CARDIFF UNIVERSITY
PRIFYSGOL CAERDYD

ORCA – Online Research @ Cardiff

This is an Open Access document downloaded from ORCA, Cardiff University's institutional repository:https://orca.cardiff.ac.uk/id/eprint/143583/

This is the author’s version of a work that was submitted to / accepted for publication.

Citation for final published version:

Rogers, David , Preece, Alun , Innes, Martin and Spasic, Irena 2022. Real-time text classification of user-generated content on social media: Systematic review. IEEE Transactions on Computational Social Systems 9 (4) , pp. 1154-1166. 10.1109/TCSS.2021.3120138

Publishers page: https://doi.org/10.1109/TCSS.2021.3120138

Please note:
Changes made as a result of publishing processes such as copy-editing, formatting and page numbers may not be reflected in this version. For the definitive version of this publication, please refer to the published source. You are advised to consult the publisher’s version if you wish to cite this paper.

This version is being made available in accordance with publisher policies. See http://orca.cf.ac.uk/policies.html for usage policies. Copyright and moral rights for publications made available in ORCA are retained by the copyright holders.

Real-time Text Classification of User-Generated Content on Social Media: Systematic Review

David Rogers, Alun Preece, Martin Innes, Irena Spasić
Cardiff University
{RogersDM1, PreeceAD, InnesM, SpasicI}@cardiff.ac.uk

***Abstract*—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. Focus is on the identification of the main characteristics of data used for training and testing; the types of text processing and normalisation that is required; the machine learning methods used most commonly; and how these methods compare to one another in terms of classification performance.**

Relevant studies were selected from subscription-based digital libraries, free-to-access bibliographies, and self-curated repositories, then screened for relevance with key information extracted and structured against the following facets: natural language processing methods, data characteristics, classification methods and evaluation results. A total of 25 studies published between 2014 and 2018 covering 15 types of classification algorithms were included in this review.

Support vector machines, Bayesian classifiers, and decision trees were the most commonly employed algorithms with recent emergence of neural network approaches. Domain-specific, API driven collection is the most prevalent origin of data sets. The re-use of previously published data sets as a means of bench-marking algorithms against other studies is also prevalent.

In conclusion, there are consistent approaches taken when normalising social media data for text mining and traditional text mining techniques are suited to the task of real-time analysis of social media.

I. Introduction

Social media provide us with vast amounts of information covering various aspects of modern life. A rich level of discourse is generated online on topics ranging from global social and political issues, to national reaction to major sporting events, to the reporting of localised news and gossip.

There is an ever-increasing abundance of social media content produced daily with an estimated 48.3% of the global population being considered a social media user in 2020, projected to rise to 56.7% by 2025 [1]. This increasing uptake brings about an abundance of social media content, within a single day 500 million tweets [2], 216 million Facebook messages, and 500 million Instagram stories are produced [3].

Accessing, interpreting and using these data effectively has become a major focus of scientific, political and commercial communities. The sheer volume, variety and velocity of data pose significant challenges to digesting social media narratives efficiently. Text mining offers an opportunity to automate this process as data gleaned from social media streams can be valuable both in real-time and post hoc analyses [4], combining methods from natural language processing (NLP), data mining and machine learning to distil information from large corpora of text data [5].

The timeliness of delivering knowledge obtained through the text of mining social media can prove crucial to businesses, government and researchers looking to build situational awareness around an event or topic [6]. Development of text mining systems that operate in (near) real-time, where derived information is presented to users within an actionable time window, is therefore essential. This can generally be in the order of seconds, minutes, hours or days dependant on the domain, scope and user needs.

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. We focus 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. Specifically, this review aims to answer the four research questions given in Table I.

II. Background

Machine learning, a general inductive process that produces an automatic text classifier by learning the characteristics of the categories of interest, has become the dominant approach to text classification over the recent decades [7]. There are a multitude of different approaches to this type of classification that can be grouped into a number of types such as *logic based algorithms*, *perceptron-based techniques*, *statistical learning algorithms*, and *Support Vector Machines (SVMs)*.

TABLE I
RESEARCH QUESTION

ID	Question
RQ1	What are the main characteristics of data used to train and test real-time text classifiers?
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?

The key types of logic based algorithms are *decision trees* and *rule-based* classifiers. Decision trees classify documents by inducing a set of feature-driven rules that can incrementally bisect the training data down into the desired categories based upon features of the data [8]. Rule-based classifiers allow for overlapping rules to be induced in order to describe and classify the data, presenting a less restrictive approach to rule generation [9]. The most useful characteristic of rule-based classifiers is their comprehensibility whereas decision trees are typically more efficient when working with larger data sets [10].

Perceprton-based classification focused upon algorithms that utilise units of mathematical models (perceptrons) likened to biological neurons in order to perform the classification task. *Neural networks* [11] are re-emerging as the state-of-the-art in perceptron-based classification techniques, driven by the emergence of easily accessible deep-learning toolkits such as TensorFlow [12] and PyTorch [13]. A key benefit of neural networks is that they tend to operate on raw or lightly processed features [14], and whilst they can be considered difficult to interpret [10], advancements have been made in recent years focused on improving both the interpretability and explainability of deep learning algorithms [15].

Statistical learning algorithms are driven by probabilistic models, with Bayesian networks and instance-based learning being the most well known representative types [10]. *Naive Bayes* classifiers operate on the assumption that features present in a model are independent of each other given the class, even though in reality features are seldom independent of class [16]. *Multinomial naive Bayes* classifiers are optimised to work with features present within a class that consist of discrete values, such as word counts, providing better performance over large sets of vocabularies [17]. The naive assumption present in these classifiers leads to a computationally efficient yet arguably robust classification algorithm, whose logic can be easily grasped by users [18]. The *k-Nearest Neighbor (kNN)* technique is an instance-based learning classifier, which is more computationally heavy but more stable when compared to other algorithms, that classifies documents based upon their calculated distances to training instances within a multidimensional feature space [10].

SVMs are a more recent form of classification algorithm that aim to identify the optimal hyperplane which produces the largest separation (optimal margin) of documents between two classes using either a linear and nonlinear function [19]. The model complexity of an SVM is unaffected by the number of features encountered in the training data making them suited to deal with learning tasks where the number of features is large with respect to the number of training instances [10]. SVMs are limited by the fact that they are heavily bound by the choice of their kernel [20].

A. Related Work

A number of literature surveys have been performed looking at the text mining of social media in recent years. Salloum et al. [21] investigated text mining techniques used on both Twitter and Facebook data, highlighting issues surrounding the potential ambiguity of words and sentences found within a social media based corpus and criticising the lack of non-English text analysis.

Irfan et al. [22] survey a small selection of studies looking at both classification and clustering techniques across platforms such as Facebook, LinkedIn and MySpace. They discuss methods of feature extraction and selection, and highlight the challenges faced in having to work with texts that are not necessarily structured according to grammatical norms.

Whilst these literature reviews provide important insights into how social media data can be mined, they do not possess the same degree of structure and rigour that can be found in the more methodological systematic reviews, which aim to integrate empirical research in order to create generalisations in the form of meta-analyses [23]. The systematic approach is time consuming with a number of bottlenecks, but it allows for a large volume of articles to be screened, and benefits from having a clearly defined scope and comparable outputs [23].

Systematic reviews have are being performed on text mining of social media, but these have been focused within the Health Science domain. Wongkoblap et al. [24] looks at the application of text mining social media to research mental health, noting that the majority of the studies focused on analysis of depression (46%) and suicide (17%), but the review did not go as far as to produce any meta-analysis of classifier performance.

III. METHODS

This systematic review follows the Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) guidelines [25], which have been adapted from the health science domain.

A. 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 [26, 27], free-to-access bibliographies that provide aggregated searching across third party digital libraries [28, 29], and self-curated repositories [30, 31] where authors are able to link published works to a third party library.

These sources were primarily identified from [32] where we selected all ’Computer Science’ and ’Broad Coverage’ search engines that allowed free access to abstracts and metadata, testing our search across the suitable sources and accepting those that returned results. Individual sources are listed in Table II.

TABLE II
DATA SOURCES

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

A Boolean query was created by combining three major facets of this topic (real-time, classification and social media), whose near-synonyms and hyponyms are given in Figure 1:

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 Twitter ...)

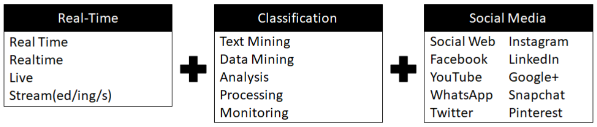

Fig. 1. Synonym Identification for Formal Query

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 [33].

B. Selection Criteria

To further refine the scope of this systematic review, we defined a set of inclusion and exclusion criteria (see Table III and Table IV).

TABLE III
INCLUSION CRITERIA

ID	Question
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 IV
EXCLUSION CRITERIA

ID	Question
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.

To ensure the rigorousness and credibility of selected studies, they were also evaluated against the quality assessment criteria defined in Table V.

TABLE V
QUALITY ASSESSMENT CRITERIA

ID	Question
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 data sets used for training.
QC4	The results were evaluated systematically in terms of accuracy, precision, recall and/or F1 score.

C. Study Selection

The search results were downloaded from the given sources (see Table II), 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.

Document screening was performed using the Rayyan QCRI, a web application and a mobile app for systematic reviews [34]. 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 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.

D. 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 VI) during the screening process.

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

2) RQ2: Text Processing and Normalisation: ***Non-linguistic Analysis*** partially covers the morphological analysis presented in Abbe et al. that covers punctuation and lowercasing [36]. 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 [36]. Popular methods that fall into this category are tokenisation and stemming, the former being the segmentation of word-like units from a text [37] and the latter is the further reduction of these words down to one heading for all variant forms which share a common meaning [38].

Syntactic Analysis is defined by Abbe et al. as methods that are used to determine the structure linking different parts of a sentence [36]. 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. 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 [36]. 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.

Dimensionality Reduction is focused upon reducing the variability present within a corpus through the use of mathematical transformation of texts. Common tech-

[1] https://git.cardiff.ac.uk/papercaddie

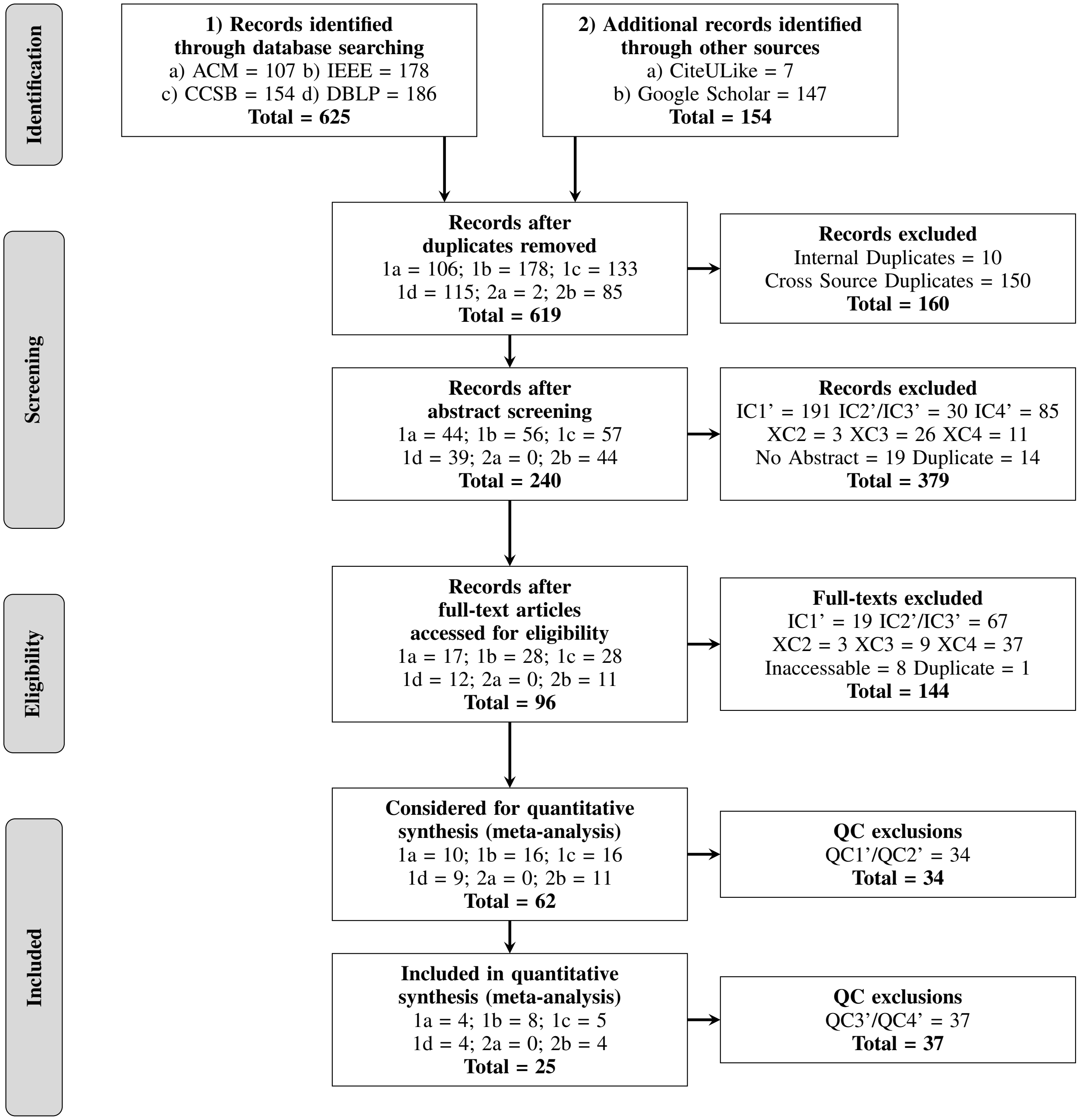

Fig. 2. PRISMA 2009 Flow Diagram

niques found in dimensionality reduction include term frequency – inverse document frequency (TF-IDF) which is used to identify the most discriminative words in a corpus of documents [39], the gain ratio which is used in decision trees to calculate the value of a document feature has for classification [40], and word embeddings which allow for words (and documents) to be represented in a low dimensional vector space [41].

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) [19], naive Bayes classifiers [16], decision trees [8], and neural networks [11].

4) RQ4: Classification Performance: ***Precision, Recall, F-Measure and Accuracy*** are standard measures

TABLE VI
DATA EXTRACTION FIELDS AND EXAMPLES

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
	Dimensionality Reduction	Abbe et al. (2016)	N-Gram Analysis, Bag-of-words, TF-IDF
Machine Learning	**Algorithm**	Supervised classification method	Naieve Bayes, RNN, Descision Tree
Results	**Precision**	$P = \frac{\text{TP}}{\text{TP+FP}}$	Numerical
	Recall	$R = \frac{\text{TP}}{\text{TP+FN}}$	Numerical
	F-Measure	$F_1 = 2 \times \frac{\text{P} \times \text{R}}{\text{P+R}}$	Numerical
	Accuracy	$A = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}}$	Numerical

used to evaluate classification performance [42]. 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. If a particular measure was not presented by the author, the value was left blank unless it could be derived from the other presented measures. A single measure of each was recorded and where these were only presented at a classification category level, overall calculations were made.

IV. RESULTS

Table VII 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 VIII 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.

A. *RQ1: Provenance, Scope and Source*

It is evident that 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 [43], Sina Weibo messages [44], and a bespoke company-based messaging service developed for IBM [45].

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. [44], which worked with exclusively with Sina Weibo data, a Chinese language platform. Neuenschwander et al. also worked with non-English data, in this case Brazilian Portuguese tweets focused on the Brazilian stock market [46]. Italian was used alongside English in two of the studies Avvenuti et al. [47] and D'andrea et al. [48]. Finally, Vincente et al. present multilingual text mining, working with data sets from Basque, English, French and Spanish [49].

TABLE VII
COMBINED STUDY RESULTS

Study	Source				Data				Algorithms		Results			
Title	**Type**	**API**	**Man**	**PrePub**	**Range (d)**	**Vol**	**Classes**	**Dist**	**Tot**	**Best**	**Acc**	**P**	**R**	**F1**
Kurniawan et al. [50]	Tweet	Y			7	35184	2	E	3	SVM	**99.77%**	99.65%	99.89%	99.77%
D'Andrea et al. [48] ∗	Tweet	Y			0.17	2660	2	E	5	SVM	**95.75%**	95.30%	96.50%	95.80%
Nguyen et al. [51]	Tweet	Y			30	5000	2	U	4	BN		94.20%	96.60%	**95.40%**
Benkhelifa and Laallam [43] §	YouTube	Y			122	10K	2	E	1	SVM	**95.30%**	95.35%	95.35%	95.35%
Behzadan et al. [52]	Tweet	Y			4	21K	2	U	1	CNN	**94.72%**		94.57%	94.62%
Alharthi et al. [53]	Tweet			[54]	111	6126	3	B	1	SVM	**93.10%**			92.66%
Subramani et al. [55]	Tweet	Y			108	618	2	U	1	LogReg	**92.50%**			
Win and Aung [56] ‡	Tweet	Y			6	1045	3	Ma	3	Linear	**92.02%**	91.20%	92.00%	91.30%
Steed et al. [57]	Tweet			[58]	80	1600K	2	E	2	NB	**90.00%**			
Michailidis et al. [59]	Tweet		Y		364	17360	2	B	4	SVM	**90.00%**	86.00%	85.00%	85.50%
Middleton and Krivcovs [60]	Tweet	Y			1	1045	2	U	5	DT		*69.00%*	*98.00%*	**88.00%**
Serban et al. [61]	Tweet	Y			256	9353	2	U	4	CNN	**85.40%**			85.20%
Karanasou et al. [62]	Tweet			[63]	30	12529	3	Ma	4	SVM	**85.10%**			
Avvenuti et al. [47] ∗	Tweet	Y				5069	2	U	1	DT	**83.50%**			
Rezaei and Jalali [64] ‡	Tweet	Y				9903	2		2	DT	**82.51%**			
Cavalin et al. [65]	Tweet				15	1910	3	Ma	1	NB	**82.00%**			
Lee et al. [66]	Tweet	Y			736	2000	2	U	4	NBM		81.10%	81.10%	**81.10%**
Yu et al. [67]	Tweet	Y			2	200	2	E	5	SVM	**77.00%**			
Golestani et al. [45] §	IBM		Y		30	130K	2	U	5	NBM		73.00%	74.00%	**73.50%**
Neuenschwander et al. [46] ∗	Tweet				153	922	2	U	3	NBM		73.40%	73.50%	**73.40%**
Mane et al. [68]	Tweet			[69]*	364	1466	3	Mi	1	NB	**72.27%**			
Vicente et al. [49] ∗	Tweet				15	12273	3	Mi	1	SVM	**70.43%**			
Vilares et al. [70]	Tweet			[71]	214	28088	7	Ma	1	Linear		69.85%	72.43%	**69.81%**
Ma et al. [44] ∗§	Sina Weibo	Y			546	160K	2K		4	RNN		67.30%	66.50%	**66.90%**
Azzouza et al. [72]	Tweet			[69]	364	3813	3	Mi	1	Rules	**55.96%**			

∗ - Non-English language corpus ‡ - 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

APIs are the most popular form of data collection; 14 studies using APIs provided by the social media platforms (Twitter, YouTube and Sina Weibo). A total of 8 studies indicated that collections were bounded by search terms [60, 67, 52, 56, 43, 44, 66, 47], 2 utilised a geospatial bounding [51, 61], 2 used a combination of search terms and geospatial bounding [48, 55] and 1 collectied using a combination of terms and specific user accounts [50].

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. [70] utilising the RepLab 2014 corpus [71] and Azzouza et al. and Karanasou et al. [72, 62] using corpora from the SemEval series [69, 63]. Interestingly, Mane et al. [68] used data from SemEval 2013.T2 corpus [69], 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. who also used the SemEval 2013.T2 corpus. Both Steed et al. and Alharthi et al. [57, 53] source their data sets from previously published articles, with the former sourcing a large 1.6 million datapoint corpus produced by Go et al. [58], and the latter citing previous work of their own where they detail the collection methodology [54]. 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.

Michalidis et al. [59] resorted to manual collection and curation of their Twitter-based data set through a third-party service FigureEight (then called Crowd-Flower), opting to bound their collection criteria through search terms. Golestani et al. [45] 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 [65, 49, 46]. The number of days covered by a data set was identified in all but two of the studies [47, 64]. 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 [57] presenting a training set size that can be considered an outlier falling outside the upper quartile by over three standard deviations. Steed et al. use 1.6 million documents within their training set which is an existing auto-generated training set produced by Go et al [58].

B. RQ2: Data Processing and Normalisation

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 [50, 53, 67, 55], replacing the username with a generic representative token [65], or stripping off the @ prefix and retaining the username, or leaving the username intact [70]. 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. [66] choosing to replace HTML instances with an HTML token and Yu et al. [67] 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 [61, 49] with one of these taking it further by employing a prefix-based space prediction algorithm [73] to break the hashtag into the minimum possible number of words when camel case splitting fails [61]. Two studies removed hashtags completely [59, 55], and the last one just removes the symbol [70]. A lot of studies chose to remove punctuation [57, 48, 52, 43, 55], non-ASCII characters [52], Unicode characters [67], non-alphanumeric content [65, 57, 48, 50, 67, 51, 52, 55], numbers [48, 43, 59, 55]; 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 [68, 65, 57, 62, 72, 49].

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 [57, 48, 64, 56, 52, 55]. 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 [57, 66].

TABLE VIII
ARTICLES INCLUDED FOR QUANTITATIVE ANALYSIS

Citation	Non-linguistic	Morphological	Syntax	Semantic	Reduction	Techniques
Vilares et al. 2014 [70]	- lowercasing - # and @ rem - url tokenising		- dependency tree - lemmatisation - PoS	- emotion dict idx	- boolean vectors - ngram (uni, bi, lemmas)	- **Linear**
Mane et al. 2014 [68]	- word norm	- stemming - stop word rem	- PoS	- emoji dict norm - sentiment dict idx - word expansion		- **NB**
Cavalin et al. 2014 [65]	- nonalphanum. rem - punctuation norm - url tokenising - user tokenising - word norm	- tokenisation - stop word rem	- proper noun tokenisation	- knowledge dict filter - knowledge dict norm		- **NB**
Neuenschwander et al. 2014 ∗ [46]		- stop word rem	- lemmatisation - PoS			- SOCAL (Rules) - NB - **NBM**
Steed et al. 2015 †¶ [57]	- lowercasing - nonalphanum. rem - punctuation rem - url tokenising - word norm	- stemming - stop word rem (modal verbs kept) - tokenisation		- word expansion	- word vectors	- **NB** - Max Entropy
Lee et al. 2015 †¶ [66]	- html tokenising - url tokenising	- stop word rem (pronouns kept)			- ngram (uni, bi, tri) - tf-idf	- NB - **NBM** - RF - SVM
Avvenuti et al. 2015 ∗ [47]	- punctuation count - retweet flag - uppercase count - url flag - user flag - word count			- vocabulary idx	- funct. (Pearson's) - information gain	- **DT**
D'Andrea et al. 2015 ∗ [48]	- nonalphanum. rem - number rem - punctuation rem	- stemming -stop word rem - tokenisation			- information gain - stem filter - tf-idf	- **SVM** - NB - DT - kNN - PART (Rules)
Middleton and Krivcovs 2016 † [60]		- tokenisation - stemming	- PoS	- NER	- ngram - tf-idf	- **DT** - kNN - NB - RF - LogitBoost
Kurniawan et al. 2016 [50]	- lowercasing - nonalphanum. rem - RT syntax rem - url rem - user rem			- word expansion	- word vectors	- **SVM** - DT - NB
Karanasou et al. 2016 [62]	- word norm	- stop word rem	- grammatical idx - negation idx - PoS	- emoji dict idx - sentiment dict idx - vocabulary norm		- **SVM** - NB - DT - Linear (SGD)
Yu et al. 2016 [67]	- html rem - nonalphanum. rem - unicode rem - url rem - user rem	- tokenisation	- lemmatisation	- knowledge dict norm	- ngram (uni, bi)	- **SVM** - NB - Nonlinear SVM - Max Entropy - kNN - DT
Nguyen et al. 2016 † [51]	- date, loc, time idx - nonalphanum. rem	- stop word rem	- lemmatisation - PoS	- knowledge dict idx		- kNN - **BN** - SVM - DT
Azzouza et al. 2017 † [72]	- word norm		- PoS	- emoji dict idx - vocabulary norm - word expansion	- tf-idf	- **Rules**
Rezaei and Jalali 2017 ‡ [64]	- lowercasing	- stemming - stop word rem - tokenisation			- funct. (Gini) - tf-idf - token filter	- **DT (McD)** - DT (H)
Win and Aung 2017 ‡ [56]	- hashtag count - url count	- stemming - stop word rem - tokenisation	- PoS	- knowledge dict idx - emotion dict idx	- ngram (uni, bi) - word vectors - information gain - word embedding	- **Linear** - SMO-SVM - RF
Behzadan et al. 2018 [52]	- lowercasing - nonalphanum. rem - non-ascii rem - punctuation rem	- stemming - stop word rem - tokenisation		- vocabulary norm	- word embedding	- **CNN**

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

TABLE VIII
ARTICLES INCLUDED FOR QUANTITATIVE ANALYSIS

Citation	Non-linguistic	Morphological	Syntax	Semantic	Reduction	Techniques
Benkhelifa and Laallam 2018 § [43]	- number rem - punctuation rem	- stemming	- interjection idx - PoS	- emotion dict idx	- tf-idf	- **SVM**
Michailidis et al. 2018 † [59]	- hashtag rem - number rem - url rem	- stemming - stop word rem				- **SVM** - NB - DT - Max Entropy
Golestani et al. 2018 § [45]	- lowercasing	- stop word rem - tokenisation			- tf-idf	- **NBM** - SVM - RF - DT - Ad. Boosting
Şerban et al. 2018 [61]	- hashtag expansion - lowercasing - url rem			- emoji dict norm	- tf-idf (NB, SVM) - word embedding (RNN, CNN)	- NB - SVM - RNN - **CNN**
Alharthi et al. 2018 †¶ [53]	- hashtag idx - Meida rem - url rem - user rem	- stemming - tokenisation		- emoji dict idx - emotion dict idx - sentiment dict idx	- ngram (uni bi, tri) - information gain - tf-idf	- **LSE SVM**
Vicente et al. 2018 ∗ [49]	- hashtag expansion - url tokenising - word norm		- grammatical idx - interjection idx - lemmatisation - PoS	- emoji dict norm - vocabulary norm	- ngram (uni)	- **SVM**
Ma et al. 2018 ∗§ [44]		- tokenisation			- attention layers (sentence, word) - funct. (softmax)	- NB - CNN - **tSAM-RNN** - SAM-RNN
Subramani et al. 2018 [55]	- hashtag rem - nonalphanum. rem - number rem - punctuation rem - user rem	- stemming - stop word rem - tokenisation	- PoS	- emotion dict idx - sentiment dict idx	- tf-idf	- **Log Reg**

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

Syntactic Analysis: This category is dominated by traditional text mining techniques, predominantly part of speech (POS) [70, 68, 46, 60, 62, 51, 72, 56, 43, 49, 55] and lemmatisation [70, 46, 67, 51, 49]. A small number of studies show interest in emphasising particular grammatical elements such as interjections [43, 49], negation and key phrases [62] and onomatopoeic tokens [49].

Semantic Analysis: Identification of emotions was used in over a third of studies, with emotion dictionaries used to index emotions in five studies [70, 56, 43, 53, 55], emoji lookup tables used in three studies [62, 72, 53], and the translation of emojis into a text representation in three others [68, 61, 49]. Interestingly the use of sentiment dictionaries [68, 62, 53, 55] is lower than that of emotion dictionaries.

A number of well-established lexicons were cited, with the SentiWordNet [74] being employed by [62, 55], and the LWIC lexicon [75] used in [70, 53] and WordNet Affect [76] by [55]. Other knowledge-specific lexicons were used in five studies to either filter out tokens [65], index key concepts [51, 56], or normalise text through disambiguation [65, 67]. Normalisation of vocabulary was also performed in a number of studies whereby either the vocabulary used was reduced [62, 72, 52, 49] or abbreviations and acronyms were expanded [68, 57, 50, 72]. Interestingly NER is only presented in one study [60].

Dimensionality Reduction: The predominant approach used for dimensionality reduction was TF-IDF [66, 48, 60, 72, 64, 43, 45, 61, 53, 55]. N-grams were used in seven studies [70, 66, 60, 67, 56, 53, 49], with several using bi-grams [70, 66, 67, 56, 53], two studies using tri-grams [66, 53] and one applying the n-gram principle to both words and their lemmas [70]. Other approaches used information gain [47, 48, 56, 53], word embeddings [56, 52, 61], Pearson's correlation coefficient [47], the Gini index [64] and the softmax function [44].

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 0.5. This is of interest as it suggests that the normalisation of social media texts has a positive effect on the classification performance.

C. 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 [50, 48, 51, 43, 52, 55, 57, 59, 60, 61, 47, 64] and ternary [53, 56, 62, 65, 66, 67, 45, 46, 68, 49, 72] classification. Only two studies used more than three classes; Vilares et al. [70] classifies into 7 classes and Ma et al. [44] 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.

Looking at the binary classifiers, a total of six were trained with evenly balanced classes [50, 48, 43, 57, 59, 67]. Four of these studies achieved even classes through under sampling of the majority class [50, 48, 43, 67], one [57] used a third party data set where a hard limit was placed on the amount of data collected for both classes [58] to produce an even data set, whilst the sixth [59] 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 [77]. SMOTE was also used in the only ternary classifier [53].

Studies employing an unbalanced data set for ternary classification performed better when there was a distinct majority class [56, 62, 65] relative to studies that had a distinct minority class [68, 49, 72]. 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. which may be suffering from smallest training data set of only 200 [67].

TABLE IX
PREVALENCE OF MACHINE LEARNING TECHNIQUES

ML Technique	Best	Papers	Versus		
			None	Self	Others
SVM	8	13	3		16
Naive 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
Baysean 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**

A wide range of machine learning methods were used to support text classification (See Table IX). 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 classification algorithms such as scikit-learn [78], Weka [79], LibLinear [80] and word2vec [41].

Three methods were used predominantly: SVMs [19], naive Bayesian learning [16] and decision trees [8]. SVMs performed best in 8 out of 13 studies, [50, 48, 43, 53, 59, 62, 67, 49]. Naive Bayes learning performed best in 3 out of 13 studies [57, 65, 68]. Decision trees performed best in 3 out of 10 studies [60, 47, 64]. SVMs were frequently compared to naive Bayes algorithms [48, 61], Decision tree algorithms [45], or both [50, 62, 67, 59], and frequently outperformed both of these algorithm types. Naive Bayes Multinomial algorithms consistently outperformed all other methods that they were compared to [66, 45, 46], including SVMs in two of the three [66, 45].

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 [60, 64]. The use of correlation coefficients for dimensionality reduction was observed in both Avvenuti et al. [47] and in Rezaei and Jalali [64], 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) [81, 11] and recurrent neural networks (RNNs) [82] due to the lowered barriers of entry provided by cloud computing platforms. CNNs demonstrated high classification accuracy in Behzadan et al. and Serban et al. [52, 61]. Ma et al. [44] also employed a number of variations of Long Short-Term Memory (LSTM) based RNNs [83] to support multi-class classification; something not commonly seen in this review.

D. 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 [57] 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. [52] and Ma et al. [44], with data set sizes of 21,000 and 160,000 respectively, both of which sit above the upper quartile value for data set size. Serban et al. [61] has a much smaller data set size relative to the other studies investigating deep learning. This study is outperformed by Behzadan et al. [52], but not by Ma et al. [44] 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 Serban et al [61].

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 [48, 51, 60, 64, 61, 49] or at bench-marking the performance of supporting architecture [62]. Only two studies highlighted a preference for speed over performance when selecting their preferred algorithms [56, 64]. Rezaei and Jalali 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% [64]. Win and Aung favoured a LibLinear [80] based classifier in their study over a SMO trained SVM [84] despite an average accuracy cost of 0.43% across several training sets, citing a faster processing time as the reason for this preference [56].

V. Discussion

As mentioned in the results, Twitter formed the largest source of data utilised by the studies, most likely due to it's highly accessible APIs. We expect a future direction of this field will see this homogeneity of data type be reduced, with more platform agnostic and language agnostic solutions being presented. This is important as the social media environment is volatile and prone to change, even monolithic platforms such as Facebook is not invulnerable to user apathy [85]. In addition to this, platforms themselves are often, in efforts to guard against data breaches, disinformation, and malicious actors, rolling back API access effecting the ability for users to extract data for analysis [86, 87].

English language only corpora were found in 80% of the studies, with 6 other languages covered in the remaining studies [48, 47, 46, 49, 44]. When compared to the distribution of language used on the web [88], we can see that English is heavily over represented in our survey set. In addition to the languages encountered in this review, studies using common languages such as Hindi and Russian [89] do exist. Furthermore, multilingual text classification is becoming more attainable through new feature learning techniques such as *word embeddings* [90, 91], and these techniques are already now being applied to the social media space [92].

As the applications of social media text classification broaden in scope, the challenge of obtaining reputable, repeatable and comparable training corpora for more niche domains increases. It was noticeable that the majority of studies sought to create their own training data, with only 6 studies using publicly available precollected data sets, and only 4 of these originated from gold-standard data sets produced as part of NLP community challenges like RepLab and SemEval, focused on reputational classification [71] and sentiment [69, 63].

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 [60], achieved a recall of 99% against a precision of 68%.

Systematic reviews are more commonly used in Medical Science and are most powerful when assessing the outputs of clinical studies [36]. These studies follow a much more rigorous and standardised means of hypothesising, candidate selection and variable control. Social Media text analysis is 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, adding to the challenge of performing this systematic, but efforts already exist that look to improve the systematic review process within the software development and computer science domain [93].

VI. 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 algorithm 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.

The integration of classification tools into large analytic platforms is increasing in importance, as the adoption of machine learning outcomes becomes more common outside of the Computer Science domain, forming a part of a much wider field of interdisciplinary research, that blends both qualitative and quantitative analysis [94, 95]. Future development in text mining of social media will be reliant on the implementation of both multilingual and platform agnostic solutions, supported by a greater diversity of gold-standard corpora.

References

[1] Statista, "Social network penetration worldwide from 2017 to 2025," 2020. [Online]. Available: https://www.statista.com/statistics/260811/social-network-penetration-worldwide/, Accessed: 2020-10-03.

[2] S. Aslam, "Twitter by the Numbers: Stats, Demographics & Fun Facts," 2020. [Online]. Avail-

able: https://www.omnicoreagency.com/twitter-statistics/, Accessed: 2020-10-03.

[3] Domo, "Media usage in an internet minute as of August 2020," 2020. [Online]. Available: https://www.statista.com/statistics/195140/new-user-generated-content-uploaded-by-users-per-minute/, Accessed: 2020-10-03.

[4] A. Preece, I. Spasić, K. Evans, D. Rogers, W. Webberley, C. Roberts, and M. Innes, "Sentinel: A codesigned platform for semantic enrichment of social media streams," *IEEE Transactions on Computational Social Systems*, vol. 5, no. 1, pp. 118–131, 2017.

[5] A. Hotho, A. Nürnberger, and G. Paaß, "A brief survey of text mining.," in *LDV Forum - GLDV Journal for Computational Linguistics and Language Technology*, 2005.

[6] H. Gao, G. Barbier, and R. Goolsby, "Harnessing the crowdsourcing power of social media for disaster relief," *IEEE Intelligent Systems*, vol. 26, no. 3, pp. 10–14, 2011.

[7] F. Sebastiani, "Machine learning in automated text categorization," *ACM computing surveys (CSUR)*, vol. 34, no. 1, pp. 1–47, 2002.

[8] J. R. Quinlan, "Induction of decision trees," *Machine learning*, vol. 1, no. 1, pp. 81–106, 1986.

[9] J. Fürnkranz, "Separate-and-conquer rule learning," *Artificial Intelligence Review*, vol. 13, no. 1, pp. 3–54, 1999.

[10] S. B. Kotsiantis, "Supervised machine learning: A review of classification techniques," *Emerging artificial intelligence applications in computer engineering*, vol. 160, no. 1, pp. 3–24, 2007.

[11] Y. Kim, "Convolutional neural networks for sentence classification," *arXiv preprint arXiv:1408.5882*, 2014.

[12] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, *et al.*, "Tensorflow: A system for large-scale machine learning," in *12th USENIX symposium on operating systems design and implementation (OSDI 16)*, pp. 265–283, 2016.

[13] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in neural information processing systems*, pp. 8026–8037, 2019.

[14] Z. C. Lipton, "The mythos of model interpretability," *Queue*, vol. 16, no. 3, pp. 31–57, 2018.

[15] G. Montavon, W. Samek, and K.-R. Müller, "Methods for interpreting and understanding deep neural networks," *Digital Signal Processing*, vol. 73, pp. 1–15, 2018.

[16] P. Domingos and M. Pazzani, "On the optimality of the simple bayesian classifier under zero-one loss," *Machine learning*, vol. 29, no. 2-3, pp. 103–130, 1997.

[17] A. McCallum, K. Nigam, *et al.*, "A comparison of event models for naive bayes text classification," *AAAI-98 workshop on learning for text categorization*, vol. 752, no. 1, 1998.

[18] I. Kononenko, "Inductive and bayesian learning in medical diagnosis," *Applied Artificial Intelligence an International Journal*, vol. 7, no. 4, pp. 317–337, 1993.

[19] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[20] C. J. Burges, "A tutorial on support vector machines for pattern recognition," *Data mining and knowledge discovery*, vol. 2, no. 2, pp. 121–167, 1998.

[21] S. A. Salloum and M. Al-Emran, "A survey of text mining in social media: facebook and twitter perspectives," *Advances in Science, Technology and Engineering Systems*, vol. 2, no. 1, pp. 127–133, 2017.

[22] R. Irfan, C. K. King, D. Grages, S. Ewen, S. U. Khan, S. A. Madani, J. Kolodziej, L. Wang, D. Chen, A. Rayes, *et al.*, "A survey on text mining in social networks," *The Knowledge Engineering Review*, vol. 30, no. 2, pp. 157–170, 2015.

[23] J. Biolchini, P. G. Mian, A. C. C. Natali, and G. H. Travassos, "Systematic review in software engineering," tech. rep., System Engineering and Computer Science Department COPPE/UFRJ, 2005. Technical Report ES 679.05.

[24] A. Wongkoblap, M. A. Vadillo, and V. Curcin, "Researching mental health disorders in the era of social media: systematic review," *Journal of medical Internet research*, vol. 19, no. 6, p. e228, 2017.

[25] D. Moher, A. Liberati, J. Tetzlaff, and D. G. Altman, "Preferred reporting items for systematic reviews and meta-analyses: the prisma statement," *BMJ*, vol. 339, 2009.

[26] Association for Computing Machinery, "ACM Digital Library," 2020. [Online]. Available: http://dl.acm.org/, Accessed: 2020-02-05.

[27] Institute of Electrical and Electronics Engineers, "IEEE Xplore," 2020. [Online]. Available: http://ieeexplore.ieee.org, Accessed: 2020-02-05.

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

[29] Database Systems and Logic Programming (DBLP), University of Trier, "DBLP Computer

Science Bibliography," 2020. [Online]. Available: http://dblp.org/, Accessed: 2020-02-05.

[30] CiteULike, "CiteULike," 2019. [Online]. Available: http://www.citeulike.org/, Accessed: 2019-01-25.

[31] Google, "Google Scholar," 2020. [Online]. Available: https://scholar.google.co.uk/, Accessed: 2020-02-05.

[32] D. Hull, S. R. Pettifer, and D. B. Kell, "Defrosting the digital library: bibliographic tools for the next generation web," *PLoS Comput Biol*, vol. 4, no. 10, p. e1000204, 2008.

[33] Wayback Machine, "CiteULike is closing down," 2020. [Online]. Available: https://web.archive.org/web/20190310145602/citeulike.org/news, Accessed: 2020-02-05.

[34] M. Ouzzani, H. Hammady, Z. Fedorowicz, and A. Elmagarmid, "Rayyan— a web and mobile app for systematic reviews," *Systematic Reviews*, vol. 5, p. 210, Dec 2016.

[35] P. Nakov, A. Ritter, S. Rosenthal, F. Sebastiani, and V. Stoyanov, "SemEval-2016 task 4: Sentiment analysis in twitter," in *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, (San Diego, California), pp. 1–18, Association for Computational Linguistics, June 2016.

[36] A. Abbe, C. Grouin, P. Zweigenbaum, and B. Falissard, "Text mining applications in psychiatry: a systematic literature review," *International journal of methods in psychiatric research*, vol. 25, no. 2, pp. 86–100, 2016.

[37] G. Grefenstette and P. Tapanainen, "What is a word, what is a sentence?: problems of tokenisation," 1994.

[38] M. F. Porter, "Snowball: A language for stemming algorithms," 2001.

[39] J. Ramos *et al.*, "Using tf-idf to determine word relevance in document queries," in *Proceedings of the first instructional conference on machine learning*, vol. 242, pp. 133–142, Piscataway, NJ, 2003.

[40] A. G. Karegowda, A. Manjunath, and M. Jayaram, "Comparative study of attribute selection using gain ratio and correlation based feature selection," *International Journal of Information Technology and Knowledge Management*, vol. 2, no. 2, pp. 271–277, 2010.

[41] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[42] C. Goutte and E. Gaussier, "A probabilistic interpretation of precision, recall and f-score, with implication for evaluation," in *European Conference on Information Retrieval*, pp. 345–359, Springer, 2005.

[43] R. Benkhelifa and F. Z. Laallam, "Opinion extraction and classification of real-time youtube cooking recipes comments," in *AMLTA* (A. E. Hassanien, M. F. Tolba, M. Elhoseny, and M. Mostafa, eds.), vol. 723 of *Advances in Intelligent Systems and Computing*, pp. 395–404, Springer, 2018.

[44] J. Ma, C. Feng, G. Shi, X. Shi, and H. Huang, "Temporal enhanced sentence-level attention model for hashtag recommendation," *CAAI Transactions on Intelligence Technology*, vol. 3, no. 2, pp. 95–100, 2018.

[45] A. Golestani, M. Masli, N. S. Shami, J. Jones, A. Menon, and J. Mondal, "Real-time prediction of employee engagement using social media and text mining," in *ICMLA*, pp. 1383–1387, IEEE, 2018.

[46] B. Neuenschwander, A. C. Pereira, W. Meira, Jr., and D. Barbosa, "Sentiment analysis for streams of web data: A case study of brazilian financial markets," in *Proceedings of the 20th Brazilian Symposium on Multimedia and the Web*, WebMedia '14, (New York, NY, USA), pp. 167–170, ACM, 2014.

[47] M. Avvenuti, F. D. Vigna, S. Cresci, A. Marchetti, and M. Tesconi, "Pulling information from social media in the aftermath of unpredictable disasters," in *2015 2nd International Conference on Information and Communication Technologies for Disaster Management (ICT-DM)*, pp. 258–264, 2015.

[48] E. D'Andrea, P. Ducange, B. Lazzerini, and F. Marcelloni, "Real-time detection of traffic from twitter stream analysis," *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 4, pp. 2269–2283, 2015.

[49] I. S. Vicente, X. Saralegi, and R. Agerri, "Talaia: a real time monitor of social media and digital press," *CoRR*, vol. abs/1810.00647, 2018.

[50] D. A. Kurniawan, S. Wibirama, and N. A. Setiawan, "Real-time traffic classification with twitter data mining," in *Information Technology and Electrical Engineering (ICITEE), 2016 8th International Conference on*, pp. 1–5, IEEE, 2016.

[51] H. Nguyen, W. Liu, P. Rivera, and F. Chen, "Trafficwatch: Real-time traffic incident detection and monitoring using social media," in *Advances in Knowledge Discovery and Data Mining - 20th Pacific-Asia Conference, PAKDD 2016, Auckland, New Zealand, April 19-22, 2016, Proceedings, Part I*, pp. 540–551, 2016.

[52] V. Behzadan, C. Aguirre, A. Bose, and W. Hsu, "Corpus and deep learning classifier for collection of cyber threat indicators in twitter stream," in *BigData*, pp. 5002–5007, IEEE, 2018.

[53] R. Alharthi, B. Guthier, and A. E. Saddik, "Rec-

ognizing human needs during critical events using machine learning powered psychology-based framework," *IEEE Access*, vol. 6, pp. 58737–58753, 2018.

[54] R. Alharthi, B. Guthier, C. Guertin, and A. El Saddik, "A dataset for psychological human needs detection from social networks," *IEEE Access*, vol. 5, pp. 9109–9117, 2017.

[55] S. Subramani, S. Michalska, H. Wang, F. Whittaker, and B. Heyward, "Text mining and real-time analytics of twitter data: A case study of australian hay fever prediction," in *HIS* (S. Siuly, I. Lee, Z. Huang, R. Z. 0001, H. W. 0002, and W. Xiang, eds.), vol. 11148 of *Lecture Notes in Computer Science*, pp. 134–145, Springer, 2018.

[56] S. S. M. Win and T. N. Aung, "Target oriented tweets monitoring system during natural disasters," in *2017 IEEE/ACIS 16th International Conference on Computer and Information Science (ICIS)*, pp. 143–148, 2017.

[57] C. A. Steed, M. Drouhard, J. Beaver, J. Pyle, and P. L. Bogen, "Matisse: A visual analytics system for exploring emotion trends in social media text streams," in *2015 IEEE International Conference on Big Data (Big Data)*, pp. 807–814, 2015.

[58] A. Go, R. Bhayani, and L. Huang, "Twitter sentiment classification using distant supervision," *CS224N project report, Stanford*, vol. 1, no. 12, p. 2009, 2009.

[59] D. Michailidis, N. Stylianou, and I. Vlahavas, "Real time location based sentiment analysis on twitter: The airsent system," in *Proceedings of the 10th Hellenic Conference on Artificial Intelligence*, SETN '18, (New York, NY, USA), pp. 21:1–21:4, ACM, 2018.

[60] S. E. Middleton and V. Krivcovs, "Geoparsing and geosemantics for social media: Spatiotemporal grounding of content propagating rumors to support trust and veracity analysis during breaking news," *ACM Transactions on Information Systems*, vol. 34, pp. 16:1–16:??, may 2016.

[61] O. Șerban, N. Thapen, B. Maginnis, C. Hankin, and V. Foot, "Real-time processing of social media with sentinel: A syndromic surveillance system incorporating deep learning for health classification," *Information Processing & Management*, 2018.

[62] M. Karanasou, A. Ampla, C. Doulkeridis, and M. Halkidi, "Scalable and real-time sentiment analysis of twitter data," in *ICDM Workshops* (C. Domeniconi, F. Gullo, F. Bonchi, J. Domingo-Ferrer, R. A. Baeza-Yates, Z.-H. Zhou, and X. Wu, eds.), pp. 944–951, IEEE, 2016.

[63] P. Nakov, T. Zesch, D. Cer, and D. Jurgens, "Proceedings of the 9th international workshop on semantic evaluation (semeval 2015)," in *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, 2015.

[64] Z. Rezaei and M. Jalali, "Sentiment analysis on twitter using mcdiarmid tree algorithm," in *2017 7th International Conference on Computer and Knowledge Engineering (ICCKE)*, pp. 33–36, 2017.

[65] P. R. Cavalin, M. d. C. Gatti, C. N. dos Santos, and C. Pinhanez, "Real-time sentiment analysis in social media streams: The 2013 confederation cup case," *Proceedings of BRACIS/ENIAC*, 2014.

[66] K. Lee, A. Agrawal, and A. Choudhary, "Mining social media streams to improve public health allergy surveillance," in *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*, ASONAM '15, (New York, NY, USA), pp. 815–822, ACM, 2015.

[67] F. Yu, M. Moh, and T. S. Moh, "Towards extracting drug-effect relation from twitter: A supervised learning approach," in *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)*, pp. 339–344, 2016.

[68] S. B. Mane, Y. Sawant, S. Kazi, and V. Shinde, "Real time sentiment analysis of twitter data using hadoop," *IJCSIT) International Journal of Computer Science and Information Technologies*, vol. 5, no. 3, pp. 3098–3100, 2014.

[69] P. Nakov, S. Rosenthal, Z. Kozareva, V. Stoyanov, A. Ritter, and T. Wilson, "SemEval-2013 task 2: Sentiment analysis in Twitter," in *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, (Atlanta, Georgia, USA), pp. 312–320, Association for Computational Linguistics, June 2013.

[70] D. Vilares, M. Hermo, M. A. Alonso, C. Gómez-Rodríguez, and J. Vilares, "Lys at clef replab 2014: Creating the state of the art in author influence ranking and reputation classification on twitter," in *CLEF (Working Notes)* (L. Cappellato, N. Ferro, M. Halvey, and W. Kraaij, eds.), vol. 1180 of *CEUR Workshop Proceedings*, pp. 1468–1478, CEUR-WS.org, 2014.

[71] E. Amigó, J. Carrillo-de Albornoz, I. Chugur, A. Corujo, J. Gonzalo, E. Meij, M. de Rijke, and D. Spina, "Overview of replab 2014: author profiling and reputation dimensions for online reputation management," in *International Conference of the Cross-Language Evaluation Forum for European

Languages, pp. 307–322, Springer, 2014.

[72] N. Azzouza, K. Akli-Astouati, A. Oussalah, and S. A. Bachir, "A real-time twitter sentiment analysis using an unsupervised method," in *Proceedings of the 7th International Conference on Web Intelligence, Mining and Semantics*, WIMS '17, (New York, NY, USA), pp. 15:1–15:10, ACM, 2017.

[73] A. V. Aho and M. J. Corasick, "Efficient string matching: an aid to bibliographic search," *Communications of the ACM*, vol. 18, no. 6, pp. 333–340, 1975.

[74] A. Esuli and F. Sebastiani, "Sentiwordnet: a high-coverage lexical resource for opinion mining," *Evaluation*, vol. 17, no. 1, p. 26, 2007.

[75] J. W. Pennebaker, M. E. Francis, and R. J. Booth, "Linguistic inquiry and word count: Liwc 2001," *Mahway: Lawrence Erlbaum Associates*, vol. 71, no. 2001, p. 2001, 2001.

[76] C. Strapparava, A. Valitutti, *et al.*, "Wordnet affect: an affective extension of wordnet.," in *Lrec*, p. 40, Citeseer, 2004.

[77] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[78] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[79] F. Eibe, M. A. Hall, I. H. Witten, and J. Pal, "The weka workbench," *Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques*, vol. 4, 2016.

[80] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "LIBLINEAR: A library for large linear classification," *Journal of Machine Learning Research*, vol. 9, pp. 1871–1874, 2008.

[81] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[82] R. J. Williams and D. Zipser, "A learning algorithm for continually running fully recurrent neural networks," *Neural computation*, vol. 1, no. 2, pp. 270–280, 1989.

[83] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[84] J. Platt, "Sequential minimal optimization: A fast algorithm for training support vector machines," Tech. Rep. MSR-TR-98-14, Microsoft, April 1998.

[85] N. Newman, R. Fletcher, A. Kalogeropoulos, and R. Nielsen, *Reuters Institute digital news report 2019*, vol. 2019. Reuters Institute for the Study of Journalism, 2019.

[86] A. Bruns, "After the 'apicalypse': social media platforms and their fight against critical scholarly research," *Information, Communication & Society*, vol. 22, no. 11, pp. 1544–1566, 2019.

[87] A. Acker and A. Kreisberg, "Social media data archives in an api-driven world," *Archival Science*, vol. 20, no. 2, pp. 105–123, 2020.

[88] Internet World Stats, "Most common languages used on the internet as of January 2020, by share of internet users.," 2020. [Online]. Available: https://www.statista.com/statistics/262946/share-of-the-most-common-languages-on-the-internet/, Accessed: 2020-10-05.

[89] N. Medagoda, S. Shanmuganathan, and J. Whalley, "A comparative analysis of opinion mining and sentiment classification in non-english languages," in *2013 International Conference on Advances in ICT for Emerging Regions (ICTer)*, pp. 144–148, IEEE, 2013.

[90] A. Mogadala and A. Rettinger, "Bilingual word embeddings from parallel and non-parallel corpora for cross-language text classification," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 692–702, 2016.

[91] W. Ammar, G. Mulcaire, Y. Tsvetkov, G. Lample, C. Dyer, and N. A. Smith, "Massively multilingual word embeddings," *arXiv preprint arXiv:1602.01925*, 2016.

[92] V. Lorini, C. Castillo, F. Dottori, M. Kalas, D. Nappo, and P. Salamon, "Integrating social media into a pan-european flood awareness system: A multilingual approach," *arXiv preprint arXiv:1904.10876*, 2019.

[93] B. Kitchenham and P. Brereton, "A systematic review of systematic review process research in software engineering," *Information and software technology*, vol. 55, no. 12, pp. 2049–2075, 2013.

[94] C. H. Yu, A. Jannasch-Pennell, and S. DiGangi, "Compatibility between text mining and qualitative research in the perspectives of grounded theory, content analysis, and reliability.," *Qualitative Report*, vol. 16, no. 3, pp. 730–744, 2011.

[95] C. Periñán-Pascual, "Bridging the gap within text-data analytics: a computer environment for data analysis in linguistic research," *LFE. Revista de Lenguas para Fines Específicos*, 2017.