ngram (uni, bi, lemmas)	- **Linear**
Mane et al. 2014	Sentiment	- word norm	- stemming - stop word removal	- PoS	- emoji dict norm - sentiment dict index - word expansion		- **NB**
Cavalin et al. 2014	Topic	- nonalphanum. removal - punctuation norm - url tokenising - user tokenising - word norm	- tokenisation - stop word removal	- proper noun tokenisation	- knowledge dict filter - knowledge dict norm		- **NB**
Neuenschwander et al. 2014 *	Sentiment		- stop word removal	- lemmatisation - PoS			- SOCAL (Rules) - NB - **NBM**
Steed et al. 2015 †¶	Emotion	- lowercasing - nonalphanum. removal - punctuation removal - url tokenising - word norm	- stemming - stop word removal (modal verbs kept) - tokenisation		- word expansion	- word vectors	- **NB** - Max Entropy
Lee et al. 2015 †¶	Topic	- html tokenising - url tokenising	- stop word removal (pronouns kept)			- ngram (uni, bi, tri) - tf-idf	- NB - **NBM** - RF - SVM
Avvenuti et al. 2015 *	Event	- punctuation count - retweet flag - uppercase count - url flag - user flag - word count			- vocabulary index	- funct. (Pearson's) - information gain	- **DT**
D'Andrea et al. 2015 *	Event	- nonalphanum. removal - number removal - punctuation removal	- stemming - stop word removal - tokenisation			- information gain - stem filter - tf-idf	- **SVM** - NB - DT - kNN - PART (Rules)
Middleton and Krivcovs 2016 †	Event		- tokenisation - stemming	- PoS	- NER	- ngram - tf-idf	- **DT** - kNN - NB - RF - LogitBoost

Bold - Best performing technique, * - Non-English language corpus, † - Presentation of user interface
¶ - Non-standard stop word removal, ‡ - Favours speed over Accuracy/F-Measure, § - Non Tweet data source

Table 2.8. Articles included for quantitative analysis (Table continued on next page)

Citation	Task	Non-linguistic	Morphological	Syntax	Semantic	Reduction	Techniques
Kurniawan et al. 2016	Event	- lowercasing - nonalphanum. removal - RT syntax removal - url removal - user removal			- word expansion	- word vectors	- **SVM** - DT - NB
Karanasou et al. 2016	Sentiment	- word norm	- stop word removal	- grammatical index - negation index - PoS	- emoji dict index - sentiment dict index - vocabulary norm		- **SVM** - NB - DT - Linear (SGD)
Yu et al. 2016	Sentiment	- html removal - nonalphanum. removal - unicode removal - url removal - user removal	- tokenisation	- lemmatisation	- knowledge dict norm	- ngram (uni, bi)	- **SVM** - NB - Nonlinear SVM - Max Entropy - kNN - DT
Nguyen et al. 2016 †	Event	- date, loc, time index - nonalphanum. removal	- stop word removal	- lemmatisation - PoS	- knowledge dict index		- kNN - **BN** - SVM - DT
Azzouza et al. 2017 †	Sentiment	- word norm		- PoS	- emoji dict index - vocabulary norm - word expansion	- tf-idf	- **Rules**
Rezaei and Jalali 2017 ‡	Sentiment	- lowercasing	- stemming - stop word removal - tokenisation			- funct. (Gini) - tf-idf - token filter	- **DT (McD)** - DT (H)
Win and Aung 2017 ‡	Event	- hashtag count - url count	- stemming - stop word removal - tokenisation	- PoS	- knowledge dict index - emotion dict index	- ngram (uni, bi) - word vectors - information gain - word embedding	- **Linear** - SMO-SVM - RF
Behzadan et al. 2018	Event	- lowercasing - nonalphanum. removal - non -ascii removal - punctuation removal	- stemming - stop word removal - tokenisation		- vocabulary norm	- word embedding	- **CNN**
Benkhelifa and Laallam 2018 §	Topic	- number removal - punctuation removal	- stemming	- interjection index - PoS	- emotion dict index	- tf-idf	- **SVM**

Bold - Best performing technique, * - Non-English language corpus, † - Presentation of user interface
¶ - Non-standard stop word removal, ‡ - Favours speed over Accuracy/F-Measure, § - Non Tweet data source

Table 2.8. Articles included for quantitative analysis - Continued

Citation	Task	Non-linguistic	Morphological	Syntax	Semantic	Reduction	Techniques
Michailidis et al. 2018 †	Sentiment	- hashtag removal - number removal - url removal	- stemming - stop word removal				- **SVM** - NB - DT - Max Entropy
Golestani et al. 2018 §	Topic	- lowercasing	- stop word removal - tokenisation			- tf-idf	- **NBM** - SVM - RF - DT - Ad. Boosting
Șerban et al. 2018	Topic	- hashtag expansion - lowercasing - url removal			- emoji dict norm	- tf-idf (NB, SVM) - word embedding (RNN, CNN)	- NB - SVM - RNN - **CNN**
Alharthi et al. 2018 †¶	Topic	- hashtag index - Media removal - url removal - user removal	- stemming - tokenisation		- emoji dict index - emotion dict index - sentiment dict index	- ngram (uni bi, tri) - information gain - tf-idf	- **LSE SVM**
Vicente et al. 2018 *	Sentiment	- hashtag expansion - url tokenising - word norm		- grammatical index - interjection index - lemmatisation - PoS	- emoji dict norm - vocabulary norm	- ngram (uni)	- **SVM**
Ma et al. 2018 *§	Topic		- tokenisation			- attention layers (sentence, word) - funct. (softmax)	- NB - CNN - **tSAM-RNN** - SAM-RNN
Subramani et al. 2018	Topic	- hashtag removal - nonalphanum. removal - number removal - punctuation removal - user removal	- stemming - stop word removal - tokenisation	- PoS	- emotion dict index - sentiment dict index	- tf-idf	- **Log Reg**

Bold - Best performing technique, * - Non-English language corpus, † - Presentation of user interface
¶ - Non-standard stop word removal, ‡ - Favours speed over Accuracy/F-Measure, § - Non Tweet data source

Table 2.8. Articles included for quantitative analysis - Continued

2.3.2 RQ2 - DATA PROCESSING AND NORMALIZATION

Non-linguistic Analysis: User-generated content from social networking platforms features prevalent use of non-linguistic content such as references to web site and other users by their identifiers that may pose difficulties to NLP algorithms that have been developed for traditionally formatted discourse. Six studies manipulated user mentions (words preceded by the @ symbol) either by removing username (Kurniawan et al., 2016, Alharthi et al., 2018, Yu et al., 2016, Subramani et al., 2018), replacing the username with a generic representative token (Cavalin et al., 2014), or stripping off the @ prefix and retaining the username, or leaving the username intact (Vilares et al., 2014). URLs were more likely to be normalised, with a 50/50 split on tokenising vs. removing for the ten studies that manipulated them. There are two instances of HTML normalisation, with Lee at al. (2015) choosing to replace HTML instances with an HTML token and Yu et al. (2016) choosing to remove any instances of HTML completely, which will also cover URLs present in the text.

Surprisingly, only five studies reported manipulating hashtags in any way. Two use expansion techniques for segmenting the hashtag, e.g., through the use of heuristic camel case word splitting (Șerban et al., 2019, Vicente et al., 2018) with one of these taking it further by employing a prefix-based space prediction algorithm (Aho and Corasick, 1975) to break the hashtag into the minimum possible number of words when camel case splitting fails (Șerban et al., 2019). Two studies removed hashtags completely (Michailidis et al., 2018, Subramani et al., 2018), and the last one just removes the symbol (Vilares et al., 2014). A lot of studies chose to remove punctuation (Steed et al., 2015, D'Andrea et al., 2015, Behzadan et al., 2018, Benkhelifa and Laallam, 2018, Subramani et al., 2018), non-ASCII characters (Behzadan et al., 2018), Unicode characters (Yu et al., 2016), non-alphanumeric content (Cavalin et al., 2014, Steed et al., 2015, D'Andrea et al., 2015, Kurniawan et al., 2016, Yu et al., 2016, Nguyen et al., 2016, Behzadan et al., 2018, Subramani et al., 2018), numbers (D'Andrea et al., 2015, Benkhelifa and Laallam, 2018, Michailidis et al., 2018, Subramani et al., 2018); reducing user-generated content down to plain text. A total of six studies used word normalisation methods, generally reducing repetitive vowels within words down to single instances to compensate for social media vernacular (Mane et al., 2014, Cavalin et al., 2014, Steed et al., 2015, Karanasou et al., 2016, Azzouza et al., 2017, Vicente et al., 2018).

Morphological Analysis: This proved to be the least diverse part of the data extraction process with tokenisation, stemming and stop word removal representing all of the methods extracted from the data set here, often with all three being used in concert (Steed et al., 2015, D'Andrea et al., 2015, Rezaei and Jalali, 2017, Win and Aung, 2017, Behzadan et al., 2018, Subramani et al., 2018). There is some interesting use of stop word removal in several studies where non-standard stop word

removal was presented. They were focusing on sentiment analysis, whose performance may be affected by removing certain stop words including modal verbs and pronouns (Steed et al., 2015, Lee et al., 2015).

Syntactic Analysis: This category is dominated by traditional text mining techniques, predominantly part of speech (POS) (Vilares et al., 2014, Mane et al., 2014, Neuenschwander et al., 2014, Middleton and Krivcovs, 2016, Karanasou et al., 2016, Nguyen et al., 2016, Azzouza et al., 2017, Win and Aung, 2017, Benkhelifa and Laallam, 2018, Vicente et al., 2018, Subramani et al., 2018) and lemmatisation (Vilares et al., 2014, Neuenschwander et al., 2014, Yu et al., 2016, Nguyen et al., 2016, Vicente et al., 2018). A small number of studies show interest in emphasising particular grammatical elements such as interjections (Benkhelifa and Laallam, 2018, Vicente et al., 2018), negation and key phrases (Karanasou et al., 2016) and onomatopoeic tokens (Vicente et al., 2018).

***Semantic Analysis*:** Identification of emotions was used in over a third of studies, with emotion dictionaries used to index emotions in five studies (Vilares et al., 2014, Win and Aung, 2017, Benkhelifa and Laallam, 2018, Alharthi et al., 2018, Subramani et al., 2018), emoji lookup tables used in three studies (Karanasou et al., 2016, Azzouza et al., 2017, Alharthi et al., 2018), and the translation of emojis into a text representation in three others (Mane et al., 2014, Şerban et al., 2019, Vicente et al., 2018). Interestingly the use of sentiment dictionaries (Mane et al., 2014, Karanasou et al., 2016, Alharthi et al., 2018, Subramani et al., 2018) is lower than that of emotion dictionaries.

A number of well-established lexicons were cited, with the SentiWordNet (Esuli and Sebastiani, 2007) being employed by Karanasou et al. (2016) and Subramani et al. (2018), the LWIC lexicon (Pennebaker et al., 2001) used in Vilares et al. (2014) and Alharthi et al. (2018), and WordNet Affect (Strapparava et al., 2004) by Subramani et al. (2018). Other knowledge-specific lexicons were used in five studies to either filter out tokens (Cavalin et al., 2014), index key concepts (Nguyen et al., 2016, Win and Aung, 2017), or normalise text through disambiguation (Cavalin et al., 2014, Yu et al., 2016). Normalisation of vocabulary was also performed in a number of studies whereby either the vocabulary used was reduced (Karanasou et al., 2016, Azzouza et al., 2017, Behzadan et al., 2018, Vicente et al., 2018) or abbreviations and acronyms were expanded (Mane et al., 2014, Steed et al., 2015, Kurniawan et al., 2016, Azzouza et al., 2017). Interestingly NER is only presented in one study (Middleton and Krivcovs, 2016).

Feature selection: The predominant approach used for feature selection was TF-IDF (Lee et al., 2015, D'Andrea et al., 2015, Middleton and Krivcovs, 2016, Azzouza et al., 2017, Rezaei and Jalali, 2017, Benkhelifa and Laallam, 2018, Golestani et al., 2018, Şerban et al., 2019, Alharthi et al., 2018,

Subramani et al., 2018). N-grams were used in seven studies (Vilares et al., 2014, Lee et al., 2015, Middleton and Krivcovs, 2016, Yu et al., 2016, Win and Aung, 2017, Alharthi et al., 2018, Vicente et al., 2018), with several using bi-grams (Vilares et al., 2014, Lee et al., 2015, Yu et al., 2016, Win and Aung, 2017, Alharthi et al., 2018), two studies using tri-grams (Lee et al., 2015, Alharthi et al., 2018) and one applying the *n*-gram principle to both words and their lemmas (Vilares et al., 2014). Other approaches used information gain (Avvenuti et al., 2015, D'Andrea et al., 2015, Win and Aung, 2017, Alharthi et al., 2018), word embeddings (Win and Aung, 2017, Behzadan et al., 2018, Șerban et al., 2019), Pearson's correlation coefficient (Avvenuti et al., 2015), the Gini index (Rezaei and Jalali, 2017) and the softmax function (Ma et al., 2018).

When looking at the pre-processing methods, in the context of classification performance, there was a positive correlation of 0.41 and 0.42 in the social media Normalisation and Morphological Analysis categories, respectively. When the two categories were to be combined into one to align with Abbe et al.'s original categories, the correlation rises to 0.50. This is of interest as it suggests that the normalisation of social media texts has a positive effect on the classification performance.

2.3.3 RQ3 - Machine Learning Algorithms

This section focuses primarily on the choice of machine learning algorithms used to support text classification. The difficulty of this task depends on the underlying classification scheme and the distribution of training data across the classes. The majority of studies focused on binary (Kurniawan et al., 2016, D'Andrea et al., 2015, Nguyen et al., 2016, Benkhelifa and Laallam, 2018, Behzadan et al., 2018, Subramani et al., 2018, Steed et al., 2015, Michailidis et al., 2018, Middleton and Krivcovs, 2016, Șerban et al., 2019, Avvenuti et al., 2015, Rezaei and Jalali, 2017) and ternary (Alharthi et al., 2018, Win and Aung, 2017, Karanasou et al., 2016, Cavalin et al., 2014, Lee et al., 2015, Yu et al., 2016, Golestani et al., 2018, Neuenschwander et al., 2014, Mane et al., 2014, Vicente et al., 2018, Azzouza et al., 2017) classification. Only two studies used more than three classes; Vilares et al. (2014) classifies into 7 classes and Ma et al. (2018) is the only outlier with a classification set of 2000. Next, we examined the class balance using a standard deviation in class volume over of 0.05 of the total volume as an indicator of class unbalance.

The focus of the classification tasks can be broken down into three main foci; event detection (Avvenuti et al., 2015, D'Andrea et al., 2015, Middleton and Krivcovs, 2016, Kurniawan et al., 2016, Nguyen et al., 2016, Win and Aung, 2017, Behzadan et al., 2018), topic based classification (Vilares et al., 2014, Cavalin et al., 2014, Lee et al., 2015, Benkhelifa and Laallam, 2018, Golestani et al., 2018, Șerban et al., 2019, Alharthi et al., 2018, Ma et al., 2018, Subramani et al., 2018), and sentiment analysis (Mane et al., 2014, Neuenschwander et al., 2014, Karanasou et al., 2016, Yu et al., 2016,

Azzouza et al., 2017, Rezaei and Jalali, 2017, Michailidis et al., 2018, Vicente et al., 2018). Classification performed by (Steed et al., 2015) was focused on emotion, which could be both related to topic classification and sentiment.

Interestingly, we notice a gap in the literature here focused on the analysis and classification of emotion. We can see above that sentiment analysis is a popular task within the studies, and within these studies five of them use emotion dictionaries as part of semantic analysis (Vilares et al., 2014, Win and Aung, 2017, Benkhelifa and Laallam, 2018, Alharthi et al., 2018, Subramani et al., 2018). But sentiment is a coarser level of classification when compared to emotion classification, which is only performed by a single study within the survey (Steed et al., 2015).

Looking at the binary classifiers, a total of six were trained with evenly balanced classes (Kurniawan et al., 2016, D'Andrea et al., 2015, Benkhelifa and Laallam, 2018, Steed et al., 2015, Michailidis et al., 2018, Yu et al., 2016). Four of these studies achieved even classes through under sampling of the majority class (Kurniawan et al., 2016, D'Andrea et al., 2015, Benkhelifa and Laallam, 2018, Yu et al., 2016), the fifth (Steed et al., 2015) used a third party data set where a hard limit was placed on the amount of data collected for both classes (Go et al., 2009a) to produce an even data set, whilst the sixth (Michailidis et al., 2018) achieved a 47:53 balanced data set by artificially inflating the volume of the minority class using Synthetic Minority Oversampling TEchnique (SMOTE) to synthetically create additional data using the existing classified content (Chawla et al., 2002). SMOTE was also used in the only ternary classifier (Alharthi et al., 2018).

ML Technique	Best	Papers	Versus		
			None	Self	Others
SVM	8	13	3		16
Naïve Bayes	3	13	2		1
Decision Tree	3	10	2	1	4
NB Multinomial	3	3			9
CNN	2	3	1		3
Linear Classifier	2	3	1		2
Rule Based	1	3	1		
RNN	1	3		1	2
Bayesian Network	1	1			3
Regression	1	1	1		
kNN		4			
Random Forest		4			
Max Entropy		3			
Boosting		2			
Nonlinear SVM		1			
	25	**67**	**11**	**2**	**40**

Table 2.9. Prevalence of machine learning techniques.

Studies employing an unbalanced data set for ternary classification performed better when there was a distinct majority class (Win and Aung, 2017, Karanasou et al., 2016, Cavalin et al., 2014) relative to studies that had a distinct minority class (Mane et al., 2014, Vicente et al., 2018, Azzouza et al., 2017). The implication is that the under-sampled class negatively effects the average performance values, resulting in the lower performance scores. Studies with an evenly distributed data set proved to be the best performing and would perform better (avg of 4.0) if not for Yu et al. (2016) which may be suffering from smallest training data set of only 200.

A wide range of machine learning methods were used to support text classification (See Table 2.9). As expected, the vast majority used supervised learning algorithms. Out of 25 studies, a total of 12 compared multiple methods and two compared different implementations of the same method. A number of studies cited common software packages that support multiple implementations of natural language processing algorithms such as scikit-learn (Pedregosa et al., 2011), Weka (Eibe et al., 2016), LibLinear (Fan et al., 2008) and word2vec (Mikolov et al., 2013).

Three methods were used predominantly: SVMs (Cortes and Vapnik, 1995), naive Bayesian learning (Domingos and Pazzani, 1997) and decision trees (Quinlan, 1986). SVMs performed best in 8 out of 13 studies (Kurniawan et al., 2016, D'Andrea et al., 2015, Benkhelifa and Laallam, 2018, Alharthi et al., 2018, Michailidis et al., 2018, Karanasou et al., 2016, Yu et al., 2016, Vicente et al., 2018), naive Bayes learning performed best in 3 out of 13 studies (Steed et al., 2015, Cavalin et al., 2014, Mane et al., 2014), and decision trees performed best in 3 out of 10 studies (Middleton and Krivcovs, 2016, Avvenuti et al., 2015, Rezaei and Jalali, 2017). SVMs were frequently compared to naive Bayes algorithms (D'Andrea et al., 2015, Șerban et al., 2019), decision tree algorithms (Golestani et al., 2018), or both (Kurniawan et al., 2016, Karanasou et al., 2016, Yu et al., 2016, Michailidis et al., 2018), and frequently outperformed both of these algorithm types. Naive Bayes Multinomial algorithms consistently outperformed all other methods that they were compared to (Lee et al., 2015, Golestani et al., 2018, Neuenschwander et al., 2014), including SVMs in two of the three (Lee et al., 2015, Golestani et al., 2018).

Studies favouring SVMs featured a heavy use of normalisation techniques, with the majority opting to remove idiosyncrasies of the social media texts or non-alphanumeric characters. Interestingly, only three studies that favoured SVMs used POS tagging and did so to identify particular lexical classes such as interjections and onomatopoeias to improve classification performance. Emotion or emoji dictionaries were used in 50% of the SVM favouring studies.

From the studies that favoured Bayesian classifiers, the better performing approaches made heavier use of normalisation techniques relative to the lower performing ones. Stop words were consistently

removed, suggesting that probabilistic models may be more sensitive to these features. However, Steed et al. and Lee et al. chose to retain parts of speech that would usually be lost with stop word removal. Steed et al. choose to keep modal verbs as they are commonly used in emotive content. Lee et al. postulated that people describing their allergies are more likely to use possessive pronouns.

There was sparse use of data processing and normalisation in the studies favouring Decision Trees. Tokenisation, stemming, and the use of TF-IDF was present in two studies (Middleton and Krivcovs, 2016, Rezaei and Jalali, 2017). The use of correlation coefficients for feature selection was observed in both Avvenuti et al. (2015) and in Rezaei and Jalali (2017), although different methods were selected with the former using *Pearson's* and the latter *Gini*.

Recent years have brought an increased use of deep learning methods, in particular convolutional neural networks (CNNs) (LeCun et al., 1998, Kim, 2014) and recurrent neural networks (RNNs) (Williams and Zipser, 1989) due to the lowered barriers of entry provided by cloud computing platforms. CNNs demonstrated high classification accuracy in Behzadan et al. (2018) and Șerban et al. (2019). Ma et al. (2018) also employed a number of variations of Long Short-Term Memory (LSTM) based RNNs (Hochreiter and Schmidhuber, 1997) to support multi-class classification; something not commonly seen in this review.

2.3.4 RQ4 - Performance

Training data size did not appear to have a major impact on the performance, with a correlation coefficient of -0.30 when outlier training sets (Steed et al., 2015) were removed. There is a stronger correlation between the time range and classification, with a correlation coefficient of -0.47. It is possible that the shorter time windows are more homogeneous leading to better classification performance, though it is not clear whether this was due to overfitting.

Deep learning relies on a large volume of training data and this is reflective in the data set sizes for both Behzadan et al. (2018) and Ma et al. (2018), with data set sizes of 21,000 and 160,000 respectively, both of which sit above the upper quartile value for data set size. Șerban et al. (2019) has a much smaller data set size relative to the other studies investigating deep learning. This study is outperformed by Behzadan et al. (2018), but not by Ma et al. (2018) which is possibly due to the fact that Ma et al. are classifying with 2,000 classes and so although they have 160,000 documents that only averages out at 80 documents per class, versus 4,676.5 documents per class in Șerban et al. (2019).

There was no consistent means of assessing the speed of an algorithm operating within a real-time environment. A number of studies performed additional experiments aimed at either assessing the

performance of their algorithms within a live experiment (D'Andrea et al., 2015, Nguyen et al., 2016, Middleton and Krivcovs, 2016, Rezaei and Jalali, 2017, Şerban et al., 2019, Vicente et al., 2018) or at bench-marking the performance of supporting architecture (Karanasou et al., 2016). Only two studies highlighted a preference for speed over performance when selecting their preferred algorithms (Win and Aung, 2017, Rezaei and Jalali, 2017). Rezaei and Jalali (2017) favoured the McDiarmid Tree algorithm; which processed documents 0.57 seconds faster than the Hoeffding Tree at the cost of decrease in accuracy by 0.08%. Win and Aung (2017) favoured a LibLinear (Fan et al., 2008) based classifier in their study over a SMO trained SVM (Platt, 1998) despite an average accuracy cost of 0.43% across several training sets, citing a faster processing time as the reason for this preference (Win and Aung, 2017).

Classification performance was reported inconsistently, with the better performing studies tending to report accuracy. Lower performing studies tended to omit accuracy favouring precision, recall and F1 score instead. It should be noted that of the eight studies that did not report accuracy, the top four all presented user interfaces, suggesting that either precision or recall was preferred by the user requirements. For instance, Middleton and Krivcovs (2016), achieved a recall of 99% against a precision of 68%.

2.4 Conclusions

We have seen that text classification results are affected by the quality of the training data, with an emphasis on the preference towards an evenly balanced data set. Larger data sets were correlated with better performance, but not to the same degree as the size of the collection window, with a smaller window correlated with better performance. Domain-specific, API driven collection is the most prevalent origin of data sets, although there is also a lot of re-use of previously published data sets as a means of bench-marking algorithms against other studies where the application domain is more generic or popular. Twitter still dominates in terms of document type, with a small number of studies exploring data from other platforms.

Consistent trends in text normalisation have been observed, with attention being paid to the non-natural language entities found in social media. Username, URI, and hashtag normalisation techniques are present within many of the studies in this review. These are key elements in enriching social media text and so it is of no surprise that this featured heavily in the review. It is also apparent that it is important to reduce documents down into plain text to assist algorithms in processing social media content. Classic NLP tasks such as tokenisation, POS tagging, lemmatisation and stemming are present throughout the review and there is a focus on lexico-semantic analysis of sentiment and emotion.

We saw three types of algorithms frequently presented in the study set: SVMs, Bayesian classifiers, and decision trees. These algorithm types were regularly tested in concert with one another, with SVMs outperforming these and other algorithms most frequently. Neural networks were present in the study set, but only in the more recent studies. This is reflective of the ease by which these algorithms are available through software packages such as the Python-based scikit-learn or the Java-based Weka, that makes comparable implementation very accessible to researchers.

It is acknowledged that systematic reviews are more commonly used within the Medical Sciences and are most powerful when assessing the outputs of clinical studies. These studies can follow a much more rigorous and standardised means of hypothesising, candidate selection and variable control. Social media text analysis is still a relatively new field when compared to clinical study, and so there was no expectation that text mining papers would follow a consistent form in experiment design or paper presentation. This added to the challenge of performing a systematic review of this kind.

Finally, we identify two potential gaps in the literature that our data analytics methodology focuses upon, namely the lack of diversity in the types of social media data and the lack of focused classification of emotion. As the public's use of social media matures, discourse is increasingly moving across platforms and is not confined to a single space but there is currently an over-reliance on Twitter as a source of data that is evident through the study. The study also highlights the popularity of sentiment analysis. Whilst this area of classification is well defined, the more nuanced analysis of emotion is not nearly as well explored in the survey. These two gaps are relevant to the rest of this thesis, with emotion classification becoming a key output of some of our earliest case study (Section 4.4.4), and our download functionality presented in Section 5.2 and topic modelling covered in Section 5.5 being driven by the need to allow users access to a range of social media data types.

CHAPTER 3: DATA COLLECTION
THE SENTINEL PIPELINE
SOURCE IDENTIFICATION, INFRASTRUCTURE DESIGN AND PIPELINE IMPLEMENTATION

3.1 INTRODUCTION

This chapter focuses on the development of the data processing pipeline component of the Sentinel Platform, called the Sentinel Pipeline. As discussed in Chapter 1, this work was undertaken as part of a co-design development process between social scientists focused on the practice and science of policing and computer scientists with research interests in data mining and decision-support (Preece et al., 2018). This chapter will address elements of the core principles that drove the overall research focus, namely the creation of a glass box platform for social media ingestion processing and storage, that is designed to be as open as possible to the integration of new components, models, data sources, and user interfaces.

3.1.1 PRIMARY OUTPUT: THE SENTINEL PIPELINE

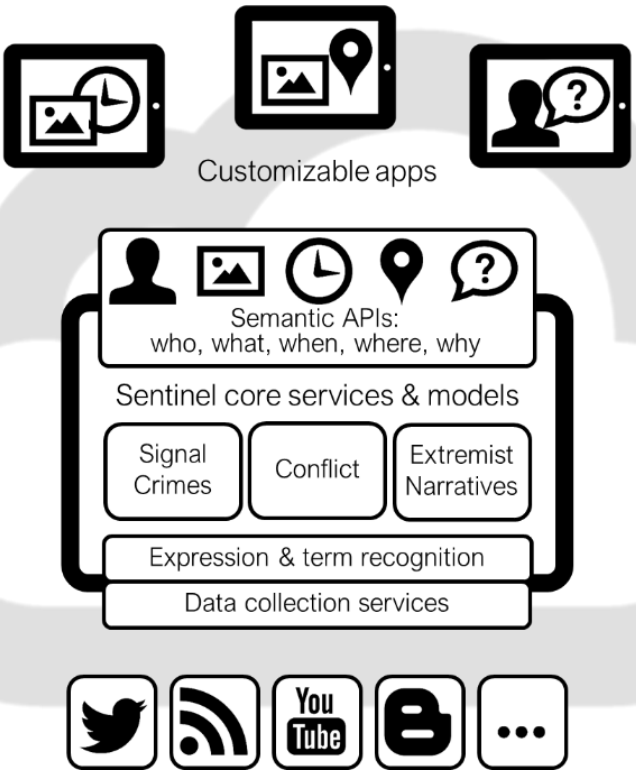


Figure 3.1: High level overview of the Sentinel Platform.

A practical vehicle for this work is through the development of our own social media analysis platform, named Sentinel ('Semantic Intelligence'), whose primary purpose is to provide a data-agnostic datastore of social media content that are enriched via a number of pluggable semantic models and classifiers, along with a number of customisable interfaces and apps that allow the data to be interrogated from multiple viewpoints and contexts (Figure 3.1).

There are several components of the Sentinel Platform that will be discussed throughout this thesis, which we present below in order to clearly differentiate these components:

- **Sentinel Platform** - The overall research tool that all other components belong to along with database and indexing services.
- **Sentinel Pipeline** - The data processing pipeline that covers all of the real-time collection, normalisation, and enrichment of social media data.
- **Sentinel Interface(s)** - The initial single-view prototype and alpha interfaces used to interrogate the data collected and processed by the Sentinel Pipeline.
- **Open-Source Communications Analytic Research (OSCAR) Hub** - The beta and production multi-view web portal that manages user access to the data, hosting all available interfaces into the collected data and processed information.

These components were designed with the intention of allowing qualitative researchers to perform situational awareness tasks as defined in the Sensemaking Loop, in order to answer questions pertaining to the 5Ws sensemaking model, using social media data as the driving source of information.

The Sentinel Pipeline is intended to be social media Agnostic, where data from any social media Platform could be imported and analysed. At its core, this comprises of platform-specific data collectors, data cleaning processes and data storage facilities that allow the agile development of web services that support researchers in building sensemaking tools.

3.1.2 Secondary Output: Prototype and Alpha Sentinel Monitoring Interfaces

A secondary output presented in this chapter is the early development of the Sentinel Web application, which provides a multi-level interface into the information extracted from the social Web through this project. The purpose of this output is to illustrate how one can determine the "Who", "What", "When", "Where" and "Why" elements of discourse surrounding narrative events. "Who" are the groups escalating events; "What" are they talking about; "When" are events unfolding and "Where" are they occurring; "Why" are these events are happening?

3.2 EVENT DRIVEN CO-DESIGN

The evolution of the Sentinel system is driven through a series of planned and spontaneous events that provided focal points for the co-design process; allowing for the peer group to collaborate on new applications and interface designs, both proactively and reactively. The choice to use live events as the core source of data, experimentation, and evaluation of the system is driven by the desire to maintain the ecological validity of findings. Ecological validity is the study of subjects' behaviour in naturally occurring situations, allowing problems and questions to emerge from within these situations (Gehrke, 2014). It requires a trade-off against the control of any case study and experimentation used for evaluation of a system, but the social media ecosystem is constantly evolving (Liu et al., 2014). Stewart and Arnold (2018) highlight that the process of social listening is dynamic due to the ever-changing nature of the digital communication landscape and state that the multidimensional nature of social listening invites opportunities for listening within events. Users and platform providers themselves are observing events through the same lens and changing their behaviour accordingly, meaning synthesised or controlled datasets may not remain representative of the current state.

Figure 3.2 presents the event driven co-design lifecycle that was operational during the development of the Sentinel system. This model captures the co-design activity engaged with by stakeholders and developers through cycles of system iteration through layered elaboration. It describes how the developers and stakeholders shift their mode of operation between responsive and reflective work, and between knowledge led and data led approaches. The workflow cycles around these two principal components resulting in four states, each of which consists of a primary driving task and co-design focused activity, along with four transitioning activities that can occur when stakeholders move between these states. This repeated return to ecologically valid data allows for tools, models, and protocols developed by co-design to be repeatedly tested and refined against real world events.

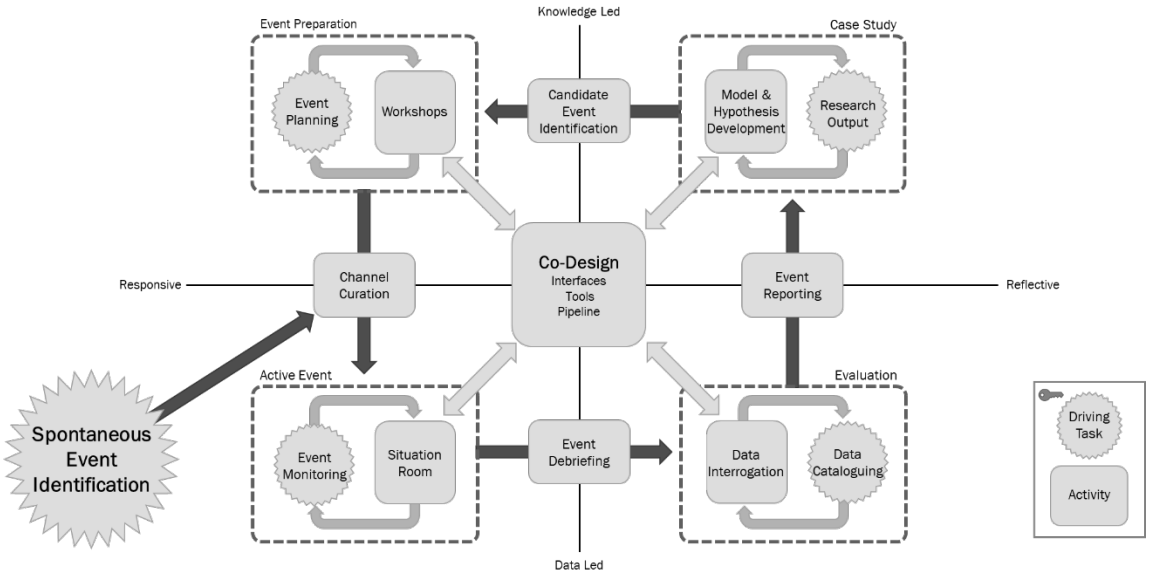


Figure 3.2: Event driven co-design lifecycle.

The knowledge led and data led component can be linked back to the sensemaking loop by Pirolli and Card (2005) that is discussed in Section 1.3.1.2; with the data led half primarily focused on *bottom-up* processes, and the knowledge led half having a greater focus on *top-down* processes. This does not mean that the processes in these halves are exclusive to these two concepts, only that a stakeholder's attention is typically more focused on one of these concepts at a particular point in the lifecycle.

The responsive half of lifecycle is primarily focused on using events in the real world to trigger activity and creativity within the co-design process. Moving through these states is typically time-critical, commencing with the identification of an event (either following the development of new models and hypotheses, or through alerts to spontaneous events) and concluding with a debriefing session undertaken by stakeholders following an event. Channel curation is an important part of the process and is either performed as part of the planning process during the co-design activity, or as part of a rapid response to a spontaneous event. The reflective half is focused on using theoretical triggers for investigation into events and the development of case studies in order to further the co-design of the system. Focus of the co-design activities is towards deeper dives into the collected data to better understand an event post-hoc; with an emphasis on Data Science within the evaluation phase, and on Social Science when performing case study.

3.2.1 Co-Design Activities

The four codesign activities are discussed below, in which we refer back to the co-design types defined by Zamenopoulos and Alexiou (2018) and the co-design enablers highlighted by Pirinen (2016) that were introduced in Section 1.1.4.

The co-design activities present in the responsive half are characterised by a greater degree of large multi-stakeholder/developer sessions, consisting of teams of stakeholders preparing for events and monitoring developments as a collective in near real-time and in a shared or virtual space:

- **Workshops** – These are a core activity in the co-design process, where stakeholders and developers can jointly explore and articulate their latent needs and jointly explore and develop solutions (Steen et al., 2011). We utilised these workshops to develop research ideas around planned events that had been identified by the members of the stakeholder group, seeking to enable the *search for mutual value* in events and to provide opportunity for the *coordination and timing of* existing co-design outputs. The output of these workshops came in the form of; the *collaborative* design of new interface designs for Sentinel aimed at supporting any real-time experimentation planned for an active event; collection channels curated though *collective* intelligence for data to be fed into the situation rooms; and engaging new stakeholders through development of parallel research strands (Preece et al., 2015) through the provision of a *co-operative* space.
- **Situation Rooms** – A novel approach to workshopping in live events. *Situation rooms* are active during the event taking place and are characterised by a combination of Computer Science and Social Science stakeholders using the Sentinel platform among other tools for investigative tasking and situational awareness. Like the workshops, this activity takes place in a shared "third space", with a number of stakeholders present. In addition to this, "field team members" may be present at the event performing supporting tasks (Preece et al., 2016). The co-design process here was driven by the need for stress testing newly developed tools and interfaces produced during the planning workshops, and on occasion rapid prototyping of new interfaces and analysis techniques to quickly respond to requests from team members both inside and outside of the situation room. This helps to facilitate new *connections* between strands of work and tools. This activity has a short duration in comparison to the other co-design activities, as it is time dependant on the focal event.

The co-design activities for the reflective half were generally performed by much smaller teams or pairs of stakeholders and developers collaborating over long periods of time in a discrete manner without the need to be in the same physical or virtual space:

- **Data Interrogation** – Over the course of a planned or unplanned event, a large volume of data can be collected. This activity is driven by the need to characterise the collected data post-event quantitatively and qualitatively and build on any *prototype development* that has taken place during the situation room activity. Furthermore, the interrogation activity allows for the identification of any post-event re-heating of an event (Collins, 2012). This is a novel

activity in that it is primarily led by Data Science practice and is performed by smaller groups of stakeholders and developers. These smaller groups enable stakeholders to *take responsibility and ownership* of co-designed components, and further develop and refine them.

- **Model & Hypothesis Design** – This activity links the co-design system back to the wider stakeholder literature. It is important in this activity that developers are embedded within the stakeholder's research process, in order to respond to any need for new system components. From a co-design perspective *continuity beyond singular projects* is the main enabler focus here, ensuring that tools and learnings developed through the event driven lifecycle are relevant to a broader scope. This activity is generally driven by a Social Science focus, with the development of new interpretative models that can link both the observed behaviours present in events being key. Hypothesis design through case study also allows stakeholders to *collectively* steer the research focus towards newly emergent theory, and the identification of suitable upcoming events or types of spontaneous events that can continue the lifecycle and iteration of the co-design process.

3.2.2 Event Timeline

Figure 3.3 presents the timeline of events utilised within the Sentinel co-design process. These planned and spontaneous events initially had an acute geospatial focus that over the course of the work broadened to cover wider and more chronic events.

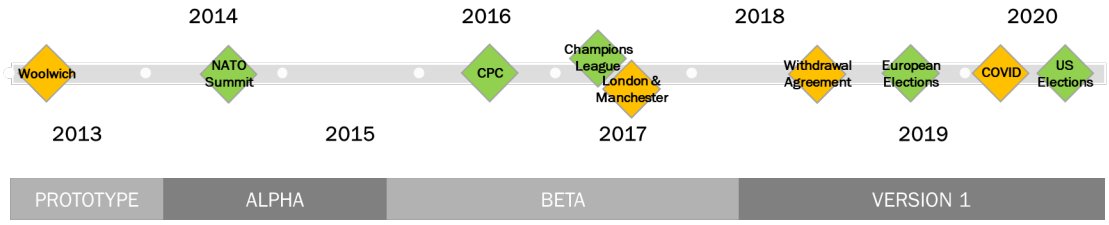


Figure 3.3: Planned Events and Spontaneous Events Used to Drive System Co-Design. Green diamonds identify planned events, amber diamonds identify spontaneous events.

3.2.2.1 Planned Events

Planned events (green, Figure 3.3) provided a structure and consistent pace to the co-design and development process and were initially focused upon events that had a geospatial focal point that had heavy policing and security concerns. Both the NATO Summit held in Newport and Cardiff in 2014, and the Birmingham based 2016 Conservative Party Conference, held shortly after the 2016 Brexit vote, provided the research team with events that had heads of state and national leaders in in attendance and the high likelihood of protests and marches occurring in the periphery of the events. The final geospatially focused planned event was the 2017 UEFA Champions League held in

Cardiff, which saw a large policing operation to accommodate the arrival of a large contingent of Italian and Spanish football fans in the city.

The latter two planned events were focused on superstate the European Union Parliamentary Elections and the United States of America Presidential and Congressional Elections. These had a much larger geographical scope and whilst both had a series of key dates associated with the elections themselves, they exhibited a large lead-up of time ahead of, and fall out following, these dates where the elections and related events heavily contributed to public discourse.

3.2.2.2 SPONTANEOUS EVENTS

It was anticipated at the beginning of this work that during the course of the research programme, the Sentinel system should be capable of supporting the rapid response to spontaneous security events. Use of spontaneous events is not seen in co-design literature likely due to the unconstrained nature of the subject matter, but through the two novel data let co-design activities we engage with them in a similar manner to a planned event. Spontaneous events (amber, Figure 3.3) regularly tested the agility and robustness of the collection capabilities of the system where need is for data to be rapidly collected to build the datasets. The aftermath of these events were also key drivers in the co-design process; through consolidatory qualitative and quantitative investigation and analysis of the collected data. Characterisation of the event datasets was often required and so the *data interrogation* activity was highly important, as there would be no opportunity to run workshops in preparation of the event, producing new questions and interpretive views of the data post-event.

This was rapidly realised when the murder of Fusilier Lee Rigby in Woolwich occurred while the prototype system was still being developed (Section 3.6.1), with a focus in the *interrogation* activities around *bottom-up* identification of key moments within the event. It also led to the development of audience behavioural models characterising public response to terrorist atrocities (Innes et al., 2018, Roberts et al., 2018), along with studies into the importance of the first few hours after an incident, and how narratives begin to build during the void of information (Innes et al., 2014). This focus on terrorist atrocities was continued into the second cluster of spontaneous events; a series of 4 terrorist attacks in London and Manchester during the summer of 2017. Again, case study of these events generated further insight and understanding of the void of information present following an atrocity with a focus on rumour and disinformation present in the aftermath (Innes, 2020, Roberts et al., 2015).

These spontaneous events had a strong influence on the research direction undertaken by the stakeholders and developers influencing the choice of subsequent planned events discussed above, and the identification of the final two spontaneous events. These were the much larger in scale,

covering the national and international reactions to the 2018 Brexit Withdrawal Agreement and to the emergence of the COVID-19 Pandemic in 2020 (the *interrogation* activity of this is covered in detail in Section 5.5).

3.3 DATA SCOPING

3.3.1 SOURCES

3.3.1.1 SOCIAL MEDIA AND "MAINSTREAM MEDIA" COMMENTS IN 2013

An investigation into the types of media available for collection was undertaken at the beginning of the project. This focused on how readily available user-generated content was. Note was also taken, where possible, of the numbers of unique users exposed to each online platform. Table 3.1 presents a summary of relevant platforms assessed in 2013.

Data Source	Type	User Reach	Access to Content
Facebook	Social Media	1.11 billion [2]	Web Crawler
YouTube	Social Media	1 billion [3]	Google Dev API
Google+	Social Media	235 million [4]	Google Dev API
Twitter	Social Media	200 million [5]	Twitter APIs
WordPress	Blog	300 million/month [6]	Web Crawler
Blogger	Blog	n/a	Google Dev API
TypePad	Blog	n/a	Web Crawler

Table 3.1: Summary of access to major online discussion and social media platforms, 2013.

Facebook was initially considered a high priority for content retrieval due to its vast user base and the composition of content found within, but due to the lack of an open API for programmatic access and, more importantly, terms and conditions prohibiting programmatic access lowered their priority within the development of Sentinel. Both Twitter and the Google provide several APIs designed to assist developers in retrieving content from their sites.

A second investigation was undertaken focused on web sites belonging to British news media. The investigation focused on the simplest way of extracting article content and comments from mainstream news sites with the results presented in Table 3.2.

[2] http://investor.fb.com/releasedetail.cfm?ReleaseID=761090
[3] https://web.archive.org/web/20131004182301/http://www.youtube.com/yt/press/statistics.html
[4] http://googleblog.blogspot.co.uk/2012/12/google-communities-and-photos.html
[5] https://blog.twitter.com/2013/celebrating-twitter7
[6] http://www.techspot.com/news/46236-wordpress-powers-60-million-blogs-300-million-unique-visitors-monthly.html

Newspaper	Online Readership [7]	Access to Comments
Daily Mail	2.449 million	JSON endpoint
The Guardian/ The Observer	2.475 million	Guardian Open Platform/JSON endpoint
The Daily Telegraph	1.848 million	DISQUS
Daily Mirror	1.123 million	HTML endpoint
The Sun	1.076 million	LiveFyre
The Independent	1.056 million	DISQUS
Financial Times	334 thousand	Paywall
Daily Express	291 thousand	JSON endpoint
The Sunday Times/ The Times	178 thousand	Paywall

Table 3.2: Summary of available access to national newspaper comments.

The Times and The Financial Times both have online access restricted via paywalls, and since the investigation took place, The Sun has also moved behind a paywall. It was decided early that paywalled sites would not be investigated further at this point, as they do not qualify as Open Source data (Glassman and Kang, 2012).

The majority of the media web sites load their comments into articles through an AJAX request to a 'comments' endpoint, with the majority of these endpoints returning the data as JSON objects. These can be easily exploited to quickly harvest article comments.

From this investigation we identified Twitter, YouTube and MailOnline comments as three data sources that we would prioritise collection methods for over the course of this research in order to provide a range of Social and Traditional media comment types, to push the development of a homogeneous open-source analysis platform.

From these three, we decided in consultation with collaborators in the Crime and Security Research Institute (CSRI) that Twitter would become the pilot data source from which we would focus initial development and case studies around. The collection strategy and methods are discussed later in this chapter in Section 3.4.3.4.

3.3.1.2 Reassessment of Collection Focus in 2017

Following the core development of the Sentinel pipeline using Twitter, we returned to the other two priority data sources and began to incorporate collection strategies for both YouTube and MailOnline into the Sentinel pipeline. Along with this, we reassessed the current social media landscape to identify any emergent platforms that would complement the collection focus which is presented in Table 3.3. From this we identified Reddit as a readily accessible data source that is

[7] http://www.pressgazette.co.uk/uk-newspapers-ranked-total-readership-print-and-online

complementary to previously considered data sources. The collection strategy for YouTube, MailOnline and Reddit is also discussed in Section 3.4.3.4 later in this chapter.

Data Source	Type	User Reach	Access to Content
Facebook	Social Media	2.13 billion/month [8]	FB Graph API
YouTube	Social Media	>1 billion [9]	Google Dev API
Instagram	Social Media	800 million [10]	Instagram API
Reddit	Social Media	330 million/month [11]	Reddit API
Twitter	Social Media	330 million/month [12]	Twitter APIs

Table 3.3: Summary of access to top social media platforms, 2017.

3.3.2 Behavioural Coverage

Social media platforms form a constantly evolving ecosystem existing and newly emerging platforms competing for users' time. Kietzmann et al. (2011) published a formalisation of seven functional building blocks that they consider to be the building blocks of social media and microblogging platforms: presence, sharing, relationships, identity, conversations, reputation and groups (Figure 3.4a).

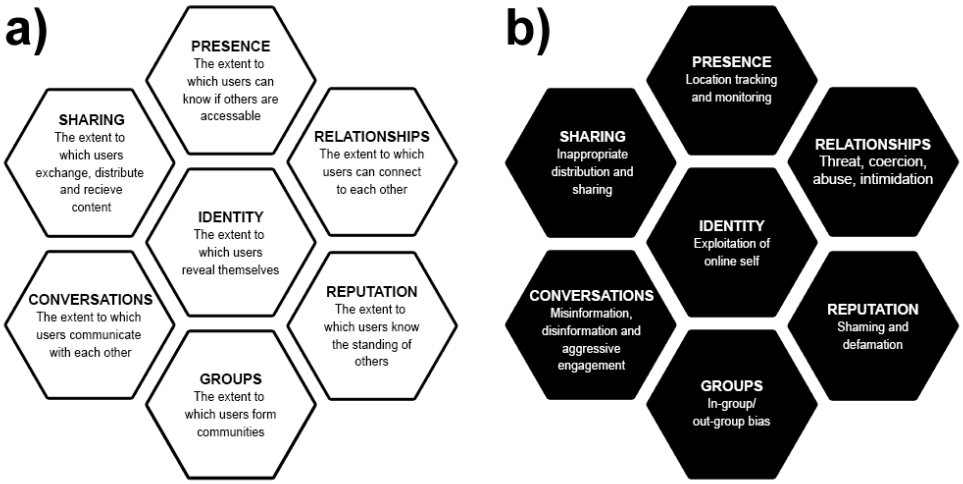


Figure 3.4: The functional building blocks of social media (Kietzmann et al., 2011) and The dark side of social media functionality (Baccarella et al., 2018).

[8] https://investor.fb.com/investor-news/press-release-details/2018/Facebook-Reports-Fourth-Quarter-and-Full-Year-2017-Results/default.aspx
[9] https://web.archive.org/web/20170815002447/https://www.youtube.com/yt/about/press/
[10] https://business.instagram.com/blog/safety-and-kindness-for-800-million/
[11] https://web.archive.org/web/20180409192132/https://www.redditinc.com/press/
[12] https://s22.q4cdn.com/826641620/files/doc_financials/2017/q3/Q3_17_Shareholder_Letter.pdf

Facebook and YouTube were among the initial platforms presented in this study and a reassessment of these and other platforms was performed by Haefner (2014) shows that these characteristics can change over time as a platform matures and attempts to retain and grow their userbase.

This evolution can be observed in Figure 3.5 in the development of Facebook into a monolithic platform that has adapted into a full information aggregation platform that attempts to fully cater for a users' social needs with article sharing, instant messaging and group pages being just some of the core features (Innes et al., 2017).

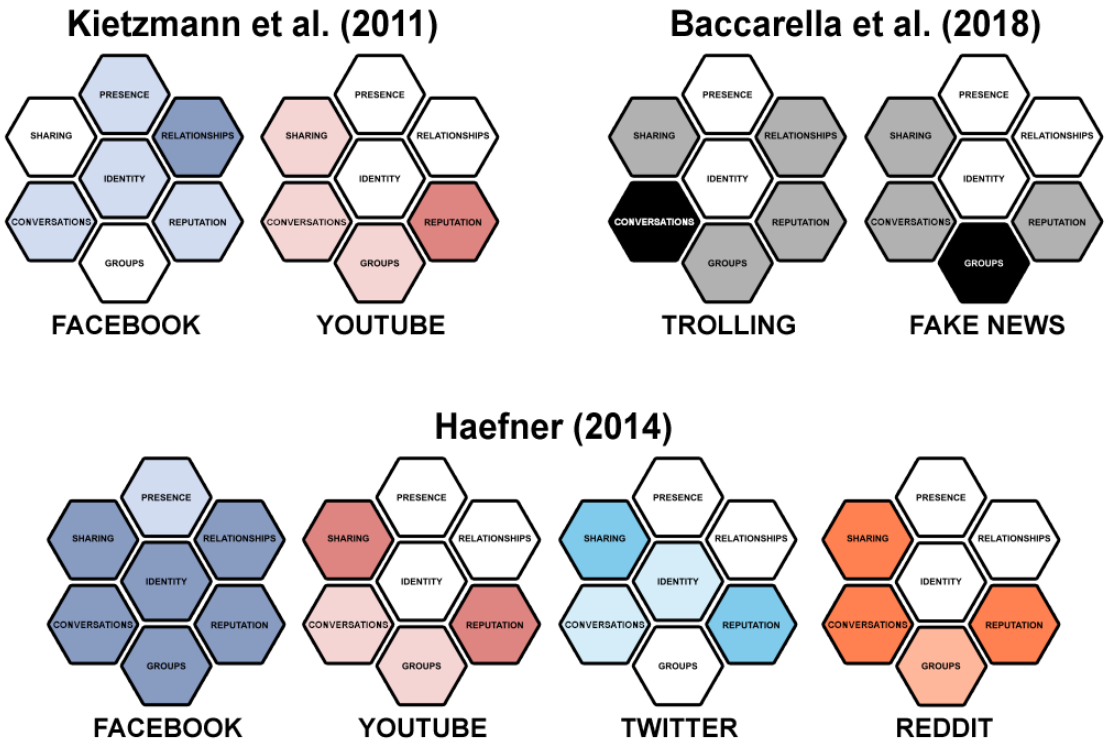


Figure 3.5: Social media honeycombs.

Also presented in Figure 3.4b is the further development the honeycomb concept by Baccarella et al. (2018). They identify means by which each honeycomb concept can be exploited by bad actors, highlighting the dangers and harms present to users and society within social media. They provide further context to these concepts with a demonstration of how *trolling* and *fake news* focus on a number of these building blocks (Figure 3.5).

There is a significant overlap between these common harmful social media activities and the main features of the three social media platforms that have been incorporated into the Sentinel pipeline. This overlap suggests that these platforms are fertile ground for bad actors to operate. Indeed example cases from all three platforms are cited by Baccarella et al. (2018) when describing the dark side functionality.

3.4 INFRASTRUCTURE DESIGN

3.4.1 RELATED WORK

As stated in Chapter 2, the processing of social media at scale is reliant on a well-managed coordination, communication, and processing infrastructure. Several big data analysis platforms and toolkits were utilised among the studies encountered whilst performing the literature review.

The GATE [13] text processing framework (Cunningham, 2002) is platform for Natural Language Processing and was utilised by Maynard et al. (2017) to build a system to monitor the UK 2015 election. The IBM InfoSphere [14] platform supports high performance stream processing, supporting structured as well as unstructured data stream processing and can be scaled to a large number of compute nodes (Biem et al., 2010). Cavalin et al. (2014) build their system on this platform, performing sentiment analysis of streaming Twitter data during a football tournament.

The Apache Software Foundation [15] (ASF) hosts many software suites that have proved useful to researchers building social media processing pipelines. Hadoop [16] is a highly popular distributed computing framework that supports scalable processing of big data via MapReduce (Ghazi and Gangodkar, 2015). One of the pillars of Hadoop is the Hadoop Distributed File System (HDFS), which was used in a number of studies to provide high-throughput access to their collected social media corpora (Mane et al., 2014, Ayvaz and Shiha, 2018, Talhaoui et al., 2018).

Trupthi et al. (2017) also employ the computational components of Hadoop, electing to store data in a NoSQL (Not only Structured Query Language) database called MongoDB [17], which is also designed for MapReduce operation. MongoDB was also used as datastore in (Angaramo and Rossi, 2018), and (Karanasou et al., 2016) who used Apache Storm [18] as a framework to develop their processing nodes.

Apache Storm is a distributed real-time computation system, and alternative to Hadoop, that is aimed at providing a scalable solution to processing streamed data rather than batch data (da Silva Morais, 2015). It was used as a core platform in several applications (Rahnama, 2014, Karanasou et al., 2016, Middleton and Krivcovs, 2016). A more recent alternative to Hadoop's MapReduce

[13] https://gate.ac.uk/
[14] https://www.ibm.com/uk-en/analytics/information-server
[15] https://www.apache.org/
[16] https://hadoop.apache.org/
[17] https://www.mongodb.com/
[18] https://storm.apache.org/

solution is Apache Spark [19], which provides an in-memory MapReduce solution. It was used by Ayvaz and Shiha (2018) and by Das et al. (2018), acting as a core processing component, both sitting on top of HDFS.

Das et al. (2018) incorporated another ASF project into their system, with Apache Flume [20] being used to move collected data into HDFS ahead of processing. In more loosely coupled systems, publish/subscribe platforms provide a scalable solution to diverse systems, where the ability to route messages to multiple consumers is vital (Dobbelaere and Esmaili, 2017). Apache Kafka [21] was used by Șerban et al. (2019) as the message broker within their system, moving data from an Apache Lucene based search engine called ElasticSearch [22] to process nodes and then on into PostgreSQL [23] databases afterwards. Middleton and Krivcovs (2016) employ RabbitMQ [24] to link Storm topologies together.

3.4.2 Requirements

Through consultation with end users and research partners, we developed four key requirements that drove the development of the Sentinel pipeline (Preece et al., 2018). These are presented in Table 3.4 with an explanation of the key motivators for each requirement.

	Requirement	Motivation
R1	Scalability	The system needs to be scalable to allow for increase in data flow, be that in short, spiked increases, or chronic growth over the course of a project.
R2	Reconfigurable	End user needs change over time, hence the system needs to be open to new development that complements existing work.
R3	Data Agnostic	New social media platforms emerge constantly, and the nature of the disinformation space means that features, analyses, and inferences that are made using any data are consistent. This requires a consistent document type, or at least the flexibility to support the generalisation of data for mass and cross-platform analysis.
R4	Agile	Development needs to be reactive to the collaborative environment and provide minimal barriers to the development of new features and services.

Table 3.4: System Requirements.

[19] https://spark.apache.org/
[20] https://flume.apache.org/
[21] https://kafka.apache.org/
[22] https://www.elastic.co/
[23] https://www.postgresql.org/
[24] https://www.rabbitmq.com/

3.4.3 Design Principles and Core Infrastructure

To address the requirements defined above, three key design principles were produced, and existing technologies that would allow the rapid development of a backbone architecture. These principles are defined in Table 3.5 and the core technologies are then discussed below.

	Principle	Requirement Addressed	Existing Technology
P1	Modular components	R1, R2, R4	Java, Python, TrianaCloud, WZeroRPC, Docker
P2	Lightweight messaging	R1, R2, R4	RabbitMQ
P3	Document-based storage	R1, R3, R4	MongoDB

Table 3.5: Design Principles.

3.4.3.1 Modular Components

Modularity is a common system design that breaks elements of a system down into manageable stateless modules that can be maintained and plugged in when and where needed. This allows the system to be scaled-to-need through the duplication of modules, and to be reconfigured through addition and removal of modules. Additionally, this allows for agile development of the system where new modules can be introduced and tested without interfering with the overall system workflow.

Throughout the lifetime of this project, different technologies were utilised for different iterations of the Sentinel pipeline. A prototype implementation was performed using TrianaCloud (Rogers et al., 2013) to manage a series of Java applets that performed the core collection, filtering, parsing and aggregation of data.

Alpha versions of these same pipeline modules were then developed to form the core *SentinelStream* Java components, designed to be deployable within self-contained virtual machine images known as Docker containers, and managed by a web-based workflow orchestration and execution environment, called WZeroRPC (Gesing et al., 2014). This was in order to create a dynamically reconfigurable application environment that adapted to the processing requirements of the pipeline (Evans et al., 2015).

In subsequent iterations of the Sentinel pipeline development of *dynamic* scalability was halted to focus on other elements discussed in this thesis; with the modules developed in the Alpha phase maintained but deployed outside of the Docker/WZeroRPC environment.

3.4.3.2 Lightweight Messaging

The lightweight message passing principle focuses on allowing more flexible communication between modules to occur within a dynamic decoupled environment (Dobbelaere and Esmaili,

2017). Within the Sentinel pipeline, this is provided by RabbitMQ [25] which is an implementation of the Asynchronous Message Passing Queue [26] (AMPQ) middleware standard. AMPQ is a networking protocol that allows for the passing of plain text messaging between agents, through a message broker.

The AMPQ model cosists of messages, exchanges and queues. Messages are published to exchanges which then distribute message copies to queues dependant of a set of routing rules that are defined when queues and exchanges are connected. Messages are then delivered to consumers subscribed to queues. When publishing a message to an exchange, message attributes may be attached, which correspond to the routing rules between queues and exchanges.

Messaging within the Sentinel pipeline comprises of JSON objects. A mix of direct queues and exchanges have been utilised. RabbitMQ provides interface implementations in different programming languages, which helps addressing the requirement for agile and modular development. AMQP also allows for the development of resilient systems as queues and exchanges can be set to a persistent state, whereby messages are maintained in system memory until a consumer becomes available (John and Liu, 2017).

3.4.3.3 Document-Based Storage

Due to the high volume and velocity of data ingested by the Sentinel pipeline, the third design principle is focused on the use of a document-based NoSQL database to store social media content. As observed in the wider systematic review studies, NoSQL database systems are emerging as an alternative to traditional relational databases, with horizontal scalability and structural flexibility being a priority, as a solution to big data management.

A document-based database is a type of NoSQL database where data are stored in a standardised document format such as XML, PDF or JSON. These structured documents are similar to records in relational databases, but with greater flexibility as each document may have similar as well as dissimilar data, due to the system's schema less nature (Nayak et al., 2013). This is crucial in supporting the requirement that the Sentinel pipeline be data agnostic, as it means any form of social media comment or post can be stored as JSON within the NoSQL database.

MongoDB is an implementation of the document-based NoSQL database management system offering an array of desirable features including geospatial processing, MapReduce, indexing, JavaScript-based querying, and driver support for many programming languages including Java,

[25] https://www.rabbitmq.com
[26] http://www.amqp.org/

Python, C++ and Ruby, the first two of which are used within components found in the Sentinel Platform. It is suitable for applications that require auto-sharding and high horizontal scalability (Gudivada et al.). Much like RabbitMQ, support of multiple programming languages and the low-maintenance nature of MongoDB helps to support agile development of new modules and features.

3.4.3.4 Data Collection and Management

Data collection is organised and managed by means of semantic *channels* that are associated with one or more social media sources. Formally, a channel is defined as a stream of data associated with a specific topic described by a set of parameters and search terms, which express the user's information need. In practice, the choice of channel's parameters tend to undergo refinement using feedback over a project's lifetime (Preece et al., 2018).

3.4.3.4.1 Twitter

The Search API [27] facilitates historical paginated searching of Twitter against recently published Tweets, with the service holding up to 7 days' worth of cached Tweets available to search against, limited to 180 queries per 15-minute window.

The Twitter Streaming API [28] allows collection of up to 1% of overall Twitter's throughput at any time. Rate limiting is handled within the streaming service, whereby message volumes exceeding 1% of current throughput are not served up by the service, without disrupting connection.

Framing Twitter stream into a channel is straightforward as the API's parameters allow for focused collection based on search terms, user IDs and/or geographical measures. Data are provided as a live stream of JSON objects, which simplifies the processing required to import the Tweet fully into the Sentinel pipeline as only a channel identifier and collection timestamp need to be appended to the JSON. Established wrappers available for both Java and Python (Twython) have been utilised by the Sentinel pipeline throughout its development.

3.4.3.4.2 YouTube

YouTube API services are provided via the Google Developers website [29]. Videos, channels, and playlists (known as resources) of interest can be retrieved, with query options allowing for keyword, locational, regional, or topical searches. Comments relating to videos, channels and playlists may be retrieved from a separate API endpoint with up to 100 comments retrieved per request, thus requiring paginated queries in order to retrieve a resource's full comment thread (Innes et al., 2017).

---

[27] https://developer.twitter.com/en/docs/tweets/search/overview
[28] https://developer.twitter.com/en/docs/tweets/filter-realtime/overview
[29] https://developers.google.com – Accessed 30/06/2015

YouTube's focal content are their videos, with comments associated with a particular video. It is possible to search video titles by keywords either across the entire catalogue or within a YouTube channel. Keyword searching is not available at the comment level, meaning that the context by which a YouTube comment falls within a *channel* is different to that of a Tweet. A Sentinel *channel's* YouTube content can be pre-framed by targeting a curated set of YouTube channels, centred around a particular topic or role.

Data are again returned in JSON format, requiring minimal transformation of fields in order to be further processed by the Sentinel pipeline. API access is limited to 10,000 *credits* per day; with a cost attributed to the core request and to additional metadata elements that can also be requested from the API endpoints. Within the Sentinel pipeline, API requests were made direct to the API endpoints and rate limit management was also handled with bespoke code.

3.4.3.4.3 Reddit

The Reddit API [30] provides access to user details, posts, comments, and live threads through OAuth protected endpoints. Much like the YouTube API and Twitter Search API, a limited number of results are returned per request and so pagination is required to access a full comment thread.

Reddit and its API are structured in the same way as YouTube, with comments bound to a focal piece of content, in this case posts and articles. Searches can be made across the entire site, or across a collection of subject based self-moderated communities commonly known as subreddits.

Like the YouTube API, keyword searching is available only at the post and article title level, not at the level of comments themselves, and so they share the same *channel* collection context as YouTube comments. We use the Python Reddit API Wrapper (PRAW) within our collection scripts, which automatically handles rate limiting and returns content in JSON format.

3.4.3.5 NOISE FILTERING

This module's primary function is the removal of significant volumes of documents with common phrases such as 'Happy birthday' or the names of celebrities which otherwise tend to dominate channels, especially the ones with a significant proportion of geospatially relevant documents.

Early versions of the filtering system were managed through a properties file, with a set of blacklisted terms. This was later moved to a MySQL managed data field. The addition of filtering via document language was also built into module as the wider research team focus shifted to a geopolitical domain.

[30] https://www.reddit.com/dev/api/

3.4.3.6 *TEXT NORMALISATION*

We observed in Chapter 2 that the normalisation of social media features present within texts was an important part of being able to successfully mine social media data. Several normalisation techniques were integrated into the Sentinel pipeline early on in development. A *normalised* version of the document text is added to the document data.

Regex	Purpose	Example
"?[rR][tT]:? @.*:? .*"?	Remove retweet text from quoted tweet.	**RT @XXXXX: "@YYYYY: TR actually talks sense, it's just all the nobs who tag along who let the EDL down"** this is exactly right
.*\bvia @\w+		**Around 40 EDL members run at police in Woolwich http://t.co/xxxxx via @itvnews** ---- and so it begins.
^(@\w+).*:.*[^\.](\.\.\.)		**@XXXXX: @YYYYY I respect your opinion on all things life but help me with this one....** The #uaf protest against the #edl fair enough
([#@]\w+ ?) {1,} [!:,.;]?$	Remove trailing hashtags and usernames.	RT @XXXXX: A rise in #Muslim attacks in the #UK. What's behind them? **#EDL #Islamophobia #Woolwich**
(?:@\w+\s*){2,}	Remove clustered usernames.	An absolutely fantastically written article **@XXXXX @YYYYY @ZZZZZ** sharp and to the point.
(https?|ftp|file)://[-a-zA-Z0-9+&@#/%?=~_|!:,.;]*[-a-zA-Z0-9+&@#/%=~_|]	Remove URLs.	RT @XXXXX: Fair point.... #woolwich **http://t.co/xxxxx**

Table 3.6: Tweet translation regular expression examples.

All hashtags were collected from the tweet data and translated into natural language versions of their text; breaking up the hashtag into tokens based on camel-case formatting (e.g., #simpleCaseOne becomes simple case one) using regular expressions. These are cached to support hashtag normalisation in other components of the system. Additionally, a series of regular expressions are applied to the tweet text in order to remove HTML and Twitter syntax/artefacts, examples are given in Table 3.6 with the affected text highlighted in bold.

3.5 System Evolution

The genesis of Sentinel, presented below in Table 3.7, highlights the four main evolutions of the pipeline infrastructure and analytical features present in each version of the system. These evolutionary steps reflect operational shifts within the wider research environment that Sentinel belongs to, highlighting the ability for Sentinel to adapt rapidly to meet end user needs.

	Prototype	Alpha	Beta	v1
Data Sources	Twitter Search API	Twitter Streaming API		
				YouTube API
				Reddit API
				MailOnline JSON
Query Management	SQLite	MySQL		
				Python
Data Collection	Java	SentinelStream		
				Python & SQLite
Processing	TrianaCloud	SentinelStream		
		Docker / WZeroRPC		
				Django
Messaging	RabbitMQ			
Storage	MongoDB			
Indexing			ElasticSearch	
User Interface	PHP	Flask	Django	
Features	FlexiTerm		SentiSum	FlexiTerm
		Sentiment		
		Ontology		
			NER	
			Anger	
			Search	
				Download
				Timelines
				Projects
				Unpacking
				ZeroDay
				Escalation
				Videos
				Clustering

Table 3.7: Infrastructure Evolution.

This Section covers the key development features of both the Prototype and Alpha versions of the system. Section 3.6 then presents the use cases and research focus that drove the development of

both the Prototype and the Alpha versions. The Beta version of the system is predominantly covered in Chapter 4, and Version 1 of the system is discussed throughout Chapter 5.

3.5.1 PROTOTYPE

Initial development was focused upon building a `proof of concept` system focused on the implementation of the core components, and the integration of these components with the supporting infrastructure described previously in Section 3.4.3.

3.5.1.1 CONFIGURATION

Initially, data collection was built on top of the Twitter Search API, with a paginated search developed in Java, utilising a set of SQLite databases to track collection progress across search terms and manage the rate limit restrictions of the API. Each SQLite database stored the channel-specific search terms along with the timestamp for the latest search performed on each term. This allows the search terms to be cycled through to ensure even term coverage within the channel. This comes at the cost of high-velocity terms, which may not be fully paginated over a 15-minute time frame when the number of tweets exceeds 18,000 per 15 minutes (i.e., 20 tweets/second). Collected tweets were inserted directly into MongoDB collections unique to each *channel* to ensure that all data are retained for further processing.

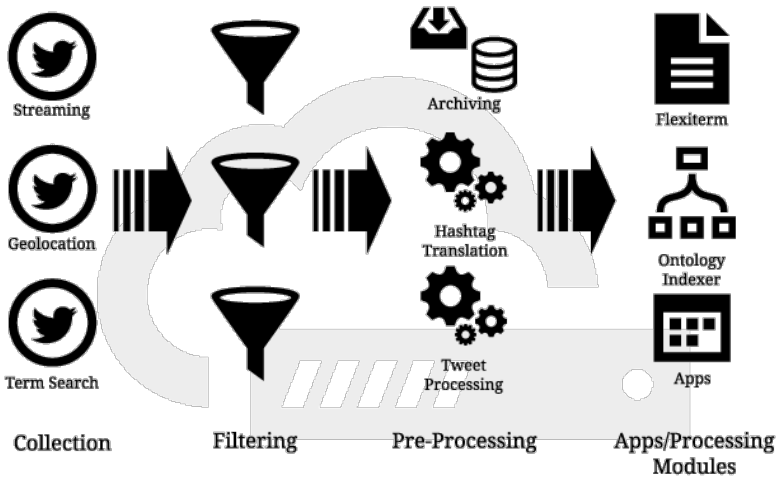


Figure 3.6: Prototype System Diagram.

The core components were developed in Java and deployed as executable JARs using TrainaCloud; a cloud based version of the Triana workflow engine (Taylor et al., 2007) that leverages the SHIWA Workflow Bundles (Rogers et al., 2013) as a means of packaging up data and executables to be passed to a pool of Triana instances distributed in a cloud environment through a RabbitMQ system (Figure 3.6).

3.5.1.2 *Feature Implementation - Bottom-up Data Interpretation*

Because of the nature of both the execution environment and the Twitter Search API, the initial text mining exercise was focused upon the use of a cross-corpus analysis tool as opposed to the use of any single-document analysis such as classification. It was, therefore, decided that an automatic term recognition algorithm called FlexiTerm (Spasić et al., 2013) would be used to derive noun-phrases present within 3-hour strata of collected data.

FlexiTerm is an open-source standalone automatic term recognition (ATR) tool that differs from other ATR techniques by being token order agnostic, treating the tokens as a bag-of-words when considering term candidates. This results in a greater amount of flexibility to term candidate comparison via approximate token matching based on lexical and phonetic similarity, indicating both semantic relationships and equivalence (Spasić et al., 2013). This approach is well suited to web data, where less formal language is used, and frequent spelling errors are encountered.

We frame this information derived directly from the data as *bottom-up* information (Pirolli and Card, 2005), as the information is derived from the corpora in an unsupervised fashion without any injection of domain knowledge. Noun phrases are often better contextualised and therefore more informative than Twitter's trending topics (e.g., the term 'armed police' rather than the words 'armed' or 'police') as well as being generally more relevant to a particular area of interest as the terms that form a *channel* intrinsically build context around the phrase (Preece et al., 2018).

3.5.1.3 *User Interface*

The first version of the Sentinel Interface was considered a priority during the first part of this project, with a proof of concept achieved using the Lee Rigby case study data (discussed in Section 3.6.1), developed in PHP, with a JQuery supported front end. Figure 3.7 presents the original view for the Sentinel Interface.

The app presents a timeline of channels belonging to a selected project, with document frequencies recorded for every three-hour period, resulting in 9 time periods being displayed on screen. Below the frequency graph is a summary of the top FlexiTerm terms for each time period, which are used to illustrate the key concepts discussed within the project at that time.

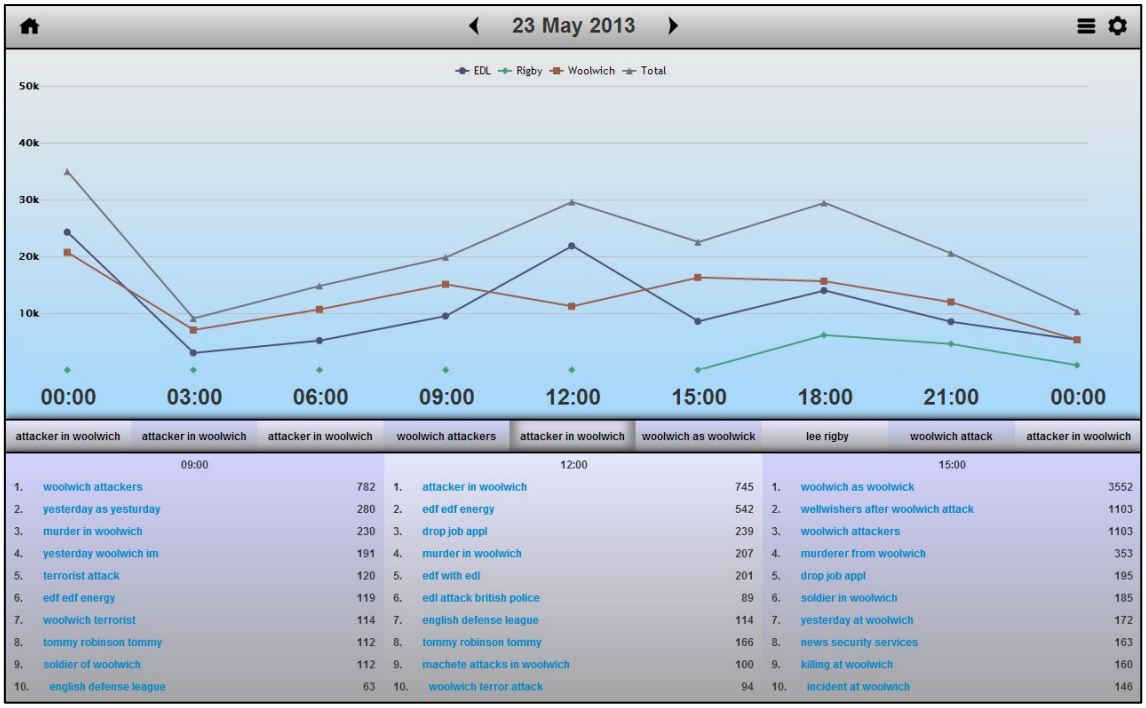


Figure 3.7: Sentinel Prototype Interface.

Only three lists are ever shown at once, with the user having the ability to horizontally scroll through these term lists, or just to a specific term list by clicking the 'top term' shown below each time value.

During the FlexiTerm run, frequencies for each term's occurrence are recorded. All term frequencies in a FlexiTerm list are checked against the preceding FlexiTerm list, in order to give the term a status of *new, rising, falling* or *non-moving* Users can filter the terms by status in order to get an impression of the emerging and fading topics within the project.

Clicking on a term will present the user with a sample of tweets containing the term, providing the user with a quick means of understanding the context in which the term is being mentioned. A further option is then presented to the user that allows them to plot the term's frequency on the project timeline graph, providing them with a different means of observing a term's lifespan within a project.

3.5.2 Alpha System

Development of the Alpha system shifted the focus towards building and testing the streaming abilities of the Sentinel pipeline, allowing the collection and processing to be performed in real-time. Building upon the *bottom-up* batch processing that was available in the Prototype system; effort was now focused upon enriching streamed data with further out-of-the-box *bottom-up* analysis and also with *top-down* domain specific knowledge through the use of Ontologies.

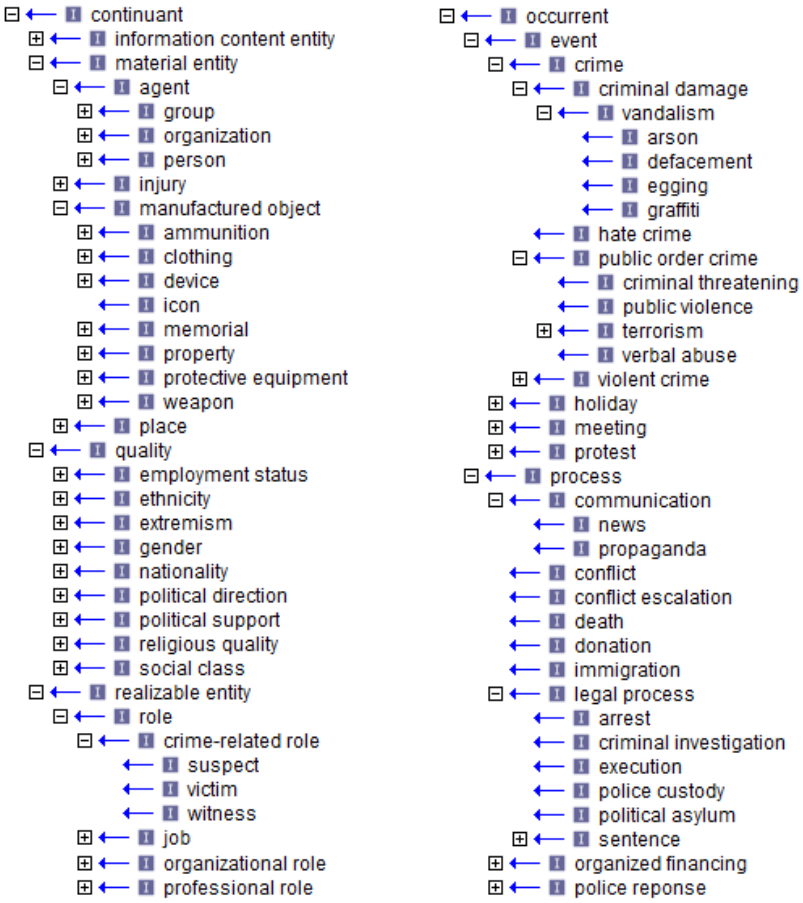


Figure 3.8: Partially expanded Sentinel ontology.

Ontologies are an explicit conceptualisation of a domain through a set of concepts definitions and their relationships (Uschold, 1996). They provide a standardised means of understanding between human and machine, facilitating information exchange (Altman et al., 1999). A bespoke ontology consisting of 479 concepts (with 389 synonyms) related to the domain of crime was developed in collaboration with social scientists for use within Sentinel (see Figure 3.8). The ontology is structured using the Basic Formal Ontology (BFO) (Arp et al., 2015), a small upper-level ontology designed for easy integration with other BFO-compatible ontologies. This decision was made with a long-term view of easy integration with other domains from which existing ontologies can be readily incorporated into the Sentinel pipeline.

3.5.2.1 CONFIGURATION

The data and *channel* term management interfaces were reimplemented using a lightweight web-app development framework called Flask [31], with the back-end management database now moved into MySQL in order to make the database accessible to any remote pipeline components. The

[31] https://flask.palletsprojects.com/en/1.1.x/

integration of the term management into the main interface allows the pipeline to be responsive to users' needs, with the ability to update *channels* collection *on the fly*.

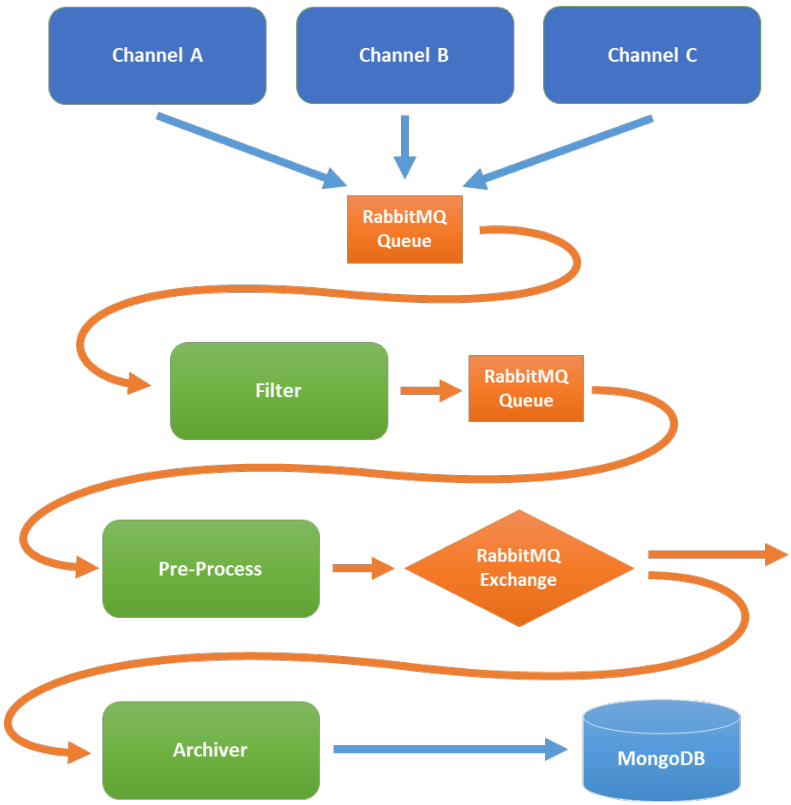


Figure 3.9: Sentinel Streaming Pipeline – Collection, Filtering and Parsing.

To facilitate real-time analysis, the second collection module was developed to incorporate the Twitter Streaming API. Along with this collection module, the core processing modules were reconfigured for deployment as executable JARs within Docker containers in order to support horizontal scaling of the system (Evans et al., 2015).

Figure 3.9 shows an infrastructure diagram for the core collection, parsing and archiving elements of the Sentinel pipeline developed in the Alpha version of the system. A number of RabbitMQ queues act as points of horizontal scalability and content is stored in a series of MongoDB collections. After text normalisation, documents are passed to an exchange that duplicates messaging so that documents can be archived in parallel with further batched data analysis with FlexiTerm described in the following section.

3.5.2.2 Feature Implementation – Real-Time FlexiTerm

In order to support real-time analysis using FlexiTerm, an aggregator module was developed. These aggregators are capable of binding onto the RabbitMQ exchange and funnelling documents into temporary SQLite databases used by FlexiTerm to perform automatic term recognition. The aggregators are controlled by a parent module AggegatorManager, which allows aggregations to be

run concurrently and to a fixed schedule for all aggregators that to belong to a single collection. By providing the frequency, granularity, and buffer values; the aggregation of documents and delivery of corpora to a pool of FlexiTerm jobs is automated, as shown in Figure 3.10.

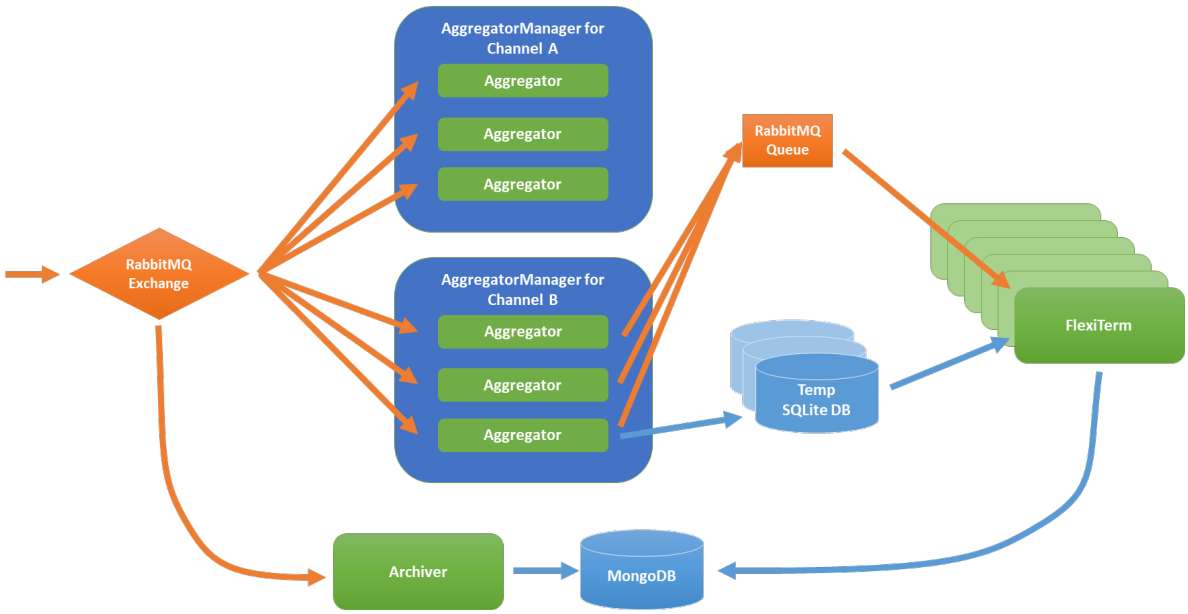


Figure 3.10: AggregatorManager and FlexiTerm Pool.

The FlexiTerm module has also been augmented by the inclusion of the Sentiment annotator packaged within the Stanford CoreNLP (Manning et al., 2014), which is used to gauge the sentiment of the collection of Tweets associated with a specific noun phrase set. This analysis is performed after FlexiTerm has run, with the resulting 5-point scale of sentiment being appended to each noun phrases' dictionary element within FlexiTerm's JSON output. The managers by default were set to collect in 60-minute windows, with a 5-minute 'buffer' either side of the hour to ensure the aggregator is pre-loaded ahead of collection and waits long enough for lagging documents to make it through the pipeline.

3.5.2.3 Ontology-based Concept Recognition

The injection of domain-specific knowledge was also incorporated into this iteration of the Sentinel pipeline, allowing for *top-down* interpretation of big data. This was achieved through the automatic marking of concepts from the *Sentinel Ontology* using a modified version of the PathNER tool that uses soft string matching to match a dictionary against free text (Wu et al., 2013). PathNER is built upon the GATE Embedded (Cunningham et al., 2013) text-engineering system, which allows for a lightweight version of the GATE system to be deployed as a toolkit for other applications to exploit. We built a Java module that hooks into the RabbitMQ to serve archived documents, pass them to an instance of PathNER and GATE, and then update the documents in MongoDB with the added ontology concepts by extending the *entities* property within document metadata.

3.5.2.4 *User Interface*

Building upon the original user interface developed in the Prototype, a geospatial view was developed to provide user with a map view of geotagged tweets. Both interfaces are overlaid with a time point selection bar to support temporal navigation through the data; with a finer grained 15-minute selector allowing the user to retrieve 15-minute blocks of documents from the database.

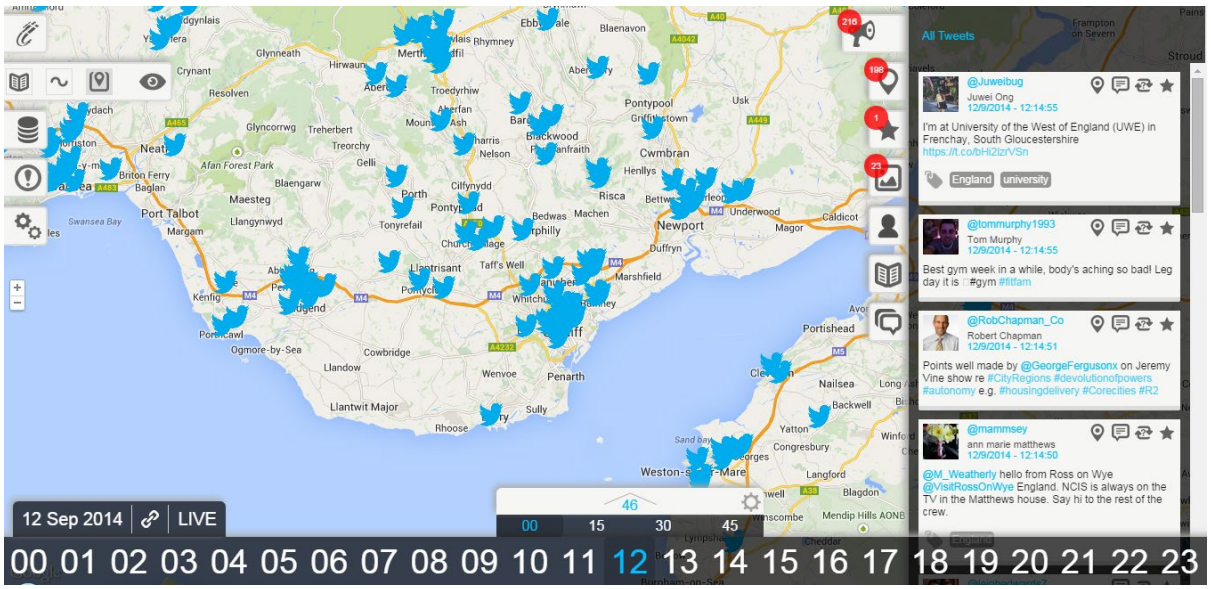


Figure 3.11: The Sentinel Web App geo spatial view, with *document drawer*.

A *document drawer* is also provided on the right-hand side of all views that by default provides a comprehensive list of all documents present within a 15-minute block. *M*atching concepts from the ontology are displayed for each document, along with flags to show if the document has a geocode or contains an image. The timeline can be also set to "live mode", whereby new documents are polled for every 15 seconds, and added to the interface. Figure 3.11 shows the d*ocument drawer* being actively used whilst in the geographic view.

A floating *FlexiTerm tab* is also present on the timeline, seen in Figure 3.12, which presents the FlexiTerm run for that hour. Each FlexiTerm is accompanied by the number of Tweets that mention that term, a colour representation of the Sentiment present within those tweets (green to red), an indicator of the FlexiTerms change in rank relative to the previous hour.

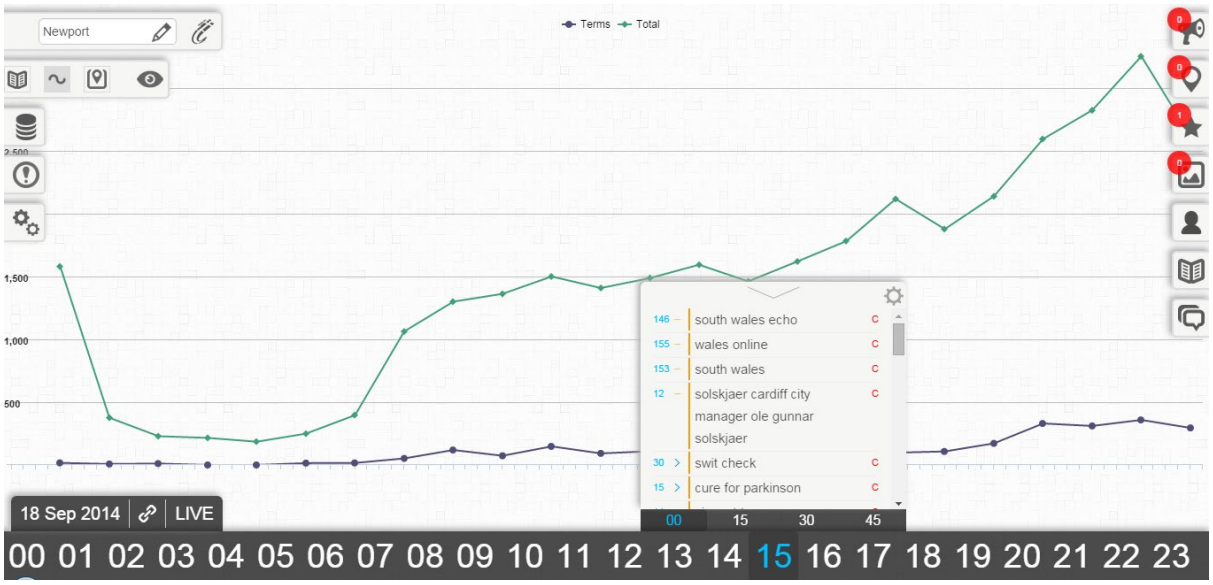


Figure 3.12: The Sentinel Web App timeline view, with *FlexiTerm tab*.

Clicking on a FlexiTerm filters the *document drawer* to show only the documents that created the FlexiTerm, as well as a historic poll for documents belonging to past FlexiTerms which share a common variant. In timeline view, as before, the FlexiTerm frequency is plotted over the day, and in the geo spatial view geotagged documents are highlighted in red.

3.6 Case Studies

3.6.1 Prototype: The Murder of Lee Rigby

An early stress test of the prototype system came in the form of the spontaneous events surrounding the murder of Fusilier Lee Rigby in Woolwich [32]. The incident took place on the 22nd of May 2013 and sparked worldwide attention due to the gruesome nature of the attack. A key aspect of this event was the early escalation of events by the Right-Wing group the English Defence League (EDL) [33]. Collection was initiated early in the event timeline, and grew significantly in the following hours and days, eventually running for 10 months from the time of initial incident, through to the trial and conviction of the perpetrators.

Three channels were created during this study through close consultation with users via the co-design process that were refined down from the running *channels*; the "EDL" *channel*, which consists of any tweet collected that matches the case insensitive regular expression "\b#?edl\b", the "Woolwich" *channel*, that matches any tweet containing the string "woolwich", and the "Rigby"

[32] http://www.bbc.co.uk/news/uk-22630303
[33] http://www.dailymail.co.uk/news/article-2329290/Woolwich-attack-More-100-English-Defence-League-supporters-gather-near-scene-killing.html

channel that matched any tweet containing "rigby". This gave us an overview of the public reaction to the event, the victim, and a specific insight into the EDL's activity during and following the incident.

3.6.1.1 OUTPUT

The data collected by Sentinel also allowed the research team to piece together the initial timeline of reporting of the incident, to authorities, on social media, and by Mainstream Media (Innes et al., 2014).

In order to identify the nature of the frequency spikes within the collected *channels*, the tweets belonging to the highest peaks of each day for *Woolwich* and *EDL* were analysed using FlexiTerm. Figure 3.13 shows the highest ranked FlexiTerm terms returned for each major three-hour spike, highlighting the main topic of discussion around the two channels at these peaks.

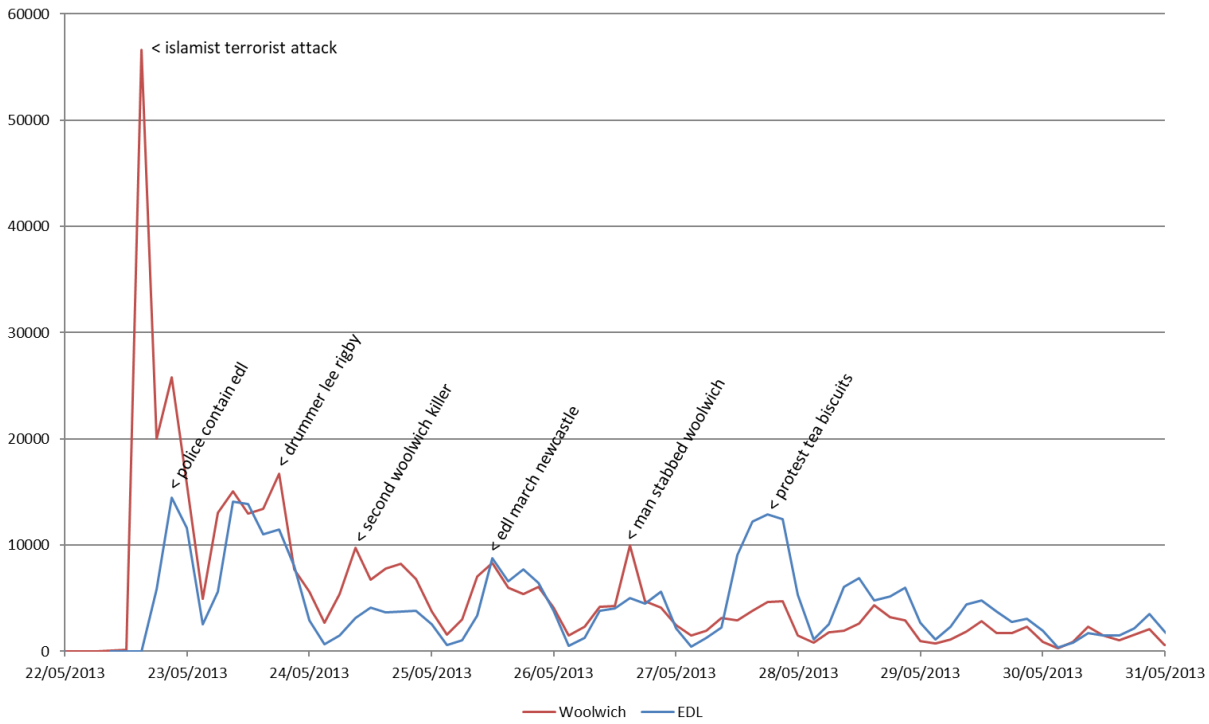


Figure 3.13: Top FlexiTerm term for major timeline spikes, 3-hour window.

This then drove the selection of candidate data sets for qualitative analysis, by ensuring that a broad range of topics could be quickly identified. These data were further reduced down through sampling across time-points to build a corpus of 17,000 tweets that were used to observe and characterise the travel of information through different forms of behavioural communication that occurs in the wake of a terrorist atrocity (Innes et al., 2018).

Figure 3.14 shows the daily frequency of tweets collected for *EDL*, *Rigby* and *Woolwich*. There was an initial spike in both *Woolwich* and *EDL* on the day following the murder (23rd of May), whilst *Rigby*

first appears the day after the incident once their name becomes public. *Woolwich then* shows a decline in activity, with the frequency of tweets first dropping below 10% of the initial spike on the 1st of June and never then returning above that threshold with the average percentage of initial spike from the 2nd of June onwards being 1.4%.

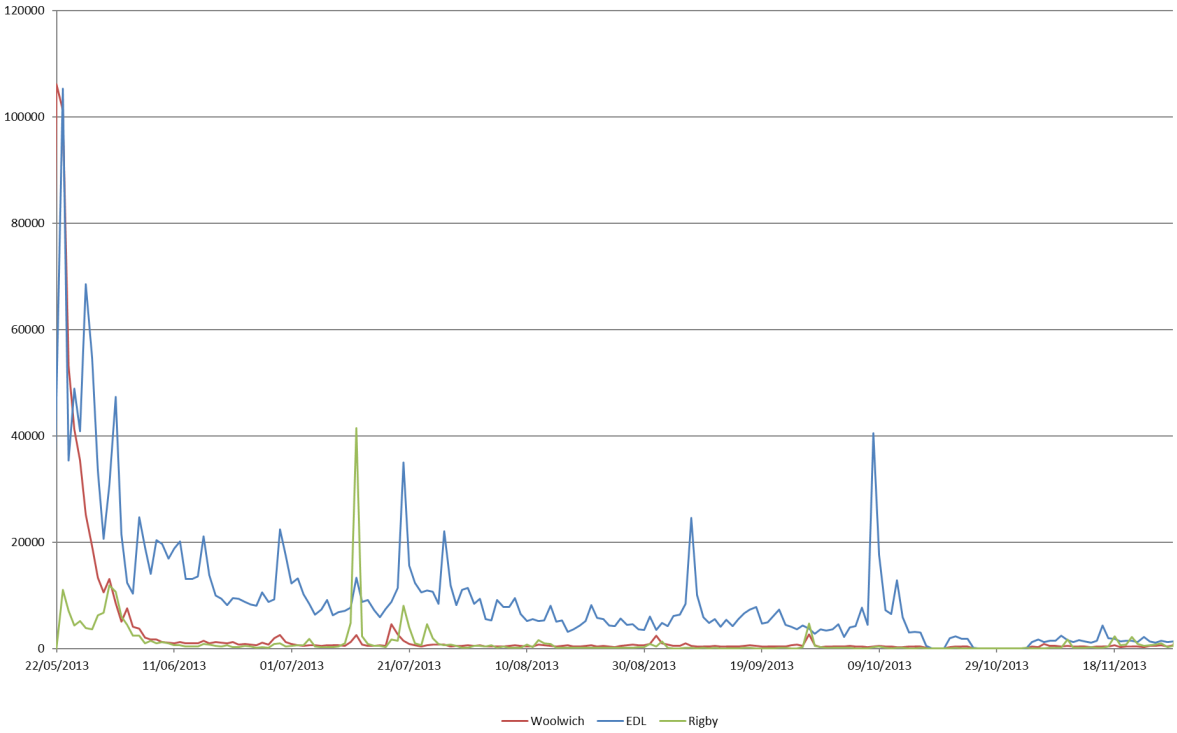


Figure 3.14: Daily frequency of tweets by search term.

Contrastingly *EDL* remains active past that date, dropping below 10% for the first time on the 18th of June but the frequency of tweets increases by approximately 10% on multiple occasions, the average percentage of initial spike from the 19th onwards being 10.1%. This quantitative analysis in addition to further qualitative analysis on the same corpus used in Innes et al. (2018) was used to test the C-escalation, D-escalation conflict dynamics theory (Collins, 2012), showing how "conflict talk" appears repeatedly through the corpus as a means of re-igniting tensions over a long period of time (Roberts et al., 2018).

3.6.2 ALPHA: 2014 NATO SUMMIT

The Alpha Sentinel Interface and Pipeline were incorporated into a field experiment performed by members of the Cardiff University School of Computer Science and Informatics, the University Police Science Institute, and colleagues from IBM Hursley, surrounding the 2014 NATO summit hosted in the cities of Newport and Cardiff. It ran for three months leading up to and including the Summit, to test the effectiveness of human-interaction methods and technologies for situational awareness.

A *channel* consisting of a geospatial search covering South Wales and South West England, along with search terms relating to protest groups and topics, was set up. These terms were combined with place names across South Wales that were anticipated as congregational point for both protestors and delegates. Additionally, a collaborative space was set up using the Slack [34] business communication platform that acted as both a means of tasking fieldworkers, and as a shared set of field notes and observations.

In the week of the Summit, a situation room was run serving as the coordination centre for researchers taking on the role of *field agents,* who performed face to face interviews with protesters, members of the public and Summit attendees. This allows for the research team to contrast qualitative findings against outputs from Sentinel, and to drive the development of new features for the Sentinel system in a series of "hackathon" sprints.

3.6.2.1 HACKATHON DEVELOPMENTS

Acting as part of the co-design process observed during planned events, a series of "hackathon" workshop sessions were undertaken in the three months preceding the start of the NATO Summit, the outputs of these are presented in this section. Firstly, a "live" version of FlexiTerm was produced for the NATO channel. This was achieved by instantiating a new Aggregator set with an hour's granularity, but with a 15-minute frequency as opposed to the normal hourly frequency. The intention was that we could identify emerging FlexiTerms much quicker through rolling updates.

Secondly, a module was developed that produces a timeline showing the sentiment of all FlexiTerms match particular search terms on a particular day. This was designed as a tool for identifying the particular mood from within the corpus relative to the search topic. Figure 3.15 shows the timeline for the search "summit", showing the exponential growth of FlexiTerms containing the word "summit" during the course of the event.

[34] https://slack.com/intl/en-gb/

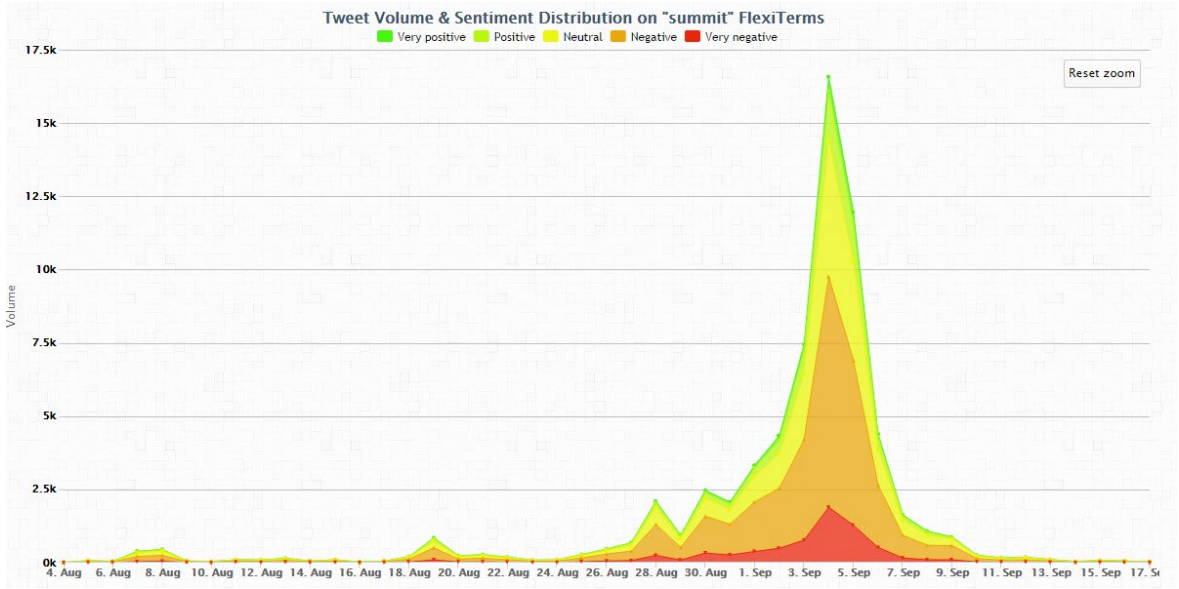


Figure 3.15: Sentiment timeline for "summit" as shown in the Sentinel interface.

3.6.2.2 OUTPUT

Sentiment scores from all tweets that constituted a FlexiTerm term (Figure 3.16) were compared against tweets belonging to FlexiTerms that contained the word “summit” (Figure 3.17). The daily scores were generated and used to derive the mean sentiment for each concept. This showed that overall, discussion around the NATO Summit was more negatively framed vs general conversation in the area (Table 3.8).

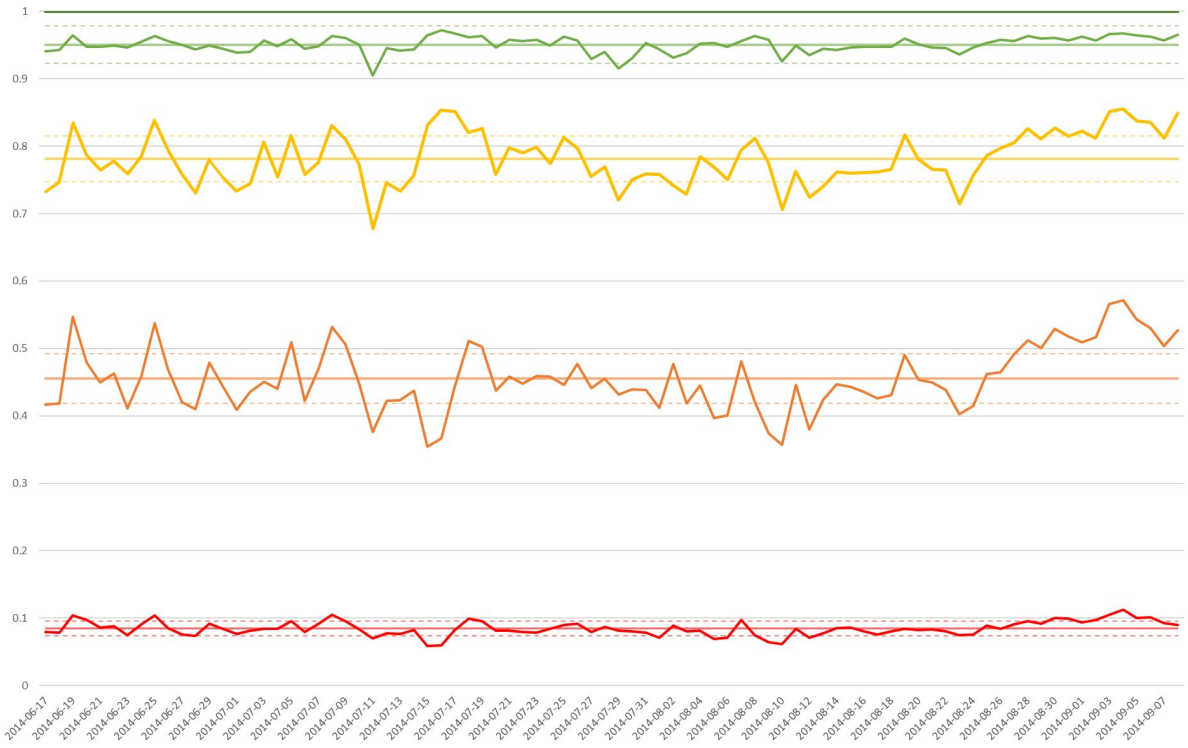


Figure 3.16: Stacked Sentiment for all FlexiTerms in NATO Summit project.

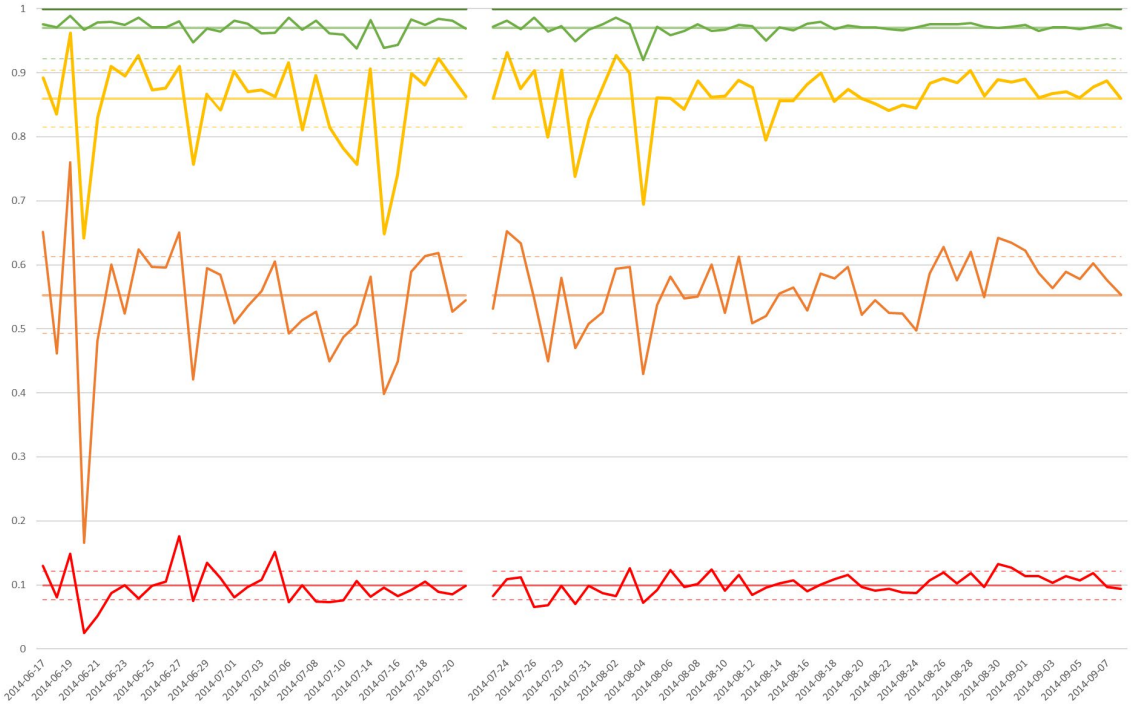


Figure 3.17: Stacked Sentiment for "Summit" FlexiTerms in NATO Summit project

A major finding from the exercise was that there was a distinct and observable difference in public perception between the physical presence of security measures in Cardiff city centre, such as the "ring-of-steel" and armed police, and the military showcase of ships and a fly-by that occurred a mile away in Cardiff Bay (Table 3.8).

FlexiTerms	Sentiment				
	V Negative	Negative	Neutral	Positive	V Positive
All	8.55%	37.15%	32.43%	16.94%	4.94%
"Summit"	10.09%	45.84%	30.19%	10.88%	2.99%
City centre					
"Armed Police"	10.12%	46.98%	28.60%	11.43%	2.87%
"Ring of Steel"	10.57%	48.51%	29.19%	9.08%	2.66%
Cardiff Bay					
"HMS Duncan"	8.50%	39.12%	31.85%	15.88%	4.65%
"Red Arrows"	6.63%	33.83%	36.42%	18.78%	4.34%

Table 3.8: Sentiment Scores for FlexiTerms.

Finally, a third party integration was undertaken incorporating a "chatbot" interface into the Sentinel interface called MOIRA (Preece et al., 2014), that uses data collected and enriched by the Sentinel pipeline and a *controlled natural language* (Braines et al., 2013) to support the research team in incorporating social media data into the sensemaking process (Preece et al., 2016).

Findings from the case study were fed back to the College of Policing and South Wales Police through a series of commissioned reports and presentations [35].

3.7 Discussion & Conclusions

This chapter focus primarily on the decisions made when initially designing the Sentinel pipeline and how to best position the architecture to support the ever-evolving needs of the end users. Because the pipeline is acting as the groundworks for qualitative and quantitative cross-discipline research, it was important for flexibility of resource and interpretability of results to be at the forefront of its design. The adoption of the event driven co-design lifecycle by the stakeholder and development team provided a methodical approach to introducing new interfaces, tools, and research foci. We have presented a high-level overview of the pipeline evolution that has taken place in the duration of the thesis, before focusing on the prototype and alpha versions of the system. The use cases covered in this chapter show how the Sentinel pipeline can support research into both spontaneous and planned events, and how those research needs also drive the co-design of the system.

This chapter's use cases evidence that the Sentinel pipeline can fulfil the objectives set out at the commencement of this work; to create an open platform that allows social and computing scientists to co-design useful analytic components and apps, able to semantically-enrich social media data in both a bottom-up and top-down manner. Furthermore, this work has fed into a number of other publications produced in collaboration with the research team both in the field of computer science (Evans et al., 2015, Preece et al., 2016) and social science (Innes et al., 2014, Innes et al., 2018, Roberts et al., 2018). It has also served as a platform for external collaborators to conduct research on (Preece et al., 2015).

The prototype version of the pipeline was never intended to be employed in a use case scenario, as it was simply meant to demonstrate that the core infrastructure operated in concert correctly. The Murder of Lee Rigby occurred early on in the system's development, and it was immediately clear that this tragic event was to have a serious and consequential effect on how discourse on social media would evolve, and so the prototype was maintained in an operative role. This is a direct benefit provided by the co-design process. The relationship built between stakeholders and developers allowed for stakeholders direct the rapid deployment of the prototype collection process in support of this event with confidence that it was capable of supporting their own digital fieldwork. This was a highly emotive subject that rapidly engaged a global audience and remained engaging for

[35] https://www.cardiff.ac.uk/research/impact-and-innovation/research-impact/using-social-media-to-manage-large-scale-events

an extended period of time, with a rise in reports of hate crimes against Muslims, and several attacks on mosques across the UK.

The collection of this rich and emotive data set is evidence that the system was already capable of proving valuable in identifying user-generated content for further qualitative analysis. The data driven *bottom-up* analysis performed using Sentinel's implementation of FlexiTerm proved key in producing effective sampling across the topic for use in (Roberts et al., 2015), Innes et al. (2018) and Roberts et al. (2018). Findings from these studies have been fed back into the Sentinel platform as part of the first *data interrogation* and *case study* activities, namely emotion classification, which are addressed in Chapter 4.

The data driving the Rigby use case was well suited to running the task at the 3-hour granularity level in the first few weeks, as the volume was high, and the topic dominated both social and mainstream media. This meant that key words such "woolwich", "attack" and "edl" were contextually closely linked to the incident, and chances of noise being introduced into the data were lower than small scale event, but even so we observed through qualitative coding that approximately 20% of the Twitter messages referred to an incident relating to the 2013 Boston Marathon Bombing (Innes et al., 2018). This was due to "shot" forming part of the collection terms, which covered news emerging from the US about a suspect related to the bombing being shot and killed by FBI agents [36].

This was taken into consideration during the planning workshops for the NATO Summit case study, with the decision made to ensure that search terms were contextualised geographically in order to reduce irrelevant content. This increased the difficulty for Sentinel to operate in a "monitoring" capacity, as the summit was relatively low velocity in comparison to the Rigby case. Nonetheless events were identifiable across the course of the study, and the integration of closed-loop tasking of field teams to obtain ground truth on events detected proved effective and valuable in situational understanding (Preece et al., 2018).

[36] https://www.theatlantic.com/national/archive/2013/05/fbi-shooting-orlando-boston-marathon/314979/

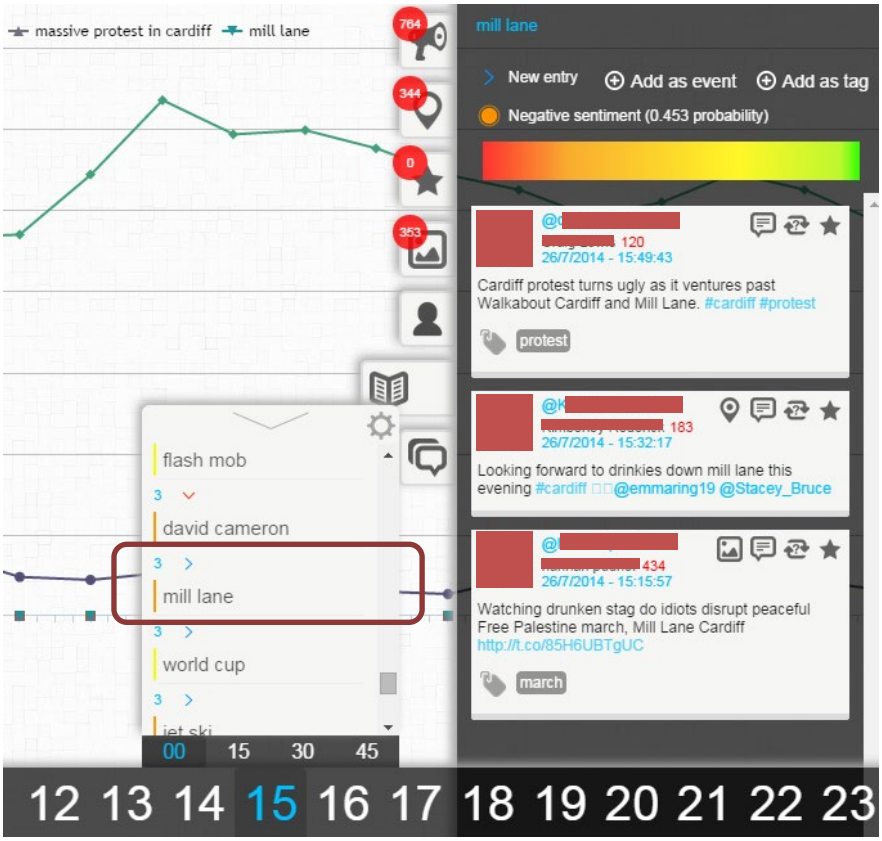


Figure 3.18: Mill Lane FlexiTerm

Preece et al. (2015) highlight a protest on the 26th of July that resulted in a confrontation between protestors and revellers at Mill Lane in Cardiff City Centre, discussing how we tasked field agents and fed information via a Controlled Natural Language, called Controlled English (Mott, 2010), into a central knowledgebase that could be interrogated through a chatbot. This event was captured within Sentinel, but as Figure 3.18 shows, this was somewhat of a serendipitous result; as FlexiTerm was tuned to require a minimum of three matched documents within a corpus before the noun phrase is recorded, and one of the constituent Tweets for the "Mill Lane" FlexiTerm was completely unrelated to the incident. This shows that the NATO Summit study was pushing the limits of data sparsity in terms of bottom-up analysis within a monitored *channel,* and that there are challenges in pulling out information when it is fixed to a single geospatial event. This, along with the dwindling number of geocoded tweets, drove us on to look at how Sentinel as a platform can be reconfigured to provide insights into more geographically disparate data.

These observations on channel framing are applicable to any data collection exercise performed with social media. Lexical disambiguation of polysemous words is key when building a set of collection terms that define the scope of a dataset. When discourse surrounds an emerging event this happens naturally, due to the likelihood of a term sharing the same sense within a single discourse (Gale et al., 1992), but the challenge is in ensuring that a collection is focused upon a single event. Case study of the Woolwich terror attack led to the identification of a number of

audience behaviours present in the immediate aftermath of an atrocity (Innes et al., 2018) two of which, “reporting” and “requesting”, are key in the audience establishing a single discourse around an event. Rapid deployment of a virtual *situation room* allows for the qualitative identification of content expressing these behaviours, which can be fed back into channel curation for corpus development.

CHAPTER 4: DATA ENRICHMENT
SEMANTIC SEARCH

4.1 INTRODUCTION

Chapter 3 presents the early developments and considerations relating to the Sentinel Pipeline and overall platform. Up to this point, Sentinel was considered a tool that was to act in a "scanning" and "monitoring" fashion, with interfaces built that allow for horizon scanning within events. The ability to apply top-down knowledge driven analysis was tested, but we found there were limitations to how well these data-driven approaches can perform with relatively sparse or loosely contextualised data.

The collection and processing of open source data to derive higher-level information products beyond data collection is fundamentally a Human-Computer Collaboration (HCC) and cannot be fully automated (Preece et al., 2016). To this end, this chapter is focused upon data enrichment in the Sentinel Pipeline with top-down knowledge: both through out-of-the-box solutions and via internally developed classification tools whose features are derived from findings in the Woolwich use-case presented in Chapter 3 and published in (Roberts et al., 2018).

This results in the development of the Semantic Search interface that highlights how this knowledge and enriched data can then be presented to the users to help them easier identification of key content within channels and corpora. We present results of an analysis of usage and user experience, derived from a usability survey and system access logs.

4.2 PRIMARY OUTPUT: SENTINEL PIPELINE BETA

In order to support the components that are to be presented within this chapter, changes to the Sentinel pipeline infrastructure were made. These form part of the beta version as shown in Table 3.7: Infrastructure Evolution. from the previous chapter. These adaptations allow for greater flexibility of the pipeline on order to support rapid interrogation of data through ElasticSearch (Section 4.2.1) and re-configuration of the pipeline via by an improved user interface, OSCAR Hub (Section 4.2.2).

4.2.1 SEARCH ENGINE

Throughout the Prototype and Alpha version of the Sentinel pipeline, searches were performed directly against the MongoDB instance, which over time slows down due to the increased volume of data (Abramova and Bernardino, 2013). In order to alleviate this scaling issue, we introduce an

ElasticSearch (ElasticSearch, 2020) service into the system architecture. This is a RESTful search and analytics platform built upon the Apache Lucene search engine, which provides rapid text-based searching in a distributed environment. Abubakar et al. (2014) show that ElasticSearch outperforms MongoDB in document searching when updating and reading records.

Key	Type	Description	Section
channel	text	The channel id document has been collected on.	3.5.1.1
id_str	text	The unique document ID.	
text	text	The full text of the document.	
created_at	date	Timestamp of document creation.	
screen_name	text	The username (and ID) of the author.	
lang	text	Document language field.	
article_title	text	Title of article document belongs to.	
article_body	text	Full text of article document belongs to.	
article_author_screen_name	text	Author username (and ID) of article document belongs to.	
entities	text	Flag list of all entities matched.	
location	text	Coordinates of document.	
ner	text	IDs of Named Entity Recognition matches.	4.3.2
ontology	text	IDs of Ontology matches.	3.5.2.3
emotion	text	IDs of Emotion word matches.	4.3.1
swearword	text	IDs of Swear word matches.	4.3.1
classification	text	IDs of Classifiers that have positively matched.	4.4.4
quote_id_str	text	ID of document quoted.	
quote_screen_name	text	Username (and ID) of author of document quoted.	
reply_id_str	text	ID of document replied to.	
reply_screen_name	text	Username (and ID) of author of document replied to.	
rt_id_str	text	ID of document retweeted.	
rt_screen_name	text	Username (and ID) of author of document retweeted.	

Table 4.1: ElasticSearch Keys.

We employ ElasticSearch as the document indexing recording tool, creating text-based indexes that our enrichment modules update, whilst maintaining MongoDB as a datastore that hosts the full document metadata of any collected data. The search in Elasticsearch is near real-time; documents are indexed immediately after they are successfully added to an index but will not appear in the search results until the index is refreshed. It is a good choice when near real-time searching that scales to terabytes of information is required (Kononenko et al., 2014).

Like MongoDB, ElasticSearch is supported by a number of language-specific interfaces that allow an ElasticSearch instance to be hooked into a system's logic. For our Java core components, both the *archiver* module and the document enrichment modules presented in Section 4.3 and Section 4.4, we produced an ElasticSearch management interface that would periodically insert and update records within ElasticSearch that ran concurrently with our MongoDB insertion logic.

Data within ElasticSearch are stored as JSON objects and recorded in a series of *indexes* defined by the administrator, these being collections of related documents. Each *index* allows for the addition of *key-value pairs* for each *document*. Table 4.1 presents the *keys* present within the ElasticSearch *indexes* used by Sentinel and where relevant, the section of this Chapter where the contents of this field are explained.

4.2.2 Web Framework

As discussed in Chapter 3, one of the key features of the Sentinel pipeline and system as a whole is the encouragement of pluggable interfaces being built on top of the pipeline. To this end, Django [37] was employed as a more robust web framework, as it allows for rapid, robust, modular development of interfaces to be built for the Sentinel system.

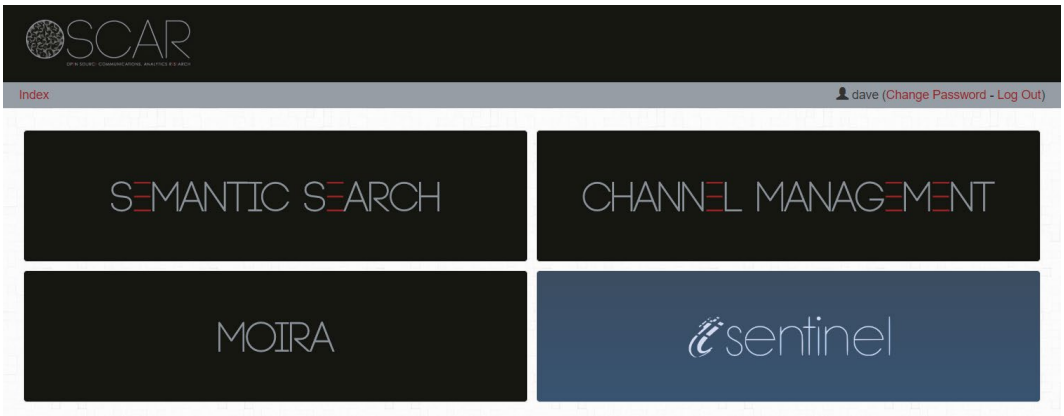


Figure 4.1: OSCAR Hub homepage as of Sentinel Beta.

Django loosely conforms to the Model View Controller (MVC) pattern, whereby the data structures, business logic, and user interfaces are maintained separately from one another within the codebase. Use of the MVC design pattern improves app development and maintenance as the "look" can be drastically changed without changing the underlying data structures and business logic, and different interfaces can be maintained easily (Leff and Rayfield, 2001).

[37] https://www.djangoproject.com/

Django adapts the MVC pattern, to take into account the nature of the HTTP protocol, by using the Model Template View (MTV) pattern consisting of a series of database model interfaces (*models.py* files), request-processing interfaces (*views.py* files), and a presentation layer (*HTML template* files) (Ravindran, 2015). These form the core elements of *modules* within Django, with a single application being made up of a number of *modules*.

The Open-Source Communication Analysis Research (OSCAR) Hub acts as a web portal into several decoupled interfaces that interact with the Sentinel pipeline and its data, providing a single point of access and single set of login credentials. Figure 4.1 presents the initial OSCAR Hub homepage, showing links to the Alpha Sentinel Interface (Section 3.5.2) and the MOIRA conversational tool (Preece et al., 2014), along with the newly developed Semantic Search Tool (Section 4.5) and the Channel Management Interface.

In order to improve the flexibility of any interfaces built within the OSCAR Hub suite, we integrated the Django REST Framework [38] into the OSCAR Hub project. The Django REST Framework provides an out-of-the box REST (Representational State Transfer) interface that maps directly onto any defined Django models and can also be extended to provide additional *logic control* elements.

The REST API was tested by deployment as an updated backend to the Alpha map-based Sentinel Interface, providing JSON serialised access to the documents stored in MongoDB via a datetime contexed endpoint.

4.3 Document Tagging

In Chapter 3 we introduced the Sentinel Ontology and how we utilised a modified version of the PathNER tool (Wu et al., 2013) to automatically tag these concepts with soft dictionary matching (Section 3.5.2.3). Within the Beta version of the Sentinel pipeline, we expand the scope of our document tagging capabilities with the introduction of swear and emotion words and swear word tagging (Section 4.3.1), Named Entity Recognition, and sentiment analysis (Section 4.3.2).

4.3.1 Dictionary Lookup

The PathNER instance is loaded with the latest version of the Sentinel Ontology in OBO format using the OBO Edit Java API (Day-Richter et al., 2007), with each concept represented with a unique numerical ID prefixed with *SENTINEL*. Along with this, we include all WordNetAffect (Strapparava et al., 2004) emotion words (prefix *EMO*) and a list of swear words taken from a freely available forum

[38] https://www.django-rest-framework.org/

swear filter list (von Ahn, 2009) (prefix *SWEAR*). PathNER uses the SoftTFIDF method from the SecondString string matching package (Cohen et al., 2003); a combination of the TFIDF weighting and Jaro-Winkler distance (Winkler, 1999) measure, accepting a match on a 95% similarity. The ease by which new ontologies and dictionaries can be plugged into the PathNER tool was a key reason for its inclusion within the pipeline.

Like with the ontology matches, we update the documents in Mongo with the added swear and emotion word matches by extending the *entities* property within Tweet metadata. The IDs of any matches are also recorded within the *emotion*, *swear word* and *ontology keys* of the relevant ElasticSearch *document*, and the contains *key* of the *document* is appended with "onto", "emo" and "swear" if any of the match types are present in the document text.

4.3.2 Linguistic Processing

A second document tagging module was also developed in order to exploit an established out-of-the-box Java based natural language processing (NLP) toolkit, Stanford CoreNLP, that provides a consistent API for a collection of common NLP techniques. The toolkit centres around a persistent *pipeline* class that is instantiated with a series of *annotators* that are able to perform NLP tasks, such as tokenisation and sentence splitting. More complex *annotators* that we use to tag documents include *annotators* focused on Part of Speech (PoS), Named Entity Recognition (NER), and Sentiment Analysis.

The Stanford PoS Tagger is an implementation of the log-linear part-of-speech algorithm (Toutanova et al., 2003) that acts as a core *annotator* within the pipeline. It uses the Penn Treebank tag set (Marcus et al., 1993) to construct a hierarchical grammar tree that other *annotators* within the configured pipeline are dependent upon [39]. Within Sentinel-specific *annotators,* we use the PoS Tagger output to identify any *personal pronouns* (I, we, you, etc.) present within the document texts. Steed et al. (Steed et al., 2015) chose to ignore personal pronouns when performing stop word removal from their corpus, and we agree that these stop words are cornerstone of emotive expressions.

We then pass the text to the NER *annotator* (Finkel et al., 2005) which uses the PoS tree and Conditional Random Field taggers trained on various corpora, such as ACE and MUC [40], to identify *people*, *places* and *organisations* present within the text. These are recorded within the *document* metadata using the corresponding IDs found within the Sentinel Ontology and indexed in the

[39] https://stanfordnlp.github.io/CoreNLP/dependencies.html
[40] https://stanfordnlp.github.io/CoreNLP/dependencies.html

relevant ElasticSearch entry via the *ontology key*. NER was used by Middleton and Krivcovs (Middleton and Krivcovs, 2016)to aid geo semantic feature extraction, allowing them to link data to geographical areas of interest.

Finally, we enrich the documents using the Sentiment Analysis tool (Socher et al., 2013) that uses Recursive Neural Networks (RNNs) to apply a 5-point sentiment probability score for each token within the PoS tree, and a cumulative score at each branch root all the way up to the full-text level (Socher et al., 2013). The use of the hierarchical tree allows for much more nuanced application of negation within sentence, where it can be focused purely on the noun phrase or statement that it collocates with. Sentiment analysis was a common NLP objective of the studies observed within Chapter 2 (Mane et al., 2014, Neuenschwander et al., 2014, Karanasou et al., 2016, Azzouza et al., 2017, Rezaei and Jalali, 2017, Michailidis et al., 2018) as a number of large training datasets such as the SemEval (Nakov et al., 2015, Nakov et al., 2016) datasets are commonly available and utilised.

4.3.3 Indexing

Each indexer is written as an instance of the Abstract class *IndexerBase* which provides and interface for the Indexers can implement. Each implementation of *IndexerBase* requires a number of initialisation methods that will create the necessary runtime environment within the machine that is running the process. A runtime environment is created on the fly to support the Indexers, in a temporary dictionary so that each individual process has its own environment, stopping cross contamination.

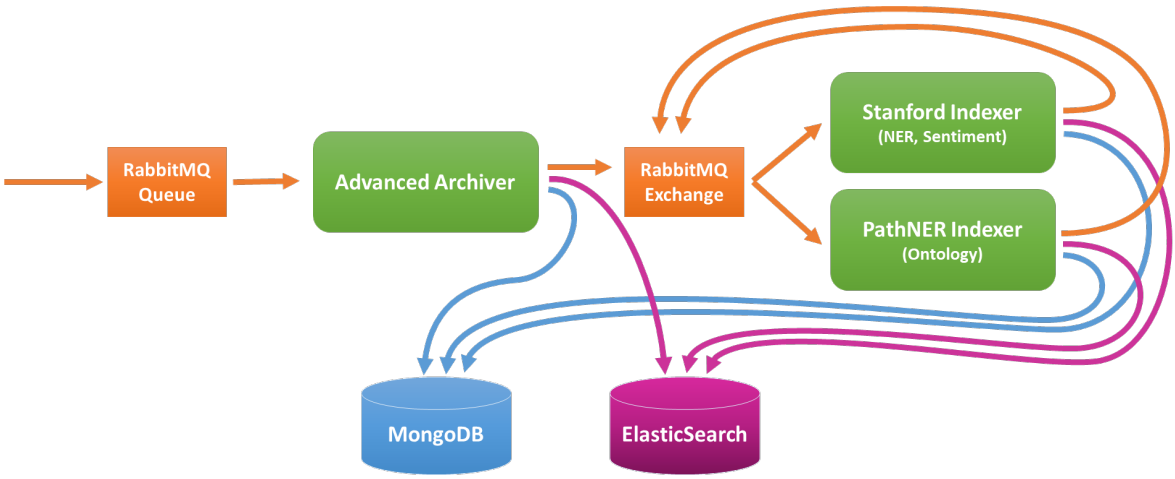


Figure 4.2: Indexing modules present within Sentinel Architecture

Indexers are managed by the *IndexerManager* class, where combinations of *IndexerBase* implementations are registered to. The *IndexerManager* acts as a broker to RabbitMQ's message queues, consuming documents and passing them on to its registered *IndexerBases*. We use a

persistent queue called *toindex* to consume all messages from the *archived* exchange using the routing key rule "index.*" (Figure 4.2).

4.4 Document Classification

The implementation of document tagging modules in the pipelines serves a primary function of enriching documents with relevant knowledge of the world. We can use these enriched data to inform and build our own machine learning tools that will allow for multi-faceted searching of the data, which we discuss in Section 4.5. This work was the first significant output of the co-design process, with the conflict score model being the result of a large series of *collaborative* sessions during *data interrogation activities* with stakeholders. This allowed us to then develop the concept further into a series of text classification tools that were incorporated into the processing pipeline.

4.4.1 Background - Conflict Score Model

As discussed in Chapter 3 an early case study of data obtained through the Sentinel Pipeline was used to test the C-escalation, D-escalation conflict dynamics theory (Collins, 2012) and to study Symbolic Social Interactionism; how "players" from within an event act off one another (Charon and Hall, 2009). To this end, an early classification scheme for modelling these two concepts was developed, and consisted of three main categories:

- *Routine* - the tone of the message being delivered in a tweet.
- *Dynamic* - the 'conflict talk' occurring within a Tweet, e.g., insulting the other side, boasting about one's own power, making threats or repeating previous perceived wrongdoing by the other side.
- *Conflict Action* - a weighted categorisation designed to provide a 'conflict score' to events being discussed within a tweet.

Sociological annotation requires a very deep and subjective exploration of the corpus, with classification groups being expanded upon when existing options did not apply to Tweets. This resulted in classification groups consisting of a broad range of sub-categories.

The development corpus was taken from the largest *EDL* channel spike on the first day (from 9pm to 12pm on the 22nd of May). It was during this time that the EDL arrived in Woolwich and began clashing with riot police. A 10% sample of the spike resulted in roughly 2500 tweets being selected. These were manually annotated, cyclically generating the categories found in Table 4.2, as well as two other minor categories; *Target Side* (the 'side' of the subject of the tweet) and *Stance* (the writer's level of support or opposition to the subject side).

Category	Options
Routine	*Rumouring; Broadcasting; Mobilising; Reacting; Reality Check; De-escalation Messaging; Eyewitness Narration; Government Statement*
Dynamic	*Threatening; Insulting Opponents; Criticising Opponents; Informing; Warning; Locating; Polarising; Sacrificing; Sympathising with Victim/Family; Reasoning; Calls for Action/Conflict; Boasting of Group Strength; Satisfaction at Side's Misfortune; Disassociating; Joke Making; Knowledge Seeking; Emotionally Stunned; Under Attack!*
Target Side	*Far Right Ex; Islamist Ex; Anti-Fascist; Moderate Muslim; Neutral; Mainstream News; SM News; All Extremists; Police; Government; Victim(s); Country; Anarchist; Religion; Joke Makers; Minorities; Eyewitness; Local Community*
Stance	*Hard Support; Soft Support; Neutral; Soft Opposed; Hard Opposed*

Table 4.2: Annotation categories from initial exercise.

Figure 4.3 shows the cross categorisation of the *Routine* and *Dynamics* of tweets. By looking at the tweet *Routine* with respect to the message *Dynamic*, it is possible to identify the primary nature of these *Routines* within this spike.

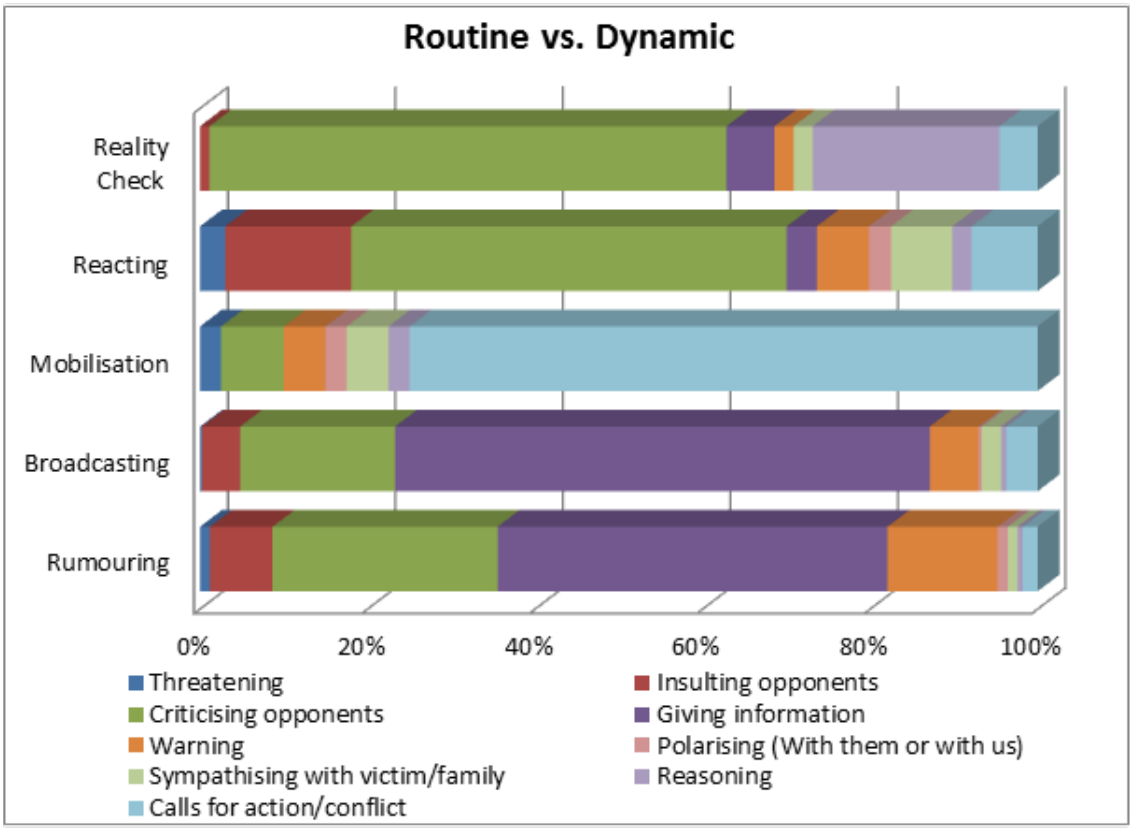


Figure 4.3: Routine vs. Dynamic categorisation.

Reality Checks primarily consist of criticism of opponents (61.7%) and reasoning with antagonists (22.2%). The two main *Dynamics* found in the *Reacting Routine* were criticising (51.9%) and insulting (15.0%) opponents. These two *Dynamics* were generally negative in their stance towards the target of their message. *Mobilisation* was the smallest of the major *Dynamics* (only 80 classifications) and was dominated by calls to action/conflict (75.0%). The major *Dynamic* of both broadcasting and rumouring was *information giving* (63.8% and 46.5% respectively).

Conflict Action	Score	Conflict Action	Score
Mass Killing	50	Policing - FC Action	2
Killing	10	Celebrating Anniversary	1
Physical Attack	7	Display of Patriotic Emblems	1
Arrests	5	Displays of Security	1
Arson/Damage Attacks	4	Donating	1
Clashing	4	Exploiting Events	1
Protesting	3	Hate Talk	1
Assembling	2	Membership Increasing	1
Deliberate Hoaxes	2	Silent Witness	1
Hate Graffiti	2	Spreading Inflammatory Rumours	1
Moving Towards	2		

Table 4.3: Conflict Action Categories.

The *Conflict Action* categories were identified during the qualitative analysis process by expert users, through the application of Grounded Theory (Corbin and Strauss, 1990). Any instance of a call for, report of, expression of, or counter to violent and aggravating acts is recorded, with any high-level concepts that are identified more than once added to the *Conflict Action* category. Once a complete analysis of the sample data was completed, agreement between annotators was reached on a relative weighting of actions and is presented in Table 4.3.

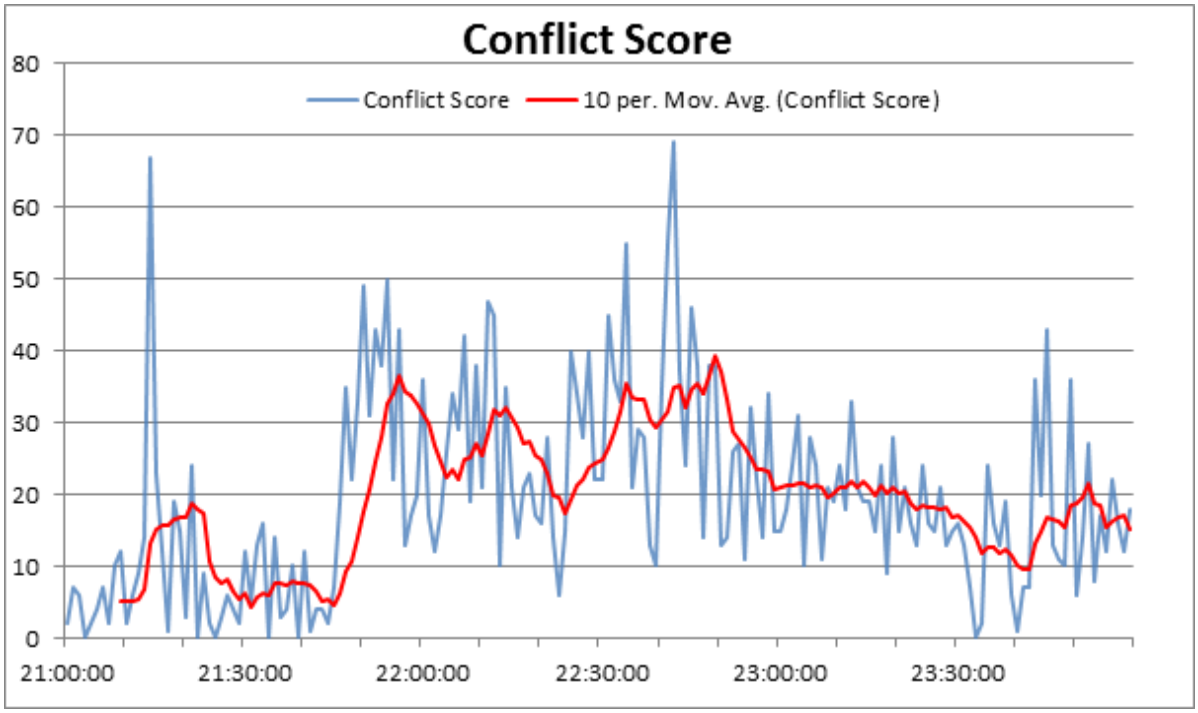


Figure 4.4: Conflict Score Timeline.

The classification and scoring of *Conflict Action* allowed for the creation of a Conflict Score timeline, illustrated by Figure 4.4. This graph shows the sharp increase in *Conflict Action* occurring between 21:40 and 22:00, this is due to the first reports of EDL members clashing with Police officers. Levels remain at that level for the following hour before the score begins to tail off.

This work allowed us to conclude that the EDL leadership sought to mobilize their members, whilst other political groupings opposing their views engaged in a counter-mobilization efforts. (Roberts et al., 2018). Furthermore, we used this first exploratory classification to then identify what key characteristics we could programmatically classify, as the classification scheme complexity is too high.

4.4.2 Annotation

The next step taken in building a streaming classifier for the pipeline was to perform a second annotation exercise focused around annotating potential feature groups in order to assist in training a classifier. It was decided that in order to facilitate the rapid annotation of a training and test corpus of an effective feature set for machine learning and classification, features would need to be normalized so that they only contained a small number of options per feature.

We decided to focus on the *Routine* feature that was developed through the first annotation exercise. The feature was reduced down to *Mobilisation* and *De-Escalation*; reflecting the rise and fall in Conflict Talk theorised by Collins (2012). A second feature selected to support *Routine* is *Emotion*, where we took the six basic emotions: *anger, disgust, fear, happiness, sadness* and *surprise* (Ekman and Friesen, 1971). It has been shown in Williams et al. (2019) that these six basic emotions compare favourably to other emotion classification schemes when tested for ambiguity between classes. In addition to these six emotions, *empathy* and *shock* were included as we anticipated these may be additional facets characteristic to the sample set.

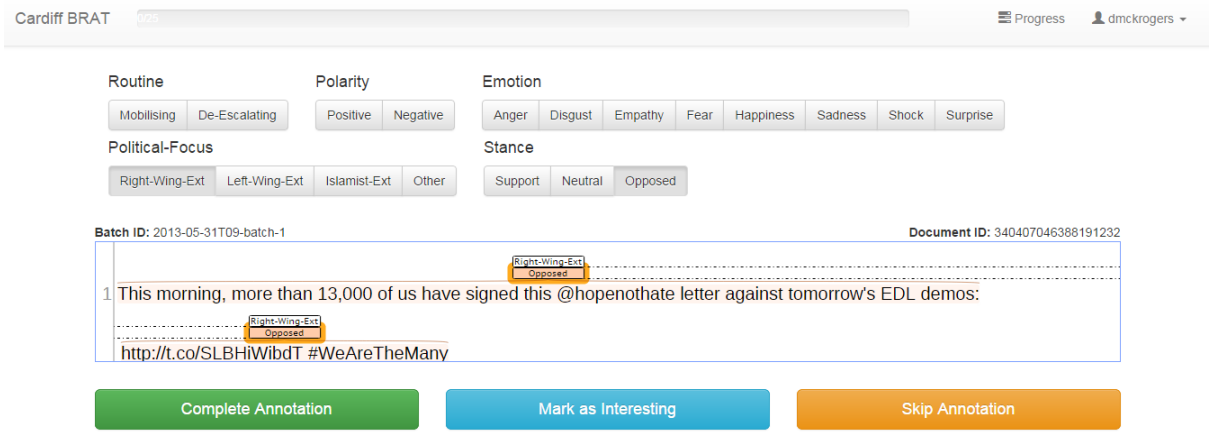


Figure 4.5: The Cardiff BRAT Rapid Annotation Tool Wrapper.

A sample of 8,797 tweets were taken from the first week following the initial incident in Woolwich, with 100 tweets being randomly sampled from each 3-hour block within the week. Figure 4.5 shows the coding interface that we implemented in order to facilitate the next round of coding. We built a small Django wrapper module around the BRAT Rapid Annotation Tool (Stenetorp et al., 2012), so that data could be managed in multiple projects (with their own annotation categories, corpora and users) and to serve documents in a random order so as to alleviate annotator fatigue.

4.4.2.1 Further Development

	All	Anger	Disgust	Fear	Happiness	Sadness	Surprise	Empathy	Shock
All	8,797	1171	60	84	162	130	14	247	21
Mobilising	982	423	24	15	11	6	1	10	1
De-Escalating	230	14	3	1	12	8	0	26	0

Table 4.4: Routine vs Emotion

Taking the annotation results and comparing documents that share both an *emotion* and a *routine* in Table 4.4, we see that *anger* and *mobilisation* are the two most populous categories in their respective features, and that they together are also the most frequently cooccurring. This observation drove further investigation into *escalation* (Section 4.4.3) and the development of an *Anger Classifier*, using the annotation set as training data (Section 4.4.4). Further motivation for this task came from the gap analysis performed in Chapter 2, which identified a lack of emotion classification present within the survey's studies.

4.4.3 Escalation Classification

The *mobilisation routine* derives from elements of conflict talk present within counter-escalation feedback loops that surround an event. Conflict talk is a combination of insulting the opponent, boasting about one's own power, and making threats, all of which in an escalating situation can be taken as real and engage emotional stimuli that strengthen group solidarity (Collins, 2012). In Roberts et al. (2018) we show a small number of messages from the Woolwich dataset that actively advocate violence from one group to another in line with the conflict talk model.

In order to capture this form of content, we focused classification rules on violence-promoting speech consisting of words of intent, will, or threat of aggression against a person or group, labelling content captured by this classifier as "escalation" content.

4.4.3.1 Rule Design

PoS tagging is syntactic analysis of the structure and composition of a sentence as a whole or in part, used to determine the grammatical context of its constituent words, producing a representative list of tags or tokens. A commonly employed tag set is the Penn Treebank tag set (Marcus et al., 1993), with both the Stanford NLP Toolkit (Toutanova et al., 2003) and the popular Python based NLTK

(Natural Language Tool Kit) package (Loper and Bird, 2002). Both packages produce a tree structure of the sentence which can be interrogated and utilised by other parts of the NLP toolkits. From these parse trees complex grammatical structures, such as noun phrases, can be defined and identified.

The rules that form this classifier are primarily driven by the interaction of several PoS types into two types of phrase rules that we define, and noun phrases. Table 4.5 presents the types of PoS that are the focus of the phrase rules used in the classifier, along with the reason for their inclusion, example words, and the corresponding Penn Treebank tags.

PoS	PT Key	Motivation	Examples
Personal Pronouns	PRP	Key to positioning groups or individuals within or without the audience.	I, we, them
Modal Verbs	MD	Expresses necessity or possibility.	Must, shall, can
Violent Verbs	VB	Representative of act of violence.	Kill, defeat, defend

Table 4.5: Parts of Speech used in Escalation Classifier.

We define two phrase rules centred around these PoS tags, and a further rule relating to noun phrases:

- **Adversarial Violent Verb (AVV)** - This is a phrase that contains two personal pronouns of differing perspectives (first, second and third person) bounding a violent verb, or preceding them.
- **Modal Violent Verb (MVV)** - This is a phrase that contains a personal pronoun, followed by a modal verb and then a violent verb.
- **Slurring Noun Phrase (SNP)** – This is a noun phrase that contains a swear word and an identity group (see Section 4.4.3.3).

Phrase Rule	Structure	Examples
Adversarial Violent Verb	<PRP>…<VB.*>+…<PRP'>…<VB.*>*	***They** will **kill us*** ***We** will **destroy them*** ***I** want **you** to **die***
Modal Violent Verb	<PRP.*><MD><VB.*>+	***We must fight***
Slurring Noun Phrase	<NN.*|JJ>+<NN.*>	***English Wanker***

Table 4.6: PoS Phrase Rules with Examples.

Table 4.6 presents shorthand versions of the regular expression rules that are used to match the phrases of interest. Both personal pronouns and modal verbs are core concepts from the Penn Treebank, but violent verbs are sub-type we define, and so we must build our PoS rules with the general verb tag (VB) and then check that the verbs within the phrases are violent verbs.

4.4.3.2 *VIOLENT VERB IDENTIFICATION*

In order to perform violent verb validation on the candidate verbs, we needed to build a dictionary of violent verbs. We sourced candidate terms from two sources; a list of words related to war sourced from an online scrabble solving site [41], and a list of criminal terms sourced from an English language learning site [42]. These were identified as potentially containing a significant number of verbs associated with violence and aggression.

Phase	Initial Retrieval	Manual Review	Synset Identification	Synset Reduction
Volume	630	108	213	147 (81)

Table 4.7: Volumes of Violent Term Synset Generation Stages.

Terms from these two dictionaries were then manually pruned by 4 reviewers to remove terms not relevant to violent speech, accepting terms only when a majority of reviewers agree. We then create an extended dictionary of WordNet *synsets,* synonym groups which can be semantically linked to one another (Miller, 1995), that our curated dictionary belong to. Because WordNet holds a semantic map of all *synsets* we can use an edge similarity measure (Wu and Palmer, 1994) to cross-map all similarities within the dictionary of *synsets* and find the average similarity for each *synset*. We rescale the similarity scores to a 0,1 scale, and then reduce the *synset* dictionary to all *synsets* that score above .5 on the normalised scale. This results in 147 *synsets* belonging to 81 distinct *senses* as shown in Table 4.7

4.4.3.3 *IDENTITY GROUP IDENTIFICATION*

The third phrase rule focuses on the use of swear words in conjunction with identity groups, deemed "slurring". The three identity groups we chose to focus on are nationality and race, religion, and sexuality. Dictionaries of nationality and race, and religion were developed from the Ethnicity and Religion of the Non-UK Born Population in England and Wales report [43] from the 2011 UK census conducted by the Office of National Statistics (ONS). We added further demonyms to the nationality and race dictionary expanding it to cover the four constituent countries of the UK, all European countries, and the members of the Five Eyes intelligence alliance: Australia, Canada, New Zealand, the United Kingdom and the United States (Pfluke, 2019). The sexuality dictionary was derived from

[41] https://words-solver.com/war-words-list/
[42] https://www.engvid.com/english-resource/vocabulary-crime-criminals/
[43] https://webarchive.nationalarchives.gov.uk/20160105172419/http://www.ons.gov.uk/ons/rel/census/2011- census-analysis/ethnicity-and-religion-of-non-uk-born-population-in-england-and-wales--2011-census/rpt.html

the Sexual Identity Topic Report [44] in the 2021 Census topic consultation also of the ONS, with the addition of transsexual.

4.4.3.4 IMPLEMENTATION

The classification logic was implemented in Python primarily using the NLTK (Loper and Bird, 2002) package to prepare and analyse documents. We tokenise each sentence within a document and then pass them through the NLTK PoS tagger. The PoS tree is then passed to the first *RegexpParser*, an NLTK class that identifies segments of a PoS tree based off regular expressions (Table 4.6), using the SNP rule. The tokens matched from this parser are then stemmed and checked against the stemmed forms of swearwords and each of the Identity Group dictionaries, with documents being flagged when matching to both a swearword and an Identity Group.

Following this, the pronoun PoS tags are checked against a fixed list of first, second and third person pronouns, with the PoS tag appended with a digit representing the perspective. This updated PoS tree is then passed through the second *RegexpParser* that is loaded with the AVV and MVV rules. Matches on this parser are again stemmed, and then checked against the violent verb dictionary. Any matches that contain a violent verb are flagged to the rule that matched it.

4.4.3.5 EXPERIMENTATION

In the summer of 2017, a series of terrorist attacks took place in the UK that varied in their methods and targets but all contributed to a shared escalating narrative (Innes et al., 2019). First, a single-perpetrator vehicle attack occurred on Westminster Bridge on the 22nd of March. This was followed by a suicide bombing outside the Manchester Evening News Arena on the 22nd of May. Shortly after on the 3rd of June a marauding attack by three individuals took place that began on London Bridge, and on the 19th of June another single-perpetrator vehicle attack took place in Finsbury Park. These resulted in a rapid online reaction, with tens of thousands of unique messages being posted to Twitter within the first 3 hours (Table 4.8).

[44] https://www.ons.gov.uk/file?uri=/census/censustransformationprogramme/consultations/the2021censusinit ialviewoncontentforenglandandwales/topicreport03genderidentity.pdf

Attack	Window	Tweets		
		All	Retweets	Original
Westminster	3 hr	303293	244164	59129
	48hr	993428	791991	201437
Manchester	3 hr	2035891	1798785	237106
	48hr	11951971	10420754	1531217
London Bridge	3 hr	527014	444382	82632
	48hr	1181324	1020807	160517
Finsbury Park	3 hr	106763	87088	19675
	48hr	601147	527275	73872

Table 4.8: Tweet volumes for 2017 Terrorist Attacks.

Original Tweets (i.e., non-retweets) were retrieved from each of the four incidents in covering both the first 3 hours and 48 hours following the event. The Escalation Classifier was run over these datasets and the volume and content of all escalating tweets were recorded.

Attack	Window	Violent Verb		Slurring Noun Phrase		
		Adversarial	Modal	Race	Religion	Sexuality
Westminster	3 hr	$1.18e^{-4}$	$3.38e^{-4}$	$0.51e^{-4}$	$2.37e^{-4}$	$0.00e^{-4}$
	48hr	$1.74e^{-4}$	$3.67e^{-4}$	$0.35e^{-4}$	$1.99e^{-4}$	$0.00e^{-4}$
Manchester	3 hr	$1.77e^{-4}$	$3.12e^{-4}$	$0.59e^{-4}$	$1.98e^{-4}$	$0.04e^{-4}$
	48hr	$2.67e^{-4}$	$4.05e^{-4}$	$0.42e^{-4}$	$1.49e^{-4}$	$0.02e^{-4}$
London Bridge	3 hr	$3.51e^{-4}$	$7.26e^{-4}$	$1.57e^{-4}$	$12.95e^{-4}$	$0.00e^{-4}$
	48hr	$5.11e^{-4}$	$9.72e^{-4}$	$1.87e^{-4}$	$10.47e^{-4}$	$0.00e^{-4}$
Finsbury Park	3 hr	$2.54e^{-4}$	$9.15e^{-4}$	$0.51e^{-4}$	$4.57e^{-4}$	$0.00e^{-4}$
	48hr	$3.25e^{-4}$	$11.91e^{-4}$	$0.41e^{-4}$	$4.20e^{-4}$	$0.00e^{-4}$

Table 4.9: Proportion of Violent Verb phrases and Race and Religion Slur phrases.

Table 4.9 presents the proportions of tweets matching a phrase rule and it can be seen that these measures are highly granular. Addressing the SNPs first we see that religious slurring is the most prevalent form of slurring present within the corpora, with smaller volumes of racial slurring and little to no sexuality slurring. This is reflective of the relationships all four attacks have with religion, with perpetrators in the initial three attacks being quickly associated to Islamic terrorism and the fourth attack targeting Muslims (Innes, 2020). In violent verb use, MVV is consistently more present than AVV, with use of MVVs increasing over the course of the four attacks.

The low latency of these measures in highly emotional corpora suggested that this classifier is best suited to supporting qualitative analysis through the production of a manageable sub-corpus that can be manually assessed. As such, it was not operationalised into the streaming element of the pipeline, and instead was incorporated into the Download services presented in Chapter 5.

4.4.4 ANGER CLASSIFICATION

Supervised machine learning is dependent upon the transformation of arbitrary data such as images or text into numerical features usable for classification (Win and Aung, 2017). Chapter 3 presented a number of pre-processing analyses performed by the studies that formed the systematic review. We have incorporated a number of these processes into the Sentinel Pipeline both in preparation for data driven analysis with FlexiTerm (Section 3.5.2.2), and as part of document tagging process described in Section 4.4.

Type	Feature	Measure	Origin
Morphological	Lowercase Characters	Count and Percentage	Anger Classifier Module
	Uppercase Characters		
	Non-Alphanumeric Characters		
	Question Marks		
	Exclamation Marks		
	Characters	Count	
	Tokens		
Syntactic	Pronouns		Stanford CoreNLP Document Tagger
Semantic	Named Entities		
	Emotion Words		PathNER Document Tagger
	Swear Words		
	Sentinel Ontology		
	Sentiment	Score	Stanford CoreNLP Document Tagger

Table 4.10: Anger Classification Features.

We are then able to re-use the additional metadata appended to the documents as they pass through the Sentinel Pipeline and utilise them as feature vectors within the *Anger Classifier* module. Table 4.10 presents the features utilised by the *Anger Classifier* and their origin within the Sentinel Pipeline.

It should be noted that there are a number of non-linguistic pre-processing steps that are taken prior to feature extraction which are performed in the *text normalisation* module covered in Section 3.4.3.6. A number of morphological counts are handled within the *Anger Classification* module itself, focused primarily on teasing out the different counts for characters, tokens, and important punctuation.

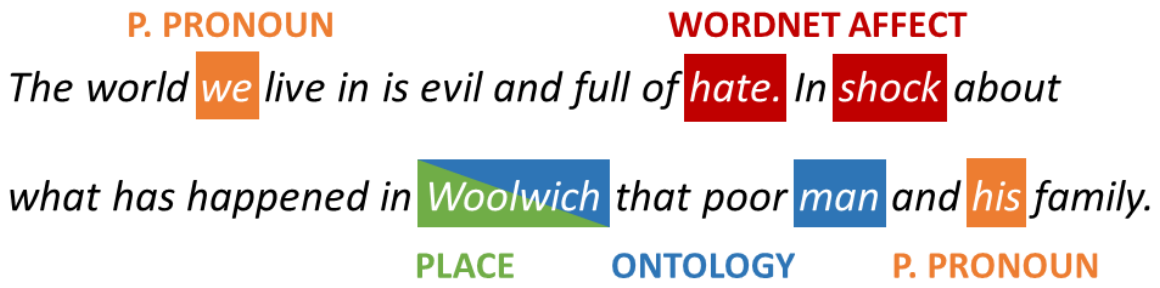


Figure 4.6: Syntactic and Semantic annotations.

The sole syntactic feature is a count of pronouns present in the text. There are a number of semantic features, taken from the output of both *document taggers* described in Section 4.3, with the 5-point sentiment score being reduced to a single integer presenting the category with the highest probability.

Figure 4.6 gives an illustration of the application of the frequency based syntactic and semantic features within a candidate text. Pronouns and emotive words share a relationship in the implications for emotional disclosure in text (Fuentes et al., 2018), and so we give weight to these entities within the text.

4.4.4.1 Performance

Implementation of the classifier is made using Weka, a Java package which provides a common interface into a multitude of NLP algorithms (Hall et al., 2009). A prototype classifier was built using the Naïve Bayes class provided by Weka, with the training data passed through the Sentinel Pipeline into a separate collection so that reprocessing was simpler as we tweaked features.

Full Dataset (P-1161/N-7636)										
Class	**All**				**Positive**			**Negative**		
Measure	**A**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
None	0.79	0.59	0.62	0.60	0.28	0.39	0.33	0.90	**0.85**	**0.88**
Count	0.79	0.60	0.63	0.60	0.29	0.40	0.33	0.90	0.85	0.87
Flat	**0.79**	0.60	0.64	0.62	0.30	0.44	0.36	0.91	0.84	**0.88**
Cascading	0.77	0.60	0.65	0.61	0.29	0.49	0.36	**0.91**	0.82	0.86
50% Dataset (P-1161/N-1161)										
Class	**All**				**Positive**			**Negative**		
Measure	**A**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
None	0.64	0.66	0.66	0.66	0.64	0.69	0.67	0.67	0.62	0.64
Count	0.66	0.66	0.66	0.66	0.64	**0.70**	**0.67**	0.67	0.61	0.64
Flat	0.67	**0.67**	**0.67**	**0.67**	0.66	0.70	0.68	0.68	0.64	0.66
Cascading	0.63	0.65	0.63	0.61	**0.71**	0.42	0.53	0.59	0.83	0.69

Table 4.11: Ontology Strategy and Category Performance for Naïve Bayes Classifier.

The classifier was trained using both the full 8,797 document training dataset (where the positive class has a share of 13% of the documents), and a balanced dataset, where the non-angry documents were randomly sampled down to 1,161 so that both classes are represented by 50%. Table 4.11 presents the initial performance analysis performed on the Naïve Bayes classifier using both data sets. Whilst the unbalanced set produces a higher accuracy, much like we saw in Chapter 2, the unbalanced set biases the accuracy towards the larger negative class, which is undesirable. The balanced dataset produces better F1 and show much higher performance within the positive class.

Table 4.11 also presents the results of using the Sentinel Ontology to tune the classification task towards our domain of interest (crime and social unrest). We apply a number of strategies when creating the *ontology* feature within the classifier, with this section exploring how this affects performance. The results were generated by performing 10-fold cross validation on the training data (N=8797 for the full dataset, and N=2322 for the balanced 50% dataset). We use macro-F1 scores to assess the overall performance of the classification models.

First, we tested performance where the ontology feature is not used at all (*none*) and where the feature is a simple count of all ontology occurrences (*count*). We then had two further strategies in which the ontology elements were each used as a vector, where an element is only counted if there is an exact match (*flat*) or where the element and all parent nodes are incremented on a match (*cascading*) thereby utilising the ontology to its fullest.

Within the evenly balanced dataset, the increasing complexity of the ontology feature produces an improved performance on the precision of the positive class, although at the point of using a *cascading* feature, recall drops off significantly suggesting that the feature becomes too narrow with this strategy. This is reflected in the overall performance, which also shows the *flat* strategy produces the best performance.

Algorithm	**Naïve Bayes**				**J48 DT**				**SMO SVM**			
Measure	**A**	**P**	**R**	**F1**	**A**	**P**	**R**	**F1**	**A**	**P**	**R**	**F1**
None	0.64	0.66	0.66	0.66	0.64	0.64	0.64	0.64	0.66	0.66	0.66	0.66
Count	0.66	0.66	0.66	0.66	0.64	0.64	0.64	0.64	0.67	0.67	0.67	0.67
Flat	0.67	**0.67**	**0.67**	**0.67**	0.67	0.67	0.67	0.67	0.69	**0.70**	0.69	0.69
Cascading	0.63	0.65	0.63	0.61	0.68	0.68	**0.68**	**0.68**	0.70	**0.70**	**0.70**	**0.70**

Table 4.12: Ontology Strategy and Overall Algorithm Performance 50% Dataset.

We then tested the feature set against two other machine learning algorithms again using Weka, a J48 Decision Tree (DT), using a 0.25 confidence factor and a minimum of 2 object per leaf, and a SMO based Support Vector Machine (SVM), with a complexity of 1.0. Table 4.12 presents the

performance results for the three algorithms using all four strategies on the 50% training set. The SVM algorithm outperforms the other algorithms in all strategies, which reflects the findings of Chapter 2 where SVM algorithms were regularly found to be the best performing algorithm in the studies. In contrast to the Naïve Bayes, the SVM algorithm performs better with the *cascading* strategy when compared to the *flat* strategy, although the *flat* strategy showed marginally higher precision in the positive class vs the *cascading* strategy (0.722 vs 0.715).

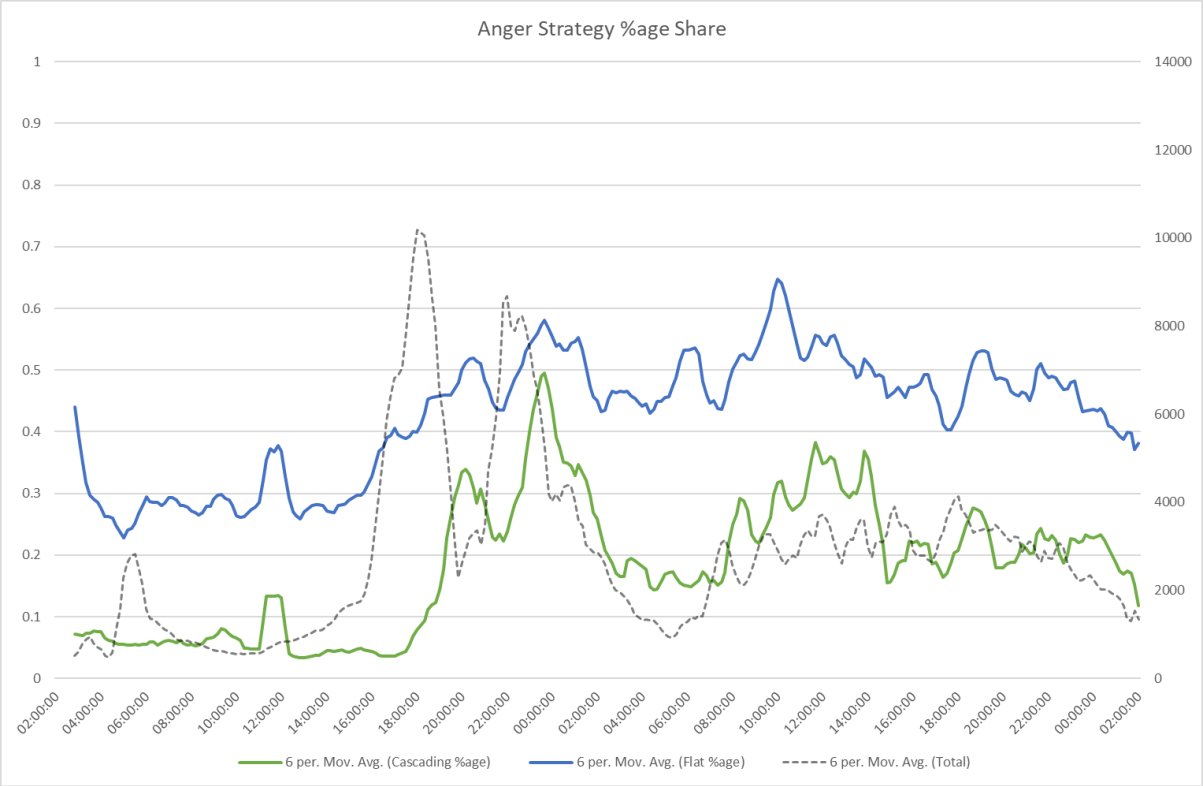


Figure 4.7: Percentage Share of Anger by Strategy for Woolwich Dataset.

Comparing these performances to the studies in Chapter 2, overall performance by any of the algorithms is relatively poor, but the output from this classifier can still be informative. Figure 4.7 shows percentage share of tweets in the first two days of the Woolwich dataset that are "angry", as classified using for the *flat* and *cascading* strategies in 10-minute windows with a 60-minute rolling average.

The *flat* strategy sits at a higher average percentage of 29.6% vs 6.3% within the hours preceding the incident. The *flat* strategy reacts faster to the event with a gradual increase in reported anger that then persists at an average of 48.4%. The *cascading* strategy shows a more pronounced shift in anger that lags behind the initial reporting but does correspond with the reaction to the event from the EDL including the protests that took place later in the evening peaking at 22:50 on the 22nd with a recording of 61.7% before dropping off overnight to an average of 16.5%. There is a second increase between 08:00 and 14:00 on the 23rd. The fact that the *cascading* strategy does not move

with the first spike but then increases with the second and morning spike suggests that this maps closely more closely to the *conflict score* trend observed in Figure 4.4.

4.4.4.2 OPERATIONALISATION

We use a management module to incorporate the anger classifiers into the Sentinel Pipeline in a similar fashion to how the document taggers were added. There is an Abstract class *ClassifierBase* for which an *Emotion* implementation exists. This classifier collects all the pre-tagged metadata and performs the remaining feature extraction ad-hoc.

The *Emotion* class can be instantiated with either a Naïve Bayes or SMO classifier which can be loaded with multiple versions of the pre-processed training data, with versions that cover both types of dataset (*full* and *balanced*) and three strategies (*none, flat* and *cascading*). Mirroring the Indexers, a *ClassifierManager* class acts as a broker to RabbitMQ and can have *Classifiers* registered to it.

Positive classifications are added to the document metadata with the classification algorithm, training set and strategy recorded. The ElasticSearch index's *classification* key is updated with the same information, and the *contains* key is updated with a shorthand record consisting of the first two characters of the algorithm, training set and strategy name. By default, the *ClassificationManager* has six classifiers registered to it: both of the algorithms operating on the balanced training set, with all three strategies.

4.5 DOCUMENT SEARCHING

In Chapter 3 we presented a number of interfaces that were developed to provide insight into the data collected by the Sentinel Pipeline. These interfaces were designed to allow researchers to scan and monitor topics and events with a fixed framing and continuous stream of data. This chapter has so far covered how we can enrich these data streams with further knowledge and semantic features.

This section covers the development of a Boolean querying interface that allows researchers to exploit the developments of this chapter and interrogate the collected and enriched data to a much finer and more targeted detail.

Development of this interface within the new OSCAR Hub portal focused upon three key principles:

- Interface with low barrier to entry and steady learning curve to more complex features.
- Ability to exploit the enriched features through complex Boolean functionality.
- Ability to re-use complex queries built by users to further inform other tools and features.

In the following section we present the integration of this system into the underlying architecture, an overview of the interface design and key interface features, before presenting a study into the usage of the search tool in a period of heightened activity supported by a user survey.

4.5.1 QUERY BUILDING

The Semantic Search tool is designed to allows users to rapidly build complex Boolean queries that can cut across collection *channels* and document types. We extended a JavaScript library that provides a dynamic query building interface, the jQuery QueryBuilder [45], with a number of custom adaptations (discussed in Section 4.5.1.1) and common functionality plugins. Using the QueryBuilder, multiple query rules can be nested together and combined with AND and OR operators.

Field	Operations	Type	Description
Text Content	equal not equal	Text	Search across text field, wildcard (*) accepted for single word terms.
Account	author mentions quotes replies to retweets	Text	See Section 4.5.1.1
Contains	in not in	Checkbox	See Section 4.5.1.1
Language	equal not equal	Text	Search for reported language using ISO 639-1 language codes.
Document Source	in	Checkbox	Filter document source by constraining indexes queried.
Date/Time Stamp	equal not equal between less less or equal greater greater or equal	Date/ Datetime	Series of query options to bound documents to time period through the created_at field.
Article Title	equal not equal	Text	Search across article_title field, wildcard (*) accepted for single word terms.
Article Text	equal not equal	Text	Search across article_text field, wildcard (*) accepted for single word terms.
Article Author	author	Text	Search across article_author field, wildcard (*) accepted for single word terms.
Ontology	in not in	Guided text	See Section 4.5.1.1

Table 4.13: Query Builder Options.

[45] https://querybuilder.js.org/index.html

Table 4.13 presents the fields made available in the QueryBuilder implementation. The majority of these are text fields that can be invoked multiple times. There is a restriction put on the Date and Time Stamp fields, allowing only one instance of each within the whole query. The QueryBuilder has an inbuild validation method to ensure all fields are valid and provides export functionality for both SQL and MongoDB serialised queries.

4.5.1.1 *Semantic Features*

Like other text fields in the field options, the *author* field allows for multi-word text encapsulated with quotation marks, and wildcard (*) use for single word terms. The “author” “mentions”, “quotes”, “replies to” and “retweets” operators are extensions made to the QueryBuilder code to provide deeper options that cover Twitter specific context, where Tweets relate to a second document through quoting, replying, or retweeting, in which case the query is applied to subject Tweet author field.

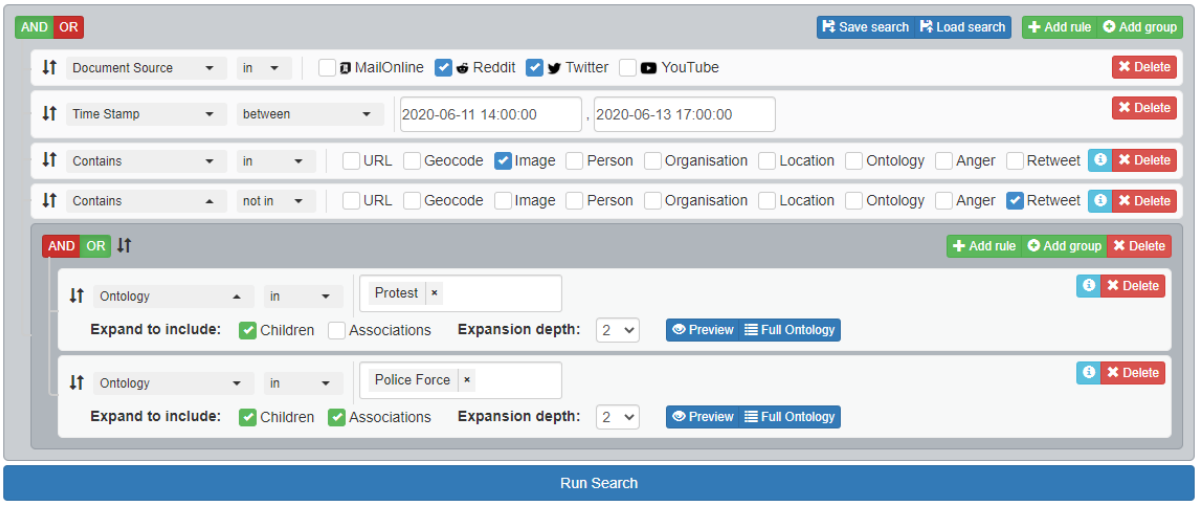


Figure 4.8: QueryBuilder interface with Semantic Example Semantic Options.

Figure 4.8 presents an example usage of several of the semantically enriched features of the QueryBuilder interface. The *contains* field allows a user to retrieve documents that contain matches from the NER and Ontology document taggers, the Anger Classifier, or in the case of Twitter documents contain an image, URL, Geocode or a Retweeted document. Multiple tag types can be selected and will automatically be ANDed within the query.

The *ontology* field consists of an autocomplete text entry pre-loaded with all ontology concepts retrieved from the same OBO file that drives the Ontology tagging presented in Chapter 3. The *ontology* input can accept up to 10 terms that are automatically OR-ed within the query. We have extended the *ontology* filed to include checkboxes that allow the ontology concept to exploit the *is_a* and *related_to* semantic links so that the query scope can be expanded vertically to child nodes

and horizontally to sibling nodes. A preview button will generate a popup (Figure 4.9) that shows the user the current ontology concepts that are retrievable by the query.

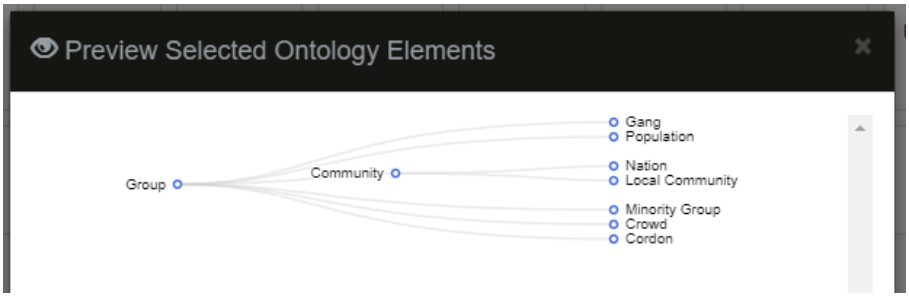


Figure 4.9: Ontology Preview Window.

4.5.2 Data Retrieval

Once a user has completed building a query, they may then submit the search via a series of calls to four endpoints within the REST API. A sequence diagram of the interactions of these endpoints with one another and the MongoDB and ElasticSearch services are presented in Figure 4.10. Interactions between the interface and the three API endpoints are all performed through AJAX requests initiated by the interface.

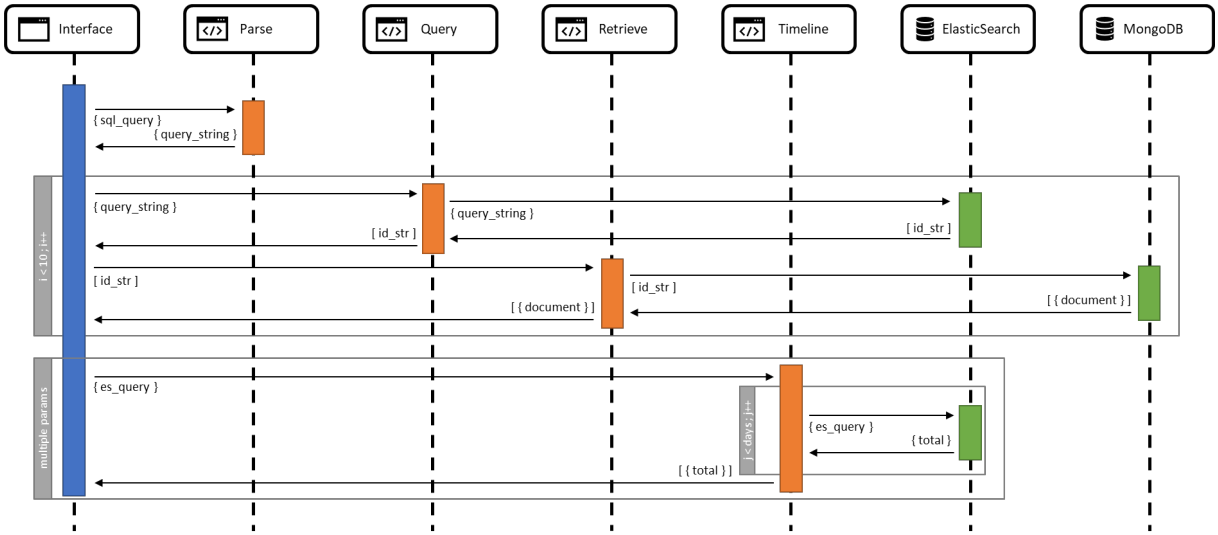


Figure 4.10: Sequence Diagram of Search Component Interaction.

The *parse* endpoint uses a complex regular expression to transforms the SQL serialised query into a valid ElasticSearch *query_string*. In the case of *ontology* parameters being present within the SQL query the *parse* endpoint will expand out ontology IDs to the relevant *expansion depth* (Figure 4.11). The *query_string* is also prefixed with the *channels* that the query will be run across. Upon receiving the translated *query_string* from the *parse* endpoint, the *interface* fires off two parallel AJAX calls to the *query* and *timeline* endpoints.

Document Types Searched: Reddit, Twitter
Query Used: (channel:20172 OR channel:20206) AND (created_at:[2020-06-11T14:00:00.000Z TO 2020-06-13T17:00:00.000Z] AND (entities:img) AND NOT (entities:rt) AND ((ontology:0000277 OR ontology:0000278 OR ontology:0000279 OR ontology:0000280 OR ontology:0000281 OR ontology:0000282 OR ontology:9900015 OR ontology:9900016 OR ontology:9900017 OR ontology:9900021 OR ontology:9900022 OR ontology:9900023 OR ontology:9900024 OR ontology:9900025 OR ontology:9900026 OR ontology:9900028 OR ontology:9900029 OR ontology:9900030 OR ontology:9900031 OR ontology:9900032 OR ontology:9900033 OR ontology:9900034 OR ontology:9900036 OR ontology:9900037 OR ontology:9900038 OR ontology:9900039 OR ontology:9900040 OR ontology:9900041 OR ontology:9900042 OR ontology:9900045 OR ontology:9900046 OR ontology:9900047 OR ontology:9900048 OR ontology:9900049 OR ontology:9900051 OR ontology:9900052 OR ontology:9900053 OR ontology:9900054 OR ontology:9900055 OR ontology:9900056 OR ontology:9900057 OR ontology:9900058 OR ontology:9900059 OR ontology:9900060 OR ontology:9900061 OR ontology:9900062 OR ontology:9900063 OR ontology:9900064 OR ontology:9900065 OR ontology:9900067 OR ontology:9900068 OR ontology:9900069 OR ontology:9900086 OR ontology:9900087 OR ontology:9900088 OR ontology:9900089 OR ontology:9900090 OR ontology:9900096 OR ontology:9900097 OR ontology:9900102 OR ontology:9900103 OR ontology:9900104 OR ontology:9900106 OR ontology:9900108 OR ontology:9900109 OR ontology:9900111 OR ontology:9900112 OR ontology:9900113 OR ontology:9900114 OR ontology:9900115 OR ontology:9900116 OR ontology:9900117 OR ontology:9900118 OR ontology:9900122 OR ontology:9900131 OR ontology:9900193) OR (ontology:0000288 OR ontology:0000272 OR ontology:0000380)))
Total Documents Found: 746

Figure 4.11: Parsed ElasticSearch Query.

The *query* endpoint uses the *query_string* to search the ElasticSearch *indexes* for matches, receiving 20 hits at a time. These are passed back to the *interface* and stored as the initial representation of any matched documents in case the documents are not yet persisted in MongoDB. The key pieces of information are the array of *id_str* fields belonging to each document and an array containing the *channel_id* where each document can be found.

Next, the *interface* passes the *id_str* and *channel_id* arrays to the *retrieve* module, that interfaces with the MongoDB service retrieving the full matching document metadata from each *channel* that contains them. If a document is present in multiple channels, it is only retrieved once. An array of all retrieved documents is then returned to the *interface* and used to update the already stored data for each document. The timestamp of the earliest retrieved is then appended to the *query_string* so that the next 20 documents can be retrieved from the *query* and *retrieve* endpoint.

In parallel to this, the *timeline* endpoint is sent a number of modified versions of the *query_string* with additional constraints added such as document type, retweet status, and image status. The *timeline* endpoint uses the modified *query_string* to make multiple time-bounded queries to the ElasticSearch *indexes*. These consist of 30 24-hour windows running back form the point of query. The *timeline* endpoint can then use the response metadata from ElasticSearch to build an array of volumes for each day, which is returned to the *interface*.

4.5.3 Data Presentation

The document data retrieved from the *retrieve* and *timeline* endpoints are presented to the user via a series of document trays. The *images*, *ontology*, *NER*, and *anger* trays provide filtered versions of the *document* tray.

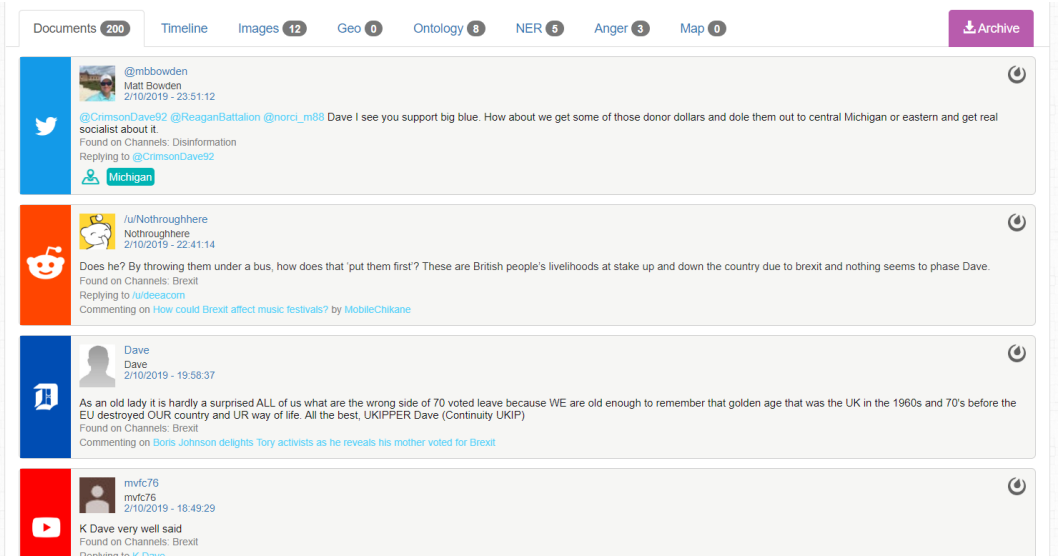


Figure 4.12: Document Tray.

Figure 4.12 presents a screenshot of the *documents* tray, highlighting the multiple document types available to users. The original source of a document is available through the timestamp link, whilst the username link opens the user account page for the document author. Ontology, NER and Anger matches are flagged with colour coded icons for quick identification, and non-English documents can be translated by clicking on the language code link.

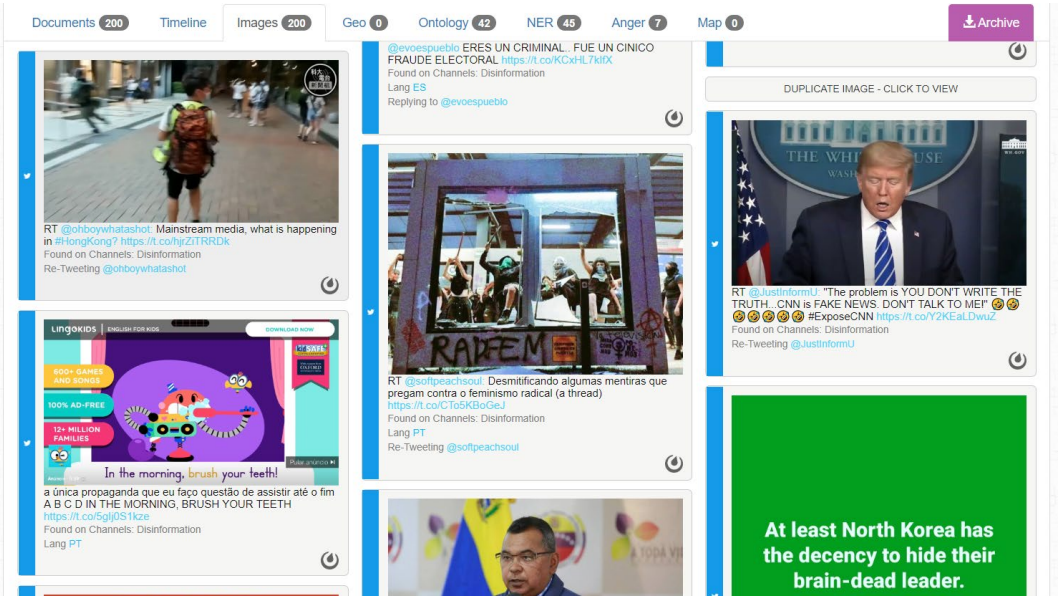


Figure 4.13: Image Document Tray.

The *image* tray (Figure 4.13) embeds the images present within a document into a three-column gallery with duplicates hidden to allow for fast scrolling and identification of pertinent content.

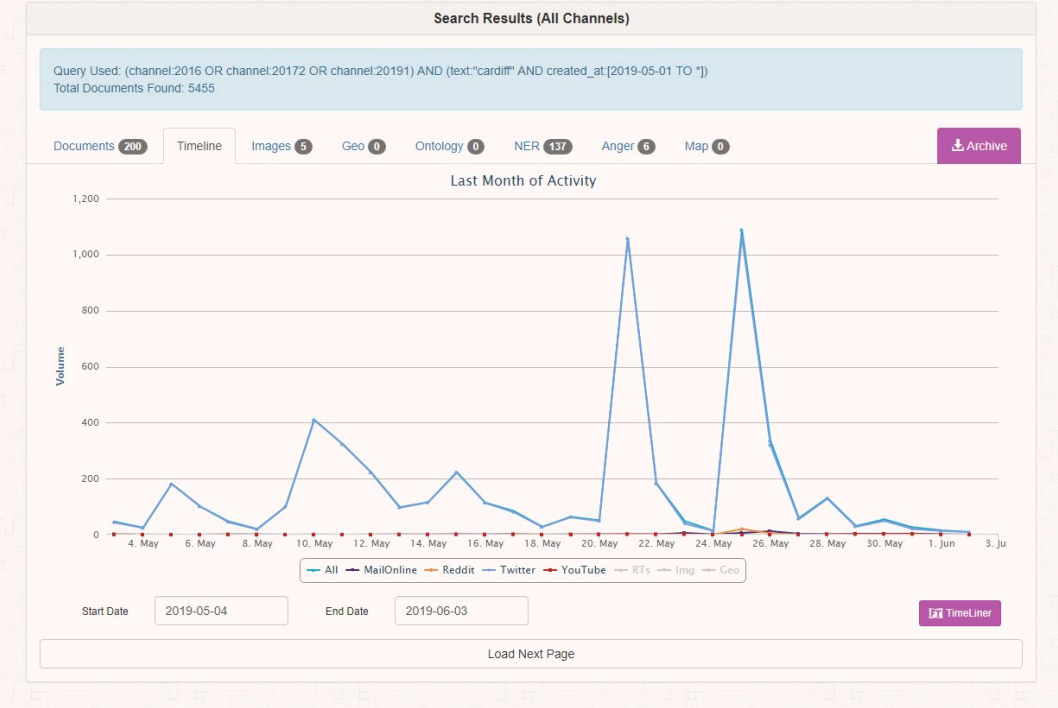


Figure 4.14: Timeline Tab.

Output from the multiple calls to the *timeline* endpoint (Figure 4.14) is loaded into a JavaScript graphing package called Highcharts [46], which produces a line graph with multiple series representing the volumes of data broken down into document type, retweet status, image status, and geotag status.

4.5.3.1 Saved Searches

Search configurations can also be saved for re-use by other users. The QueryBuilder module is able to ingest SQL queries and convert them back into a full configuration of the interface. The SQL serialisation of the query is saved to the MySQL database belonging to the Django app so that it can be loaded back in via an API request.

These *saved searches* along with the *parse*, *query* and *retrieve* endpoints can also be used to support other applications within the OSCAR Hub portal. The SentiSum interface (Figure 4.15) was designed for pre and post event analysis using the FlexiTerm tool, output from Stanford Sentiment Analysis, and from the Anger Classifier. It uses the *saved searches* to filter documents into sub-corpora for the batch processing tasks to be run over, giving the user the ability to drill down into predefined sub-categories within the SentiSum interface.

[46] https://www.highcharts.com/

This interface was designed and deployed as part of the co-design exercises that surrounded the 2016 Conservative Party Conference that was held in Birmingham. The tool was used within the *situation room* to monitor sentiment and anger shifts in the discussions surrounding the event, with a series of *saved searches* used to segment the collected data from the channel into extreme-left, left, right, extreme-right, and policing sub-corpora.

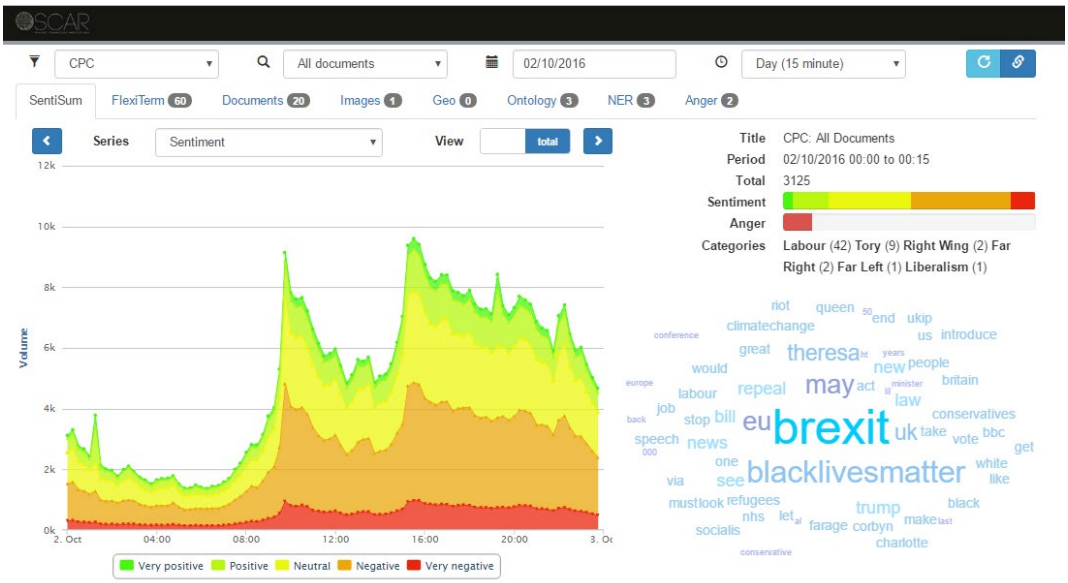


Figure 4.15: SentiSum Interface.

FlexiTerm was still used within the new SentiSum interface, with it being run in parallel to the SentiSum process and presented within the same interface. Users are able to navigate through a timeline by clicking on an individual 15-minute time point. They are presented with a summary of Tweets collected over that time window showing the aggregate sentiment, a word cloud generated from all content, all FlexiTerms within the corpus and the same Document Tray that is presented to the user when using Semantic Search (which provides the user with a familiar method of exploring the Tweets themselves). Output from the Anger Classifier was also integrated into the SentiSum interface, with the data being presented as an alternate view of the timeline (Figure 4.16), with each version of the ontology *strategies* presented in Section 4.4.4.1 selectable.

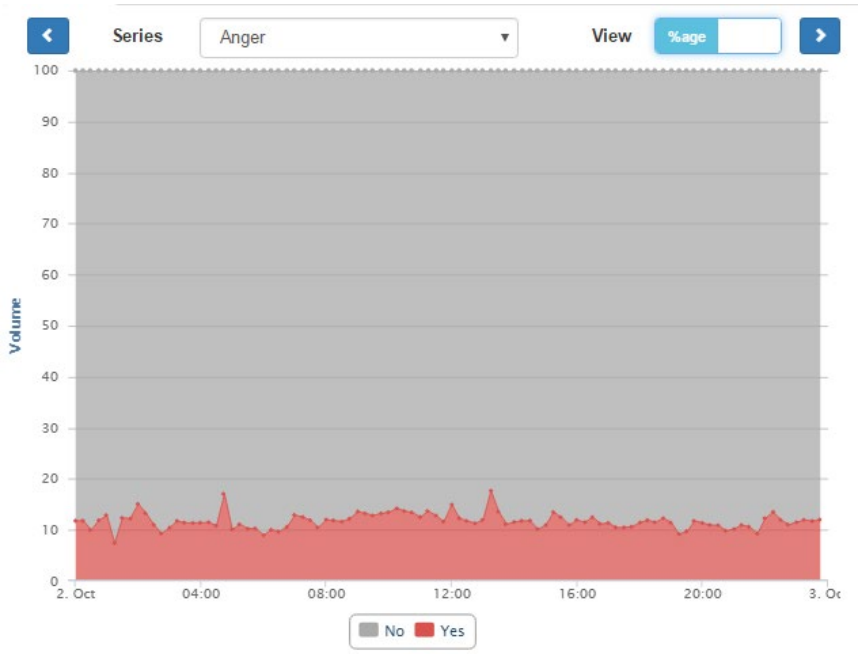


Figure 4.16: Anger View for SentiSum Timeline.

4.5.4 User Study and Usage Assessment

In order to assess how the Classification, Tagging and Semantic Search developments are being used, we record all SQL serialised queries submitted to the *parse* endpoint by the Semantic Search *interface.* We analysed the submission logs of all end-users between January 2019 and May 2020 (1789 in total submissions, across 19 users) using the same regular expression we use to parse the queries to a matrix of feature usage within queries.

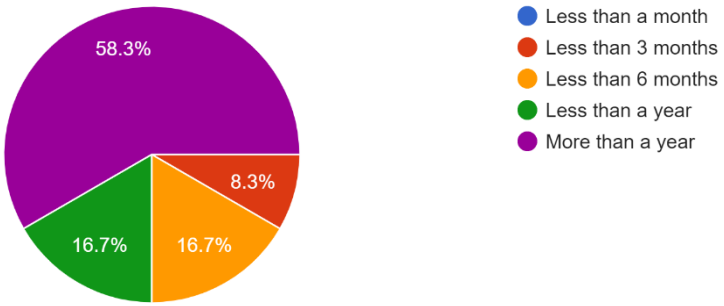


Figure 4.17: Experience with Sentinel from Survey Respondants.

Additionally, in May 2020, we surveyed 12 of our users on their usage, experience, familiarity and understanding of a series of Sentinel Interfaces. This survey consisted of multiple Likert scales (Vagias, 2006) focused upon the main interfaces of OSCAR Hub at the time. The majority of this survey is presented in Chapter 5, but we take the Semantic Search survey results into consideration within this section. The majority of the 12 survey respondents have been using the Sentinel interfaces for over a year, as shown in Figure 4.17.

We cluster query submissions together into sessions of activity, where there is no more than 10 minutes between two consecutive query submissions. Table 4.14 presents the query feature usage across the entire logging set, listing the average number of instances, total number of submissions, total number of submissions containing multiple instances of a feature, and the percentages of these two measures. This breakdown is also presented for the *final query* submissions of sessions, in order to see what query features users generally settle on during a session.

We see that the most commonly used feature is the *text* field, present in 82% of query submissions, and the only non-operation feature that averages over one use per query, rising to over two uses per query within the *final query* subset suggesting users may expand to cover synonyms as they refine their searches. Second to the *text* feature, users bound their query with a temporal feature 34% of the time, with just under a quarter of all submissions using two temporal features to query within a time window.

		All Queries					Final Query				
		Avg.	**Total**	**Multi**	**Tot%**	**Multi%**	**Avg.**	**Total**	**Multi**	**Tot%**	**Multi%**
	AND	1.10	962	528	54%	30%	1.13	257	143	53%	29%
	OR	0.72	259	195	14%	11%	1.25	96	73	20%	15%
	Doc Type	0.03	41	7	2%	0%	0.06	21	5	4%	1%
	Date	0.50	533	363	30%	20%	0.56	157	115	32%	24%
	Time	0.08	74	68	4%	4%	0.05	12	11	2%	2%
	Text	1.71	1474	455	82%	25%	2.21	397	131	82%	27%
	Lang	0.13	228	8	13%	0%	0.09	44	1	9%	0%
	Onto	0.00	0	0	0%	0%	0.00	0	0	0%	0%
User	**Author**	0.06	96	3	5%	0%	0.06	28	3	6%	1%
	Retweet	0.00	4	0	0%	0%	0.00	0	0	0%	0%
	Quote	0.00	2	0	0%	0%	0.00	0	0	0%	0%
	Reply	0.01	12	0	1%	0%	0.01	4	0	1%	0%
	Mention	0.00	7	0	0%	0%	0.01	3	0	1%	0%
Contains	**RT**	0.16	288	0	16%	0%	0.20	99	0	20%	0%
	URL	0.01	13	0	1%	0%	0.01	3	0	1%	0%
	Geo	0.00	8	0	0%	0%	0.00	2	0	0%	0%
	Img	0.01	13	0	1%	0%	0.01	3	0	1%	0%
	Per	0.00	2	0	0%	0%	0.00	0	0	0%	0%
	Org	0.00	1	0	0%	0%	0.00	0	0	0%	0%
	Loc	0.00	2	0	0%	0%	0.00	0	0	0%	0%
	Onto	0.00	1	0	0%	0%	0.00	0	0	0%	0%
	Anger	0.00	2	0	0%	0%	0.00	1	0	0%	0%
Article	**Title**	0.00	6	0	0%	0%	0.00	1	0	0%	0%
	Text	0.00	2	0	0%	0%	0.00	1	0	0%	0%
	Author	0.00	1	0	0%	0%	0.00	0	0	0%	0%

Table 4.14: Feature Usage 2019 & 2020.

The other two main features used are the *language* filter and *retweet containing* flag, the first of which reduces in use within the *final query,* whilst the second increases in use by 5%. With the *language* feature, this may be because users wish to expand to catch any uses of a word outside of its language of origin. The increase in the *retweet containing* flag is likely an indication of users adjusting their queries to reduce the noise of repetitive messages, allowing them to access more actionable content.

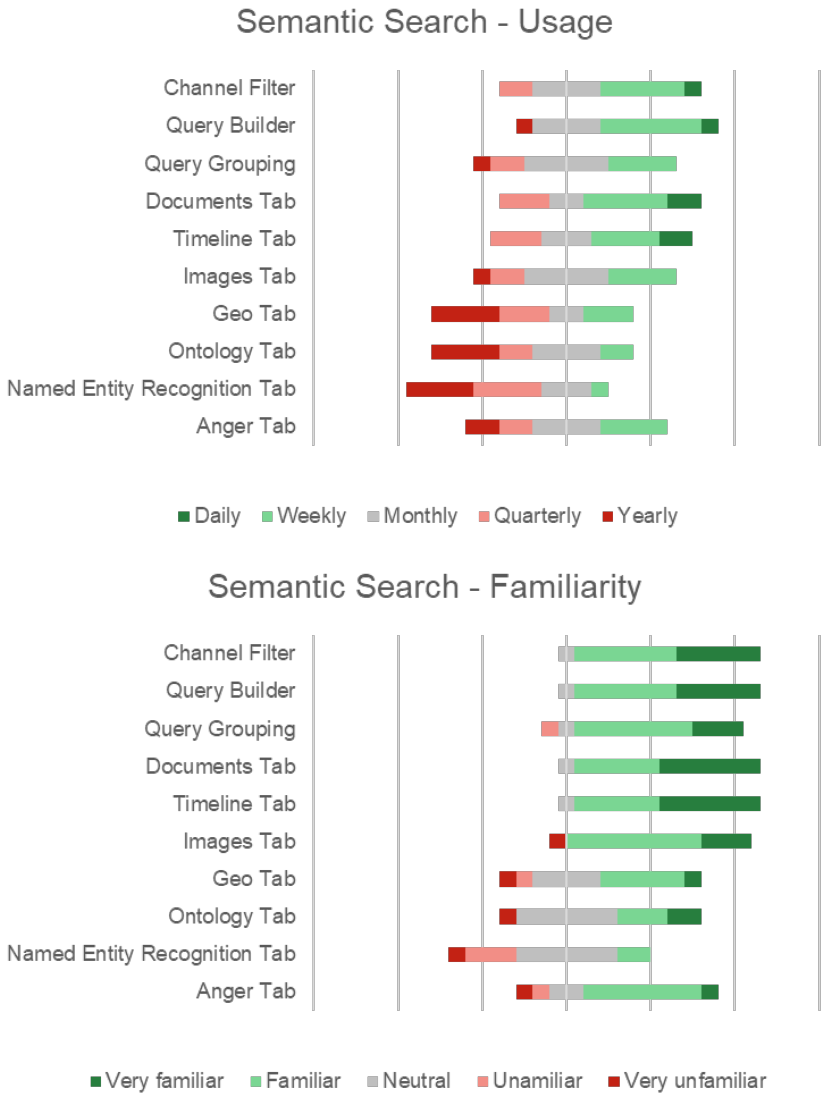


Figure 4.18: User Feedback on Interface Elements.

Disappointingly, features linked to the semantic enrichments presented in this chapter are not used in any significant manner, with the *ontology* field not used at all within the study, and the *contains* flags used sparingly. Figure 4.18 presents answers from the survey relating to usage and familiarity of interface components. It is also seen here that the use of the semantically driven components (the ontology, NER and anger tabs) are among the least regularly accessed components of the

interface, which is reflected in the user's lower familiarity with these tools. However, the survey results do suggest that some users may still find these features useful, as 50% of users access the *ontology* tab at least once a month, rising to 66% access to the *anger* tab at least once a month.

Table 4.15 breaks the usage logs down by user, linking this information to averaged scores from their survey responses, with a score of 5 being given to the most positive responses (Daily/Very Familiar) and 1 to the lease positive (Yearly/Very Unfamiliar). We have ordered the data based on their reported usage of the Semantic Search tool, followed by the number of sessions recorded. We have broken down the query types identifying two subroutines that occur within a session, paging, and refining. Paging occurs when a user resubmits an identical query to the *parse* endpoint and accounts for 11% of all query submissions recorded, whilst refining occurs when a query differs from its predecessor in a session occurring 61% of the time. The remaining 28% therefore are the first query sent in a session. The use of Boolean operators is present in 55% of all submitted queries, with 20% of these queries (12% overall) consisting of complex queries that use both types of Boolean operator within the query.

	Survey			Usage				Complexity		
User	**Exp.**	**Fam.**	**Use.**	**Sessions**	**Queries**	**Paging**	**Refining**	**Basic**	**Boolean**	**Complex**
Avg.	-	3.9	3.5	25.3	93.5	10.2	57.1	41.5	40.4	11.6
Pct.	-	-	-	-	-	11%	61%	44%	43%	12%
A	<6M	2.9	4.2	43	129	17	69	45	58	26
B	>Y	4.3	4	43	265	14	202	48	205	12
C	<3M	3.7	4	5	6	0	1	5	1	0
D	>Y	4	3.9	33	127	15	79	27	97	3
E	>Y	3.5	3.8	75	342	14	248	334	8	0
F	>Y	4.3	3.7	14	52	12	23	33	19	0
G	>Y	4.1	3.6	25	66	3	37	16	43	7
H	>Y	4.7	3.4	29	174	30	111	47	85	42
I	<Y	2.9	2.8	13	31	13	7	24	4	3
J	<6M	3.8	2.8	11	21	4	7	18	3	0
K	>Y	4.2	2.7	20	74	11	45	13	35	26
L	<Y	3.8	2.7	16	30	1	11	6	14	10
M	-	-	-	49	119	4	63	54	50	15
N	-	-	-	47	138	28	64	6	60	72
O	-	-	-	20	66	8	38	53	13	0
P	-	-	-	16	44	3	25	26	18	0
Q	-	-	-	10	59	13	37	28	31	0
R	-	-	-	8	26	3	14	4	18	4
S	-	-	-	4	8	0	4	2	6	0

Table 4.15: Query Routine and Complexity by User.

Two of our most inexperienced users report the highest frequency of use out of the survey group, which is to be expected as they could not have been using any components of Semantic Search less

frequently than the time that they have had access to the system. Despite this User A's log data shows that they already have a higher-than-average usage of the Search tool, and that they are using complex Boolean searches regularly.

We observe that some users heavily favour the use of basic querying, where a single feature is searched over at one time. User E is the greatest example of this, 97% of all their queries consisting of basic queries. They also exhibit the highest number of total overall sessions and queries suggesting their use of the search tool is more focused on the regular scanning of a set of single topics. Their reporting of a low familiarity of the Semantic Search features relative to other survey respondents suggests this is a key reason for the lack of more complex querying.

We test this hypothesis by looking at the correlations between reported familiarity and use of types of queries. We see a correlation of 0.53 when comparing reported familiarity to the use of Booleans within a query using Spearman's Rank Correlation Coefficient (Gautheir, 2001). We also see a weak correlation of 0.25 when comparing reported familiarity against the number of queries performed. This jumps to 0.52 when we remove User E from the dataset, a statistical outlier when looking at query volume, showing there is a general growth of understanding among the user group as usage increases.

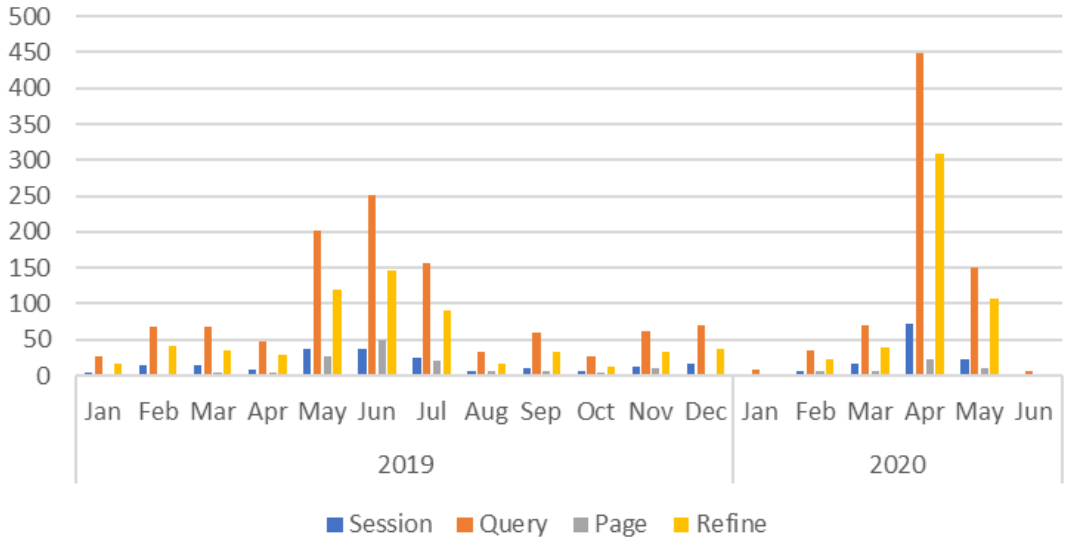


Figure 4.19: Usage Frequency from Logs.

Figure 4.19 presents the usage frequency over the 2019/2020 logging period, highlighting the "bursty" nature of the tool's use. This was driven by two operational foci: the 2019 European elections and the emergence of COVID-19. This shows the nature of engagement from the Sentinel user base, whereby activity is not consistent across a year. The recent burst of activity in the first half of 2020 may account for some of the discrepant reporting of frequency of use within the survey, as

the survey was performed shortly thereafter. But this also adds credibility to the survey, as all respondents have been actively engaged with the system just prior to questioning.

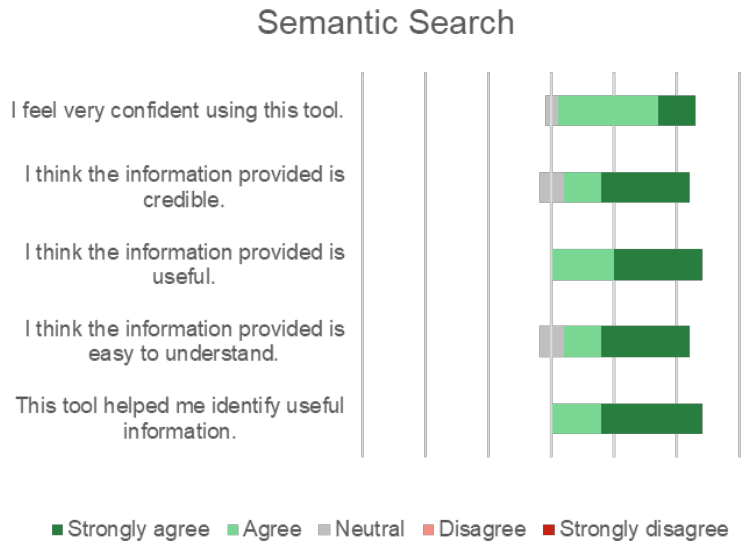


Figure 4.20: User Feedback on Semantic Search Tool.

We can see through the user feedback survey and submission logs that the semantic features available within the search tool have a higher barrier of entry and understanding. Figure 4.20 presents the results from a feedback question of the survey asking users to assess their confidence in the information returned to them via the Semantic Search tool. Whilst there may be features within the search tool that users do not feel confident with using, it can be seen that there is a strong satisfaction within the user base that the developments made in this Chapter do provide users with credible information, that is understandable and useful to their research goals.

4.6 Discussion & Conclusions

This chapter covers the evolution of the Sentinel Pipeline and its supporting interfaces away from the scan and monitor functionality of the prototype and alpha versions presented in Chapter 3, and into a beta system that gives users the ability to delve into the collected data using their own knowledge and experience as guidance. We have presented a number of semantic enrichment modules that take out-of-the box NLP techniques and effectively incorporate them into the pipeline, again highlighting the versatility of the architecture design presented in Chapter 3. These enrichments also successfully provided us with the necessary tools to explore and develop our own Machine Learning models that can exploit expert knowledge present within the wider research team.

The Escalation rule classification proved interesting, and it is unfortunate that we have not been able to develop this classifier further into a streamed environment. We also feel that these rules could be

folded back into the Anger classification tool as features as a means of boosting accuracy and allowing the classifier to better represent the *conflict action* model that preceded it. The anger classifier itself showed an acceptable degree of accuracy, to the point where it can be used diagnostically within an event to show shift-change in mood but is less suited to confident single-document attribution when comparing it to performance scores observed in Chapter 2. The creation of a new training corpus sampled from within the collected data may improve the performance of the classifier and will be performed in future work.

The work represented in this chapter reflects on how iterations of the co-design lifecycle allow stakeholders to build upon previous developments and to dictate new directions of research. It is at this stage of development, as initial tools and interfaces are being fed back into event planning through *workshops*, that we see the focus of robustness move away from data retention and quality and towards functionality and data accessibility. Usability also becomes more of a focus of the co-design efforts at this point. Stakeholders become more confident in how they wish to interrogate datasets and find their own idiosyncrasies in doing so, as evidenced by the varying but repetitive behavioural patterns seen in the Semantic Search usage logs.

The Semantic Search tool again highlights the flexibility of the architecture design, showing that the Sentinel platform can provide multi-modal access to the underlying collected data. We were able to build a robust interface that supports a large number of users within the stakeholder group, allowing them to easily build complex Boolean queries. The usage statistics presented in Section 4.5.4 indicated that the enrichments were underutilised within the search feature. Better utilisation of these enrichments became a driving motivation during the next iteration of *workshops*, the developments of which are presented in the next Chapter which is focused on the downloading and batch processing of search results.

The anger and escalation classification tools were developed after the identification of a gap in the literature surrounding emotion in Chapter 2. Sentiment analysis is a robust and well established area of research, with a number of studies in our surveys engaging with emotion dictionaries as part of their sentiment analysis process (Vilares et al., 2014, Win and Aung, 2017, Benkhelifa and Laallam, 2018, Alharthi et al., 2018, Subramani et al., 2018). But, we saw very little in the way of actual classification of emotion within the systematic review, with only one study performing focusing on emotion (Steed et al., 2015). Case study of the Woolwich terrorist attack provided the platform and insight to develop the anger classification tool, while the series of terrorist attacks observed in 2017 provided idea datasets to examine our novel approach to escalation content, through the identification of adversarial violent verb phrases.

CHAPTER 5: CORPUS CREATION *DOWNLOAD AND PROJECTS*

5.1 INTRODUCTION

Chapter 4 described implementation of a semantically enriched search tool that allows social science researchers to express their information needs and efficiently identify relevant documents. We showed in Chapter 3 that the data collected via the Sentinel Pipeline has been repeatedly used in qualitative research targeted at understanding the dynamics and content present in social media.

In Chapter 1 we discussed how we want to be able to engage users in explanatory and exploratory sequential analysis (Creswell et al., 2011), where the information that the Sentinel Platform provides allows a researcher to navigate around the Sensemaking Loop of Situational Awareness (Pirolli and Card, 2005) when performing social media analysis.

In this chapter, we build upon the developments of Chapter 4 to produce a means of rapidly collecting corpora for further qualitative research. We focus upon both packaging up relatively small datasets for qualitative analysis and employing rapid pre-processing and data analysis to produce information that can help guide an analyst's understanding of a corpus of social media documents.

This allows us to incorporate an unsupervised learning technique known as Topic Modelling into the Sentinel Pipeline. Topic Modelling uses word distribution to identify a set of underlying topics present in a corpus along with each document's affinities to these topics, and can assist researchers in identifying documents with high association with specific topics in order to perform further in-depth study (Nikolenko et al., 2017).

We present two instances of topic modelling within this Chapter. First, we use it as an automated component of the download workflow, which acts as a heuristic topic modelling service to assist in further rapid qualitative analysis with a fixed topic count and a fixed set of corpora. Second, we present a more measured attempt at topic modelling social media content relating to a particular case study.

5.1.1 PRIMARY OUTPUT: SENTINEL PIPELINE PRODUCTION

One of the main outputs from this chapter is the initial production level version of the Sentinel Pipeline and OSCAR Hub. In addition to the Semantic Search interface presented in Chapter 4, we provide the user with two new interfaces; the Download Manager interface presented in Section 5.3.3 and the Project Interfaces that are covered by Section 5.3.4.

The core processing elements of the Download function and services are implemented using Django Commands. These are extensions to an interface which is designed to support repetitive or complex tasks run via command line invocation of the *manage.py* core script, so that the logic is fully supported by all of the context and configurations present within the Django app. The commands are all integrated into a RabbitMQ index allowing for requests to be quickly consumed and processed by the relevant chain of commands.

We also develop additional interfaces that plug into the OSCAR Hub web portal, giving users the flexibility to manage and explore datasets and the output information derived from the download execution pool. These interfaces are covered in the user feedback study first discussed in Chapter 4, along with an overall assessment of the Sentinel Platform.

5.1.2 COVID-19 CASE STUDY

The final part of this chapter showcases how we can advance our analysis methods and features through reflective analysis of an emerging topic, moving in a *bottom-up* direction through the sensemaking loop. We also use the case study as an opportunity to assess the functionality and usability of the major components in the OSCAR Hub web portal.

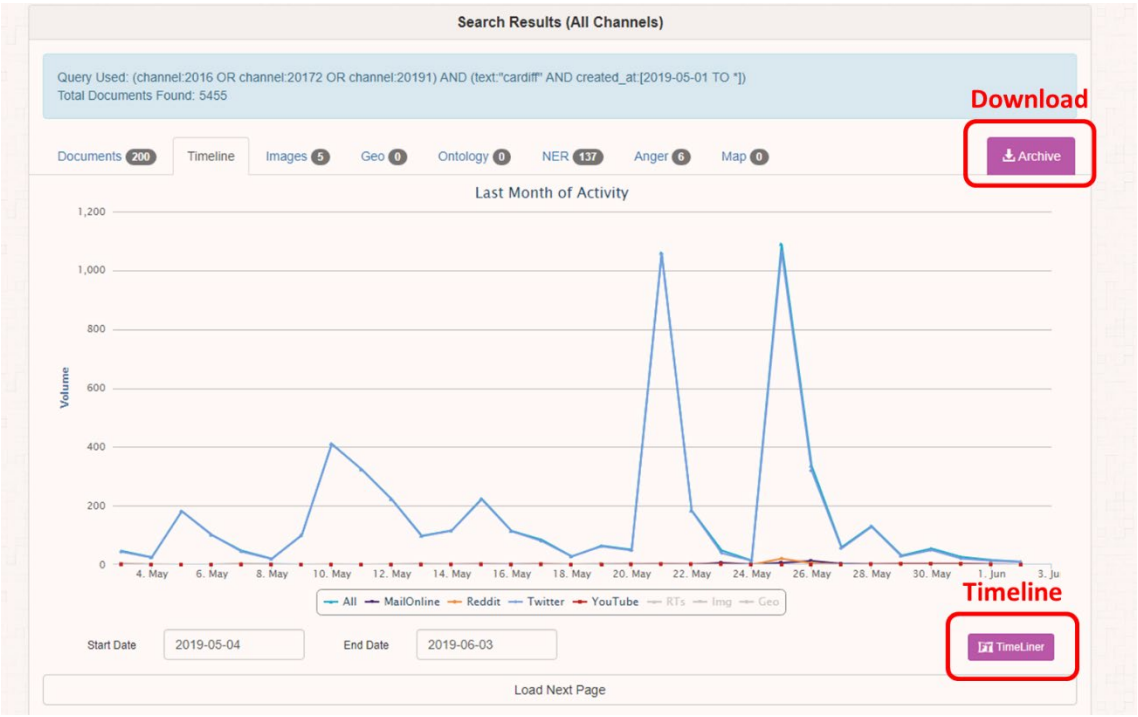


Figure 5.1: Semantic Search Download Buttons.

5.2 DOWNLOAD WORKFLOW

The download service is built to complement the Semantic Search tool presented in Chapter 4, allowing users to archive a corpus of documents matched via a semantic search query. In order to manage processing capacity, the download option is only offered when a semantic search query returns fewer than 10,000 documents. Upon clicking the download button (top right of Figure 5.1) the user is prompted to name the corpus before the search query is posted to the *commands* exchange to be picked up by the download command.

Where a Semantic Search returns a result set of over 10,000 documents, users are given the option of submitting a timeline download request (bottom right of Figure 5.1). In this case, the user is also prompted to input a date, this will form the end-date of the timeline which is sent along with the *query_string* to the *timeline command.*

The timeline command generates a series of download requests covering 3-hour windows for a 7-day period. These are all passed on to the download command and processed in the same way as any other download. Figure 5.2 presents the workflow components that make up the Download service.

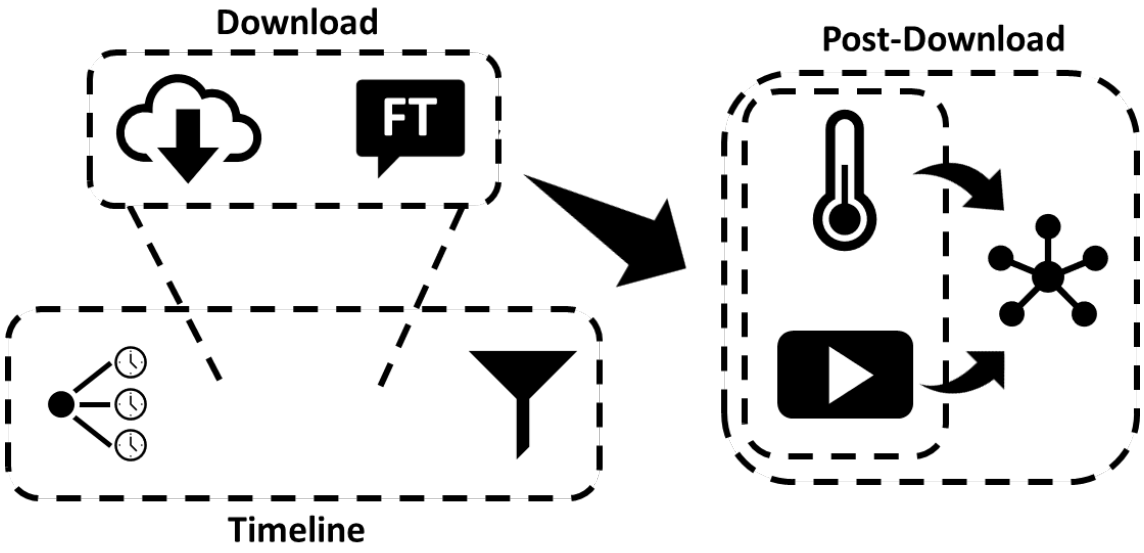


Figure 5.2: Download Workflow Components. The Download block provides the core download functionality and comprises of the download module (left) and the FlexiTerm module (right). The Timeline block consists of the timeline module (left) and aggregation module (right) that wrap a series of Download requests. The post-download block covers the additional analysis that is done once the core download processing has been completed and consists of the escalation module (top left) and the YouTube module (bottom left) that feed into the LDA clustering module (right).

5.2.1 DOWNLOAD

The download command's primary role is to serialise the relevant documents to a JSON file that can be passed to other commands and third-party tools as a complete archive of the corpus metadata. This is done using the same processes contained within the *query* and *retrieve* endpoints presented in Section 4.5.2, with the *query_string* being used to interrogate ElasticSearch, and the resultant list of document IDs and *channel* IDs used to retrieve the documents from MongoDB.

In addition to the generation of the JSON file, the supporting documents are created which are used to both highlight semantic features found within the corpus and to provide the end user with additional insight into the corpus via a series of summary web pages that are described in Section 5.3.1.

We perform entity extraction on all Tweets included in a download corpus, creating a series of Comma Separated Variable (CSV) files that list the frequency of user mentions, hashtags used, URLs present, and domains referenced in the URLs present.

A CSV file is also generated that covers all Twitter authors present in the corpus containing the total number of matched documents belonging to the author, the age of the account relative to the first and last of their corpus Tweets, and the number of followers and accounts they are following relative to their first and last Tweets in the corpus.

In addition to Entity Identification, we also generate an Excel Open XML Spreadsheet (XLSX file) that contains four separate sheets that together cover abridged metadata for all the documents present within the corpus:

- **User Sheet** – All Twitter user accounts. Information includes account name, handle, id, and description. Along with this information are the same computed fields that were generated for the user CSV file.
- **Articles Sheet** – All Reddit articles, Mail Online articles, and YouTube videos that the documents in the corpus belong to. Fields in this sheet include article title, article author, and number of comments belonging to the article.
- **Tweet Sheet** – All Tweet documents from the corpus. Fields cover document ID, user, text, author age, follower and following counts at time of posting, links, ontology elements, and timestamp.
- **Document Sheet** – All other documents from the corpus. Included in the fields document ID, user, text, and timestamp.

5.2.2 FLEXITERM & AGGREGATION

We adapt the FlexiTerm tool (Spasić et al., 2013) to plug into the RabbitMQ *command* exchange to receive notification of the completion of a download. Once notification is received, the JSON file is loaded into FlexiTerm and analysed, with the result files added to the download's folder.

Following completion, notification is passed to an aggregation module that, in the case of a set of Timeline downloads, listens for the final download to be processed by FlexiTerm before creating top-level versions of all the CSV produced in Section 5.2.1. Additionally, the aggregation module modifies the output HTML file by appending the data from the CSV files in order to form one of the output tabs presented in Section 5.3.1.

5.2.3 POST DOWNLOAD PROCESSES

At this point the core processing for a download is considered complete, and the data is made accessible to users. Notification of this completion is passed back to the *command* exchange, where a further series of post processing modules are run over the JSON corpus.

The **Escalation Tagging** command applies the Escalation Classifier described in Chapter 4 to the collected data and creates an XLSX file containing all of the documents present in the corpus that contain either an Adversarial Violent Verb phrase, Modal Violent Verb Phrase, or a Slurring Noun Phrase. Matches are presented in the same fashion as the tweet sheet from the Download XLSX with additional columns that list the PoS tags that matched and the violent verb or identity group and swearword that caused a rule to match. The matching text is also highlighted within the spreadsheet for user convenience (Figure 5.3).

@XXXXXXXXXX @XXXXXXXXXX They are in on it! They wanted to keep us locked down, but we fought for Freedom. So now they will burn our businesses. All deaths should be on them along with Soros who paid them $$ & Obama who brought the people in to riot.	[{"synset_value": "burn", "tokens": "they will burn", "pos": "PRP3 MD VB", "synset_name": "burn.v.15"}]	
if you're protesting peacefully and you see a police helicopter following you go the other way. today they followed us and then tear gassed us while being peaceful. that's when violence started but mainstream media will tell you other wise #BLACK_LIVES_MATTER #GeorgeFloydProtests		[{"synset_value": "tear", "tokens": "they followed us and then tear gassed", "pos": "PRP3 VBD PRP1 CC RB VB VBD", "synset_name": "tear.v.03"}]
These rioting criminal thugs are going to go down hard if they do not stop. Law abiding citizen are going to start defending themselves and it won't be with machetes. Is what Soros is paying them worth dying for? Acting out in Hatred & Anger will not end well for anyone.	[{"synset_value": "defending", "tokens": "start defending themselves", "pos": "VB VBG PRP3", "synset_name": "maintain.v.08"}]	
@XXXXXXXXXX The #TrumpRegime is loosing it's cool with peaceful protesters. Today I am calling on the nations of the world to stand with the peaceful protesters in the #US, who fighting for their life. We must be arming & funding them to defend themselves against the dictator #Trump.	[{"synset_value": "defend", "tokens": "defend themselves", "pos": "VB PRP3", "synset_name": "maintain.v.08"}]	
THE MASONIC TRUTH ABOUT POLICE BRUTALITY, RIOTS, & THE NEW WORLD ORDER A... https://t.co/wx45JIRTm0 via @YouTube they want us to fight just stop demand the real arrests of hrs Obama Bill Gates George Soros we need to do this right		[{"synset_value": "fight", "tokens": "they want us to fight", "pos": "PRP3 VBP PRP1 TO VB", "synset_name": "fight.v.02"}]

Figure 5.3: Example Content from Escalation XLSX File.

The **YouTube Retrieval** command takes the list of shared links produced as part of Section 5.2.1 and identifies any YouTube videos present within the list. It then polls the YouTube API in order to collect the title, description, author, and view count.

5.2.4 TOPIC MODELLING

The final workflow component is a topic modelling command, that performs rapid cluster analysis on several of the sub-corpora that are identified throughout the download process in order to assist users in identifying core topics present within a downloaded set of documents. The three sub-corpora of *targeted text* used are:

- Tweets from accounts that have an age of 0 within the download (named **Zero-Day accounts**). Derived from data identified in Section 5.2.1.
- **Escalation Tweets** identified by the Escalation command.
- **YouTube Video descriptions** retrieved using the YouTube collection command.

We use a Latent Dirichlet Analysis (LDA) method from the unsupervised semantic modelling package Gensim (Řehůřek and Sojka, 2010), to perform topic modelling on all three sub-corpora individually and collectively. LDA is a generative probabilistic model of a corpus where documents are clustered to a pre-set number of latent topics which are characterized by word probabilities (Jelodar et al., 2019).

In order to perform LDA analysis on a sub-corpus, we clean the text by removing URLs, social media specific syntax, alphanumeric removal, and stop word removal. We then tokenise and lemmatise all remaining text into a dictionary of words and their frequencies. Next, we add all bi-grams and tri-grams to the dictionary before filtering the dictionary down by removing all words and n-grams that have a frequency of less than 20 or are present in more than 50% of the documents. LDA clustering is then run on a bag-of-words representation of the dictionary.

Importantly we do not remove single character words or numbers from the dataset during pre-processing due to a co-design driven decision, motivated by the user group's research interests (Innes et al., 2021). The emergent 'QAnon' conspiracy theory is having significant effect on US and international politics (Garry et al., 2021), and as such the term 'Q' can be considered a valid token, especially when incorporated into bi-grams, Both the emergence of the Covid 19 pandemic (Van Bavel et al., 2020), and instances of far-right symbology such as '1488' that combines reference to David Lane's '14 words' and 'Heil Hitler' ('H' being the 8th letter of the alphabet) (Conway et al., 2019), dictate that numbers should also remain valid tokens within the bag-of-words.

The maximum number of available documents to cluster will be 100,000, the number of sentences in a sub-corpus is considerably lower as highlighted by Figure 5.4, which presents the proportion of sub corpora to the overall corpora sizes for 204 downloads run by users between 08/02/2019 and 03/08/2021. The bottom cluster of points show downloads where there is a 1:1 ratio between download volume and sub-corpora volume, which occurs when a sub-criterion forms part of the

download query (i.e., a query only retrieving posts from zero-day accounts). When discarding these sub-criterion specific downloads, we see that the average size of the sub-corpus being topic modelled is 0.75% of the total volume.

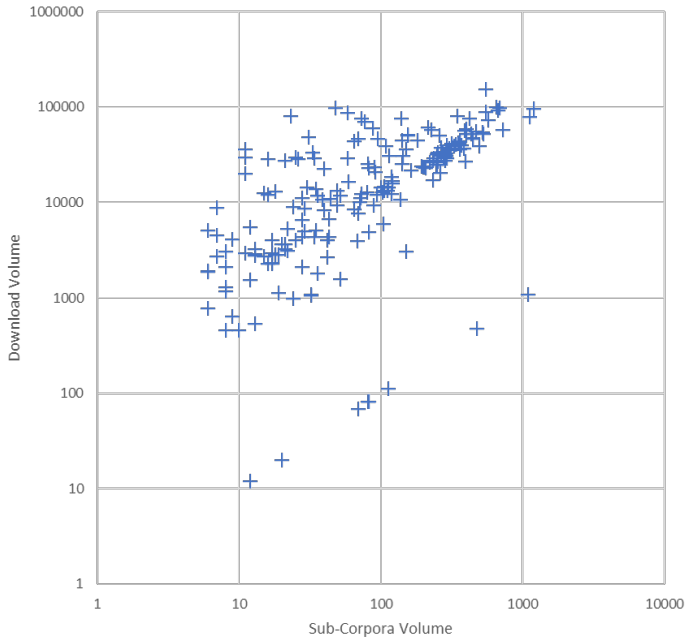


Figure 5.4: Download Volume vs Sub-Corpora Volume.

LDA clustering requires the number of topics to be defined before the algorithm can be run. Miller (1956) explains that human judgment and memory impose limitations on the amount of information that they are able to receive suggesting that seven, plus or minus two, is the ideal number of options or concepts to present to the user. We take the upper bound of this heuristic as the maximum number of topics and run 5 iterations of LDA clustering across a range of 2 to 9 topics, selecting the most coherent model as the final model presented to the user. Topic coherence can be measured by a series of measures (Newman et al., 2010). We used the C_V coherence measure, which has been shown to perform best correlated to human perception of coherence (Röder et al., 2015) and is one of the supported coherence measures found in the Gensim semantic modelling package (Řehůřek and Sojka, 2010).

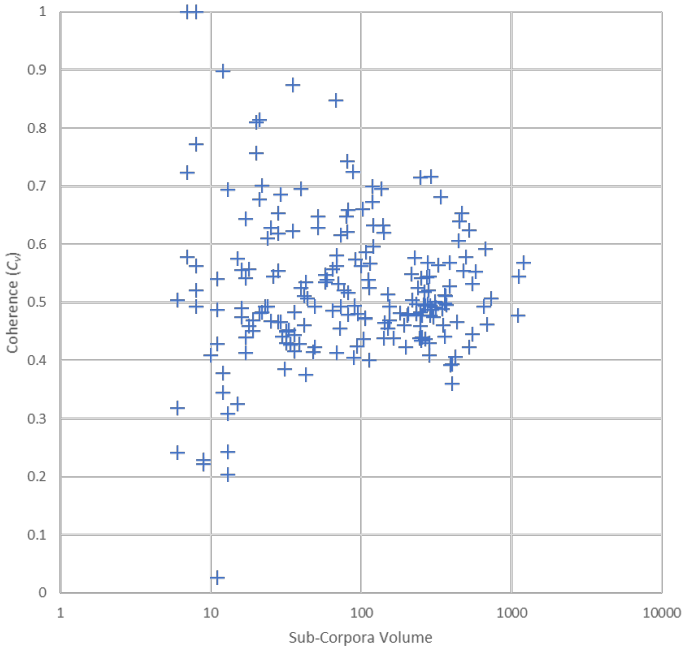


Figure 5.5: Coherence (C_v) vs Sub-Corpora Volume.

Figure 5.5 shows us the coherence (C_v) values for each of the sub-corpora's best performing models, again for downloads run between 08/02/2019 and 03/08/2021 by system users. It can be seen that while the majority of the models have an acceptable coherence, with an average across all best performing models being a C_v of 0.52, the models built on a low volume of documents have a higher degree of variation. We consider this acceptable as the LDA clustering is here being deployed as a means of characterising medium to large datasets, and so these low volume datasets can be interrogated manually.

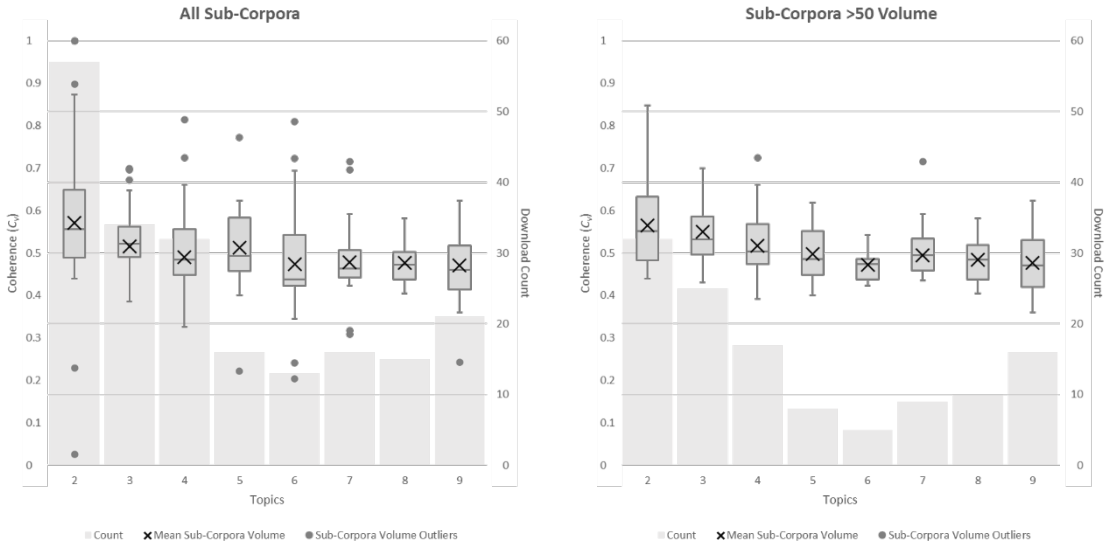


Figure 5.6: Coherence (C_v) vs Topic Count.

Figure 5.6 presents the best performing model coherences against the number of topics that model was run against, with all results presented on the left, and only those of model run on a corpus size of greater than 50 on the right. We see that the smaller topic counts are favoured, although the number drops once the sub-50 document models are removed, but models appear to perform acceptably regardless of the number of topics selected.

The final model with the best performing coherence is finally passed to a visualisation tool called pyLDAVis, that produces an embeddable interface that allows users to actively interrogate the model (Sievert and Shirley, 2014). We repeat this process for all three sub-corpora, and a combined corpus of all *targeted text*. This output is presented as part of the Download Manager interface presented in the following section.

5.3 INTERFACE

In order to move the overall Sentinel System into a production state, we developed two new sections of the OSCAR Hub; the Download Manager and a series of Project Pages, which are both presented in this section. Both of these interfaces provide ways for the user to navigate through the collections of downloaded datasets and the information created by the download workflow that relate to each dataset through multiple interactive tabs.

There are two types of download page that reflect the two types of download request, with the Search Based Download (SBD) page covering single search downloads that are made directly from the download module, and the Timeline Based Download (TBD) page covering requests made through the timeline module. We first present the download pages and their tabs, before discussing the two new OSCAR Hub sections.

5.3.1 SEARCH BASED DOWNLOAD

The core interfaces developed to present the information derived from the Download Workflow are covered within the SBD page, Figure 5.7 presents its top-level components. The breadcrumb menu (top left) that links the user back to the OSCAR Hub homepage, the Download Manager page (Section 5.3.3), and when applicable the download's parent folder. The management menu (top right) allows users to share access to a download, move the download to a different folder, and delete the download. The *document tray*s (bottom right) are initially loaded with the most recent documents that match the download's query, retrieved via the same endpoints used in Section 4.5.2. There are some differences to the Semantic Search *document trays* in that the NER and Ontology *document trays* have been merged to make space, we have added a Zero-Day *tray*, and the anger *tray* has been expanded to contain Escalation documents also. Finally, the core download tabs

(bottom left) present the computed information derived from the modules presented in Section 5.2.1.

Figure 5.7: Search Based Download, Tabs and Menu Options.

The *summary* tab that acts as the default tab within the Search Based Download interface. It is broken down into a number of sections that provide quick summarisations of information generated, and access to more detailed files that same information.

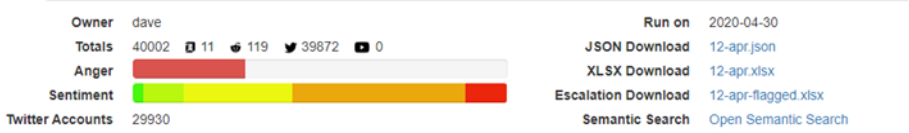


Figure 5.8: Summary Tab, Aggregate Summary and Documents Pane.

Figure 5.8 shows the *summary pane* and the available downloadable document links that form the topmost part of the *summary* tab. The *summary pane* passes extended queries to the *query* endpoint to bring back counts for document type matches and anger percentage, and uses aggregated information generated in Section 5.2.1 to present the overall sentiment and the number of unique accounts in the corpus.

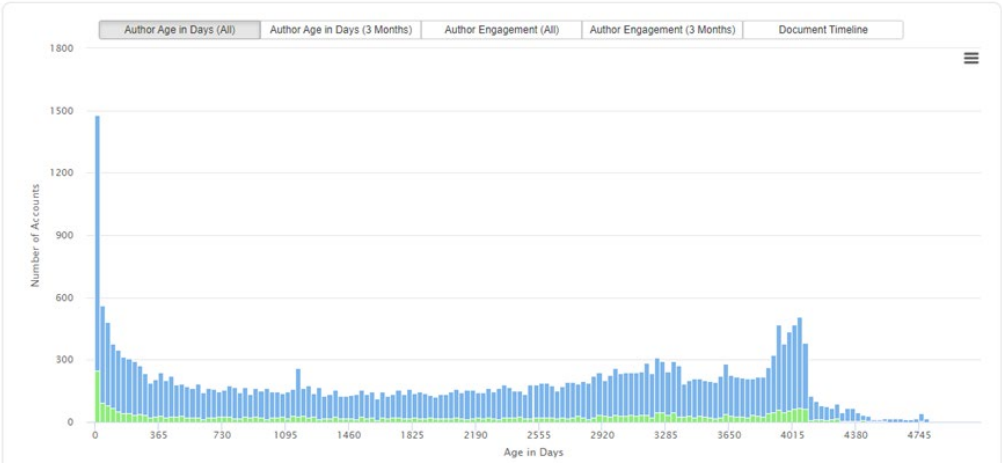


Figure 5.9: Summary Tab, User Histogram.

The user histogram visualises the distribution of author age (at time of first post) and can be broken down into 1-month bins covering the full userbase, as shown in Figure 5.9, and into daily bins covering the accounts younger than 91 days. Accounts that have multiple documents present in the download are also listed in a second series (green). Each of the bins are selectable and will re-load the documents shown in the *document trays* with documents belonging to the subset of users.

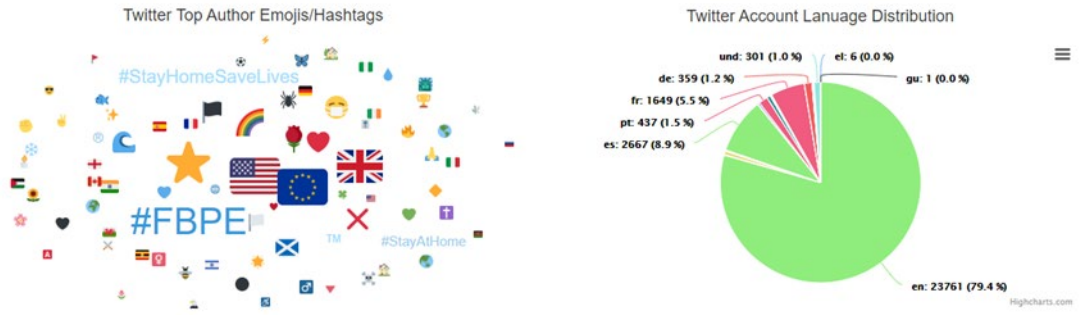


Figure 5.10: Summary Tab, Word Clouds and Language Pie Charts.

Figure 5.10 gives examples of the word clouds and language pie charts that are present on the *summary* tab. These provide the user with quick insight into some of the key entities and demographics present within a download. For the hashtag and author account word clouds, users are able to load related documents by selecting a word from the cloud, these are loaded into the *document trays*. The information present within the word clouds is reflected in more detail with raw numbers in the *details* tab along with the FlexiTerm output.

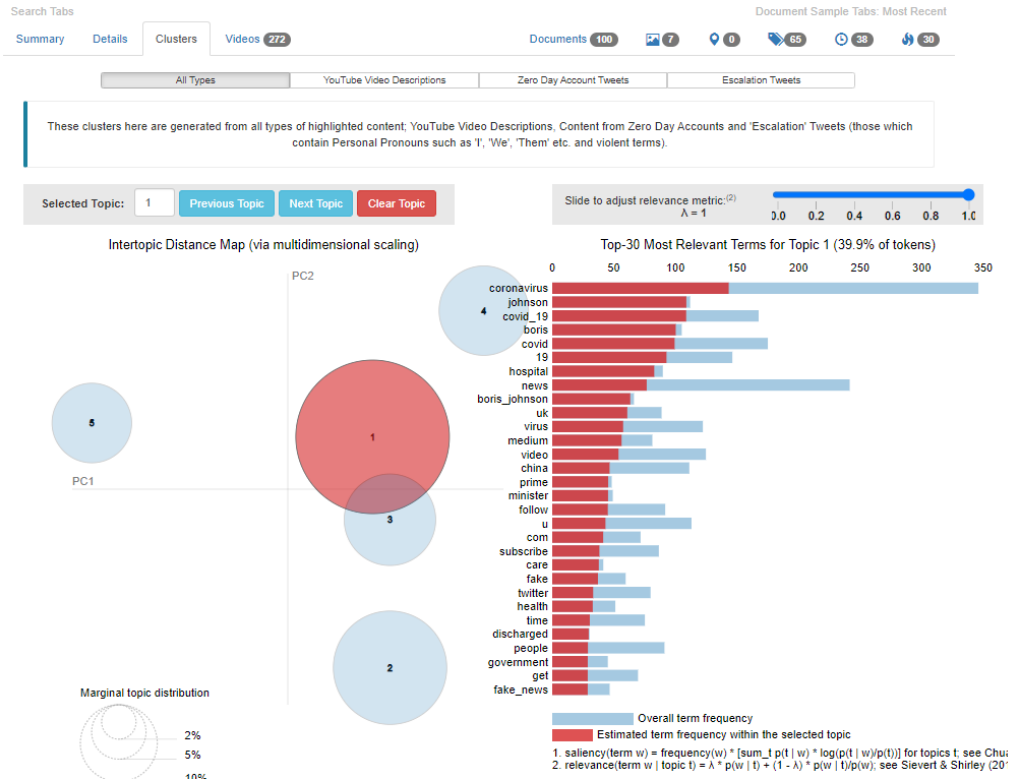


Figure 5.11: Topic Model Tab.

Figure 5.11 shows the *clusters* tab that presents the output from the LDA topic modelling that is performed in Section 5.2.4. Users are able to switch between the 4 LDA models that are generated for each of the downloads in, with the information presented using the pyLDAVis visualisation interface (Sievert and Shirley, 2014). The models are stored in the JSON output that pyLDAVis generates, and then loaded back into pyLDAVis using jQuery.

The pyLDAVis interface presents users with a two-dimensional Principal Component perspective of the topic clusters and allows users to interact with each topic, retrieving the most *relevant* terms for each topic. *Relevance* is a weighted average of the logarithms of a term's probability and its *lift* (derived from Taddy (2012)) with the weighting parameter λ being dynamically adjustable within the interface. Sievert and Shirley (2014) report that the "optimal" value for λ was found to be roughly 0.6 that resulted in an estimated 70% chance of current topic identification by users. We extended the pyLDAVis JavaScript code so that related Tweets are loaded into the *document drawers* when a user selects that topic in order to further aid in topic identification.

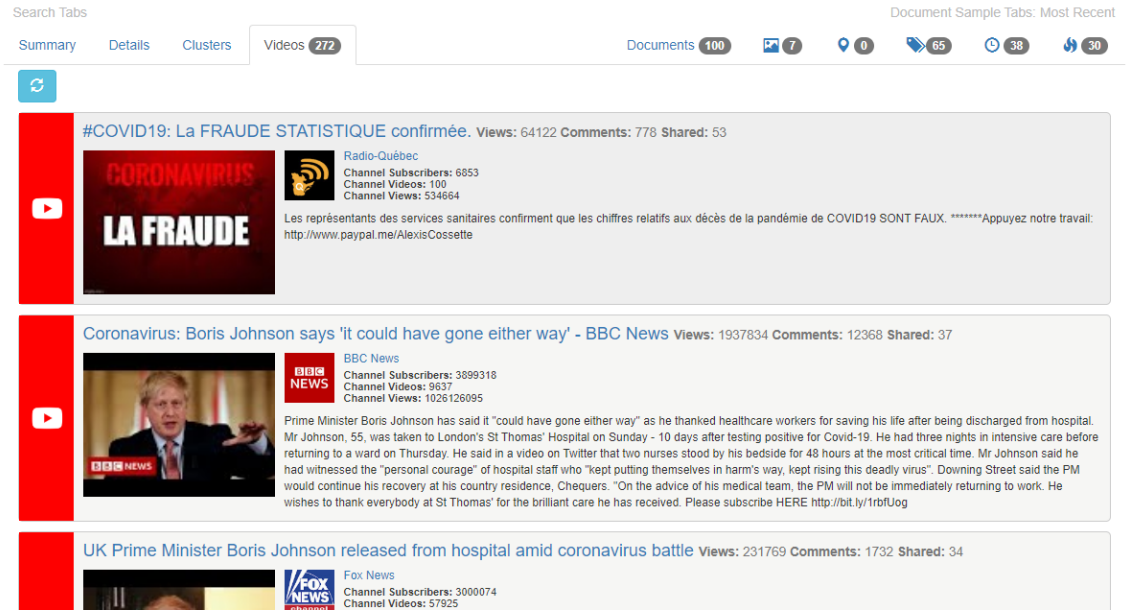


Figure 5.12: Videos Tab.

The *videos* tab (Figure 5.12) presents all the YouTube content that has been shared within the download dataset, listing the views, comments description and author name at the time that the YouTube Data retrieval was performed (Section 5.2.1). Each video also has the number of times the video appears within the document set.

5.3.2 Timeline Based Download

The TBD interface presents much of the same information as the SBD with some alterations to account for the fact that a TBD actually covers multiple iterative SBDs. The *summary*, *details*, *clusters,* and *videos* tabs remain the same, with an additional *graph* tab that is shown in Figure 5.13.

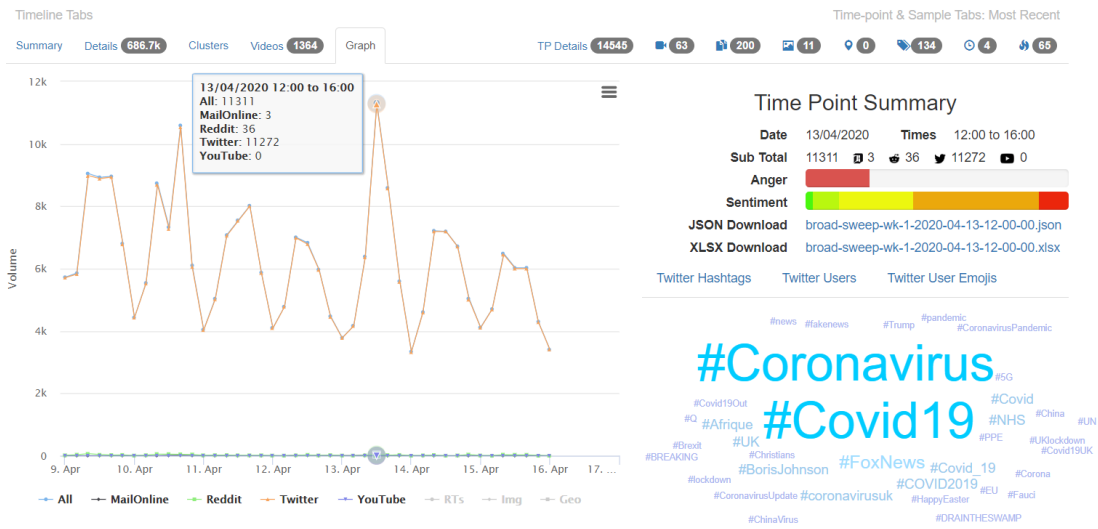


Figure 5.13: Graph Tab.

The graph tab uses the *timeline* endpoint from Section 4.5.2 to produce and interactive line graph. Users can click on any of the time points to load the same information that is found in the summary and documents pane, specific for the 4-hour window. Two extra tabs are found to the left of the *document trays* that provide specific a *details* tab and *videos* tab for the 4-hour window, and the *document trays* themselves are updated with the 4-hour window specific documents.

5.3.3 Download Manager

In order for users to be able to manage and access all of the processed downloads the Download Manager section of the OSCAR Hub was created. In here, users have access to all downloads ran by themselves, and any downloads and folders that have been made public by other users. SBDs and TBDs are distinguished by icons and a short description, as shown in Figure 5.14

Figure 5.14: Download Manager.

5.3.4 PROJECTS

This section of the OSCAR Hub is focused on exploiting downloads in a research project specific space to improve how the teams can monitor and inspect data collected through canned queries. Projects consist of a collection of Download folders, each of which contain TBDs run at weekly intervals, resulting in consistent collections of data contextualised via known query sets. Projects are found on the OSACR Hub homepage (Figure 5.15) and are visible only to members of each project.

Figure 5.15: Project Links on OSCAR Hub.

Clicking through to a project leads to the project dashboard (Figure 5.16), giving access to all the weekly downloads that are being performed for the project (accessible via the download folders), and a timeline summary graph that provides quick views of the social media volumes present within the project. These are a mixture of real-time queries made to ElasticSearch via the *timeline* endpoint from Section 4.5.2, and from aggregated information produced by the Download Workflow that is contained within each download.

In addition to the timelines that are generated from queries to the *timeline* module and from computed data, we have incorporated a timeline driven by the Global Data on Events, Location and Tone (GDELT) service, an online resource that provides API access to more than 200-million news articles (Leetaru and Schrodt, 2013). We query the endpoint for news article matches on any keyword elements present in the *query_string* of a download, receiving back a daily percentage of news articles present in the GDELT dataset that match our keywords.

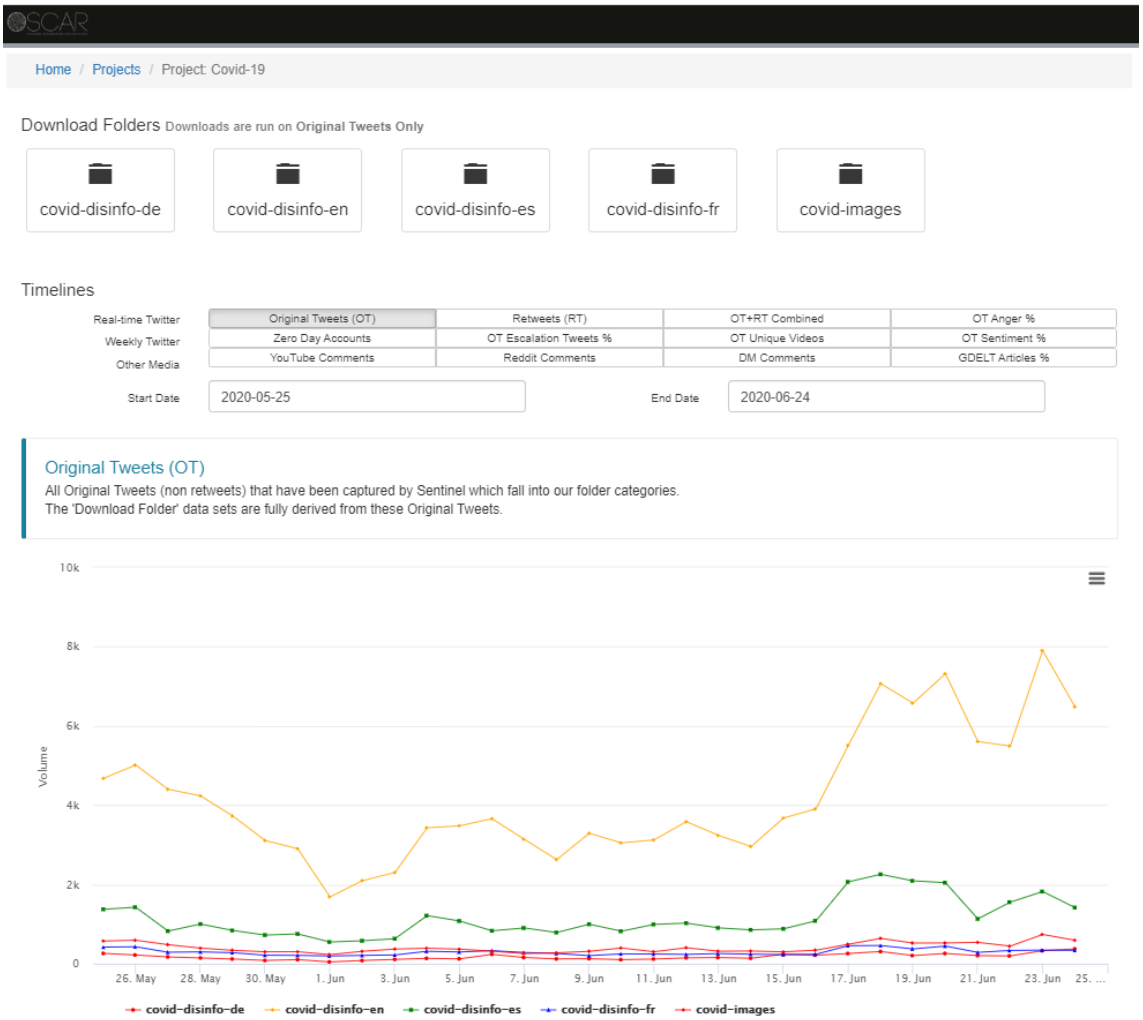


Figure 5.16: Project Dashboard.

Clicking on a timeline button will update the graph with the relevant timelines for each of the download folder queries, with an explanation of the data used to generate the timeline is available.

Selecting a folder will bring up a second dashboard that contains a volumetric timeline and a percentage-based timeline (Figure 5.17). These are populated with all the available timelines found on the project dashboard driven by the single search query that belongs to the folder's downloads.

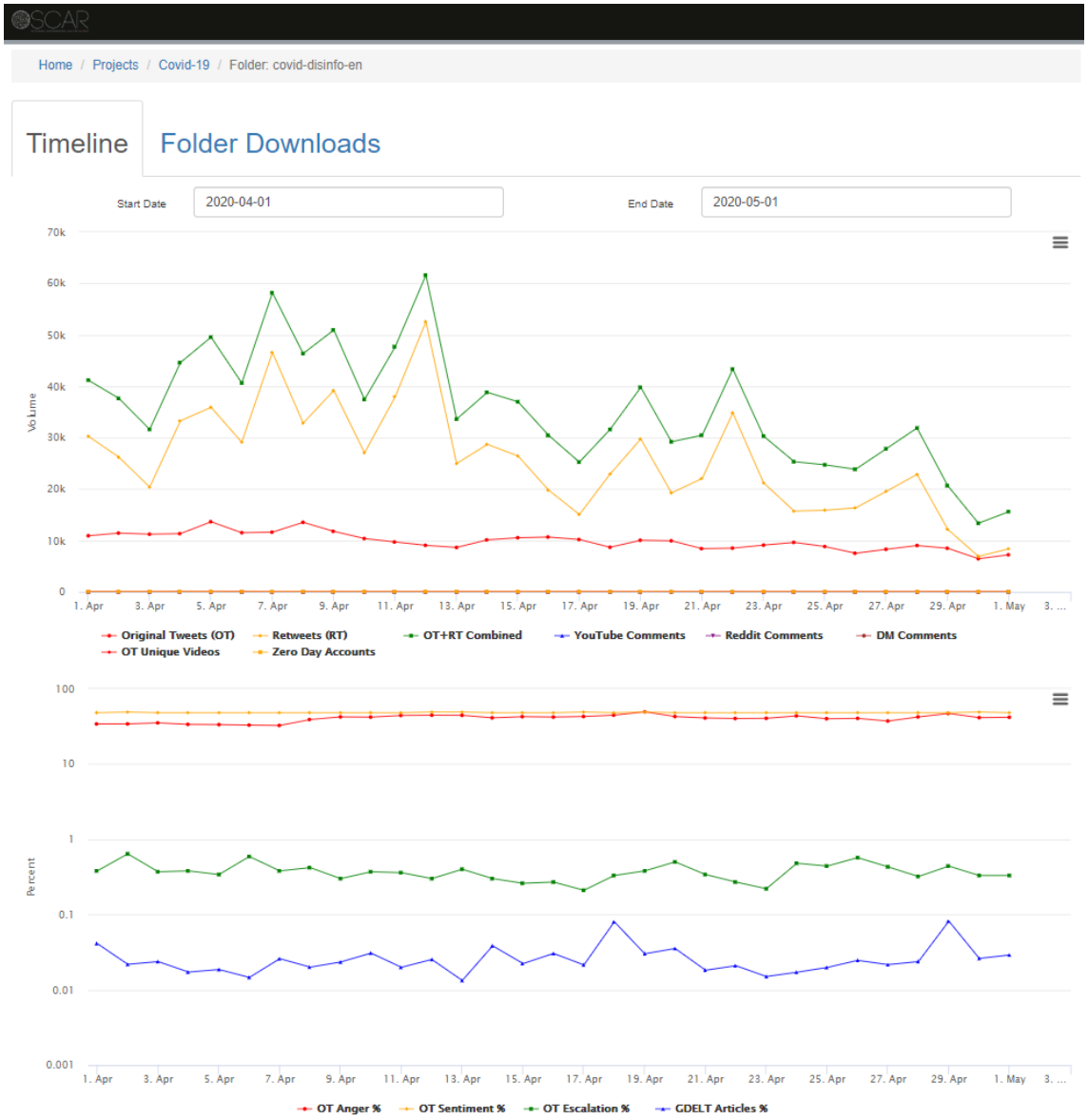


Figure 5.17: Project Folder Dashboard.

There is also a *folder downloads* tab that gives a list of all the downloads present in the folder, which will take a user through to the TBD pages of those downloads.

5.4 Usability Study

As reported in Section 4.5.4, we ran a Likert scale (Vagias, 2006) based survey on usage, experience, familiarity and understanding of the Sentinel Platform, with 12 stakeholders of varying experience (Figure 4.17), covering the majority of the userbase in May 2020. We focused the survey on the three main OSCAR Hub component interfaces that cover the existing interface functionality, and on how the Sentinel Platform as a whole relates to the 5W and Sensemaking Loop models discussed in Chapter 1.

Figure 5.18 presents the responses to six questions based around the 5W (*who*, *what*, *when*, *where*, *why*) framework presented in Roberts et al. (2015) which acted as a framing of how users wished to use Sentinel. Q1 and Q2 collectively relate to *who*, with Q3, Q4, Q5, and Q6 relating to *what*, *when*, *where,* and *why* respectively.

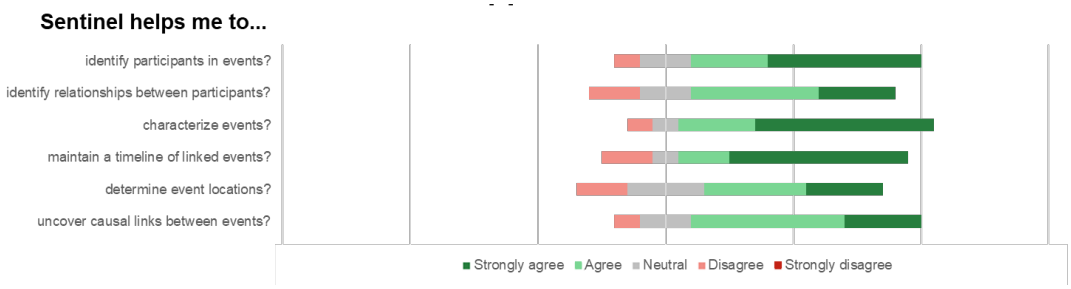


Figure 5.18: User Feedback on the 5Ws.

Responses suggest that Sentinel is perceived to perform best at identifying *what* events are happening and *when* they happen, and that it performs well at identifying *who* is participating in said events. These responses also highlight the gaps in Sentinel's performance with the poorer responses indicating that Sentinel is not as accurate at identifying relationships between those *who* participate in events, and characterising *where* events are occurring.

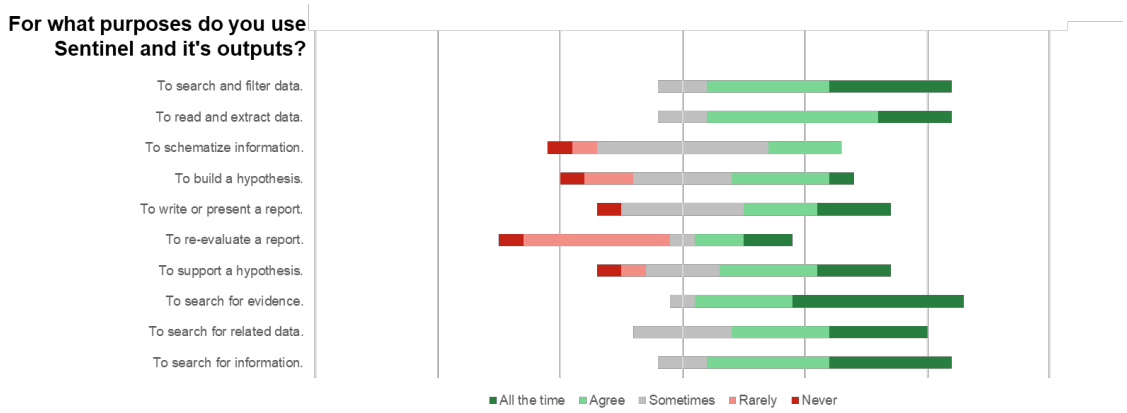


Figure 5.19: User Feedback on the Sensemaking Loop.

Figure 5.19 displays the questions asked to users that relate to how often users perform tasks defined in the Sensemaking Loop model derived by Pirolli and Card (2005) and presented in Figure 1.1. Q1 to Q5 link to the bottom-up process that takes data through to theory, and Q6 to Q10 link to the top-down process where theory is validated with data. As mentioned in Chapter 1, within the Sensemaking Loop model, there are a pair of sub-loops that bisect the two processes across the data/theory boundary. Answers to the questions in Figure 5.19 show that Sentinel's strength lies in supporting the *foraging loop sub-loop* of sensemaking, providing tools that aid in the development and validation of *data shoeboxes* and *evidence files*.

Figure 5.20 breaks down the 10 questions in Figure 5.19 across the three OSCAR Hub tools, asking the users which tools in particular they use to perform each task. The Semantic Search tool performs as expected and forms the core support of both bottom-up and top-down *foraging loop* activity. The Download Manager tool is also used regularly within the *foraging loop*, being the main process in developing *evidence boxes*, but interestingly it is also employed as part of the *presentation* process showing that Sentinel can support processes at both ends of the *structure* and *effort* axes. Unsurprisingly, the Project Pages are used sparingly across all tasks as they only become operational for institute level research, whereas the Semantic Search and Download Manager allow for more independent research.

	To search and filter data.	To read and extract data.	To schematize information.	To build a hypothesis.	To write or present a report.
Semantic Search	100.00%	50.00%	33.33%	50.00%	33.33%
Download Manager	41.67%	75.00%	33.33%	41.67%	58.33%
Project Pages	25.00%	33.33%	25.00%	41.67%	33.33%
	To re-evaluate a report.	To support a hypothesis.	To search for evidence.	To search for related data.	To search for information.
Semantic Search	25.00%	25.00%	91.67%	83.33%	100.00%
Download Manager	16.67%	41.67%	41.67%	41.67%	50.00%
Project Pages	25.00%	33.33%	33.33%	33.33%	41.67%

Figure 5.20: Reported Utilisation of Components Relative to the Sensemaking Loop.

Figure 5.21 presents the user responses relating to how users rate their agreement against a number of questions covering confidence in the tool, credibility of information, usefulness of the information, ease of understanding, and whether or not the users have identified useful information from the OSCAR Hub component tools.

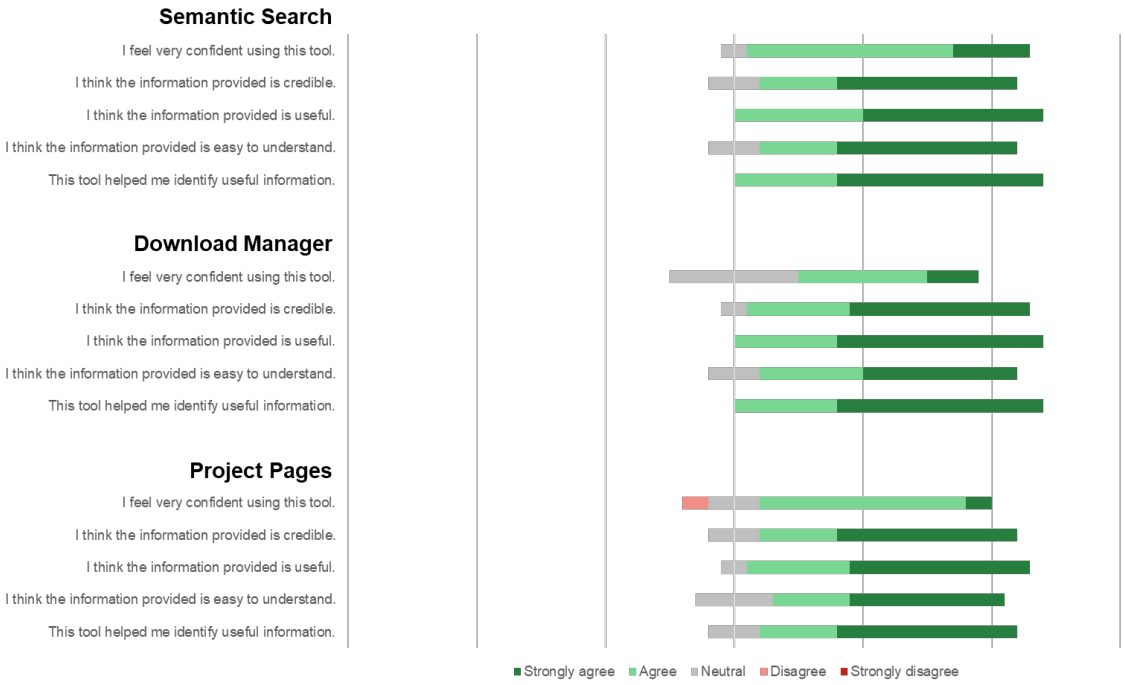


Figure 5.21: User Feedback on OSCAR Hub Components.

In addition to the Likert scale questions, users were provided with open-ended questions focused on the three surveyed components. Table 5.1 presents responses from users in relation to these open-

ended questions. These responses show users' comfort with both the Semantic Search tool and Download Manager tool within their work, highlighting in a number of responses again show the strength in engagement with the *foraging loop* of the *sensemaking loop* derived by Pirolli and Card (2005).

Do you have any final comments regarding the Semantic Search tool?
"A very useful tool but needs to be used regularly in order to become familiar enough with it to get best benefit."
"It's a really handy tool especially one that is purposefully built for analysing disinformation especially when a lot of the other programmes available are more geared toward brand and marketing analysis."
"Sentinel Semantic Search is what I use before a task, while doing research for the task, and as a revision at the end. It is the best data extraction source for all stages of my projects so far."
"Worth saying that the rapid sense making analytics support has improved significantly recently."
Do you have any final comments regarding the Download Manager tool?
"Very useful in conjunction with Semantic Search once you have settled on good search parameters for your purpose."
"The new features of the download manager have been super helpful in finding accounts of interest."
"Word clouds and graphics are very useful for getting a *(sic)* overview very quickly."
"This allowed me to extract my own data in a familiar Excel format and analyse it slow time later (at my own pace). This facilitates data coding and quantitative analysis. Not to mention it completely removes the need for slow and laborious manual data collection and analysis."
"The ability to create any number of bespoke datasets is really valuable in supporting a multi-method/platform approach."
Do you have any final comments regarding the Project Pages tool?
"Helpful if working on projects jointly with other team members."
"I have actually never used the project pages tool, I have always just used the download manager."

Table 5.1: Selected Feedback from Usability Study Open-Ended Questions.

Overall, the responses are clearly positive or very positive and there is strong reporting from stakeholders that all three tools provide credible, useful, and easy to understand information. Confidence in using the tool shows the most variability among answers, with confidence in the Semantic Search tool understandably the most positively reported which is likely due to it being developed and in use earlier than the other two tools.

5.5 USE CASE: APRIL 2020 CORONAVIRUS DISINFORMATION

On the 30th of January 2020, the World Health Organisation (WHO) declared a Public Health Emergency of Global Concern over the emergent COVID-19 Severe Acute Respiratory Syndrome coronavirus (World Health Organization, 2020a). By the 11th of March, the outbreak had grown to effect more than 100 countries and the WHO was forced to categorise the outbreak as a pandemic (World Health Organization, 2020b), with the UK Government announcing regulations pertaining to a partial lockdown of the country, coming into effect on the 26th of March (Public Health England, 2020).

This rapid series of developments, the unknown nature of the virus, and the unprecedented scale of multi-national lockdowns meant that the information ecosystem surrounding this event was highly volatile, with conspiracy theories, misinformation, and fake news proliferated widely on social media (Van Bavel et al., 2020). Sentinel has been regularly used by the CSRI research team to develop theory pertaining to rumours, propaganda, and conspiracy theories acting as "soft facts" surrounding spontaneous and chronic events (Dobreva et al., 2019, Innes et al., 2019, Innes, 2020). It was therefore inevitable that the COVID-19 pandemic would rapidly become focal to the CSRI research programme.

This section covers the building of a collection channel focused on disinformation surrounding COVID-19, how we used data from this channel to move our analysis beyond the automated features of the Download Workflow, and how user behaviour changed within Sentinel, during the month of April 2020.

5.5.1 CHANNEL ATTRIBUTES

The initial corpus of Tweets used to observe Zero-Day activity was built using collection terms focused upon disinformation ("fake news", "disinformation", "misinformation", etc.), prominent UK politicians (e.g., "Boris Johnson", "Kier Starmer", "Nicola Sturgeon"), and subsidiary Russian media outlets ("sputnik", "RT"). The objective of this collection is to focus on the calling out of fake news by Twitter users. Sakaki et al. (2010) describe users as being able to act as *social sensors*, where the tweets are regarded as *sensory information*, and so we take assumption that users will be tagging and replying to content with the statement that something is "fake news", "disinformation", "lies", etc. with a focus on topics relating to UK politics and the UK and Russian state media's focus on them in.

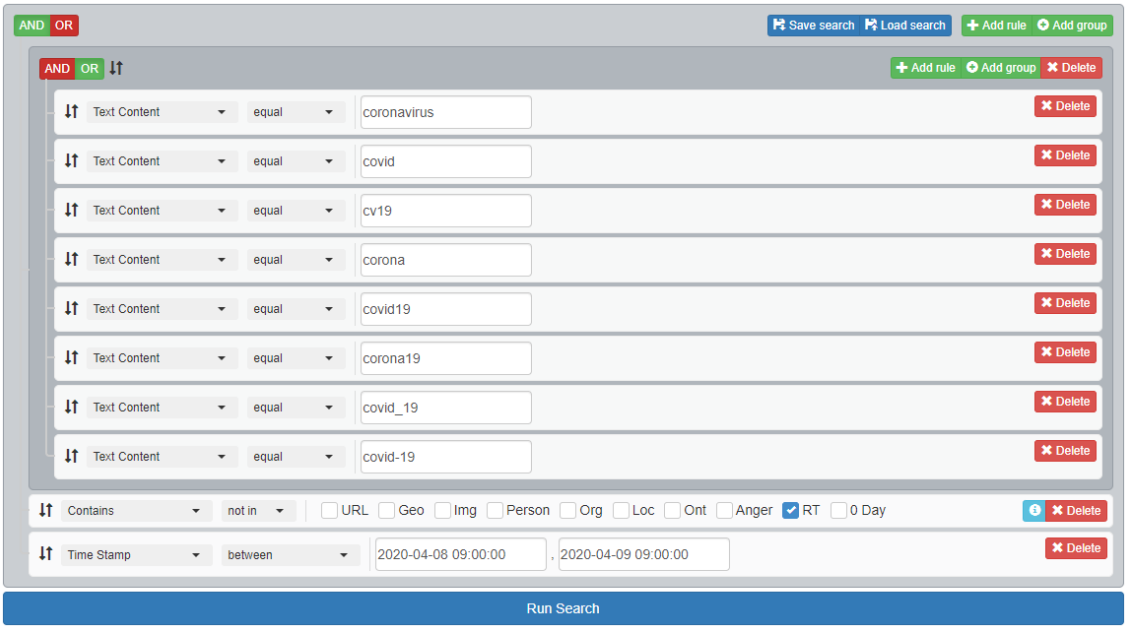


Figure 5.22: COVID-19 Query Configuration.

From this collected corpus, we retrieved any original Tweets mentioning "coronavirus" or synonyms of. These were broken down into 24hr windows (9am – 9am) that would allow for daily rapid assessment of key COVID-19 discussions on social media and were run daily across the month of April 2020. Figure 5.22 presents the query loaded into the Semantic Search query builder interface, whilst Figure 5.23 presents the timeline of Tweet documents retrieved from this query.

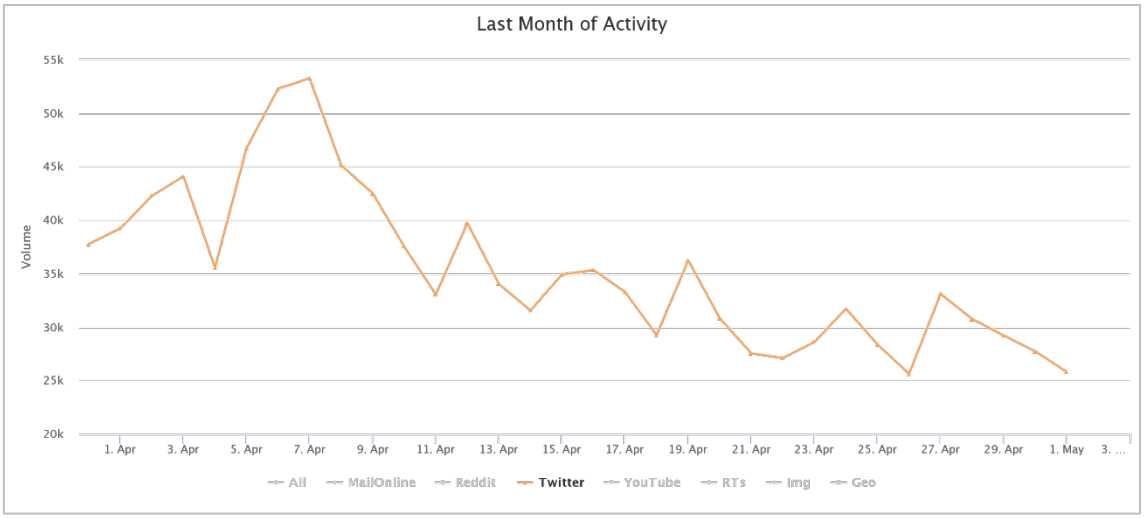


Figure 5.23: COVID-19 Query Timeline, April 2020.

5.5.2 Video Content Analysis

We took this opportunity to expand our understanding of YouTube content present within a corpus, and to expand our methods of analysis of YouTube data, with the intention of feeding any developments back into future versions of the Sentinel Pipeline. In Chapter 2 we saw only one study that used YouTube data, using SVMs to perform opinion mining on recipe videos (Benkhelifa and

Laallam, 2018). Thelwall et al. (2012) highlight that videos relating to news and politics provide natural sources of discussion, with a third of the top 100 videos in their study ranked by comment density falling into this category. Susarla et al. (2012) investigate the proliferation of videos through the YouTube network based on different forms of networked social interactions, showing how subscriber and friend networks influence the rate at which a video gains exposure. We feel that this use case provides an opportunity to investigate the types of news and politics videos that are proliferated outside of the YouTube network.

Over the course of the month of collection we identified 4807 distinct YouTube videos shared by the Twitter corpus. Of these 4,648 (96.69%) remained available long enough for us to retrieve their related metadata at 10am each day when the daily download was performed.

Whilst we perform some basic topic modelling on all of the downloads by default via the Download Workflow presented in Section 5.2, this use case gives us an opportunity to further analyse the content of the shared YouTube videos with a much larger dataset and to diversify the content used to cluster videos.

In order to better understand the content of the videos that were identified in the Apr2020 corpus, we retrieved all available closed caption texts for each video via the Google *timedtext* endpoint [47]. Captions can be added to a YouTube video by the video author and are generally provided for accessibility. We also chose to look at the reaction to videos, and to do this we collected up to 1000 comments belonging to the Apr2020 corpus via the YouTube API.

5.5.2.1 Corpora Collection

Figure 5.1 shows the steps taken to reduce the set of videos down so that we have three parallel corpora of descriptions, captions and comments covering the same set of videos. Each video needs to have a multi-line description, 20 or more comments, and English language captions available.

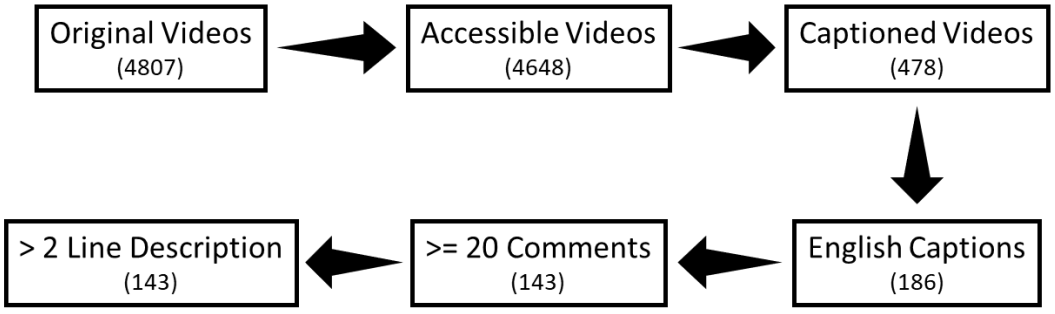


Figure 5.24: Video Candidate Reduction.

[47] http://video.google.com/timedtext

We then perform the same pre-processing methods as described in Section 5.2.4; text normalisation, lemmatising, n-gram identification, dictionary reduction. Table 5.2 summarises the unique word count and final dictionary count for each of the three corpora.

	Words	Unique Tokens	Final Tokens
Descriptions	15525	5223	1118
Captions	132877	14092	5022
Comments	153688	25234	5902

Table 5.2: Word and Token Counts for Apr2020 YouTube Corpora.

5.5.2.2 Topic Coherency

In order to determine the ideal number of topics for each of the three sub-corpora, we ran the LDA algorithm multiple times varying the number of topics required, ranging from 2 topics up to 50 topics, measuring topic coherence for each run.

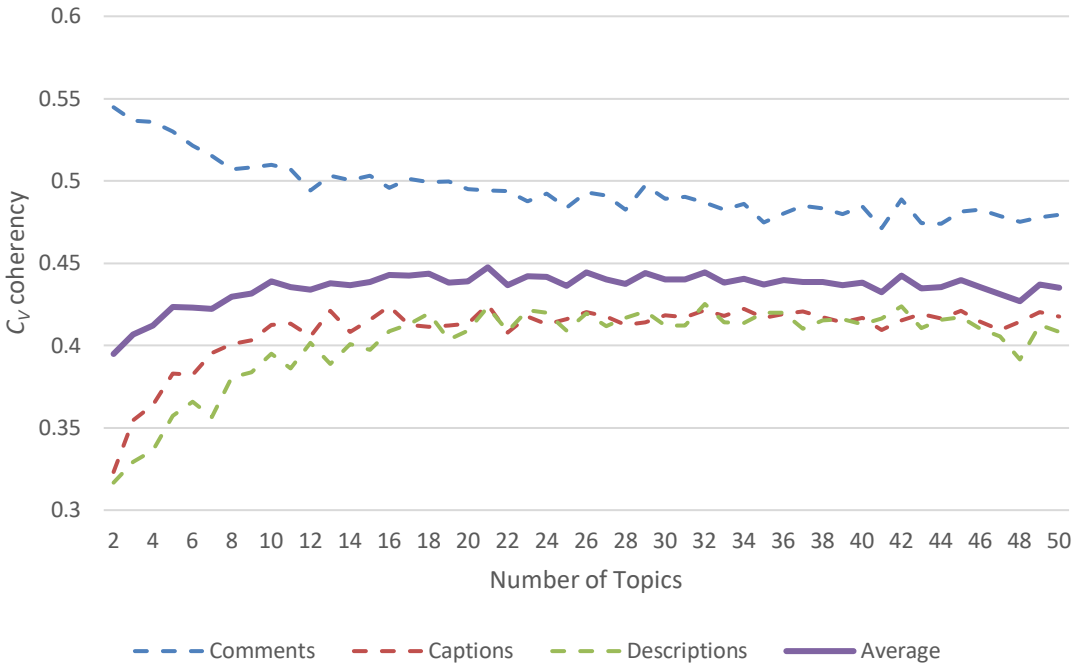


Figure 5.25: Coherence of YouTube LDA Models.

We repeat the model generation and evaluation for each of the sub-corpora and topic count to account for LDA's probabilistic nature causing instability in the model (Mantyla et al., 2018) and generate an average coherence for each configuration. Figure 5.25 presents the average C_V score for each configuration, as well as the average score at each topic count to attempt to identify an optimal number of topics with respect to their coherence across all sub-corpora.

The topic coherence for comments can be seen to decrease as numbers of topics rise, starting from a C_V of 0.54 for two topics, sharply dropping to 0.49 by the 12-topic point where after C_V decreases at a shallower rate as topic count increases. Both captions and descriptions show an increase in C_V at

topic numbers grow, with both starting from a C_V of 0.32, and both improving to a rough plateau of 0.41. Captions achieves this plateau at an earlier stage, at around the 10-topic mark, compared to descriptions plateauing at around the 18-topic point. This suggests that the audience are more focused on a small number of talking points regardless of the content of the videos which have a broader range of topics as suggested by the improving coherence of captions and descriptions.

5.5.2.3 Cross Corpora Topic Similarity

By comparing the similarity of topics found within each of the three corpora, we are able to understand how these three representations of a video; what the author wants to present (the description), what the author actually says (captions), how the audience respond (comments), relate to one another. We identified that the models produced with 21 topics have the best shared average coherence and also observe that that the average coherence begins to plateau at 9 to 12 topics. With this in mind, we chose to observe topic similarity for models at the 21-topic level and also chose to look at the 9-topic level, due to the seven plus or minus two rule defined in (Miller, 1956).

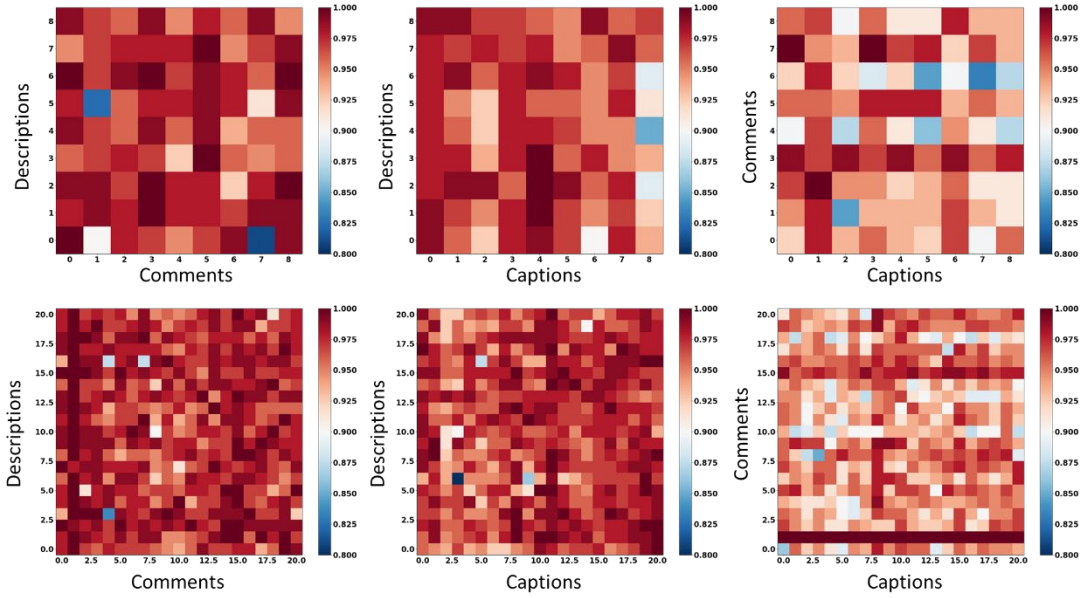


Figure 5.26: Topic Jaccard Distance Heat Maps for Best C_V Models.

Figure 5.26 presents similarity heat maps covering the three combinational pairs of corpora, with a lower difference (thus high similarity) coloured blue, indicating which topics within a model share more keywords with one another. We calculate the difference between the top 50 keywords present within each model's topic using the Jaccard distance (J_δ) which measures the dissimilarity between two sets, the invert of this being the Jaccard similarity index (J) (Niwattanakul et al., 2013). The models compared in Figure 5.26 are taken from the models with the highest C_V present within each the 10 runs made for both the 9-topic and 21-topic levels.

We observe that overall, topic similarity is low, with the highest topic pair sharing a J_δ of 0.795, found within the comments and descriptions 9-topic model, indeed this is the only topic pair that has a J_δ of less than 0.8. We can observe that there are a greater number of similar topics present between the captions and comments models at both the 21-topic and 9-topic level, than in the other two model combinations.

Corpora		Topics	Best C_V Model		All Models	
X	Y		J_δ Avg.	$J_\delta < 0.9$	J_δ Avg.	$J_\delta < 0.9$
Comments	Descriptions	9	0.970	2.47%	0.970	1.75%
Captions	Descriptions		0.963	3.70%	0.962	3.00%
Captions	Comments		0.940	14.81%	0.941	13.73%
Comments	Descriptions	21	0.976	0.68%	0.978	0.67%
Captions	Descriptions		0.979	0.68%	0.970	1.02%
Captions	Comments		0.946	6.58%	0.951	5.36%

Table 5.3: Average Jaccard Distances Accross Models.

In order to validate this observation, we calculate the average J_δ for the pair of models. We also calculate the percentage of topic pairs that have a J_δ of less than 0.9. We can then calculate the average score across all 10 models for each corpus, obtaining an average similarity across all identified topics and is presented in Table 5.3. The captions and comments models are shown to have the most similar topic sets, averaging a similarity of 0.946 and 0.940 found in the 21-topic and 9-topic models. We also see that the captions and comments have the largest number of topic pairs that exhibit less than 0.9 J_δ.

There is an 8.37% drop in the percentage share of $J_\delta < 0.9$ pairs when moving from the 9-topic models (13.73%) to the 21-topic models (5.36%). This drop also occurs in the two other models and is to be expected due to the number of topic pairs increasing at a rate of n^2 alongside the decreasing number of words present within a topic reducing as topic count increases.

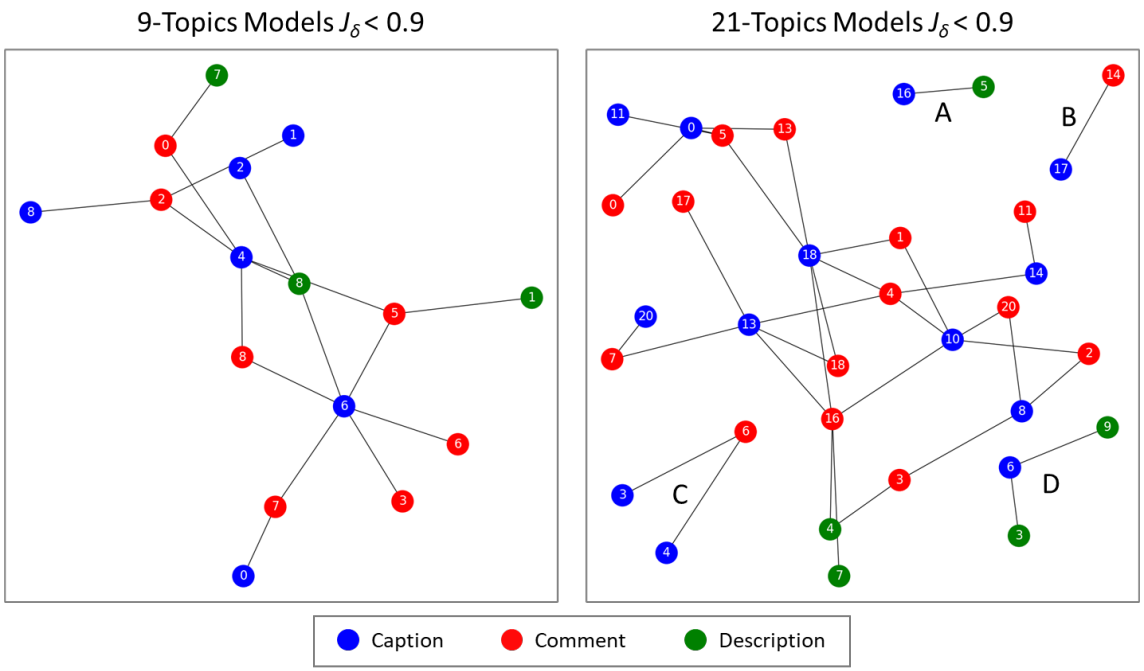


Figure 5.27: Overlapping Topic Network for Captions, Comments and Descriptions.

Figure 5.27 presents a network diagram of overlapping topics for the most coherent models from both the 9-topic and 21-topic sets, both of which highlight the predominance of the Comments (blue) and Captions (red) overlap. In the 9-tpoic model, we see all overlapping topics belong to the same graph, with Captions 4 and Captions 6 acting as central nodes with 5 and 6 edges, respectively. In the 21-topic model, we can see that in addition to the central graph, we have four orphaned graphs that do not interact with the majority of nodes.

5.5.2.4 Top Overlapping Topic Pairs

Table 5.4 highlights the topic pairs with the lowest J_δ (and thus most similar pairs) for both the 9-topic most coherent model sets, covering the top 10 overlapping terms for each pair and the top 10 *relevant* terms for each of the topics. For the sake of discussion in this section, we name each pair using the topic level as a prefix, and a concatenation of the corpora titles and topic numbers as the core identifier. For example, *9ComDes6/7* refers to the 6th Comments topic and the 7th Description topic for the 9-topic models.

Interestingly the top two pairs (*9ComDes6/7* and *9ComDes5/1*) in the 9-topc set are derived from the Comments and Descriptions corpora despite this corpora pair presenting the highest average J_δ in Table 5.3. The *9ComDes0/7* pair looks to be centred around some of the core disinformation subjects, namely the stories that 5G telecommunication towers are spreading COVID-19 [48] and that

[48] https://www.gov.uk/guidance/5g-and-coronavirus-covid-19 (accessed 02-07-20)

hydroxychloroquine can be taken as a cure to COVID-19 [49], with the Descriptions focused on the links to both the China and the USA and the descriptions heavily relating to news reporting. We see that *9ComDes5/1* is also focused on the hydroxychloroquine topic, but the nature of the Comments and Descriptions suggest that this pair is more focused on a medical framing than news.

J_δ	Overlapping Terms	Topic X *Relevant* Terms (λ = 0.6)		Topic Y *Relevant* Terms (λ = 0.6)	
0.810	viral hydroxychloroquine covid19 pandemic update epidemic prevention rate spread 5g	COM 0	china trump us chinese ccp president there america truth usa	DES 7	guardian youtube news network support today owen jones sport football
0.824	use hydroxychloroquine study dr medical news hospital zinc health doctor	COM 5	song tom chocolate sugar fat ironic kid eat movie feel	DES 1	dr berg health covid news hydroxychloroquine study new medical good
0.832	saying us corona come being lab news china always long	CAP 6	ray can god yeah matt yes death fear james jan	COM 7	update epidemic 5g pandemic rate prevention treatment alarmist graduate infection
0.846	us hydroxychloroquine study being immune body believe help health using	CAP 1	china chinese health wuhan can us hong kong yes outbreak	COM 2	dr berg zinc hydroxychloroquine study drug doctor vitamin vaccine treatment
0.846	whole being us come real fact around war trump president	CAP 6	ray can god yeah matt yes death fear james jan	COM 5	song tom chocolate sugar fat ironic kid eat movie feel

Table 5.4: Top 5 Overlapping Terms for Best Coherence 9-Topic Models.

The 5th ranked pair in the 9-topic model *9CapCom6/5* contains the same Comments topic as *9ComDes5/1*, which in this case is paired with a Caption topic whose *relevant* terms suggest this is clustered around interview style videos due to the presence of names and affirmations. The overlapping terms in *9CapCom6/5* do not appear in the top *relevant* terms of either topic but show there is overlap focused on terms relating to the USA. The 3rd ranked pair *9CapCom6/7* is also part of this chain of topics, with the interview style Captions topic being linked to another a Comments topic which contains the 5G story, with the overlapping terms highlighting China and laboratories as shared themes within the topics.

The only topic pair from the 9-topic top 5 that is not directly chained together is *9CapCom1/2* whose Captions topic is geographically focused covering China, Wuhan, and Hong Kong, whilst both the *relevant* terms in the Comments topic and the overlapping terms cover contain medical terms and

[49] https://www.gov.uk/government/news/mhra-suspends-recruitment-to-covid-19-hydroxychloroquine-trials (accessed 02-07-20)

more specifically hydroxychloroquine. Interestingly, Figure 5.27 shows that there is not a strong enough overlap between *9Des2* and *9Com2* even through their topmost *relevant* terms have a 40% overlap.

J_δ	Overlapping Terms	Topic X *Relevant* Terms (λ = 0.6)		Topic Y *Relevant* Terms (λ = 0.6)	
0.795	still said information covid korea outbreak china pandemic public health	CAP 6	energy music green wind can coal plant gas yeah laughing	DES 3	boris covid johnson uk video part comedy response people watch
0.837	used hydroxychloroquine medical dr information news zinc health doctor cure	COM 3	dr berg news zinc there study hydroxychloroquine drug believe vaccine	DES 4	video china patreon get therapy use facebook music people instagram
0.851	ccp already us lab outbreak china pandemic wuhan market party	CAP 8	can matt james vaccine yes saying men facility today important	COM 3	dr berg news zinc there study hydroxychloroquine drug believe vaccine
0.864	being us immune long body 5g health system energy healthy	CAP 0	can yes number covid infected everyone testing help flu system	COM 0	trump us there president news china being vote russia american
0.864	still covid science china testing pandemic point public health wuhan	CAP 6	energy music green wind can coal plant gas yeah laughing	DES 9	late china virus show watch click twitter covid vox follow

Table 5.5: Top 5 Overlapping Terms for Best Coherence 21-Topic Models.

Table 5.5 presents the same information as Table 5.4, but for the 21-topic pairs. We can see that the pair with the lowest observed J_δ is *21CapDes6/3* whose Caption terms is seen to relate to green energy and whose Description terms point to a UK government focus, but the overlapping terms highlight Korea and China and general public health. This same Caption topic is also present in the 5th ranked *21CapDes6/9* sharing a similar set of terms with a Description topic whose *relevant* terms are more focused on China and on social media sharing.

The Description topic within the 2nd ranked *21ComDes3/4* also exhibits similar social media sharing terms along with mention of China. The overlapping terms of *21ComDes3/4* show that again it is concerned with hydroxychloroquine and its medical impact. The *relevant* terms for the Comments topic are very similar to that of *9Com2* from the 9-topic models, suggesting that these are the same topic being expressed through both topic levels.

This Comment topic is also present in the 3rd ranked *21CapCom8/3*, where it is paired with a Caption topic that mirrors the *9Cap6* topic present in the 9-topic models, this time overlapping heavily on concepts relating to China, such as their leading party the CCP, Wuhan, and "market" that likely refers to the suspected source of the virus.

Finally, *21CapCom0/0* is not directly linked to any of the other top-ranking pairs of the 21-topic models, and contains a Captions topic focused on virus testing, Comments discussing both China,

Russia, and the USA, and overlapping topics that centre on the 5G telecommunication conspiracy. Interestingly *21Com0* seems to be the same topic as *9Com0*, both focusing on China and the USA, and in both of their overlapping relationships, the 5G conspiracy is present.

5.5.2.5 Orphaned Topic Graphs

As seen in Figure 5.27, there are a small number of relationships that do not interact with the core network structure, Table 5.6 covers their overlaps and *relevant* terms for the member pairs.

Graph A consists of a single topic pair *21CapDes16/5* which looks to cover comedic interpretations of both the hydroxychloroquine and 5G subjects, with the former being found in both Captions and Descriptions, and the latter found to be relevant to the Caption topic.

Graph B also consists of a single pair *21CapCom17/14* that focuses on the 5G conspiracy theory, with the Captions topic relating the pair back to China and possibly some religious connotations, whilst the related Comments topic discuss quarantine measures and former US President Barack Obama.

	J_δ	Overlapping Terms	Topic X *Relevant* Terms (λ = 0.6)		Topic Y *Relevant* Terms (λ = 0.6)	
A	0.876	covid hydroxychloroquine medical immune testing system infection may case care	CAP 16	5g conspiracy can corona question covid radiation being frequency theory	DES 5	show daily facebook twitter follow instagram comedy central news trevor
B	0.876	cause theory being us dangerous 5g using conspiracy data away	CAP 17	star can chinese god attention ancient important name blind china	COM 14	update song ironic pandemic epidemic obama treatment barry spread quarantine
C	0.889	john already being us flu always fact another trump	CAP 3	power america president being dream said society became politics can	COM 6	doctor being there health hope body amazing immune system ventilator
	0.889	already being us always health doctor cure trump bad	CAP 4	china chinese wuhan hong kong health ccp sars outbreak human	COM 6	doctor being there health hope body amazing immune system ventilator
D	0.795	still said information covid korea outbreak china pandemic public health	CAP 6	narrator putin russia trump president clinton election disinformation american hilary	DES 3	boris covid johnson uk video part comedy response people watch
	0.864	still covid science china testing pandemic point public health wuhan	CAP 6	narrator putin russia trump president clinton election disinformation american hilary	DES 9	late china virus show watch click twitter covid vox follow

Table 5.6: 21-Topic Model, Orphaned Graphs Topic Paris.

Graph C is formed from two topic pairs, *21CapCom3/6* and *21CapCom4/6*, with the shared Comments topic containing *relevant* terms linked to healthcare surrounding COVID-19. This topic is linked to the two Caption topics, the first of which *21Cap3* is focused on American politics, and the second *21Cap4* covers Chinese politics and Hong Kong's past relationship with the SARS outbreak that occurred in 2002.

Graph D is centred around the *21Cap6* topic, which is characterised by the 2016 US Presidential Elections and Russia's relationship with it. This is linked to a UK politics driven Description topic via *21CapDes6/3* with overlapping terms relating to China, Korea, and public health. China and public health are also overlapping terms in the *21CapDes6/9* topic pair linking to a Description topic that too is characterised by China, along with social media sharing terms.

5.5.2.6 Observations

We see a number of subjects that appear across the topic pairs, and the *relevant* terms of constituent topics. Unsurprisingly, two of the most prolific pieces of rumour in this time period, the hydroxychloroquine as a cure and 5G as a cause feature significantly. Interestingly several nations and their leaders or leading parties repeatedly appear, UK, China, USA. It is likely due to the nature of collection that the UK is present, as we have collection terms that relate specifically to UK politics.

It is of note that all topic pairs in the orphaned graphs contain a Caption topic, suggesting that captions contain the most diverse content. Descriptions on the other hand, had the lowest unique token count which resulted in less overlap with other topics and low coherence in their generated models. Comments are as rich in unique token count as Captions and seem to be the most susceptible to topic pairing.

5.5.3 Zero Day Account Analysis

We took this use case as an opportunity to look at the content and behaviour of Zero-Day accounts within the COVID-19 corpus at a large scale. Figure 5.28 shows a histogram of the ages of all accounts within the April corpus, grouped to 30.42-day bins (the average length of a month).

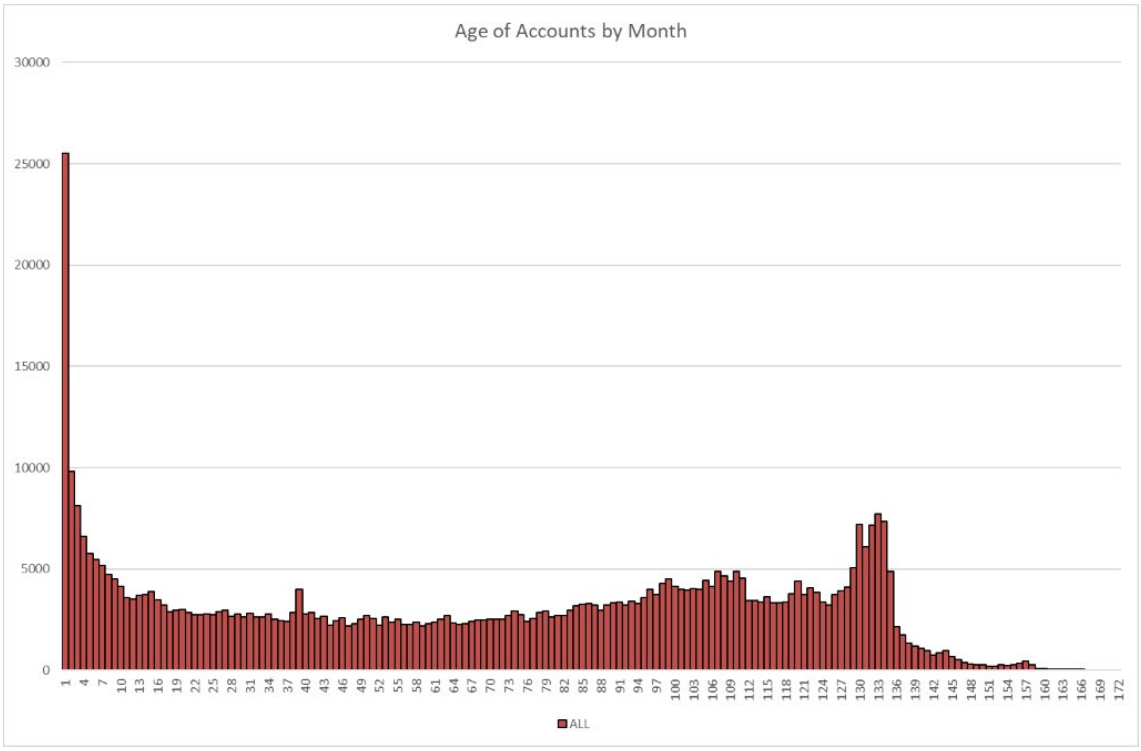


Figure 5.28: Age by Month of accounts in corpus.

5.01% of all accounts within the April set were tweeting within a month of creation, and 9.37% of these “First-Month” accounts (0.79% of “all” accounts), engaged in the topic within the first 24 hours of creation (see Figure 5.29).

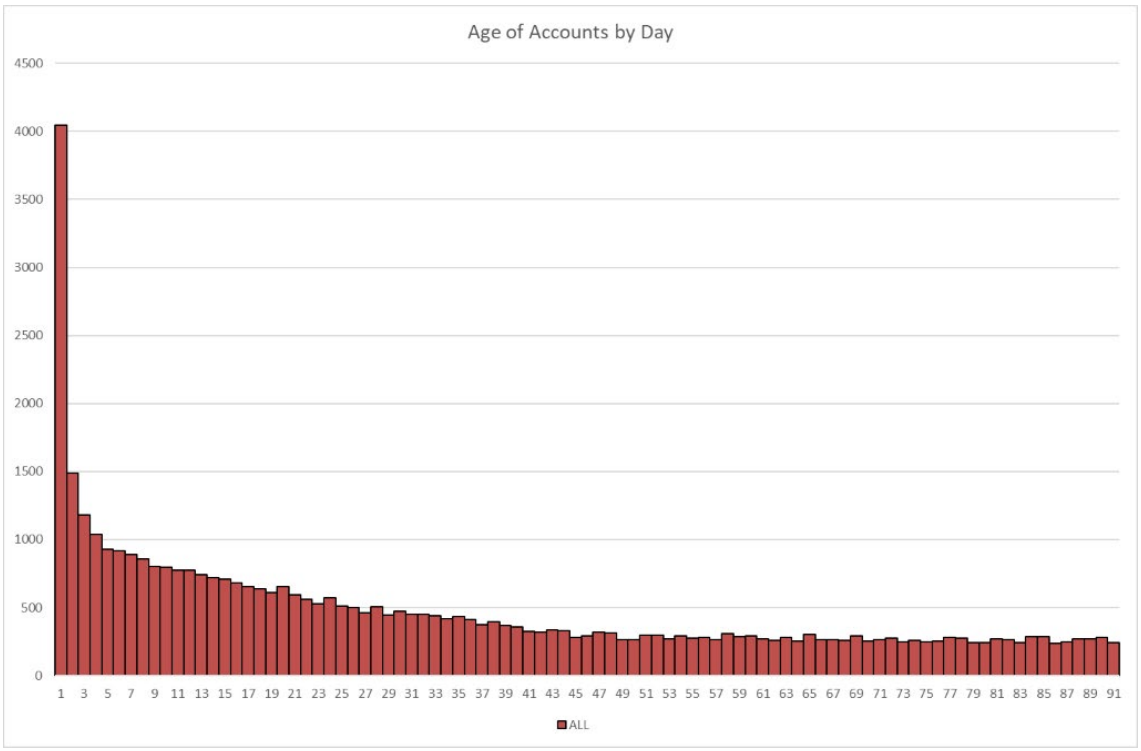


Figure 5.29: Age by Day of accounts in 90-day corpus.

We contrasted the percentage breakdown of First Month and Zero-Day accounts against three single topic downloads run in January on the same *channels* as a means of baselining April against a less volatile time, presented in Table 5.7. This shows that engagement from "young" accounts is higher within the focused April corpus compared to the baseline corpora, although the "Boris Johnson" corpus does show similar levels of engagement from Zero-Day accounts to the April corpus.

	First Month	Zero Day
April	5.01%	0.79%
Disinformation	2.58%	0.21%
General Election	2.63%	0.33%
BorisJohnson / Boris Johnson	3.52%	0.61%

Table 5.7: Prevalence of First Month and Zero Day accounts.

We then analyse some of the metadata surrounding the tweets of both the full April corpus, and the corpus of tweets belonging only to the Zero-Day accounts. We focus on the hashtags used and the accounts mentioned in each Tweet as a means of gauging the general focus of the sub-group of accounts. Table 5.8 shows the word clouds associated with both hashtags and user mentions, relating to all the content from April, just the Zero-Day content (both weighted by frequency), and finally the hashtags and mentions that see the biggest shift in prominence when drilling down to the Zero-Day accounts (weighted by the change in percentage relative to all hashtags/mentions).

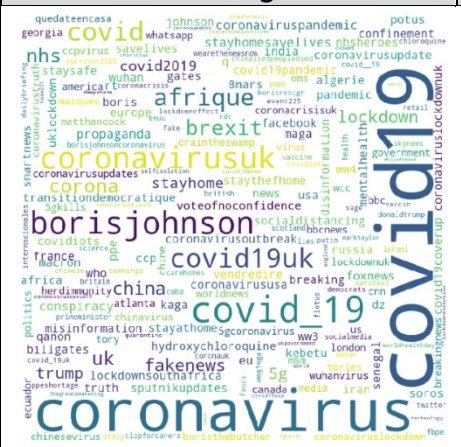
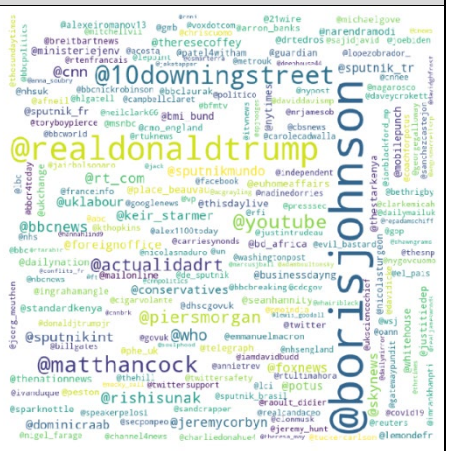
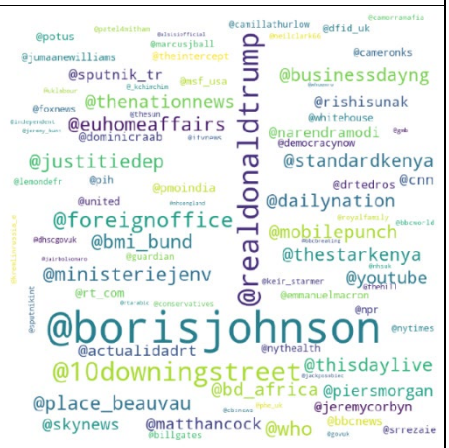
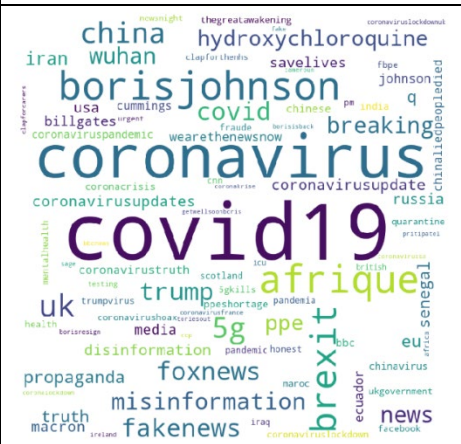
	Hash Tags	User Mentions
All Content		
Zero Day Content		
%Change All -> Zero		

Table 5.8: Word Clouds for April dataset.

The hashtag word clouds are dominated by the coronavirus filtering terms, which is to be an expected artefact of the filtering process. It is interesting to note how the three sets of collection terms ("UK politicians", "disinformation", and "Sputnik/RT") influence the composition of these metadata-based word clouds. The "UK politicians" terms dominate user mentions, showing that a body of the corpus is focused on reacting to political action. The "disinformation" and "sputnik/RT" terms appear more prominently in the final row of Table 5.8, showing that there is a higher focus of engagement from Zero-Day accounts to these terms.

Hashtags				Mentions			
Term	**All**	**Zero**	**% Diff**	**Term**	**All**	**Zero**	**% Diff**
covid19	84785	434	0.034	borisjohnson	52958	536	-0.019
coronavirus	71215	406	0.019	realdonaldtrump	22348	172	0.000
covid_19	18146	146	-0.006	10downingstreet	12482	149	-0.008
borisjohnson	12604	62	0.006	matthancock	9542	47	0.004
coronavirusuk	11908	102	-0.005	youtube	7294	56	0.000

Table 5.9: Top 5 Hashtags and Mentions.

Table 5.9 shows the top 5 hashtags and mentions present within the full April corpus, along with their frequency within the Zero-Day accounts and the difference in percentage share those hashtags have in the Zero-Day corpus. This highlights how the Zero-Day accounts are using the "covid19" and "coronavirus" hashtags to a greater degree than the general population of the corpus.

5.5.3.1 Characteristic Terms

Further analysis of the April corpus was performed using the ScatterText analysis package (Kessler, 2017) that identifies distinguishing terms within corpora. The Zero-Day messages were separated from the rest of the April tweets to form the two distinct corpora. The tweets were then cleaned to remove any clusters of user mentions (to account for the same message being sent to different sets of accounts) and trailing hashtags (to retain only the core message of the tweet) before duplicate removal to prevent any repetitive accounts dominating either of the corpora.

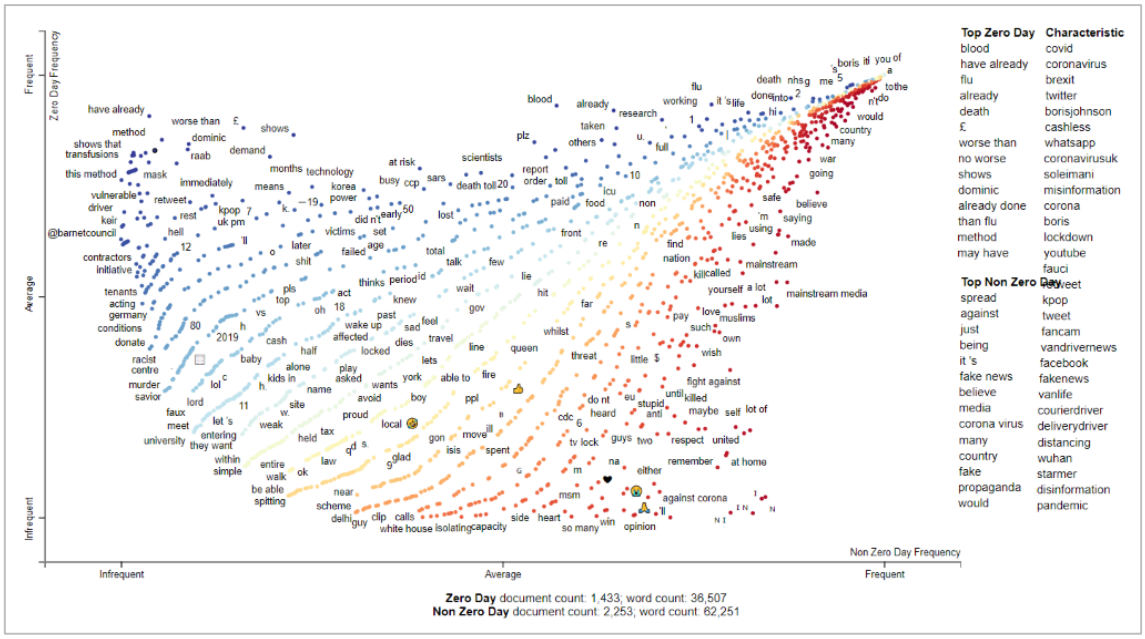


Figure 5.30: ScatterText Output.

Figure 4.5 contains the ScatterText output, showing the frequency of terms and phrases relative to both the Zero-Day (x-axis) and non-Zero-Day (y-axis) content in the scatter graph, along with the most *characteristic* (f1 measure) terms for each corpus and the combined corpora presented in a table on the right-hand side.

Sally59702386
Like why they are recreating the wheel when the research has been done.
Stanford **have already** done the research which shows that Covid is no worse than flu and that majority may have already been asymptotic

Sally59702386
Stanford **have already** done the research which shows that Covid is no worse than flu and that majority may have already been asymptotic We Don't lockdown for the flu so why corona?

Sally59702386
Stanford **have already** done the research which shows that Covid is no worse than flu and that majority may have already been asymptotic We Don't lockdown for the flu so why corona?

Sally59702386
Stanford **have already** done the research which shows that Covid is no worse than flu and that majority may have already been asymptotic

Figure 5.31: Repeated modified messaging.

The non-Zero-Day have a large number of disinformation terms present in its *characteristic* term set, whilst the Zero-Day *characteristic* terms seem to be dominated by a series of connected phrases, driven by a single account producing repeated, but slightly different, messaging (Figure 5.31). The small modifications are likely an attempt at dodging spam filtering and bot detectors. Another interesting *characteristic* term is "£", which suggests that some of the Zero-Day accounts may be being used to promote and sell products, or Forex (foreign exchange) trading.

5.5.3.2 *Topic Modelling*

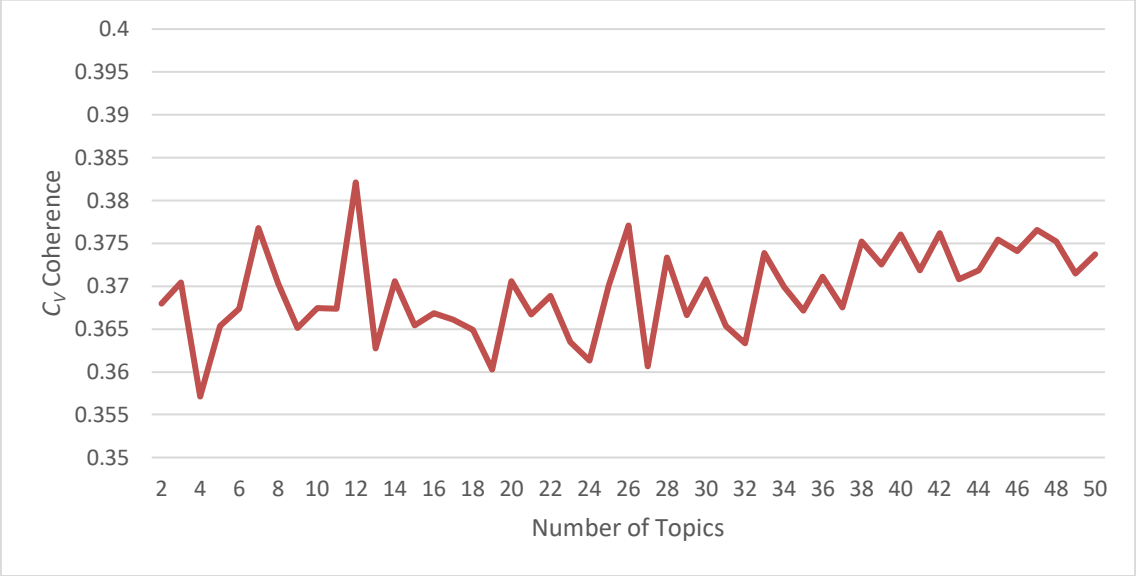


Figure 5.32: C_V Coherence Scores for Zero-Day Topic Models.

We run the same LDA modelling process that we did in Section 5.5.2.2, using the de-duplicated data from Section 5.5.3.1. Figure 5.32 shows the C_V coherence values for the Zero-Day topic models and identifies 12 topics as producing the most coherent models with a score of 0.38. It should be noted that coherence overall is lower than that of the YouTube Comments, and of the Captions and Descriptions once they begin to stabilise above the 7-topic point.

To get an understanding of what is being discussed by Zero-Day accounts, we take the top scoring model from the 12-topic set and retrieve the top 10 most *relevant* terms (λ = 0.6) the for each topic, presented in Table 5.10. From this, we can see that there is a much greater focus on the United Kingdom within the Zero-Day Tweets when compared to the top topics found in the YouTube content, with topics 2, 4 to 7, and 12 all containing UK specific *relevant* terms. Topic 4 appears to be focused on highlighting the spread of misinformation on YouTube while, topics 2 and 5 relate to the admission of Prime Minister Boris Johnson to hospital with COVID-19 symptoms [50], although Topic 5 also touches on the 5G conspiracy also seen in the YouTube content.

Topic 1	**Topic 2**	**Topic 3**	**Topic 4**	**Topic 5**	**Topic 6**
news	boris	people	misinformation	5g	stay
fake	johnson	help	spreading	conspiracy	state
virus	minister	get	stop	there	sage
trump	prime	go	youtube	boris	deep
india	intensive	want	daily	hospital	home
sir	uk	going	new	uk	pandemic
people	care	make	information	johnson	take
spreading	pm	need	spreading	admitted	borisjohnson
fox	hospital	stay	borisjohnson	theory	due
like	world	petition	share	pandemic	pleasure
Topic 7	**Topic 8**	**Topic 9**	**Topic 10**	**Topic 11**	**Topic 12**
blood	already	society	china	method	crisis
brexit	flu	cashless	virus	immune	testing
government	worse	towards	propaganda	system	uk
demand	research	bible	please	cov	picture
matt	done	spoke	via	there	serious
british	stanford	jesus	save	watch	covid_19
people	china	conspiracy	700k	among	beat
face	majority	future	explosive	treatment	coronavirusuk
virus	may	know	scheme	issue	fight
via	study	theory	history	clip	view

Table 5.10: Relevant Terms for Zero-Day 12-Topic Model.

Of the other nations seen in the YouTube topics, we only see US related terms in Topic 1 and China appearing as terms only in Topics 8 and 10. Topic 1 also contains India as a term, which did not appear in the top overlapping topics from the YouTube data. The terms of Topic 3 suggest that this covers petitioning Tweets, with Zero-Day accounts attempting to solicit signatures in support of political petitions, something that is not present within the YouTube data.

Topic 9 is interesting in that it is covering a conspiracy theory not seen in the YouTube topics, and is focused on the longstanding conspiracy theory that the emergence of a "cashless society" is part of a biblical apocalyptic scenario (Barkun, 2013) and that the COVID-19 pandemic is discouraging

[50] https://www.bbc.co.uk/news/uk-52192604 (accessed 02-07-2020)

people's use of cash for financial transactions. Figure 5.33 shows how this topic, along with Topic 2 are the most distinctive topics present in the Zero-Day model along the two principal component axes.

Figure 5.33: pyLDAVis Output for Zero-Day 12-Topic Model.

Even though the majority of these topics are tightly clustered across the principal components, the *relevant* terms found in the Zero-Day topics are much easier to categorise than those found in the YouTube models. This is likely due to the fact that the selected topic models for each of the YouTube corpora were taken from the best average C_V across all three corpora and so the coherence within each of the models may not be ideal.

What is most interesting when comparing the two data sources, is that the Zero-Day topics are more dominated by the channel collection terms with a greater focus on the UK political climate, whereas the YouTube content is more geopolitical in nature, potentially because there is an extra degree of separation between the content due to it being shared via a Tweet, and not directly collected.

5.5.4 Usability Study: Operational Shift

In addition to an exploration into further analysis of a corpus beyond what Sentinel provides, we were also able to include questions in the user survey about the change of usage during the month of April 2020. This allowed us to gain an understanding of how the research team's use of Sentinel

changes during a period of high activity for the research team. Figure 5.34 presents both the reported usage and the change in frequency, for each of the sub-components of the main OSCAR Hub tools covered in the user study. Figure 5.34 presents the responses to questions regarding users' current usage and usage change experienced in April 2020 for the Semantic Search, the Download Manager, and the Project Pages.

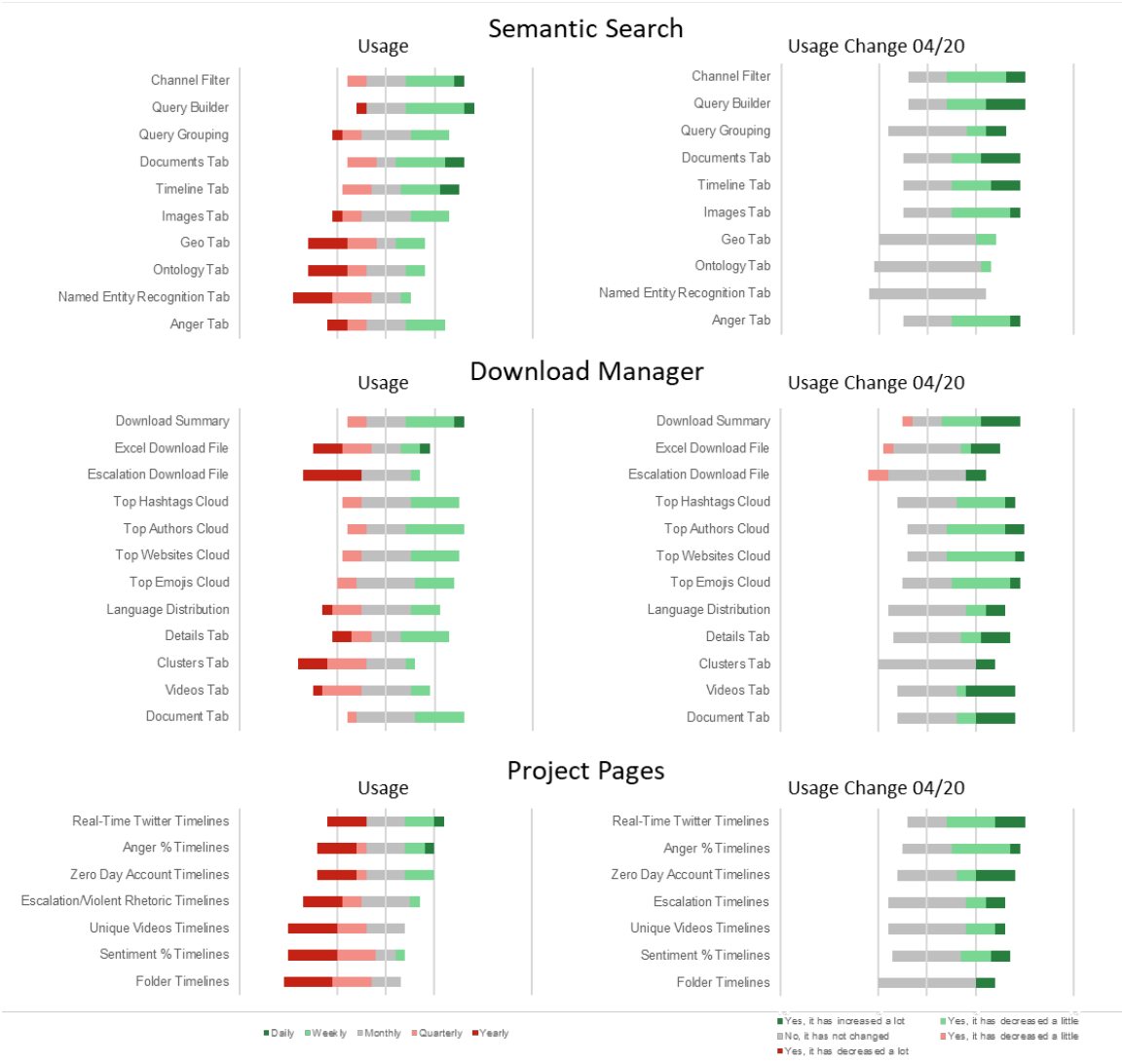


Figure 5.34: Survey Responses for Component Usage and Usage Change 04/20.

We have discussed usage of Semantic Search in Section 4.5.4 already, but here we are able to reflect on the usage and see how that shifts in an active use-case. The usage of Semantic Search's core functionality was reported as increasing most across all tools, but there was no change in use for the semantically enriched tabs which is unsurprising as findings from Chapter 4 suggest users aren't prioritising these, especially so when working at pace.

It is evident that the Escalation content presented in Section 4.4.3 is still niche, even when moving it to a passive feature of the Download Manager. The Clusters tab too is showing low usage which may

be due to it being a more recent additions to the system as users reported a lower familiarity (2.91) for this component, with only the Escalation download having a lower reported familiarity (2.42) within the Download Manager components.

Use of the Project Pages are rather specialist, with only a few users being assigned to each project normally, so experience with it is limited. It does, however, show that the Semantic enrichment from Section 4.4.4 is useful to users as evidenced by the use of the Anger and Zero-Day percentage timelines. We can see that several users have shown an increased interest in the number of Zero-Day accounts interacting with Project related searches. Interestingly, the Project Pages are the only tool where all components show an increased engagement by users without any of the users reporting a decrease in use.

Within the Download Manager components, we see that some users have reduced their use of the summary pane and the downloadable files. This coupled with the increase in the Semantic Search core component use suggests that users shifted to a "scanning" style of use during April. "Scanning" would explain the drop in download file use as users not having have time to dig deeper into a corpus to perform qualitative analysis.

Across all components, the Videos tab shows the greatest number of users that report a significant increase of use during April 2020. This highlights the interest our stakeholders have in video corpora, justifying the decision to develop further analysis methods from Section 5.5.2.

5.6 Discussion & Conclusions

This chapter has focused upon the final back end and front-end updates made to the Sentinel Platform, bringing the platform up to a production level service that supports our research team in identifying candidate datasets in order to support qualitative analysis of social media content.

This chapter highlights how through the co-design lifecycle stakeholders have been able to negotiate more control over the handling of data within the system. The Download Pool and Download Manager interface brings together a number of semantic enrichments presented in previous chapters to further support the Foraging loop. This means that the *data interrogation* co-design activity becomes more accessible to stakeholders; the ability to export data into other research environments that users are also expert in allows for faster and more detailed characterisation of the data. Additionally, the Download Manager provided us with an opportunity to consolidate our initial work into analysis of aggregated data via FlexiTerm (Spasić et al., 2013) and to further the available aggregated analysis by implementing automated topic modelling through LDA.

The Project Pages then build upon the Download Manager concept by providing a single point of access to multiple project related Downloads, sharing effort across the userbase. This allowed for the virtualisation of the *situation room* activities that form part of the co-design lifecycle. The COVID-19 use case has proved a valuable dataset, that has allowed us to further investigate targeted types of social media data, along with a small behavioural study around operational shift in a period of heightened activity.

Our analysis of YouTube content and their peripheral texts highlighted some interesting relationships between comments, captions, and descriptions. This was driven by the identification of a gap in the diversity of data source as seen in the systematic review of Chapter 2. This work provided us with an insight into the value found in cross-linked social media data; where document identification was performed using Twitter data, but the subject data is from YouTube.

Taking this work forward, we will look to extend the topic modelling analysis by introducing the candidate videos as extra vectors in the relationships between topics. This will provide both a means of weighting the importance of a topic within a corpus, and also allow us to understand how broad or narrow on average a video's subject matter is. This is also an example of how the co-design lifecycle has allowed developers to engage more with the knowledge led co-design activities as part of case study.

Further investigation is also required into whether we necessarily need parity in the number of topics used between models when building our topic overlap models. One of the reasons we presented two model sets is due to the Comments models presenting a degrading C_V as the number of topics increased, whilst both Captions and Descriptions C_V improved as topic count rose. If we were to operationalise any of this analysis though, we would likely have to take a heuristic approach to selecting the number of topics, as performing a multi topic number sweep, and then checking for best coherence is time and processing intensive.

In addition to this, it should be noted that we treated all corpora equal in during pre-processing. The question arises as to whether we should have bespoke stop-words for each type of data? The descriptions topics showed a number of *relevant* terms that were related to media dissemination, either encouraging users to like and share videos, or encouraging users to engage with them on other social media platforms. Whilst this information is interesting and would prove useful for network mapping media accounts across platforms, it is inherently noisy to the description corpus. We also saw interjections in some of the caption's topics, which could also be considered noisy.

The research area of spam and bot detection has seen a large increase in recent years, with a number of open source tools becoming available to researchers (Wu et al., 2018, Alothali et al.,

2018). Alothali et al. (2018) discuss how there is an ever-increasing sophistication in behaviour by bot accounts in an attempt to avoid detection by existing tools and methods, while Beatson et al. (2021) highlight the disagreements between these tools. The analysis of the Zero-Day accounts provides a novel approach to disinformation related account analysis, whereby focus is purely on the age of the account.

With the ZeroDay analysis we treat it as a homogenous category, but we need to split up behaviours into different types of account, may then tease out more relevance. Kang et al. (2013) show that decreased anonymity results in a discouragement of malicious activity, and so we feel this is an important sub-group of accounts, as brand-new accounts engaging in a controversial topic benefit from a level of anonymity as they do not yet have a reputation associated with themselves. Also, in the case of genuine new users, we are interested in what topics have energised a user enough that they decide to create a new social media account to discuss the issue.

In both the YouTube and ZeroDay analysis using the COVID-19 case study, it should be noted that the interpretation of topic models is subjective and relies on the investigator's knowledge of the subject matter to make informed observations. There is a need for the establishment of ground truths through empirical evidence, that can be used as a point of reference to the identified topics. The COVID-19 investigation was performed as part of the virtual *situation room* co-design activity set up for the emergent month of the pandemic.

This meant that this analysis was not performed in a vacuum, with the work being supported by virtual fieldwork by stakeholders. This work was performed in parallel to the development of these methods with a number of users engaging in qualitative analysis of COVID-19 misinformation discourse, directed towards tweets and videos of interest that belonged to particular topics. Stakeholders were able to identify a number of key stories present in topics such as Boris Johnson's admittance to hospital, hydroxychloroquine being touted as an early treatment for COVID-19 symptoms, and the "cashless society" conspiracy, that provided validation to the interpretations of topic models.

These questions again highlight the rich research potential present in the corpora we retrieve from the Sentinel Platform. We have shown through the user feedback study that Sentinel has been successful in supporting users in answering the 5Ws (Preece et al., 2018) that we presented in Chapter 1. The Semantic Search tool is integral to the users experience, giving the users the ability to perform data focused tasks present in the Foraging sub-loop of the Sensemaking Loop defined by Pirolli and Card (2005).

CHAPTER 6: DISCUSSION

6.1 INTRODUCTION

In this thesis we have explored how co-designing a social media analysis platform can improve the degree to which Social Web data can be converted into actionable intelligence. We have done this through iterative implementation driven by event driven co-design activities, where both Computer Scientists and Social Scientists collaborate to design and test the platform via a number of live experiments and research outputs. The suitability has been evaluated using three conceptual aspects: robustness, agility, and usability. As a result, we implemented a system that provides users with a suite of tools and options to assist them in developing both the investigation and validation process within social media text analysis. We have observed several benefits of a co-design approach across these three aspects and beyond.

In this chapter we look at the key findings covering each of the three aspects, then discuss what these observations tell us about how the co-design process has shaped the system evolution. We also discuss what shortcomings may arise if any one of these aspects was not supported by the development process and final system. Following this, we discuss the limitations of this work, looking at both the functional limitations encountered during the development of the system, and imitations of the work from a broader scientific perspective. Finally, we cover future work that can continue to develop the functionality of co-designed system and provide further academic interrogation of the effectiveness of this co-design approach to answering the research question.

6.2 DISCUSSION

Over the course of this work, we utilised an event driven lifecycle of co-design activities and principles to drive the development of the analysis platform. Within this lifecycle we introduced the novel use of *situation rooms* to engage co-design in live events, including in response to spontaneous events, to actively challenge stakeholders. The performance and outputs of this system, along with its integration into the routine research methodology of its users, were used to investigate three operational qualities: robustness, agility, and usability. Robustness refers to the ability of a system to continue to operate correctly across a range of operational conditions, failing gracefully outside of that range (Gribble, 2001). Agility's main definitions focus on continuous delivery, design simplicity, and regular user engagement (Fowler and Highsmith, 2001). Usability refers to the extent to which a product is used to achieve goals with effectiveness, efficiency, and satisfaction (Bevan, 2001). Table 6.1 presents the key findings and evidence in this thesis in their

relation to the aspects of robustness, agility, and usability, broken down into the three research chapters.

6.2.1 Robustness

The co-design approach to system development has afforded a number of robustness-based benefits that we have identified during the course of this project. Developer-stakeholder *cooperation* in the design and scoping phases of building the Sentinel pipeline allowed us to identify the importance to stakeholders for persistence in both storage and collection of any dataset used in qualitative analysis. The decomposition of the processing pipeline down into small modules of filtering, translation, indexing, classification, and storage, meant that the pipeline could handle single component faults gracefully. The failure of a module during the development phase did not result in the loss of collected documents, as they are cached in memory by the message passing service when processing queues have no consumer. This meant that errors could be identified and resolved before the pipeline process was spun back up.

With the introduction of the Semantic Search tool, we were able to provide information robustness in the form of the ElasticSearch text indexes. These allowed users to access the stored documents faster than through our more persistent MongoDB-based datastore, with the information presented to the user via index caches when said information had not yet been updated in the datastore. This more robust means of storing and presenting the data meant analysis could be performed in a much more reactive manner, with users having access to collected documents in real-time, as opposed to the 15-minute lag found in earlier interfaces.

Engagement with users also allows for rapid understanding of the quality of data collected, producing a robust dataset that can be comprehensively analysed. A number of case studies and datasets have been reliably collected for several years, spanning over a billion pieces of social media data, and used in several publications by the stakeholders. Significant case studies focused on the Woolwich terror attack (Innes et al., 2014, Roberts et al., 2015, Innes et al., 2018, Roberts et al., 2018), the NATO Summit in Newport (Preece et al., 2016), Brexit (Dobreva et al., 2019), the 2017 UK terrorist attacks (Innes et al., 2019, Innes, 2020), the dissemination of Islamist Extremist magazines (Macdonald et al., 2019), and the COVID-19 pandemic (Antypas et al., 2021, Tuxworth et al., 2021).

		Robustness	Agility	Usability
Chapter 3 - Collection	**Findings**	○ Early engagement with case studies breeds a focus upon key infrastructure robustness by challenging data management components. ○ Engagement with stakeholders also allow for rapid understanding of quality of data collected, producing a robust dataset that can be comprehensively analysed.	○ Agile channel framing is achieved through regular assessment and refinement of channel scope. ○ Both spontaneous and planned events can form viable case study corpora. ○ The co-design approach means stakeholders are receptive to interpreting information from rapidly developed interfaces. ○ Risk of ambiguity increases as duration of collection increases (2.3.1).	
	Evidence	○ Multiple case studies and datasets have been reliably collected for several years spanning over a billion pieces of social media data showing both a persistence of collection and a persistence of data storage(3.6, 4.4.3.5, 0). ○ Several quality publications have been produced from corpora collected by the pipeline, suggesting quality is robust (6.2.1).	○ Rapid corpus generation of case studies achieved through regular and consistent collection on terms (3.6.1). ○ Case studies were undertaken early on in the system development process, providing valuable feedback (3.6). ○ Hackathon interface creation shows how the design of the system supports agile development (3.6.2.1).	

Table 6.1: Findings and Evidence from Research Chapters.

		Robustness	Agility	Usability
Chapter 4 - Enrichment	**Findings**	○ Lightweight parallel enrichment processes to update document records at point of processing provide robust searchability from any point within the enrichment pipeline. ○ Separation of index and datastore provides increased robustness of access to information.		○ Users favour searches across raw metadata rather than via data enrichment features. ○ Semantically enriched content was engaged to some degree with post-search to further filter and refine content.
	Evidence	○ Ability of classification and tagging of documents to be recovered in event of component down-time, showing robustness of infrastructure (4.3.3). ○ Minimal interruption of data access from classification tools, whereby data not yet accessible from MongoDB is still available in search interface (4.5.2).		○ Usage log analysis indicates a preference of users to only engage with basic search functionality (4.5.4 - Table 4.14). ○ Feedback from users showed engagement with anger tab in Semantic Search interface (4.5.4 - Table 4.15)

Table 6.1: Findings and Evidence from Research Chapters. – Continued.

		Robustness	Agility	Usability
Chapter 5 - Download	**Findings**		○ Co-design approach allowed for feedback to produce an agile solution to lack of engagement with semantic content. ○ Download functionality presents a tool that improves agility in mixed methods analysis, providing fast and flexible insights into a user defined corpus.	○ A system built via co-design must cater for both parties allowing data to be accessed by either developers or stakeholders in familiar formats. ○ Giving users ownership over their data, and the ability to export their data into other workflows enriches the overall value of a platform or service.
	Evidence		○ The development of the download functionality that sits on top of the Semantic Search endpoint (5.2). ○ Feedback from usability study (5.4 - Table 5.1) showing users reporting that they are exporting datasets for work in their own environments.	○ Use of by publications such as Dobreva et al. (2019) where data from sentinel used for both big data network analysis, and qualitative analysis. ○ Feedback from usability study (5.4 - Table 5.1) showing users reporting that they are exporting datasets for work in their own environments.

Table 6.1: Findings and Evidence from Research Chapters. – Continued.

Working with these case studies has led to a better understanding of the strengths and weaknesses of social media data analysis. One limitation of the studies that formed the systematic review performed in Chapter 2, was that all algorithms were focused on a single form of social media data. It is imperative that researchers are capable of moving rapidly across platforms in order to derive more robust findings. Tools that can summarise and collect data across multiple platforms thus provide an agile means of traversing a heterogenous data landscape.

We see in the work focused on the COVID-19 data interrogation and case study that narratives and topics span multiple social media platforms. The case study shows how a corpus of Tweets can yield a wealth of additional content found in the out-links pointing to other social media platforms and news sites. We chose to focus on the shared YouTube content within the case study, using this to explore how the different forms of texts within said content related to discourse found directly on Twitter. The finding of this work suggests that these secondary sources, produce contextually relevant content that is less influenced by the collection terms.

6.2.2 AGILITY

Part of effective engagement with stakeholders is reliant on the ability to both manually and automatically ingest user feedback into the collection process in a timely fashion. Because stakeholders were heavily engaged in the design and development process of Sentinel due to the co-design approach, a better understanding of the type of social media data that can be best exploited by the Sentinel platform is reached. This reduces the implementation burden on developers when bringing new types of data into an analysis system, providing developers with an understanding of any new constructs and features that are of importance to the end users, and improving stakeholders' understanding of the information derived from the social media data. The supporting evidence comes from the integration of Reddit and YouTube data into the Sentinel collection platform and the implementation of new search functionality that allows users to search across both the comments and the main post.

Responding to automatic collection feedback should come in the form of the user being able to manage and refine the terms and parameters for any available collection stream. When coupled with an agile means of assessing the quality of data, in the case of Sentinel this is via the search and download functionality, this allows the user to identify noisy and irrelevant terms. The importance of lexical disambiguation is key when building a set of collection terms that define the scope of a dataset. Polysemy describes words or phrases that hold multiple meanings, and it has been shown that when a polysemous word appears two or more times in a discourse, it is highly likely that they will all share the same sense (Gale et al., 1992).

Therefore, care must be taken when selecting terms to ensure that the vocabulary present within the collected data belongs to a single discourse, to ensure effective outputs from any text mining algorithms employed. This can be evidenced by findings from the systematic review performed in Chapter 2, where we observed that within the candidate studies, a tighter collection window correlates with a higher degree of accuracy within text classifiers. This tighter time scope reduces lexical ambiguity by improving the likelihood that an instance of word or phrase belongs to the same discourse, and consequently the intended meaning.

For longer running collection channels, this agility in collection provides a means by which users can maintain and introduce new concepts and relevance into a channel as the scope and narratives evolve. This makes corpora derived from both planned and spontaneous events viable subjects for case study analysis. In the case of spontaneous events, following an initial information gathering and reporting phase where there may not be any coherence of voice or vocabulary, there tends to be a convergence towards a small number of emerging terms, entities, or hashtags that hold significance to the event (Innes et al., 2018). These can be quickly identified through initial scoping of a collection and added to the terms early on to improve the coverage of collection for analysis.

Operationally, the research programme driving the stakeholders was able to engage early with the data via the Woolwich case study that became available to users whilst the Sentinel platform was still in prototype stage. One of the main steps in the co-design process is prototyping of systems and tools (Spinuzzi, 2005), which conditions users to be more receptive to information from experimental interfaces, allowing the research programme as a whole to act in an agile manner. This provides ample opportunity for feedback into the development process. Such feedback was crucial to the identification of a lack of engagement with the semantic content via the Semantic Search interface, thus forcing the development of the Download Manager tool, which repurposed and reframed the information derived from the classification and analysis tools.

Responses from users to open-ended questionnaire in the usability study were positive towards the Download Manager tool, with the data summarisation elements aiding users in obtaining an overview of data and in the identification of accounts of intertest. Furthermore, they cite the ability to create bespoke datasets for analysis in a “multi-method/platform approach” as a tangible benefit of the Download Manager, with another user also reporting that the tool “facilitates data coding and quantitative analysis” allowing them to analyse data at their own pace. The ability of this tool to integrate into the wider mixed-methods analysis process helps to provide fast and flexible insights into a user defined corpus.

6.2.3 USABILITY

The integration of the platform into the researcher's wider analysis process is key to highlighting the usability of the Sentinel platform, and its constituent components. The co-design process provides a means of upskilling stakeholders, achieving a stronger understanding of the underlying sources of social media data, along with the developed analysis software.

This strengthens the relationship between users and developers, building developer trust in the stakeholders' capability to work with rawer forms of data. This allows development to be made in a more agile manner as already discussed, but it also gives users a greater ownership of their data whereby they can incorporate the advantages of the developed software into their pre-existing investigative workflows. Spinuzzi (2005) state that a goal of co-design is the improvement of quality of life for workers, empowering the worker both in their control over their own workflows (democratic), and in the ease by which they can perform their tasks (functional).

Within Sentinel, the ability to export data in both human-readable and machine-readable formats provides users with such control. From a mixed methods analysis standpoint, this gives the users the freedom to perform deeper dives into the data via Computer Assisted Qualitative Data Analysis Software (CAQDAS) tools such as ATLAS.ti (Muhr, 1997), MAXQDA (Kuckartz, 2007) or NVivo (Richards, 1999), or through network analysis tools such as Gephi (Bastian et al., 2009). A key example of this being done by stakeholders is in the work presented in Dobreva et al. (2019) where datasets were collected and curated using Sentinel, before further qualitative analysis into the types of rumour and conspiracy, and the networks propagating these rumours, was performed via third party software.

When developing an analysis platform that forms part of a wider research agenda, it is important to be able to identify the scope of platform functionality and gaps present within a user's investigative workflow. Within this thesis, we began with a broad target of supporting researchers bi-directional movement across the *sensemaking loop* (Pirolli and Card, 2005), and the understanding that this would narrow as the co-design process evolved. This resulted in the development of a platform that heavily favours the *foraging loop* component of the model, but that provides the flexibility for users to integrate the platform into a wider research methodology, through data ownership and understanding, in order to operate across the entirety of the *sensemaking loop* model.

It is through this negotiated scoping that the platform becomes usable, preventing the platform from becoming unwieldly by forcing the user into a monolithic mode of operation that replicates function from established tools they are already familiar with. This is the key benefit of adopting a

co-design approach when building an adaptable toolset that must be capable of providing actionable information in a variety of modes.

6.2.4 Co-Design as an Driver of Robustness, Agility, and Usability

We defined our research question as *"how does the application of a co-design approach to system design improve the degree to which Social Web data can be converted into actionable intelligence within an analysis platform, with respect to robustness, agility, usability?"*

Focusing the co-design process around planned and spontaneous events allowed stakeholders and developers to quickly develop a shared research scope grounded in ecological validity that enabled sustainable co-design value to stakeholders throughout the system development. We can see from the findings above that the system that was borne out of this process contained a number of tools and features that supported agility, robustness, and usability.

The question remains as to whether these qualities were expressed because of the co-design process. We would argue that the Woolwich case study (Section 3.6.1) is an excellent example of how the co-design process very early on enabled the identification by stakeholders that data persistence and collection robustness would form a keystone characteristic of the system. Because a trusting relationship was established early between stakeholders and developers, stakeholders were confident in taking a leadership role in identifying the importance of the event, driving the development of the collection and storage components. Both *trust through making together* and *taking responsibility of co-design* are enablers for co-design as identified by Pirinen (2016). In addition to driving requirements of robustness in the design process, this event also highlighted the need for and pushed the development of agile means of channel curation in order to react to emerging narratives.

The development of the underlying architecture that feeds into the Semantic Search and Download functionality occurred through a number of iterations of the events driven co-design lifecycle, as the developers and stakeholders built upon their understanding of how they can work with the system, and how they wish to access the data. This is another example of co-designed robustness and indeed usability; this time improving the accessibility and re-usability of collected data, achieved over multiple *data interrogation* and *case study* co-design activities.

We have also been able to observe on the micro scale, what shortcomings become apparent when one of the three aspects is not supported by a tool, system, or interface. During the NATO Summit *situation room* co-design activity, we saw that it was only through serendipitous circumstances that a key event was identified by the *bottom-up* analysis. This highlights how the lack of a robust dataset put heavy strain on the monitoring capabilities of the Sentinel system at that time. We saw through

the *case study* activities surrounding the spontaneous events linked to terrorist atrocity that the need for agility is key in being able to collect data on an event as fast as possible, as some of the key narrative dynamics occur within the first few hours following an incident (Innes et al., 2018, Innes et al., 2014, Roberts et al., 2015). A system that cannot react in an agile way to an emerging event can potentially miss key data and interactions that are essential for building situational awareness. Finally, we saw clearly from the challenges of getting stakeholders and users to engage with output from the anger classification module when using the Semantic Search tool, that features lacking usability will not be engaged with. We see from the usability study that users tended towards the more established tools provided by the system that support both bottom-up and top-down *foraging loop* activities that supported independent research.

In terms of reproducibility to broader class of application, whilst the process of event driven co-design will likely not result in the same system or tools being produced, we do believe that the co-design behaviours of stakeholders and developers will follow a similar path with respect to the prioritisation between developing system robustness, agility, and usability. Agility and robustness are key development targets in the early stage of development, with the focus on developing robust datasets and reacting fast in both data collection and event debrief. Usability becomes more relevant as the stakeholders' research objectives and the system itself mature. We see this in our development of the Download Management tool, which is a culmination of lifecycle iterations that drove more control of the *data interrogation* towards users and stakeholders.

6.3 LIMITATIONS

The practice of co-design places stakeholders and developers together as peers, with equal influence over the direction and scope of the areas of study that the system supported. Development of the Sentinel system was performed by an interdisciplinary group of Social Scientists and Computer Scientists as part of the group's wider research into social sensing. The Social Science members of the stakeholder group had expertise in ethnographical approaches to social research, and so required the grounding of any case study and experimentation in live events, in order to pursue an ecologically valid (Gehrke, 2014) research agenda. This produces limitations on the ability to evaluate elements of the system in a controlled manner, as interactions with the active system were predominantly in a responsive information seeking manner. This does however mean that study of user performance within the system can be considered ecologically valid.

6.3.1 Evaluative Limitations

Limitations also come from the size of the stakeholder group, and the time critical requirements of their roles. These both play into the fact that it was not possible to develop any form of ablation study on the systems as all users were engaged with the system in a fully operational manner. To deny users features and capabilities of the system would be detrimental to the research output of the stakeholder group.

This limitation also comes into play when performing the usability and usage log studies within this thesis. In both instances, the studies cover the entirety of the active userbase at the time of assessment, meaning results are very sensitive to individual behaviour and opinion. In the case of the usage log study, this was controlled for through the dissemination of training materials that explained the functionality of the Semantic Search tool, ensuring that all users had a base level understanding of the system. Additionally, the feedback received from the user group will inherently be positively biased, seeing as the users themselves are stakeholders in the design of the system, and so likely take some pride and ownership over the system.

6.3.2 Functional Limitations

Operationally, one key limitation of the Sentinel platform is that it relies on openly available APIs that are bound by rate limiting constraints. In the case of Twitter, this limits collection to 400 terms, and caps any streamed collection at 1% of the total throughput of the platform (Sampson et al., 2015, Twitter, 2020). YouTube runs a quota system that limits daily collection to 10,000 tokens worth of requests to the API service, whereby metadata snippets have a small token cost (YouTube, 2020). Whilst these limitations prevent complete collection of a topic, this forced the development of the agile channel curation and refinement functionality found within the system. This also required users to be mindful of case studies they wish to pursue through the use of Sentinel, and to be aware of any polysemous search terms they may wish to use to form case study channels.

Furthermore, many other popular platforms do not host API services that provide comprehensive access to their data. Facebook significantly reduced the capabilities of their graph API service in both 2015 and 2018 (Freelon, 2018), meaning access to this platform is limited to partner applications such as CrowdTangle (CrowdTangle, 2020) or commercial products such as Brandwatch (Brandwatch, 2020). Beyond social media, ingesting content from news articles is challenging due to the lack of a uniform point of collection. Users have engaged with the news aggregation tool GDELT which provides access to a multitude of text mining outputs applied across a diverse range of news sources (Leetaru and Schrodt, 2013). But the tool does not grant access to the text content of collected news articles, preventing any bespoke text analysis from being developed and performed.

This limits the number of suitable data sources that can be incorporated fully into the platform. Therefore, supporting of mixed methods research and cross platform compatibility through features like export functionality became very important, with the intention of freeing the user to engage with these third-party platforms within their research workflow.

The stakeholders' research focus evolved from geospatial online communities to planned and unplanned events, and finally misinformation and disinformation, over the course of the development of Sentinel. This transition towards more global concepts began to highlight limitations in the semantic enrichment modules' ability to process non-English and multi-lingual content to the same degree as English language content. Despite this, data collection through API endpoints remained language agnostic, and the Semantic Search functionality was not limited to the Latin alphabet, allowing non-English content to be collected, searched, and exported. Once again, this highlights that the Sentinel platform formed part of a wider mixed methods research workflow and should not be seen as a complete solution to social media research.

Finally, it should be noted that social media should not be considered representative of the general population; with a demographic bias towards a younger more liberal membership (Mellon and Prosser, 2017). Care must be taken when using information derived from social media to make conclusions relating to public opinion and population mood. Once more, information taken from a platform like Sentinel should only form part of the situational awareness process.

6.4 Future Work

The evaluative limitations highlighted above show that it is important the system be tested in a controlled manner, in order to be able to evaluate the performance of the system against similar tools and traditional methods. This can be in the form a controlled set of tasks to be performed by the users with and without Sentinel, or with particular interfaces removed from the system. Any such experimentation should focus on evaluating the ability to perform tasks that are reflective of stakeholders' research behaviour during the data led activities found in the event driven co-design lifecycle (*situation rooms* and *data interrogation*).

As discussed in the evaluative limitations, the system is driven by desire for analysing ecologically valid data, but this may not be practical when attempting to evaluate the system. There is value in knowing the ground truth of an event if tasking users to perform research through Sentinel in a controlled manner. The disinformation related dataset that formed part of the use case analysis performed in Section 5.5 could potentially act as a suitable training corpus for such evaluation. The dataset has shown versatility in that a number of separate events have been traceable over the

course of two years, including the European Elections of 2019 (Tuxworth et al., 2021) and the COVID-19 pandemic (Antypas et al., 2021), which could be used as part of the evaluation process.

The Escalation rule classification proved interesting, and it is unfortunate that we have not been able to develop this classifier further into a streamed environment yet. Incorporating a streamed version of the Escalation classifier would also allow these features to be folded back into the Anger classification tool as features as a means of boosting accuracy and allowing the classifier to better represent the conflict action model that preceded it. Additionally, the creation of a new training corpus sampled from within the collected data may improve the performance of the classifier and will be performed in future work.

Our analysis of YouTube content and their peripheral texts highlighted some interesting relationships between comments, captions, and descriptions. Taking this work forward, we will look to extend the topic modelling analysis by introducing the candidate videos as extra vectors in the relationships between topics. This will provide both a means of weighting the importance of a topic within a corpus, and also allow us to understand how broad or narrow on average a video's subject matter is. Further investigation is also required into whether we necessarily need parity in the number of topics used between models when building our topic overlap models.

Taking this further, there is importance in the point of crossover between two social media platforms with the COVID-19 YouTube content showing that secondary content can be used to introduce much richer information sources into a discourse; A 30-minute video can cover a lot more information than a 280-character Tweet. There has been a rise in the emergence of "alternate" social media platforms, such as Telegram, Gab, Parler, and Voat, which have grown in popularity following the removal of a number of controversial figureheads from more mainstream social media platforms (Rogers, 2020). This suggests that linking to other platforms could also be a way of bypassing deplatforming and engaging users via these new platforms. These are certainly areas to explore in conjunction with general research into fake news and disinformation links being shared on mainstream social media platforms. There should also be a focus in any of this future work to be able to track narratives and discourse within a multi-lingual space. Recent advances in word embedding based text analysis has begun to explore multi-lingual modelling of social media data within a single vector space (Camacho-Collados et al., 2019), that present a possible solutions to this problem.

6.5 Conclusions

It is important to acknowledge that the concepts of robustness, agility, and usability are somewhat overlapping. A robust system relies on the ability to react in an agile manner either through automation or redesign. A robust system also builds user trust in system performance which in itself encourages users to push and test the system thus resulting in greater usability over time. An agile system can be more reactive to changes in user needs and workflow, inherently improving usability, but relies on a robust infrastructure that can be adapted to the required changes. Finally, a usable system encourages usable feedback which in turn identifies components within the infrastructure that can be developed to further improve the system's robustness and agility. This thesis documents how these concepts can be incorporated into a social media analysis platform in an incremental and effective manner; first focusing on developing robust services, then challenging this system to act in an agile way, before integrating the system into the qualitative researcher's workflow.

Our literature survey demonstrates that social media are already well established as a valuable source of data suitable for text mining, but we have also shown that by co-designing a collection and processing pipeline we can build robust datasets for case study. This is done through challenging data management components to produce a system that supports agile and rapid tuning of collection terms by a human in the loop. We show that if this happens early in the development of an analysis platform, it allows for future developments to be driven by the data and findings of early research output. These iterative findings themselves become more robust as new features, methodologies and understandings are developed within and alongside the analysis platform.

Feedback from our usability studies and logging also indicated that users preferred the information derived from text mining of social media to be used in a subjunctive manner rather than an indicative manner. Users did not engage with semantically enriched features to search for data, instead preferring to engage with this information post corpus-identification. Once the user has narrowed down their dataset, they were much more comfortable engaging with the semantically enriched content, using it to accelerate their understanding of discourse present within a corpus.

Whilst social media provides a rich source of information and discourse, it does not exist in a vacuum. As the research objectives of the stakeholders developed in line with the evolution of the analysis platform, it became evident that social media only captures a fraction of narratives and discourse surrounding a topic or event. Many social media platforms allow for the linking to content on other platforms and mediums such as blogs, news sites, and other social media platforms. What this led to, was an understanding that the human-in-the-loop element of the co-designed system

was crucial to developing an agile system, where context may only be available outside of the scope of data available to the system.

The co-design process was important in upskilling stakeholders by improving their understanding of both the data and the text mining methods and tools applied to the data, along with the ability to identify when and where context and evidence must be sought outside of the analysis platform environment. This means users become much more capable of switching between modes and tools when analysing the data, and when coupled with the ability to export data for use in third party tools produced a highly agile means of situational awareness research. As such, any developed situational awareness platform should not be employed in isolation and should be developed to complement and improve a researcher's investigative workflow.

Prior to this project, semi-automated sensemaking of social media with applications in the crime and security field was limited to post-hoc analyses of singular case studies using bespoke tools and/or commercial products originally developed to support marketing and brand management. We have demonstrated that a co-designed analysis platform can provide an effective means of in-depth analysis of social media data within the realm of crime and security. In particular, a number of agile entry points into analysis of an event or topic is enabled by key co-designed features such as classification of anger and escalation, ZeroDay accounts, and platform switching. Thanks to these, stakeholders have been able to pursue an emerging research agenda rapidly across a broad range of spontaneous and planned geopolitical and security-based events. During the lifetime of this project we have witnessed social media emerging as a key broadcasting platform in domestic and international politics, with an increased focus on the propagation of misinformation and disinformation brought forth by events such as the 2016 US Presidential Election (Bovet and Makse, 2019), and the 2020 Coronavirus Pandemic (Pennycook et al., 2020). The benefits of a co-designed approach to social media sensemaking increased the general understanding of the uses and limitations of traditional qualitative research involving social media and how they can be rectified as illustrated by 13 mixed method research studies published by the users of Sentinel (Preece et al., 2015, Roberts et al., 2015, Preece et al., 2016, Innes et al., 2017, Innes et al., 2018, Roberts et al., 2018, Dobreva et al., 2019, Innes et al., 2019, Macdonald et al., 2019, Innes, 2020, Antypas et al., 2021, Innes et al., 2021, Tuxworth et al., 2021). Through this thesis, we have seen that these benefits have manifested via the development of a robust, agile, and usable platform of tools and services.

REFERENCES

1. ABBE, A., GROUIN, C., ZWEIGENBAUM, P. & FALISSARD, B. 2016. Text mining applications in psychiatry: a systematic literature review. *International Journal of Methods in Psychiatric Research,* 25**,** 86-100.

2. ABRAMOVA, V. & BERNARDINO, J. NoSQL databases: MongoDB vs cassandra. Proceedings of the international C* conference on computer science and software engineering, 2013. 14-22.

3. ABUBAKAR, Y., ADEYI, T. S. & AUTA, I. G. 2014. Performance evaluation of NoSQL systems using YCSB in a resource austere environment. *Performance Evaluation,* 7**,** 23-27.

4. ACHILLES, A.-C. 2020. *Collection of Computer Science Bibliographies (CCSB)* [Online]. Available: http://liinwww.ira.uka.de/bibliography/ [Accessed 2020-02-05].

5. AHO, A. V. & CORASICK, M. J. 1975. Efficient string matching: an aid to bibliographic search. *Communications of the ACM,* 18**,** 333-340.

6. ALHARTHI, R., GUTHIER, B., GUERTIN, C. & EL SADDIK, A. 2017. A dataset for psychological human needs detection from social networks. *IEEE Access,* 5**,** 9109-9117.

7. ALHARTHI, R., GUTHIER, B. & SADDIK, A. E. 2018. Recognizing human needs during critical events using machine learning powered psychology-based framework. *IEEE Access,* 6**,** 58737-58753.

8. ALLCOTT, H. & GENTZKOW, M. 2017. Social media and fake news in the 2016 election. *Journal of economic perspectives,* 31**,** 211-36.

9. ALLCOTT, H., GENTZKOW, M. & YU, C. 2019. Trends in the diffusion of misinformation on social media. *Research & Politics,* 6**,** 2053168019848554.

10. ALOTHALI, E., ZAKI, N., MOHAMED, E. A. & ALASHWAL, H. Detecting social bots on Twitter: A literature review. 2018 International conference on innovations in information technology (IIT), 2018. IEEE, 175-180.

11. ALTMAN, R. B., BUDA, M., CHAI, X. J., CARILLO, M. W., CHEN, R. O. & ABERNETHY, N. F. 1999. RiboWeb: An ontology-based system for collaborative molecular biology. *IEEE Intelligent Systems and Their Applications,* 14**,** 68-76.

12. AMIGÓ, E., CARRILLO-DE-ALBORNOZ, J., CHUGUR, I., CORUJO, A., GONZALO, J., MEIJ, E., DE RIJKE, M. & SPINA, D. 2014. *Overview of replab 2014: author profiling and reputation dimensions for online reputation management*, Springer.

13. AMOORE, L. & PIOTUKH, V. 2015. Life beyond big data: Governing with little analytics. *Economy and Society,* 44**,** 341-366.

14. ANGARAMO, F. & ROSSI, C. Online clustering and classification for real-time event detection in Twitter. ISCRAM, 2018.

15. ANTYPAS, D., CAMACHO-COLLADOS, J., PREECE, A. & ROGERS, D. 2021. COVID-19 and Misinformation: A Large-Scale Lexical Analysis on Twitter. *The Joint Conference of the 59th*

Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing.

16. ARP, R., SMITH, B. & SPEAR, A. D. 2015. *Building ontologies with basic formal ontology*, Mit Press.

17. ASSOCIATION FOR COMPUTING MACHINERY. 2020. *ACM Digital Library* [Online]. Available: http://dl.acm.org/ [Accessed 2020-02-05].

18. ASUR, S. & HUBERMAN, B. A. Predicting the future with social media. 2010 IEEE/WIC/ACM international conference on web intelligence and intelligent agent technology, 2010. IEEE, 492-499.

19. AVVENUTI, M., VIGNA, F. D., CRESCI, S., MARCHETTI, A. & TESCONI, M. Pulling information from social media in the aftermath of unpredictable disasters. 2015 2nd International Conference on Information and Communication Technologies for Disaster Management (ICT-DM), 2015 Rennes. 258-264.

20. AYVAZ, S. & SHIHA, M. O. A Scalable Streaming Big Data Architecture for Real-Time Sentiment Analysis. Proceedings of the 2018 2Nd International Conference on Cloud and Big Data Computing, 2018 New York, NY, USA. ACM, 47-51.

21. AZZOUZA, N., AKLI-ASTOUATI, K., OUSSALAH, A. & BACHIR, S. A. A real-time twitter sentiment analysis using an unsupervised method. WIMS'17: Proceedings of the 7th International Conference on Web Intelligence, Mining and Semantics, 2017 Amantea Italy. Association for Computing Machinery, 1-10.

22. BACCARELLA, C. V., WAGNER, T. F., KIETZMANN, J. H. & MCCARTHY, I. P. 2018. Social media? It's serious! Understanding the dark side of social media. *European Management Journal,* 36**,** 431-438.

23. BARKUN, M. 2013. *A culture of conspiracy: Apocalyptic visions in contemporary America*, Univ of California Press.

24. BASTIAN, M., HEYMANN, S. & JACOMY, M. 2009. Gephi: an open source software for exploring and manipulating networks. *Icwsm,* 8**,** 361-362.

25. BEATSON, O., GIBSON, R., CUNILL, M. C. & ELLIOT, M. 2021. Automation on Twitter: Measuring the Effectiveness of Approaches to Bot Detection. *Social Science Computer Review***,** 08944393211034991.

26. BECK, K., BEEDLE, M., VAN BENNEKUM, A., COCKBURN, A., CUNNINGHAM, W., FOWLER, M., GRENNING, J., HIGHSMITH, J., HUNT, A. & JEFFRIES, R. 2001. Manifesto for agile software development.

27. BEHZADAN, V., AGUIRRE, C., BOSE, A. & HSU, W. Corpus and deep learning classifier for collection of cyber threat indicators in twitter stream. 2018 IEEE International Conference on Big Data (Big Data), 2018 Seattle, WA. 5002-5007.

28. BENKHELIFA, R. & LAALLAM, F. Z. Opinion extraction and classification of real-time YouTube cooking recipes comments. *In:* HASSANIEN, A. E., TOLBA, M. F., ELHOSENY, M. & MOSTAFA, M., eds. The International Conference on Advanced Machine Learning Technologies and Applications (AMLTA2018). 2018. Springer, 395-404.

29. BEVAN, N. 2001. International standards for HCI and usability. *International journal of human-computer studies,* 55**,** 533-552.

30. BIEM, A., BOUILLET, E., FENG, H., RANGANATHAN, A., RIABOV, A., VERSCHEURE, O., KOUTSOPOULOS, H. & MORAN, C. IBM infosphere streams for scalable, real-time, intelligent transportation services. Proceedings of the 2010 ACM SIGMOD International Conference on Management of data, 2010. 1093-1104.

31. BOVET, A. & MAKSE, H. A. 2019. Influence of fake news in Twitter during the 2016 US presidential election. *Nature communications,* 10**,** 1-14.

32. BRAINES, D., MOTT, D., LAWS, S., DE MEL, G. & PHAM, T. Controlled english to facilitate human/machine analytical processing. Next-Generation Analyst, 2013. International Society for Optics and Photonics, 875808.

33. BRANDWATCH. 2020. *Brandwatch* [Online]. Available: https://www.brandwatch.com/ [Accessed 2020-11-18].

34. BRUNS, A. & BURGESS, J. E. The use of Twitter hashtags in the formation of ad hoc publics. Proceedings of the 6th European consortium for political research (ECPR) general conference 2011, 2011.

35. CALIANDRO, A. 2018. Digital methods for ethnography: Analytical concepts for ethnographers exploring social media environments. *Journal of Contemporary Ethnography,* 47**,** 551-578.

36. CAMACHO-COLLADOS, J., DOVAL, Y., MARTÍNEZ-CÁMARA, E., ESPINOSA-ANKE, L., BARBIERI, F. & SCHOCKAERT, S. 2019. Learning cross-lingual embeddings from twitter via distant supervision. *arXiv preprint arXiv:1905.07358*.

37. CARLEY, K. M., MALIK, M., LANDWEHR, P. M., PFEFFER, J. & KOWALCHUCK, M. 2016. Crowd sourcing disaster management: The complex nature of Twitter usage in Padang Indonesia. *Safety science,* 90**,** 48-61.

38. CASSA, C. A., CHUNARA, R., MANDL, K. & BROWNSTEIN, J. S. 2013. Twitter as a sentinel in emergency situations: lessons from the Boston marathon explosions. *PLoS currents,* 5.

39. CAVALIN, P. R., GATTI, M. A. D. C., DOS SANTOS, C. N. & PINHANEZ, C. Real-time sentiment analysis in social media streams: the 2013 confederation cup case. Proceedings of BRACIS/ENIAC 2014, 2014 São Carlos, SP, Brazil.

40. CHARON, J. M. & HALL, P. 2009. Symbolic interactionism: An introduction, an interpretation, an integration.

41. CHAWLA, N. V., BOWYER, K. W., HALL, L. O. & KEGELMEYER, W. P. 2002. SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research,* 16**,** 321-357.

42. CHEN, H., THOMS, S. & FU, T. Cyber extremism in Web 2.0: An exploratory study of international Jihadist groups. Intelligence and Security Informatics, 2008. ISI 2008. IEEE International Conference on, 2008. IEEE, 98-103.

43. CHEONG, M. & LEE, V. C. 2011. A microblogging-based approach to terrorism informatics: Exploration and chronicling civilian sentiment and response to terrorism events via Twitter. *Information Systems Frontiers,* 13**,** 45-59.

44. CITEULIKE. 2019. *CiteULike* [Online]. Available: http://www.citeulike.org/ [Accessed 2019-01-25].

45. COHEN, W., RAVIKUMAR, P. & FIENBERG, S. A comparison of string metrics for matching names and records. KDD Workshop on Data Cleaning and Object Consolidation, 2003. 73-78.

46. COLLINS, R. 2012. C-escalation and D-escalation: A Theory of the Time-dynamics of Conflict. *American Sociological Review,* 77**,** 1-20.

47. CONBOY, K. 2009. Agility from first principles: Reconstructing the concept of agility in information systems development. *Information systems research,* 20**,** 329-354.

48. CONWAY, M., SCRIVENS, R. & MCNAIR, L. 2019. Right-wing extremists' persistent online presence: History and contemporary trends.

49. CORBIN, J. M. & STRAUSS, A. 1990. Grounded theory research: Procedures, canons, and evaluative criteria. *Qualitative sociology,* 13**,** 3-21.

50. CORTES, C. & VAPNIK, V. 1995. Support-vector networks. *Machine Learning,* 20**,** 273-297.

51. CRESWELL, J. W., KLASSEN, A. C., PLANO CLARK, V. L. & SMITH, K. C. 2011. Best practices for mixed methods research in the health sciences. *Bethesda (Maryland): National Institutes of Health,* 2013**,** 541-545.

52. CROWDTANGLE. 2020. *CrowdTangle | Content Discovery and Social Monitoring Made Easy* [Online]. Available: https://www.crowdtangle.com/ [Accessed 2020-11-18].

53. CUNNINGHAM, H. 2002. GATE, a general architecture for text engineering. *Computers and the Humanities,* 36**,** 223-254.

54. CUNNINGHAM, H., TABLAN, V., ROBERTS, A. & BONTCHEVA, K. 2013. Getting more out of biomedical documents with GATE's full lifecycle open source text analytics. *PLoS computational biology,* 9.

55. D'ANDREA, E., DUCANGE, P., LAZZERINI, B. & MARCELLONI, F. 2015. Real-time detection of traffic from twitter stream analysis. *IEEE Transactions on Intelligent Transportation Systems,* 16**,** 2269-2283.

56. DA SILVA MORAIS, T. Survey on frameworks for distributed computing: Hadoop, spark and storm. Proceedings of the 10th Doctoral Symposium in Informatics Engineering-DSIE, 2015.

57. DAS, S., BEHERA, R. K., RATH, S. K. & OTHERS 2018. Real-Time Sentiment Analysis of Twitter Streaming data for Stock Prediction. *Procedia Computer Science,* 132**,** 956-964.

58. DATABASE SYSTEMS LOGIC PROGRAMMING - UNIVERSITY OF TRIER. 2020. *DBLP Computer Science Bibliography* [Online]. Available: http://dblp.org/ [Accessed 2020-02-05].

59. DAY-RICHTER, J., HARRIS, M. A., HAENDEL, M. & LEWIS, S. 2007. OBO-Edit—an ontology editor for biologists. *Bioinformatics,* 23**,** 2198-2200.

60. DIETVORST, B. J., SIMMONS, J. P. & MASSEY, C. 2018. Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management Science,* 64**,** 1155-1170.

61. DINGSØYR, T., NERUR, S., BALIJEPALLY, V. & MOE, N. B. 2012. A decade of agile methodologies: Towards explaining agile software development. Elsevier.

62. DOBBELAERE, P. & ESMAILI, K. S. Kafka versus RabbitMQ: A comparative study of two industry reference publish/subscribe implementations: Industry Paper. Proceedings of the 11th ACM international conference on distributed and event-based systems, 2017. 227-238.

63. DOBREVA, D., GRINNELL, D. & INNES, M. 2019. Prophets and Loss: How "Soft Facts" on Social Media Influenced the Brexit Campaign and Social Reactions to the Murder of Jo Cox MP. *Policy & Internet*.

64. DOMINGOS, P. & PAZZANI, M. 1997. On the optimality of the simple Bayesian classifier under zero-one loss. *Machine Learning,* 29**,** 103-130.

65. EIBE, F., HALL, M. A., WITTEN, I. H. & PAL, J. 2016. The WEKA workbench. *Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques".* Fourth ed.: Morgan Kaufmann.

66. EKMAN, P. & FRIESEN, W. V. 1971. Constants across cultures in the face and emotion. *Journal of personality and social psychology,* 17**,** 124.

67. ELASTICSEARCH. 2020. *Elasticsearch: The Official Distributed Search & Analytics Engine* [Online]. Available: https://www.elastic.co/elasticsearch/ [Accessed 2020-06-12].

68. ENDERT, A., HOSSAIN, M. S., RAMAKRISHNAN, N., NORTH, C., FIAUX, P. & ANDREWS, C. 2014. The human is the loop: new directions for visual analytics. *Journal of intelligent information systems,* 43**,** 411-435.

69. ESULI, A. & SEBASTIANI, F. 2007. SentiWordNet: a high-coverage lexical resource for opinion mining. *Technical Report 2007-TR-02.* Istituto di Scienza e Tecnologie dell'Informazione.

70. EVANS, K., JONES, A., PREECE, A., QUEVEDO, F., ROGERS, D., SPASIĆ, I., TAYLOR, I., STANKOVSKI, V., TAHERIZADEH, S. & TRNKOCZY, J. Dynamically reconfigurable workflows for time-critical applications. Proceedings of the 10th Workshop on Workflows in Support of Large-Scale Science, 2015. 1-10.

71. FAN, R.-E., CHANG, K.-W., HSIEH, C.-J., WANG, X.-R. & LIN, C.-J. 2008. LIBLINEAR: a library for large linear classification. *Journal of Machine Learning Research,* 9**,** 1871-1874.

72. FIELDING, N. 2001. Ethnography. *In:* GILBERT, N. (ed.) *Researching Social Life.* 2nd ed.: SAGE Publications.

73. FINKEL, J. R., GRENAGER, T. & MANNING, C. Incorporating non-local information into information extraction systems by gibbs sampling. Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, 2005. Association for Computational Linguistics, 363-370.

74. FISHER, D., DELINE, R., CZERWINSKI, M. & DRUCKER, S. 2012. Interactions with big data analytics. *interactions,* 19**,** 50-59.

75. FOLMER, E. & BOSCH, J. 2004. Architecting for usability: a survey. *Journal of systems and software,* 70**,** 61-78.

76. FOWLER, M. & HIGHSMITH, J. 2001. The agile manifesto. *Software Development,* 9**,** 28-35.

77. FREELON, D. 2018. Computational research in the post-API age. *Political Communication,* 35**,** 665-668.

78. FUENTES, A. M. M., KAHN, J. H. & LANNIN, D. G. 2018. Emotional disclosure and emotion change during an expressive-writing task: Do pronouns matter? *Current Psychology***,** 1-8.

79. FURLOW, R. B. & GOODALL, H. 2011. The war of ideas and the battle of narratives: A comparison of extremist storytelling structures. *Cultural Studies↔ Critical Methodologies,* 11**,** 215-223.

80. GALE, W. A., CHURCH, K. & YAROWSKY, D. One sense per discourse. Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992, 1992.

81. GAO, H., BARBIER, G. & GOOLSBY, R. 2011. Harnessing the crowdsourcing power of social media for disaster relief. *IEEE Intelligent Systems,* 26**,** 10-14.

82. GARRY, A., WALTHER, S., RUKAYA, R. & MOHAMMED, A. 2021. QAnon Conspiracy Theory: Examining its Evolution and Mechanisms of Radicalization. *Journal for Deradicalization***,** 152-216.

83. GAUTHEIR, T. D. 2001. Detecting trends using Spearman's rank correlation coefficient. *Environmental forensics,* 2**,** 359-362.

84. GAYO-AVELLO, D. 2013. A meta-analysis of state-of-the-art electoral prediction from Twitter data. *Social Science Computer Review,* 31**,** 649-679.

85. GEHRKE, P. J. 2014. Ecological validity and the study of publics: The case for organic public engagement methods. *Public understanding of science,* 23**,** 77-91.

86. GESING, S., ATKINSON, M., FILGUEIRA, R., TAYLOR, I., JONES, A., STANKOVSKI, V., LIEW, C. S., SPINUSO, A., TERSTYANSZKY, G. & KACSUK, P. Workflows in a dashboard: a new generation of usability. 2014 9th Workshop on Workflows in Support of Large-Scale Science, 2014. IEEE, 82-93.

87. GHAZI, M. R. & GANGODKAR, D. 2015. Hadoop, MapReduce and HDFS: a developers perspective. *Procedia Computer Science,* 48**,** 45-50.

88. GLASSMAN, M. & KANG, M. J. 2012. Intelligence in the internet age: The emergence and evolution of Open Source Intelligence (OSINT). *Computers in Human Behavior,* 28**,** 673-682.

89. GLEASON, B. 2013. # Occupy Wall Street: Exploring informal learning about a social movement on Twitter. *American Behavioral Scientist,* 57**,** 966-982.

90. GO, A., BHAYANI, R. & HUANG, L. 2009a. Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford,* 1.

91. GO, A., BHAYANI, R. & HUANG, L. 2009b. Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford***,** 1-12.

92. GOLESTANI, A., MASLI, M., SHAMI, N. S., JONES, J., MENON, A. & MONDAL, J. Real-time prediction of employee engagement using social media and text mining. 17th IEEE International Conference on Machine Learning and Applications (ICMLA), 2018 Orlando, FL. IEEE, 1383-1387.

93. GOOGLE. 2020. *Google Scholar* [Online]. Available: https://scholar.google.co.uk/ [Accessed 2020-02-05].

94. GOUTTE, C. & GAUSSIER, E. 2005. A probabilistic interpretation of precision, recall and F-score, with implication for evaluation. *In:* LOSADA, D. E. & FERNÁNDEZ-LUNA, J. M. (eds.) *Advances in Information Retrieval. ECIR 2005. Lecture Notes in Computer Science.* Berlin, Heidelberg: Springer.

95. GRAY, M., BADRE, A. & GUZDIAL, M. Visualizing usability log data. Proceedings IEEE Symposium on Information Visualization'96, 1996. IEEE, 93-98.

96. GREFENSTETTE, G. & TAPANAINEN, P. What is a word, what is a sentence?: problems of Tokenisation. Papers in computational lexicography, COMPLEX '94, 1994 Budapest, Hungary.

97. GRIBBLE, S. D. Robustness in complex systems. Proceedings eighth workshop on hot topics in operating systems, 2001. IEEE, 21-26.

98. GUDIVADA, V. N., RAO, D. & RAGHAVAN, V. V. NoSQL systems for big data management. 2014 IEEE World congress on services, 2014. IEEE, 190-197.

99. GUEST, G., MACQUEEN, K. M. & NAMEY, E. E. 2012. Introduction to applied thematic analysis. *Applied thematic analysis,* 3**,** 20.

100. HAEFNER, N. 2014. Selling & the Social Media Honeycomb. *NIBL* [Online]. [Accessed 02/07/2015 2015].

101. HALL, M., FRANK, E., HOLMES, G., PFAHRINGER, B., REUTEMANN, P. & WITTEN, I. H. 2009. The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter,* 11**,** 10-18.

102. HARTSWOOD, M., PROCTER, R., SLACK, R., VOB, A., BUSCHER, M., ROUNCEFIELD, M. & ROUCHY, P. 2002. Co-realisation: towards a principled synthesis of ethnomethodology and participatory design. *Scandinavian Journal of Information Systems,* 14**,** 2.

103. HENDERSON-SELLERS, B. & SEROUR, M. 2005. Creating a dual-agility method: The value of method engineering. *Journal of Database Management (JDM),* 16**,** 1-24.

104. HOCHREITER, S. & SCHMIDHUBER, J. 1997. Long short-term memory. *Neural Computation,* 9**,** 1735-1780.

105. HORNBÆK, K. 2006. Current practice in measuring usability: Challenges to usability studies and research. *International journal of human-computer studies,* 64**,** 79-102.

106. HOTHO, A., NÜRNBERGER, A. & PAAß, G. A Brief Survey of Text Mining. Ldv Forum, 2005. 19-62.

107. HUGHES, A. L. & SHAH, R. Designing an application for social media needs in emergency public information work. Proceedings of the 19th International Conference on Supporting Group Work, 2016. 399-408.

108. INNES, M. 2001. Exemplar: Investigating the investigators–studying detective work. *In:* GILBERT, N. (ed.) *Researching Social Life.* 2nd ed.: SAGE Publications.

109. INNES, M. 2020. Techniques of disinformation: Constructing and communicating “soft facts” after terrorism. *The British Journal of Sociology,* 71**,** 284-299.

110. INNES, M., DOBREVA, D. & INNES, H. 2019. Disinformation and digital influencing after terrorism: spoofing, truthing and social proofing. *Contemporary Social Science*, 1-15.

111. INNES, M., INNES, H., ROBERTS, C., HARMSTON, D. & GRINNELL, D. 2021. The normalisation and domestication of digital disinformation: on the alignment and consequences of far-right and Russian State (dis) information operations and campaigns in Europe. *Journal of Cyber Policy*, 1-19.

112. INNES, M., ROBERTS, C., PREECE, A. & ROGERS, D. 2017. Of instruments and data: Social media uses, abuses and analysis. *The SAGE Handbook of Online Research Methods*, 108-124.

113. INNES, M., ROBERTS, C., PREECE, A. & ROGERS, D. 2018. Ten "Rs" of social reaction: Using social media to analyse the "post-event" impacts of the murder of Lee Rigby. *Terrorism and Political Violence,* 30, 454-474.

114. INNES, M., ROBERTS, C. & ROGERS, D. 2014. Critical Timing: Social Media and the Golden Hour. *Police Professional Magazine,* 16.

115. INSTITUTE OF ELECTRICAL ELECTRONICS ENGINEERS. 2020. *IEEE Xplore* [Online]. Available: http://ieeexplore.ieee.org [Accessed 2020-02-05].

116. JELODAR, H., WANG, Y., YUAN, C., FENG, X., JIANG, X., LI, Y. & ZHAO, L. 2019. Latent Dirichlet Allocation (LDA) and Topic modeling: models, applications, a survey. *Multimedia Tools and Applications,* 78, 15169-15211.

117. JEN, E. 2005. Stable or robust? What's the difference? *Robust Design: a repertoire of biological, ecological, and engineering case studies, SFI Studies in the Sciences of Complexity*, 7-20.

118. JOHN, V. & LIU, X. 2017. A survey of distributed message broker queues. *arXiv preprint arXiv:1704.00411*.

119. JOHNSON, R. B., ONWUEGBUZIE, A. J. & TURNER, L. A. 2007. Toward a definition of mixed methods research. *Journal of mixed methods research,* 1, 112-133.

120. KANG, R., BROWN, S. & KIESLER, S. Why do people seek anonymity on the internet? Informing policy and design. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 2013. 2657-2666.

121. KARANASOU, M., AMPLA, A., DOULKERIDIS, C. & HALKIDI, M. Scalable and real-time sentiment analysis of twitter data. *In:* DOMENICONI, C., GULLO, F., BONCHI, F., DOMINGO-FERRER, J., BAEZA-YATES, R. A., ZHOU, Z.-H. & WU, X., eds. 2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW), 2016 Barcelona. IEEE, 944-951.

122. KAREGOWDA, A. G., MANJUNATH, A. S. & JAYARAM, M. A. 2010. Comparative study of attribute selection using gain ratio and correlation based feature selection. *International Journal of Information Technology and Knowledge Management,* 2, 271-277.

123. KARRAY, F., ALEMZADEH, M., ABOU SALEH, J. & ARAB, M. N. 2017. Human-computer interaction: Overview on state of the art. *International journal on smart sensing and intelligent systems,* 1.

124. KAY, S., ZHAO, B. & SUI, D. 2015. Can social media clear the air? A case study of the air pollution problem in Chinese cities. *The Professional Geographer,* 67**,** 351-363.

125. KESSLER, J. S. 2017. Scattertext: a browser-based tool for visualizing how corpora differ. *arXiv preprint arXiv:1703.00565*.

126. KIETZMANN, J. H., HERMKENS, K., MCCARTHY, I. P. & SILVESTRE, B. S. 2011. Social media? Get serious! Understanding the functional building blocks of social media. *Business horizons,* 54**,** 241-251.

127. KIM, Y. 2014. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*.

128. KONONENKO, O., BAYSAL, O., HOLMES, R. & GODFREY, M. W. Mining modern repositories with elasticsearch. Proceedings of the 11th working conference on mining software repositories, 2014. 328-331.

129. KONTIO, J., LEHTOLA, L. & BRAGGE, J. Using the focus group method in software engineering: obtaining practitioner and user experiences. Proceedings. 2004 International Symposium on Empirical Software Engineering, 2004. ISESE'04., 2004. IEEE, 271-280.

130. KRISTENSEN, M., KYNG, M. & PALEN, L. Participatory design in emergency medical service: designing for future practice. Proceedings of the SIGCHI conference on Human Factors in computing systems, 2006. 161-170.

131. KUCKARTZ, U. 2007. MAXQDA: Qualitative data analysis. *Berlin: VERBI software*.

132. KURNIAWAN, D. A., WIBIRAMA, S. & SETIAWAN, N. A. Real-time traffic classification with twitter data mining. 2016 8th International Conference on Information Technology and Electrical Engineering (ICITEE), 2016 Yogyakarta. 1-5.

133. LAWRENCE, E., SIDES, J. & FARRELL, H. 2010. Self-Segregation or Deliberation? Blog Readership, Participation, and Polarization in American Politics. *Perspectives on Politics,* 8**,** 141-157.

134. LAZER, D., KENNEDY, R., KING, G. & VESPIGNANI, A. 2014. The parable of Google Flu: traps in big data analysis. *Science,* 343**,** 1203-1205.

135. LECUN, Y., BOTTOU, L., BENGIO, Y. & HAFFNER, P. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE,* 86**,** 2278-2324.

136. LEE, G. & XIA, W. 2010. Toward agile: an integrated analysis of quantitative and qualitative field data on software development agility. *MIS quarterly,* 34**,** 87-114.

137. LEE, K., AGRAWAL, A. & CHOUDHARY, A. Mining social media streams to improve public health allergy surveillance. 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), 2015 Paris. 815-822.

138. LEETARU, K. & SCHRODT, P. A. Gdelt: Global data on events, location, and tone, 1979–2012. ISA annual convention, 2013. Citeseer, 1-49.

139. LEFEBVRE, R. K. & ARMSTRONG, C. 2018. Grievance-based social movement mobilization in the# Ferguson Twitter storm. *New Media & Society,* 20**,** 8-28.

140. LEFF, A. & RAYFIELD, J. T. Web-application development using the model/view/controller design pattern. Proceedings fifth ieee international enterprise distributed object computing conference, 2001. IEEE, 118-127.

141. LIU, S. B., PALEN, L., SUTTON, J., HUGHES, A. L. & VIEWEG, S. In search of the bigger picture: The emergent role of on-line photo sharing in times of disaster. Proceedings of the information systems for crisis response and management conference (ISCRAM), 2008. Citeseer, 4-7.

142. LIU, Y., KLIMAN-SILVER, C. & MISLOVE, A. The tweets they are a-changin': Evolution of twitter users and behavior. Eighth International AAAI Conference on Weblogs and Social Media, 2014.

143. LOPER, E. & BIRD, S. 2002. NLTK: the natural language toolkit. *arXiv preprint cs/0205028*.

144. LOWE-CALVERLEY, E. & GRIEVE, R. 2018. Thumbs up: A thematic analysis of image-based posting and liking behaviour on social media. *Telematics and Informatics,* 35**,** 1900-1913.

145. LUKE, A. 1997. Theory and practice in critical discourse analysis. *International encyclopedia of the sociology of education,* 8**,** 50-57.

146. MA, J., FENG, C., SHI, G., SHI, X. & HUANG, H. 2018. Temporal enhanced sentence-level attention model for hashtag recommendation. *CAAI Transactions on Intelligence Technology,* 3**,** 95-100.

147. MACDONALD, S., GRINNELL, D., KINZEL, A. & LORENZO-DUS, N. 2019. Daesh, Twitter and the Social Media Ecosystem: A Study of Outlinks Contained in Tweets Mentioning Rumiyah. *The RUSI Journal,* 164**,** 60-72.

148. MANE, S. B., SAWANT, Y., KAZI, S. & SHINDE, V. 2014. Real time sentiment analysis of twitter data using hadoop. *International Journal of Computer Science and Information Technologies,* 5**,** 3098-3100.

149. MANNING, C. D., SURDEANU, M., BAUER, J., FINKEL, J., BETHARD, S. J. & MCCLOSKY, D. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations***,** 55-60.

150. MANTYLA, M. V., CLAES, M. & FAROOQ, U. Measuring LDA topic stability from clusters of replicated runs. Proceedings of the 12th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, 2018. 1-4.

151. MARCUS, M., SANTORINI, B. & MARCINKIEWICZ, M. A. 1993. Building a large annotated corpus of English: The Penn Treebank.

152. MAYNARD, D., LI, Y. & PETERS, W. 2008. NLP Techniques for Term Extraction and Ontology Population. *Proceedings of the 2008 conference on Ontology Learning and Population: Bridging the Gap between Text and Knowledge.* IOS Press.

153. MAYNARD, D. G., ROBERTS, I., GREENWOOD, M. A., ROUT, D. & BONTCHEVA, K. 2017. A Framework for Real-time Semantic Social Media Analysis.

154. MCCORMICK, T. H., LEE, H., CESARE, N., SHOJAIE, A. & SPIRO, E. S. 2017. Using Twitter for demographic and social science research: Tools for data collection and processing. *Sociological methods & research,* 46**,** 390-421.

155. MELLON, J. & PROSSER, C. 2017. Twitter and Facebook are not representative of the general population: Political attitudes and demographics of British social media users. *Research & Politics,* 4**,** 2053168017720008.

156. MICHAILIDIS, D., STYLIANOU, N. & VLAHAVAS, I. Real time location based sentiment analysis on twitter: the AirSent system. SETN '18: Proceedings of the 10th Hellenic Conference on Artificial Intelligence, 2018 Patras Greece. Association for Computing Machinery, 21:1-21:4.

157. MIDDLETON, S. E. & KRIVCOVS, V. 2016. Geoparsing and geosemantics for social media: spatio-temporal grounding of content propagating rumours to support trust and veracity analysis during breaking news. *ACM Transactions on Information Systems,* 34**,** 1-27.

158. MIKOLOV, T., CHEN, K., CORRADO, G. & DEAN, J. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

159. MILLER, G. A. 1956. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological review,* 63**,** 81.

160. MILLER, G. A. 1995. WordNet: a lexical database for English. *Communications of the ACM,* 38**,** 39-41.

161. MOHER, D., LIBERATI, A., TETZLAFF, J. & ALTMAN, D. G. 2009. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *BMJ,* 339**,** b2535.

162. MOTT, D. 2010. Summary of ITA controlled english. *ITA Technical Paper, http://www. usukita. org*.

163. MUHR, T. 1997. *ATLAS. ti: The knowledge workbench: Visual qualitative data, analysis, management, model building: Short user's manual*, Scientific Software Development.

164. MULLER, M. J. & DRUIN, A. 2012. Participatory design: the third space in human–computer interaction. *The Human–Computer Interaction Handbook***,** 1125-1153.

165. MULLER, M. J. & KUHN, S. 1993. Participatory design. *Communications of the ACM,* 36**,** 24-28.

166. MURTHY, D. 2008. Digital ethnography: An examination of the use of new technologies for social research. *Sociology,* 42**,** 837-855.

167. MURTHY, D. 2011. Twitter: Microphone for the masses? *Media, culture & society,* 33**,** 779-789.

168. NAKOV, P., RITTER, A., ROSENTHAL, S., SEBASTIANI, F. & STOYANOV, V. SemEval-2016 task 4: Sentiment analysis in twitter. Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016), June 2016 San Diego, California. Association for Computational Linguistics, 1-18.

169. NAKOV, P., ZESCH, T., CER, D. & JURGENS, D. Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015). 2015 Denver, CO. Association for Computational Linguistics.

170. NAYAK, A., PORIYA, A. & POOJARY, D. 2013. Type of NOSQL databases and its comparison with relational databases. *International Journal of Applied Information Systems,* 5**,** 16-19.

171. NEUENSCHWANDER, B., PEREIRA, A. C. M., MEIRA, J., WAGNER & BARBOSA, D. Sentiment analysis for streams of web data: a case study of Brazilian financial markets. Proceedings of the 20th Brazilian Symposium on Multimedia and the Web, 2014 New York, NY. ACM, 167-170.

172. NEWMAN, D., LAU, J. H., GRIESER, K. & BALDWIN, T. Automatic evaluation of topic coherence. Human language technologies: The 2010 annual conference of the North American chapter of the association for computational linguistics, 2010. 100-108.

173. NGUYEN, H., LIU, W., RIVERA, P. & CHEN, F. 2016. *TrafficWatch: real-time traffic incident detection and monitoring using social media*, Springer.

174. NIELSEN, J. Usability inspection methods. Conference companion on Human factors in computing systems, 1994. 413-414.

175. NIKOLENKO, S. I., KOLTCOV, S. & KOLTSOVA, O. 2017. Topic modelling for qualitative studies. *Journal of Information Science,* 43**,** 88-102.

176. NIWATTANAKUL, S., SINGTHONGCHAI, J., NAENUDORN, E. & WANAPU, S. Using of Jaccard coefficient for keywords similarity. Proceedings of the international multiconference of engineers and computer scientists, 2013. 380-384.

177. NOY, N. F. & MCGUINNESS, D. L. 2001. Ontology development 101: A guide to creating your first ontology. Stanford Knowledge Systems Laboratory Technical Report KSL-01-05

178. O'CONNOR, B., BALASUBRAMANYAN, R., ROUTLEDGE, B. R. & SMITH, N. A. 2010. From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series. *ICWSM,* 11**,** 122-129.

179. OH, O., AGRAWAL, M. & RAO, H. R. 2011. Information control and terrorism: Tracking the Mumbai terrorist attack through twitter. *Information Systems Frontiers,* 13**,** 33-43.

180. OUZZANI, M., HAMMADY, H., FEDOROWICZ, Z. & ELMAGARMID, A. 2016. Rayyan - a web and mobile app for systematic reviews. *Systematic Reviews,* 5**,** 210.

181. PAK, A. & PAROUBEK, P. Twitter as a Corpus for Sentiment Analysis and Opinion Mining. LREC, 2010.

182. PEARY, B. D., SHAW, R. & TAKEUCHI, Y. 2012. Utilization of social media in the east Japan earthquake and tsunami and its effectiveness. *Journal of Natural Disaster Science,* 34**,** 3-18.

183. PEDREGOSA, F., VAROQUAUX, G., GRAMFORT, A., MICHEL, V., THIRION, B., GRISEL, O., BLONDEL, M., PRETTENHOFER, P., WEISS, R., DUBOURG, V., VANDERPLAS, J., PASSOS, A., COURNAPEAU, D., BRUCHER, M., PERROT, M. & DUCHESNAY, E. 2011. Scikit-learn: machine learning in Python. *Journal of Machine Learning Research,* 12**,** 2825-2830.

184. PENNEBAKER, J. W., FRANCIS, M. E. & BOOTH, R. J. 2001. *Linguistic inquiry and word count: LIWC 2001*, Erlbaum Publishers, Mahwah, NJ.

185. PENNYCOOK, G., MCPHETRES, J., ZHANG, Y. & RAND, D. 2020. Fighting COVID-19 misinformation on social media: Experimental evidence for a scalable accuracy nudge intervention. *PsyArXiv Preprints,* 10.

186. PFLUKE, C. 2019. A history of the five eyes alliance: possibility for reform and additions: a history of the five eyes alliance: possibility for reform and additions. *Comparative Strategy,* 38**,** 302-315.

187. PIRINEN, A. 2016. The barriers and enablers of co-design for services. *International Journal of Design,* 10**,** 27-42.

188. PIROLLI, P. & CARD, S. The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. Proceedings of international conference on intelligence analysis, 2005. McLean, VA, USA, 2-4.

189. PLATT, J. 1998. Sequential minimal optimization: a fast algorithm for training support vector machines. *Technical Report No. MST TR 98(14).* Microsoft Research.

190. PORTER, M. F. 2001. *Snowball: A language for stemming algorithms* [Online]. Available: http://snowball.tartarus.org/texts/introduction.html [Accessed 2020-02-04].

191. POURSEPANJ, H., WEISSBOCK, J. & INKPEN, D. uottawa: System description for semeval 2013 task 2 sentiment analysis in twitter. Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013), 2013. 380-383.

192. PREECE, A., GWILLIAMS, C., PARIZAS, C., PIZZOCARO, D., BAKDASH, J. Z. & BRAINES, D. Conversational sensing. Next-Generation Analyst II, 2014. International Society for Optics and Photonics, 91220I.

193. PREECE, A., ROBERTS, C., ROGERS, D., WEBBERLEY, W., INNES, M. & BRAINES, D. From open source communications to knowledge. Next-Generation Analyst IV, 2016. International Society for Optics and Photonics, 98510K.

194. PREECE, A., SPASIĆ, I., EVANS, K., ROGERS, D., WEBBERLEY, W., ROBERTS, C. & INNES, M. 2018. Sentinel: A Codesigned Platform for Semantic Enrichment of Social Media Streams. *IEEE Transactions on Computational Social Systems,* 5**,** 118-131.

195. PREECE, A., WEBBERLEY, W. & BRAINES, D. Tasking the tweeters: Obtaining actionable information from human sensors. Ground/Air Multisensor Interoperability, Integration, and Networking for Persistent ISR VI, 2015. International Society for Optics and Photonics, 946402.

196. PROCTER, R., CRUMP, J., KARSTEDT, S., VOSS, A. & CANTIJOCH, M. 2013. Reading the riots: What were the police doing on Twitter? *Policing and society,* 23**,** 413-436.

197. PUBLIC HEALTH ENGLAND. 2020. *The Health Protection (Coronavirus, Restrictions) (England) Regulations 2020* [Online]. Available: http://www.legislation.gov.uk/uksi/2020/350/contents/made/data.htm [Accessed 2020-06-27].

198. QUINLAN, J. R. 1986. Induction of decision trees. *Machine Learning,* 1**,** 81-106.

199. RAHNAMA, A. H. A. Distributed real-time sentiment analysis for big data social streams. 2014 International conference on control, decision and information technologies (CoDIT), 2014. IEEE, 789-794.

200. RAMOS, J. Using tf-idf to determine word relevance in document queries. Proceedings of the first instructional conference on machine learning, 2003. 133-142.

201. RAVINDRAN, A. 2015. *Django Design Patterns and Best Practices*, Packt Publishing Ltd.

202. ŘEHŮŘEK, R. & SOJKA, P. Software framework for topic modelling with large corpora. In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, 2010. Citeseer.

203. REUTER, C., HUGHES, A. L. & KAUFHOLD, M.-A. 2018. Social media in crisis management: An evaluation and analysis of crisis informatics research. *International Journal of Human–Computer Interaction,* 34**,** 280-294.

204. REZAEI, Z. & JALALI, M. Sentiment analysis on Twitter using McDiarmid tree algorithm. 2017 7th International Conference on Computer and Knowledge Engineering (ICCKE), 2017 Mashhad. 33-36.

205. RICHARDS, L. 1999. *Using NVivo in qualitative research*, Sage.

206. RIESSMAN, C. K. 1993. *Narrative analysis*, Sage.

207. ROBERTS, C., INNES, M., PREECE, A. & ROGERS, D. 2018. After Woolwich: Analyzing open source communications to understand the interactive and multi-polar dynamics of the arc of conflict. *The British Journal of Criminology,* 58**,** 434-454.

208. ROBERTS, C., INNES, M., PREECE, A. & SPASIĆ, I. 2015. Soft facts and spontaneous community mobilisation: the role of rumour after major crime events. *Data for Good: How big and open data can be used for the common good, P. Baeck, ed***,** 37-43.

209. ROBINSON, K. M. 2001. Unsolicited narratives from the Internet: A rich source of qualitative data. *Qualitative Health Research,* 11**,** 706-714.

210. RÖDER, M., BOTH, A. & HINNEBURG, A. Exploring the space of topic coherence measures. Proceedings of the eighth ACM international conference on Web search and data mining, 2015. 399-408.

211. ROGERS, D., HARVEY, I., HUU, T. T., EVANS, K., GLATARD, T., KALLEL, I., TAYLOR, I., MONTAGNAT, J., JONES, A. & HARRISON, A. 2013. Bundle and pool architecture for multi-language, robust, scalable workflow executions. *Journal of grid computing,* 11**,** 457-480.

212. ROGERS, R. 2020. Deplatforming: Following extreme Internet celebrities to Telegram and alternative social media. *European Journal of Communication***,** 0267323120922066.

213. ROSCHELLE, J., PENUEL, W. & SHECHTMAN, N. 2006. Co-design of innovations with teachers: Definition and dynamics.

214. SAKAKI, T., OKAZAKI, M. & MATSUO, Y. Earthquake shakes Twitter users: real-time event detection by social sensors. Proceedings of the 19th international conference on World wide web, 2010. 851-860.

215. SALZMANN-ERIKSON, M. & HIÇDURMAZ, D. 2017. Use of social media among individuals who suffer from post-traumatic stress: A qualitative analysis of narratives. *Qualitative health research,* 27**,** 285-294.

216. SAMPSON, J., MORSTATTER, F., MACIEJEWSKI, R. & LIU, H. Surpassing the limit: Keyword clustering to improve Twitter sample coverage. Proceedings of the 26th ACM conference on hypertext & social media, 2015. 237-245.

217. SANDERS, E.-N. 2000. Generative tools for co-designing. *Collaborative design.* Springer.

218. SANDERS, E. B.-N. & WESTERLUND, B. 2011. Experiencing, exploring and experimenting in and with co-design spaces. *Nordes*.

219. ȘERBAN, O., THAPEN, N., MAGINNIS, B., HANKIN, C. & FOOT, V. 2019. Real-time processing of social media with SENTINEL: A syndromic surveillance system incorporating deep learning for health classification. *Information Processing & Management,* 56**,** 1166-1184.

220. SHEPHERD, A., SANDERS, C., DOYLE, M. & SHAW, J. 2015. Using social media for support and feedback by mental health service users: thematic analysis of a twitter conversation. *BMC psychiatry,* 15**,** 29.

221. SHETH, A. 2009. Citizen sensing, social signals, and enriching human experience. *IEEE Internet Computing,* 13**,** 87-92.

222. SIEVERT, C. & SHIRLEY, K. LDAvis: A method for visualizing and interpreting topics. Proceedings of the workshop on interactive language learning, visualization, and interfaces, 2014. 63-70.

223. SNELSON, C. L. 2016. Qualitative and mixed methods social media research: A review of the literature. *International Journal of Qualitative Methods,* 15**,** 1609406915624574.

224. SOCHER, R., PERELYGIN, A., WU, J., CHUANG, J., MANNING, C. D., NG, A. Y. & POTTS, C. Recursive deep models for semantic compositionality over a sentiment treebank. Proceedings of the 2013 conference on empirical methods in natural language processing, 2013. 1631-1642.

225. SPASIĆ, I., GREENWOOD, M., PREECE, A., FRANCIS, N. & ELWYN, G. 2013. FlexiTerm: A flexible term recognition method. *Journal of Biomedical Semantics,* 4**,** 27.

226. SPINUZZI, C. 2005. The methodology of participatory design. *Technical communication,* 52**,** 163-174.

227. STEED, C. A., DROUHARD, M., BEAVER, J., PYLE, J. & BOGEN, P. L. Matisse: A visual analytics system for exploring emotion trends in social media text streams. 2015 IEEE International Conference on Big Data (Big Data), 2015 Santa Clara, CA. 807-814.

228. STEEN, M. 2013. Co-design as a process of joint inquiry and imagination. *Design Issues,* 29**,** 16-28.

229. STEEN, M., MANSCHOT, M. & DE KONING, N. 2011. Benefits of co-design in service design projects. *International Journal of Design,* 5.

230. STENETORP, P., PYYSALO, S., TOPIĆ, G., OHTA, T., ANANIADOU, S. & TSUJII, J. I. BRAT: a web-based tool for NLP-assisted text annotation. Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics, 2012. Association for Computational Linguistics, 102-107.

231. STEWART, M. C. & ARNOLD, C. L. 2018. Defining social listening: Recognizing an emerging dimension of listening. *International Journal of Listening,* 32**,** 85-100.

232. STRAPPARAVA, C., VALITUTTI, A. & OTHERS. WordNet affect: an affective extension of WordNet. Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04), 2004 Lisbon. European Language Resources Association (ELRA), 1083-1086.

233. SUBRAMANI, S., MICHALSKA, S., WANG, H., WHITTAKER, F. & HEYWARD, B. 2018. *Text mining and real-time analytics of twitter data: a case study of australian hay fever prediction*, Springer.

234. SUSARLA, A., OH, J.-H. & TAN, Y. 2012. Social networks and the diffusion of user-generated content: Evidence from YouTube. *Information Systems Research,* 23**,** 23-41.

235. TADDY, M. On estimation and selection for topic models. Artificial Intelligence and Statistics, 2012. 1184-1193.

236. TALHAOUI, M. A., EL BOUR, H. A., MOULOUKI, R., NKIRI, S. & AZOUAZI, M. 2018. An Improved Social Media Analysis on Three Layers: A Real Time Enhanced Recommendation System. *statistics,* 9.

237. TAYLOR, I., SHIELDS, M., WANG, I. & HARRISON, A. 2007. The triana workflow environment: Architecture and applications. *Workflows for e-Science.* Springer.

238. TERVEEN, L. G. 1995. Overview of human-computer collaboration. *Knowledge-Based Systems,* 8**,** 67-81.

239. THELWALL, M., BUCKLEY, K. & PALTOGLOU, G. 2011. Sentiment in Twitter events. *Journal of the American Society for Information Science and Technology,* 62**,** 406-418.

240. THELWALL, M., SUD, P. & VIS, F. 2012. Commenting on YouTube videos: From Guatemalan rock to el big bang. *Journal of the American Society for Information Science and Technology,* 63**,** 616-629.

241. TÖRNBERG, A. & TÖRNBERG, P. 2016. Muslims in social media discourse: Combining topic modeling and critical discourse analysis. *Discourse, Context & Media,* 13**,** 132-142.

242. TOUTANOVA, K., KLEIN, D., MANNING, C. D. & SINGER, Y. Feature-rich part-of-speech tagging with a cyclic dependency network. Proceedings of the 2003 conference of the North American chapter of the association for computational linguistics on human language technology-volume 1, 2003. Association for Computational Linguistics, 173-180.

243. TRUPTHI, M., PABBOJU, S. & NARASIMHA, G. Sentiment analysis on twitter using streaming API. 2017 IEEE 7th International Advance Computing Conference (IACC), 2017. IEEE, 915-919.

244. TUXWORTH, D., ANTYPAS, D., ESPINOSA-ANKE, L., CAMACHO-COLLADOS, J., PREECE, A. & ROGERS, D. 2021. Deriving Disinformation Insights from Geolocalized Twitter Callouts. *Workshop On Deriving Insights From User-Generated Text, KDD2021.* ACM.

245. TWITTER. 2020. *POST statuses/filter | Docs | Twitter Developer* [Online]. Available: https://developer.twitter.com/en/docs/twitter-api/v1/tweets/filter-realtime/api-reference/post-statuses-filter [Accessed 2020-11-18].

246. USCHOLD, M. Building ontologies: Towards a unified methodology. 16th Annual Conf. of the British Computer Society Specialist Group on Expert Systems, Cambridge, UK, 1996. Citeseer.

247. VAGIAS, W. M. 2006. Likert-type scale response anchors. *Clemson International Institute for Tourism & Research Development, Department of Parks, Recreation and Tourism Management. Clemson University*.

248. VAN BAVEL, J. J., BAICKER, K., BOGGIO, P. S., CAPRARO, V., CICHOCKA, A., CIKARA, M., CROCKETT, M. J., CRUM, A. J., DOUGLAS, K. M. & DRUCKMAN, J. N. 2020. Using social and behavioural science to support COVID-19 pandemic response. *Nature Human Behaviour***,** 1-12.

249. VICENTE, I. S., SARALEGI, X. & AGERRI, R. 2018. Talaia: a real time monitor of social media and digital press. *ArXiv,* abs/1810.00647.

250. VILARES, D., HERMO, M., ALONSO, M. A., GÓMEZ-RODRÍGUEZ, C. & VILARES, J. U., S. LyS at CLEF RepLab 2014: creating the state of the art in author influence ranking and reputation classification on twitter. *In:* CAPPELLATO, L., FERRO, N., HALVEY, M. & KRAAIJ, W., eds. Proceedings of the 5th International Conference of the CLEF Initiative (RepLab 2014), 2014 Sheffield. CEUR-WS.org, 1468-1478.

251. VON AHN, L. 2009. Offensive/profane word list. *Retrieved June,* 24**,** 2018.

252. WALSH, G., DRUIN, A., GUHA, M. L., FOSS, E., GOLUB, E., HATLEY, L., BONSIGNORE, E. & FRANCKEL, S. Layered elaboration: a new technique for co-design with children. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 2010. 1237-1240.

253. WAYBACK MACHINE. 2020. *CiteULike is closing down* [Online]. Available: https://web.archive.org/web/20190310145602/citeulike.org/news [Accessed 2020-02-05].

254. WEIMANN, G. 2010. Terror on facebook, twitter, and youtube. *The Brown Journal of World Affairs,* 16**,** 45-54.

255. WILLIAMS, H. T., MCMURRAY, J. R., KURZ, T. & LAMBERT, F. H. 2015. Network analysis reveals open forums and echo chambers in social media discussions of climate change. *Global environmental change,* 32**,** 126-138.

256. WILLIAMS, L., ARRIBAS-AYLLON, M., ARTEMIOU, A. & SPASIĆ, I. 2019. Comparing the utility of different classification schemes for emotive language analysis. *Journal of Classification,* 36**,** 619-648.

257. WILLIAMS, R. J. & ZIPSER, D. 1989. A learning algorithm for continually running fully recurrent neural networks. *Neural Computation,* 1**,** 270-280.

258. WIN, S. S. M. & AUNG, T. N. Target oriented tweets monitoring system during natural disasters. 2017 IEEE/ACIS 16th International Conference on Computer and Information Science (ICIS), 2017 Wuhan. 143-148.

259. WINKLER, W. E. The state of record linkage and current research problems. Statistical Research Division, US Census Bureau, 1999. Citeseer.

260. WORLD HEALTH ORGANIZATION. 2020a. *Statement on the Second Meeting of the International Health Regulations. Emergency Committee regarding the outbreak of novel coronavirus (2019-nCoV); 2005* [Online]. Available: https://www.who.int/news-room/detail/30-

01-2020-statement-on-the-second-meeting-of-the-international-health-regulations-(2005)-emergency-committee-regarding-the-outbreak-of-novel-coronavirus-(2019-ncov) [Accessed 2020-06-27].

261. WORLD HEALTH ORGANIZATION. 2020b. *WHO Director-General's opening remarks at the media briefing on COVID-19-11 March 2020* [Online]. Available: https://www.who.int/dg/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020 [Accessed 2020-06-27].

262. WU, C., SCHWARTZ, J.-M. & NENADIC, G. 2013. PathNER: a tool for systematic identification of biological pathway mentions in the literature. *BMC systems biology,* 7**,** S2.

263. WU, T., WEN, S., XIANG, Y. & ZHOU, W. 2018. Twitter spam detection: Survey of new approaches and comparative study. *Computers & Security,* 76**,** 265-284.

264. WU, Z. & PALMER, M. Verbs semantics and lexical selection. Proceedings of the 32nd annual meeting on Association for Computational Linguistics, 1994. Association for Computational Linguistics, 133-138.

265. YOUTUBE. 2020. *YouTube Data API Overview* [Online]. Available: https://developers.google.com/youtube/v3/getting-started [Accessed 2020-11-18].

266. YU, F., MOH, M. & MOH, T. S. Towards extracting drug-effect relation from twitter: a supervised learning approach. 2016 IEEE 2nd International Conference on Big Data Security on Cloud (BigDataSecurity), IEEE International Conference on High Performance and Smart Computing (HPSC), and IEEE International Conference on Intelligent Data and Security (IDS), 2016 New York, NY. 339-344.

267. ZAMENOPOULOS, T. & ALEXIOU, K. 2018. *Co-design as collaborative research*, Bristol University/AHRC Connected Communities Programme.