big data and cognitive computing

MDPI

Article

Topic Modelling: Going beyond Token Outputs

Lowri Williams [1,*], Eirini Anthi [1], Laura Arman [2] and Pete Burnap [1]

[1] School of Computer Science & Informatics, Cardiff University, Cardiff CF24 4AG, UK; anthies@cardiff.ac.uk (E.A.)
[2] School of Social Sciences, Cardiff University, Cardiff CF10 3NN, UK; armanl@cardiff.ac.uk
* Correspondence: williamsl10@cardiff.ac.uk

Abstract: Topic modelling is a text mining technique for identifying salient themes from a number of documents. The output is commonly a set of topics consisting of isolated tokens that often co-occur in such documents. Manual effort is often associated with interpreting a topic's description from such tokens. However, from a human's perspective, such outputs may not adequately provide enough information to infer the meaning of the topics; thus, their interpretability is often inaccurately understood. Although several studies have attempted to automatically extend topic descriptions as a means of enhancing the interpretation of topic models, they rely on external language sources that may become unavailable, must be kept up to date to generate relevant results, and present privacy issues when training on or processing data. This paper presents a novel approach towards extending the output of traditional topic modelling methods beyond a list of isolated tokens. This approach removes the dependence on external sources by using the textual data themselves by extracting high-scoring keywords and mapping them to the topic model's token outputs. To compare how the proposed method benchmarks against the state of the art, a comparative analysis against results produced by Large Language Models (LLMs) is presented. Such results report that the proposed method resonates with the thematic coverage found in LLMs and often surpasses such models by bridging the gap between broad thematic elements and granular details. In addition, to demonstrate and reinforce the generalisation of the proposed method, the approach was further evaluated using two other topic modelling methods as the underlying models and when using a heterogeneous unseen dataset. To measure the interpretability of the proposed outputs against those of the traditional topic modelling approach, independent annotators manually scored each output based on their quality and usefulness as well as the efficiency of the annotation task. The proposed approach demonstrated higher quality and usefulness, as well as higher efficiency in the annotation task, in comparison to the outputs of a traditional topic modelling method, demonstrating an increase in their interpretability.

Keywords: topic modelling; keyword extraction; natural language processing; text mining; Latent Dirichlet Allocation

Citation: Williams, L.; Anthi, E.; Arman, L.; Burnap, P. Topic Modelling: Going beyond Token Outputs. *Big Data Cogn. Comput.* **2024**, *8*, 44. https://doi.org/10.3390/bdcc8050044

Academic Editors: Zuchao Li and Min Peng

Received: 22 March 2024
Revised: 18 April 2024
Accepted: 23 April 2024
Published: 25 April 2024

1. Introduction

The increase in textual user-generated content on online platforms and social networks has become a key source for a wealth of information. Such platforms are often used to share information such as news, brands, political discussions, and more [1]. Nevertheless, the large volume of text data that are broadly available on the web increase the challenge of identifying the most relevant information in real time. Text mining is a promising solution to the problem of information overload associated with summarising and understanding vast amounts of unstructured text originating from diverse sources. More specifically, topic modelling aims to identify and extract salient concepts or themes, also known as "topics", distributed across a collection of documents [2].

Many studies have focused on extracting topics expressed in several domains, such as online social networks (e.g., [3,4]), particularly to identify influential individuals on a social media platform (e.g., [5]) and detecting signs of depression in related language on Twitter

(e.g., [6]), clinical applications such as triaging patients based on referral letters (e.g., [7]), extracting scientific topics in academic journals (e.g., [8]), and most recently, understanding the public's opinion of governments during the COVID-19 pandemic (e.g., [9]).

The output from applying topic modelling to a collection of documents is commonly presented as a set of the top most co-occurring terms appearing in each topic [10,11]. Manual effort is often associated with interpreting a topic's description from a set of isolated tokens, where a topic is often given a title or a name to reflect the understanding of its underlying meaning [12]. For example, "car, power, light, drive, engine, turn" may infer topics surrounding Vehicles, and "game, team, play, win, run, score" may infer Sports. However, one of the key concerns with topic models lies with the subjectivity surrounding this task as well as how well human readers can understand the topics, otherwise referred to as topic interpretability [12]. More often than not, topic modelling outputs may not adequately describe the meaning of the topic itself. Problems include generic topic descriptions with too many words [13], topics with disparate or poorly connected words [14], misaligned topics [15], and multiple nearly identical topics [13]. Subsequently, this misinterpretation leads to ineffective information retrieval and decision making.

In this case, several studies have attempted to automatically extend topic descriptions or select the best label for a given topic as a means of enhancing the interpretation of topic models and reducing the human in the loop. Such approaches often rely on external textual sources (e.g., Wikipedia, WordNet) or sophisticated Large Language Models (LLMs) (e.g., ChatGPT [16]). In recent advancements, the development and maintenance of machine learning models and resources in privacy-aware settings have seen significant improvements, making the utilisation of external resources more feasible and secure than ever. However, despite these advancements, relying solely on such resources can still pose inherent risks and limitations. Firstly, the availability and stability of these resources remain a concern. Although less frequent, outages or restrictions in access to these external resources can disrupt the continuity and reliability of research relying on them. Moreover, the relevance and currency of information from these sources are not always guaranteed. In rapidly evolving fields, external resources might not be updated promptly, leading to the generation of outputs based on outdated information. Secondly, while privacy-aware machine learning models have made strides, they are not entirely foolproof. There are ongoing concerns regarding data breaches and the potential for these models to inadvertently memorise and reveal sensitive data. This is particularly crucial in contexts dealing with confidential or proprietary information, where the inadvertent leakage of data through these models could have serious ramifications.

Given these considerations, this paper proposes an approach which aims to strike a balance between leveraging the advancements in privacy-aware machine learning and mitigating the risks associated with dependency on external resources. By developing a methodology that primarily relies on internal data processing and analysis, this method aims to maintain a high degree of control over data relevance and privacy, ensuring that the model remains robust and reliable. More specifically, this paper proposes:

- A novel approach towards extending the output of traditional topic modelling methods beyond token outputs. The initial experiments presented in this paper are tested using Latent Dirichlet Allocation (LDA), a traditional approach for distributing text segments into topics. However, given the positive findings of the experiments, and to demonstrate the generalisation of the proposed method, the approach is also applied when two of the latest state-of-the-art topic modelling methods, BERTopic and Top2Vec, are used.
- An evaluation of how such extended outputs improve the interpretability of topic descriptions from a human perspective.
- An evaluation of how such an approach may be generalised using other topic modelling methods as the underlying models and when using a heterogeneous unseen dataset.

- A comparative analysis of how such extended outputs benchmark against the results generated by state-of-the-art LLMs.
- A demonstration of how the proposed approach performs using an unseen, much larger dataset.

The study was designed as shown in Figure 1: (1) pre-process text responses using traditional Natural Language Processing (NLP) techniques, (2) apply the topic modelling algorithm, (3) for each topic in 2, extend the token list output by applying keyword extraction to the original text segments, (4) map token outputs from 2 to the extended outputs in 3, and (5) evaluate the ease of interpreting outputs from 2 and 4 by human readers.

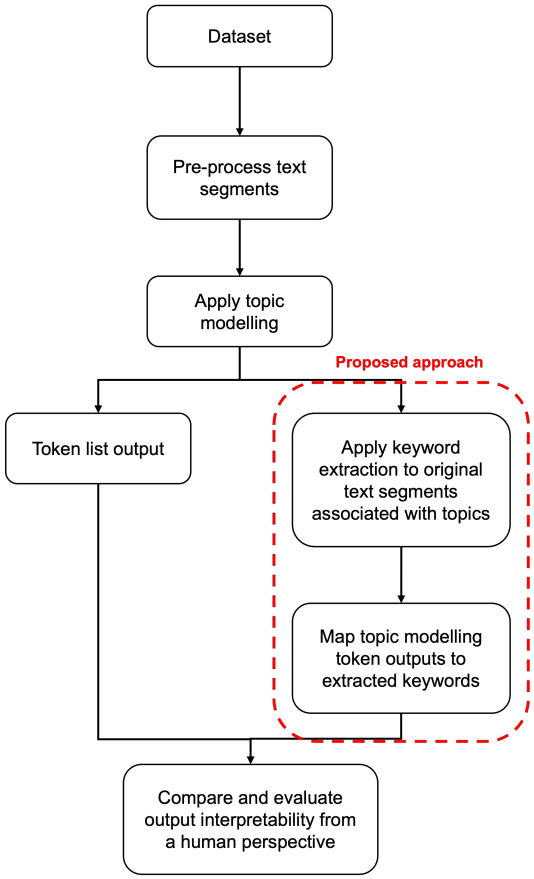

Figure 1. An overview of the study design.

The remainder of this paper is structured as follows: Section 2 presents the related work, Section 3 discusses the selection of text corpora used to support the experiments herein and the techniques used to prepare the data for such experiments, Section 4 discusses topic modelling and how it was applied to the datasets, Section 5 discusses the proposed methodology of using keyword extraction to extended topic modelling outputs, Section 6 evaluates the ease of interpreting such outputs by human readers, Section 7 discusses the generalisation of the proposed method by using other approaches as the underlying topic modelling methods, Section 8 compares the descriptors generated by the proposed method with those generated by state-of-the-art LLMs, Section 9 discusses the generalisation of the proposed method when using new data, and finally, Section 11 concludes the paper.

2. Related Work

Several current works that have adopted topic modelling often rely on the manual approach of interpreting and determining the topic's main descriptive label based on the token results generated by the topic modelling approach (e.g., [17,18]). However, previous work has shown that using top terms are not enough for interpreting the coherent meaning

of a topic following topic modelling [19,20]. Such studies include Lee et al. [21], who qualitatively evaluated how readers understand, assess, and refine topics. They demonstrate a disconnect between how readers perceive topic model outputs and the operations that current interactive systems support. In this case, several studies have attempted to automatically extend topic descriptions as a means of enhancing the interpretation of topic models. Recent works have explored the use of external sources, such as Wikipedia, WordNet, or other ontologies for supporting the automatic labelling of topics by deriving candidate labels using lexical-based (e.g., [19,22–25]) or graph-based algorithms applied on these sources (e.g., [26,27]). For example, recently, both Allahyari et al. [28] and Kinariwala and Deshmukh [29] propose an ontological approach for generating topic labels. Both approaches use the semantic relatedness of LDA outputs and their ontological categories to determine topic labels. However, limitations of this approach include the reliance of an ontology of terms, which, in the case of [29], is only 500 words, which are categorised and limited to only four main domains (crime, environment, politics, and sports). Others, such as Bhatia et al. [24], have used word embeddings to map topic modelling outputs to Wikipedia article titles using cosine similarity and relative ranking measures, whilst others, such as Bechara et al. [30], propose a transfer learning approach for topic labelling which uses domain-specific codebooks to automatically label topics.

Although the aforementioned works propose methods of automatically assigning more informative and more generic labels based on topic model outputs, they lack emphasis on whether outputs from such approaches are interpretable and meaningful to human readers. To address this, other studies have evaluated their outputs by asking human annotators to label or score them based on how semantically related they were to a topic, how meaningful they are, and their usefulness and coherence. Wan and Wang [31] propose an algorithm to extract text summaries that are much longer than keywords or phrases for describing topics. Aletras et al. [32] compare three different topic representations in a document retrieval task and report that readers find phrased labels easier to interpret in comparison to a list of terms. Lau et al. [22] also use phrases as topic labels and propose a supervised approach for ranking candidate labels. In their work, candidate labels include the top five topic terms and noun chunks extracted from Wikipedia articles. Kou et al. [33] propose another method of semantically mapping topics to candidate labels using word vectors and letter trigram vectors. Mei et al. [19] propose to use n-grams for topic labelling and approach the problem of labelling as an optimisation problem which involves minimising the Kullback–Leibler divergence between word distributions and maximising the mutual information between a label and a topic model. Chang and Boyd-Graber [34] and Morstatter and Liu [12] use crowdsourcing to measure topic interpretability using a numerical score. They validate topic models using word intrusion, an approach which requires participants to study the top words within a topic and to identify the words that do not belong. In addition, they also validate topic models using topic intrusion, an approach similar to word intrusion, where participants study the topic probabilities for a document and evaluate how relevant they are to their understanding of the text.

The work cited above focuses on extending topic modelling outputs, and more importantly, evaluating how interpretable such outputs are to human readers. However, such approaches use external textual sources that may be at risk of becoming unavailable. The size of such resources may also be relatively large, and therefore, generating results from them in real time may be ineffective and computationally intense. There has yet to be an investigation into a lightweight approach that uses the textual data themselves to extend topic modelling outputs for systems that require the almost immediate presentation of interpretable topics to humans.

3. Text Corpora and Data Preparation

To conduct the experiments described herein, the data were provided by our industrial research collaborator, a software company that focuses on the concept of crowdsourcing solutions to strategic business challenges. The eight datasets consist of short informal re-

sponses to a range of different questions, propositions, and requests for ideas. For example, dataset ID = 1 asked its participants, "Who does the channel reach, what is the purpose of the channel, and what information does the channel share?", whereas dataset ID = 4 asked, "What is your supply chain going to be, who are your most important customers, and how will you generate income?". Table 1 reports the distribution of responses across each dataset, with Table 2 reporting some examples of the text responses. By utilising these datasets, each representing unique variations and characteristics, the generality of the proposed approach of expanding topic modelling outputs can be tested, enhancing its reliability and real-world applicability.

Table 1. Distribution of responses across datasets.

ID	Total Responses	Min # of Tokens	Max # of Tokens
1	84	1	48
2	277	1	165
3	156	1	102
4	1028	1	1154
5	149	1	105
6	100	1	104
7	149	1	109
8	85	1	66

Table 2. Pre-processing text responses for topic modelling.

ID	Original Text Responses	Pre-Processed Text Responses
1	all staff	'staff'
2	improve staff moral and knowledge of ongoing issues specifically within specialist crime.	'improv', 'staff', 'moral', 'knowledg', 'ongo', 'issu', 'specialist', 'crime'
3	the army doesn't really have a lot of equipment for the cold weather	'armi', 'lot', 'equip', 'cold', 'weather'
4	most work would be submitted electronically-I would need a computer, screen, internet connection, etc.	'work', 'submit', 'electron', 'comput', 'screen', 'internet', 'connect'
5	mental health is a very important aspect of staff wellbeing, and it affects a lot of people.	'mental', 'health', 'aspect', 'staff', 'wellb', 'lot', 'peopl'
6	working remotely, particularly in transit, we need quick access to colleagues contacts	'working', 'remot', 'transit', 'quick', 'access', 'colleagu', 'contact'
7	would require some planning and availability to be able to facilitate such events.	'requir', 'plan', 'avail', 'facilit', 'event'
8	really beneficial to distance from your work for a short moment and reset again!	'benefici', 'distanc', 'work', 'short', 'moment', 'reset'

The data preparation and analysis in this study was conducted using Python (version 3.7.2). For text pre-processing, the traditional NLP techniques were applied, including:

- Converting text to lowercase.
- Removing additional white spacing using Python's regular expression package, RegEx (version 2020.9.27).
- Removing punctuation, digits, and emojis using RegEx.

- Tokenising text using Python's natural language package, Natural Language Toolkit (NLTK) (version 3.4.1).
- Removing stop words as part of the NLTK package.
- Stemming tokens using the Porter Stemmer as part of the NLTK package.

Table 2 shows examples of the original text responses against how they are represented following pre-processing.

4. Topic Modelling

There are various methods by which texts can be distributed into topics, with some of the most traditional methods being probabilistic Latent Semantic Analysis (pLSA) [35], Latent Semantic Analysis (LSA) [36], and LDA [37]. More recently, however, with the recent advancement in the NLP field, newly developed algorithms, such as BERTopic [38] and Top2Vec [39], are continuing to attract attention.

Such methods vary in complexity, the representation of text segments taken as the model's input (e.g., bag of words, word embeddings), their computational speed, and their performance. For example, the LDA algorithm proposes a fixed number of topics in a collection of documents and assumes that each document reflects a combination of those topics. When a collection of documents is analysed under these assumptions, probabilistic inference algorithms reveal an embedded thematic structure, allowing for large collections of documents to be quickly summarised, explored, and searched [40]. In general, the LDA algorithm calculates the probability that a word within a document will be included in each topic. A topic may be described by extracting words with the highest probabilities which correspond to such topic. That is, LDA analysis finds the latent topic corresponding to the words in any given document [2]. A document is determined to address a topic by calculating a probability distribution over a range of topics for each document [3] and selects the topic with the highest probability as the main topic description.

However, BERTopic and Top2Vec use word embeddings. That is, the vectorisation of text data makes it possible to locate semantically similar words, sentences, or documents within spatial proximity. As word vectors that emerge closest to the document vectors are considered as being the best description of topic of the document, the number of documents that can be grouped together represents the number of topics [41]. Whereas BERTopic uses the Hierarchical Dirichlet Process (HDP) to cluster vectors into topics, Top2Vec employs a combination of document clustering and word embedding techniques. It uses the Doc2Vec algorithm to generate document embeddings [42]. Unlike traditional topic modelling methods, Top2Vec does not require specifying the number of topics in advance. Instead, it identifies topic clusters based on the density of document embeddings and extracts representative keywords and documents for each topic.

Before applying the proposed approach to state-of-the-art models, initial experimentation was conducted using the traditional topic modelling method, LDA. By validating the method with LDA, which is a well-established and widely used technique, it is possible to verify that the approach functions as intended and produced satisfactory results.

For each pre-processed dataset described in Section 3, the topic modelling approach was applied using the 'Latent Dirichlet Allocation' [43] class available as part of the Scikit-learn package. A key hyperparameter of the LDA algorithm is the number of topics [7]. To calculate the optimal number of topics across a collection of documents, a good indication is the number of topics with which the model best predicts the data. For topic models such as the LDA, a common indicator to measure the optimal number of topics is perplexity, a measure of how well a probabilistic model predicts a sample [37]. Yet, recent studies have shown that predictive likelihood (in this case, perplexity) and human judgement are often not correlated [34]. In this case, and to retrieve a comparable number of topics, for each dataset, text responses were distributed as one of ten topics. To maintain the comparability across each dataset, the default values of the remaining hyperparameters of the LDA algorithm were set [43]. Once topics were identified as a probability distribution of words, each text sample is denoted as a probability distribution over topics. The topic with

the highest probability was therefore applied to each text segment. In addition, for each topic, the top ten tokens with the highest probabilities were selected as the description of the topic.

Dataset ID = 5 is used as an ongoing example to illustrate the results of the traditional and extended topic modelling outputs discussed in this paper. However, the results are applicable across all datasets. Table 3 reports the token list output produced by the LDA approach. Evidently, based on the content of the textual responses, several words are extracted from topics that may infer its description. For example, blood donation service may be inferred as the description for topic 1, staff mental health may infer topic 4, and email signatures may infer topic 8. Other topic descriptions highlight the possible challenges surrounding the inferences of topic descriptions to human readers as they include inconsistent tokens that may be ambiguous without context. Such examples include topic 3, which includes the words piano, kitchen, bond, etc.

Table 3. Token outputs across topics for dataset ID = 5.

Topic	LDA Output
1	'offic', 'blood', 'donat', 'welsh', 'servic', 'tree', 'clinic', 'onenot', 'help', 'milk'
2	'address', 'request', 'item', 'power', 'group', 'receiv', 'start', 'environ', 'real', 'differ'
3	'provis', 'piano', 'kitchen', 'bond', 'easi', 'stress', '100k', 'music', 'boost', 'consid'
4	'train', 'aid', 'health', 'mental', 'team', 'staff', 'provid', 'quarter', 'roll', 'recycl'
5	'calcul', 'app', 'mileag', 'adjust', 'height', 'repay', 'loan', 'trip', 'catcher', 'work'
6	'learn', 'yoga', 'repair', 'lesson', 'hand', 'sanitis', 'review', 'theme', 'improv', 'cafe'
7	'starter', 'leav', 'annual', 'advanc', 'warn', 'board', 'notic', 'digit', 'encourag', 'comm'
8	'signatur', 'email', 'electron', 'railcard', 'dbw', 'allow', 'benefit', 'flexi', 'lgbt', '30'
9	'mail', 'list', 'websit', 'vacanc', 'distribut', 'christma', 'data', 'breach', 'alli', 'budget'
10	'post', 'gener', 'polici', 'john', 'role', 'refer', 'friendli', 'social', 'messag', 'display'

5. Extending Topic Modelling Outputs

To reiterate, this paper proposes a new method towards extending the outputs of traditional topic modelling methods. The initial experiments presented herein are tested using the LDA method described in Section 4. The proposed approach includes the following steps:

- Mapping the original text responses in each dataset (Table 2) to the most dominant topic expressed in such texts by selecting the topic with the highest relevance score produced by the topic model, in this case, the LDA.
- For each topic, apply a keyword extraction approach to the original text responses.
- For each topic, map the aforementioned extracted keywords to the LDA's output (Section 4) by selecting the keywords with the highest number of intersecting tokens with those produced by the LDA model. Subsequently, of those further refined keywords phrases, the top-scoring ones were assigned as the topic's main description.

Keywords, which are often defined as a sequence of one or more words, provides a short description of the content of a text document [44]. Such keywords may be useful entries for building an automatic indexing system for a document collection or can be used to classify text. In the context herein, keywords may serve as a concise label for a given topic.

Several keyword extraction libraries exist as part of the Python programming language (e.g., Gensim [45], PyTextRank [46], YAKE (Yet Another Keyword Extractor) [47], etc.). A popular keyword extraction algorithm is Rapid Automatic Keyword Extraction (RAKE) [44], an unsupervised, domain- and language-independent tool for extracting keywords from documents. In this paper, to demonstrate the proposed method of extending topic descriptors beyond tokens, the RAKE algorithm was used due to several features that make it a favourable keyword extractor, including its known computational

efficiency and speed, its precision despite its simplicity, and its ability to work on individual documents [44].

RAKE identifies stop words and phrase delimiters to split the document into candidate keywords, which are sequences of content words as they occur in the text [48]. Firstly, RAKE tokenises the document text by specific word delimiters. Using phrase delimiters and the positions of stop words, tokenised texts are then split into sequences of continuous words. Words within a sequence are assigned the same position in the text and together are considered a candidate keyword [49]. Once each candidate keyword is identified, a score is calculated for each and is defined as the sum of its member word scores. Such scores are calculated using one of three methods: the degree of a word in the matrix (i.e., the sum of the number of co-occurrences the word has with any other content word in the text), the word frequency (i.e., the number of times the word appears in the text), or as the degree of the word divided by its frequency. To find specific keywords, RAKE searches for pairs of keywords that adjoin one another at least twice in the same document and in the same order. A new candidate keyword is then created using a combination of the extracted keywords. The score for the new keyword is the sum of its member keyword scores. As RAKE splits candidate keywords by stop words, interior stop words are not present in the extracted keywords.

After applying topic modelling (Section 4), the original text responses in each dataset (see Table 2) were mapped to the most dominant topic expressed in such texts by choosing the topic with the highest relevance score. For example, for dataset = 5, topic 8 was assigned to the text response "Implementation of electronic signatures on contracts and offer letters. This will support with process improvement and digitisation" as a consequence of achieving the highest relevant score of 0.94, whereas topic 4 was assigned to "Mental health is a very important aspect of staff wellbeing, and it affects a lot of people" as a consequence of achieving the highest relevant score of 0.94. For each topic, RAKE was applied to the original text responses, and the keywords, as well as their scores, were extracted.

The keywords extracted by RAKE were pre-processed following the techniques discussed in Section 3 and mapped to those extracted as part of the LDA's output. To facilitate this mapping, after applying direct string matching, the keywords with the highest number of intersecting tokens with those produced by the LDA model were extracted. Subsequently, of those further refined keyword phrases, the top-scoring ones were assigned as the topic's main description. Topics with more than one top-scoring keyword were concatenated and separated by the '/' delimiter.

For example, as a consequence of including the highest number of intersecting tokens, i.e., aid, train, mental, and health (see Table 3), as well as achieving the highest score produced by the RAKE algorithm (see Table 4), topic 4 in dataset ID = 5 was assigned Aid Training Courses/Mental Health Issues as the topic's descriptor. The remaining keywords may be useful in describing other topics discussed in the text responses (e.g., Mental Health Illness, Aid Training and Training Providers).

Table 4. Top scoring keyword phrases for topic 4, dataset ID = 5.

RAKE Keyword Phrases	Score
Aid Training Courses	9
Mental Health Issues	9
Mental Health Illness	8.33
Mental Health	5.33
Aid Training	4
Training Providers	4

Figure 2 reports the extended outputs across each of the ten topic extracted for all datasets. In very few examples (e.g., Free Blood Pressure Testing Carries (dataset ID = 6)

and Dedicated File Preparation Officers Located (dataset ID = 7)), extended outputs may lack minor subject–verb agreement and coherence. However, in comparison to the token list output produced by the LDA topic modelling approach (Table 3), the resulting extended method presents more cohesive and contextualised topic descriptions.

Topic	Dataset ID = 1
1	Social Media
2	Support Police Staff 24Hrs Professional Service Income Generation
3	Force Training Days Comms Opportunity
4	Wellbeing Champions Work / Knowledge Hub Area / Good Practice Portal
5	Good News Initiatives Learning
6	Weekly Podcast Ted Talks
7	General Broadcast Address Controlled
8	Digital Info Screens
9	QR Code Info Station
10	Key Issues Subjects

Topic	Dataset ID = 2
1	Dedicated Mental Wellbeing Support Group
2	Commissioning Covert Services Meeting
3	SMS Text Message Facilities
4	Alternatively DFCCU Staff Permanently Attached
5	Standard Cyber Crime Unit Uniform
6	Covert Communication Equipment
7	Embed Career Development / Cyber Protect Role
8	Police Issue Inclement Weather Clothing
9	Vulnerable Individual Involved
10	DMI Software Update Superusers

Topic	Dataset ID = 3
1	Difficult TP Transfer
2	Cover Full Body Lengths
3	Waterproof Trousers Compartment
4	Award Expects Students
5	Oil Rig Business
6	Outdoor Activities / High Impact
7	Stay Safe / Safe Space
8	Emergency Service Personnel / Emergency Service Staff
9	Construction Based Companies
10	Local Community

Topic	Dataset ID = 4
1	Boar Sports Media / Blackonblack Business Canvas / Reach Sports Club / Freedom Seeker Finopa
2	MSO Bidi Font Family
3	Customer Pays Lipa Baadaye
4	Mobile Money Users
5	Require Documents Translated
6	UK International Students Reporting Mental Health Issues
7	Virtual Law Firms Attract Customers
8	Willow Bakeries Food Hygiene Rating
9	Fashion Business Making Unique Handmade Artisanal
10	Social Media Web Site Climbing Gyms Multisport Centre Sports Gear Shops

Topic	Dataset ID = 5
1	Live Progress Tracker
2	Current Induction Process
3	Full DBW Staff Directory Transferred
4	Multi Faith Prayer Meditation Space
5	Free Blood Pressure Testing Carried
6	Meeting Scheduling Software
7	Customer Referral Programme
8	Offer Letter Storage
9	Remember Previous Journeys Undertaken
10	2021 Wales Rugby League

Topic	Dataset ID = 6
1	Welsh Blood Service Donation Clinic
2	Group Email Address
3	Kitchen Team Bonding
4	Aid Training Courses / Mental Health Issues
5	Adjustable Height Work Stations Attached
6	Repair Cafe
7	Digital Notice Board
8	Email Signatures
9	Vacancy Mailing List Mail Distribution Lists
10	General Flexi Time Policies

Topic	Dataset ID = 7
1	Developing Rounded Officers
2	Ring Fenced Admin Day
3	Dedicated File Preparation Officers Located
4	Practical Portfolio Incorporating
5	CID Awareness Days
6	CID Posts / Accredited Detectives / Job Advert / Definitely Encourage
7	Good Detectives Lost / Good Suspect Interview
8	Plain Clothes Allowance CID Specialist Crime Officers
9	Tutor Mentor
10	Case Specific Submissions

Topic	Dataset ID = 8
1	Online Enrichment Portal
2	Fancy Dress Costumes
3	Meal Snack Ideas
4	Bringing Fun
5	White Ribbon Campaign Charity
6	Mental Health Drop / MS Teams Drop
7	Daily 5 Minute Stretch Exercise
8	Recipe Swap Group
9	Changing Backgrounds
10	Email Signature Template

Figure 2. Extended topic outputs for all datasets.

6. Evaluating Interpretability from a Human Perspective

Inspired by several studies discussed in Section 2 (e.g., [31,33,34]), to gain an insight into whether the proposed method of extending topic modelling outputs increases the interpretability of topic descriptions from a human's perspective, independent annotators were asked to label both topic modelling outputs according to their:

- Quality—how easy is it to extract meaning from the text?
- Usefulness—how relevant or helpful is the text in providing information about a topic?
- Efficiency—how efficient is the annotation task?

An LDA-generated topic characterised by tokens such train, aid, health, mental, team, and staff would score medium to high in quality, usefulness, and efficiency, as these words may collectively convey a theme centred around health services or medical assistance, particularly with a focus on mental health. However, a topic represented by tokens such as piano, kitchen, bond, easy, stress, music and 100k would score low in all metrics due to the lack of a coherent narrative thread, therefore diminishing its usefulness for any insightful discourse. The efficiency of the annotation task might be impeded by such incoherence, as annotators spend additional time trying to decipher a viable connection between the terms. On the other hand, topics such as Welsh Blood Donation Clinic, Live Progress Tracker, and Dedicated Mental Wellbeing Support Group (see Figure 2) are examples of extended topic descriptors that would score high in all three metrics as they present cohesive and specific themes that are immediately recognisable and understandable. Each of these topic descriptions demonstrates efficiency as they allow annotators to quickly understand and label them without confusion or the need for extensive background knowledge. This efficiency ensures that the annotation task can be completed effectively. These contrasting examples reflect the potential range of interpretability that independent annotators may encounter and underscore the importance of our method's ability to enhance the clarity and relevance of topics for human understanding.

To facilitate the annotation task, a bespoke web-based annotation platform was implemented. This reduced any installation overhead and widened the reach of annotators as it was accessible using a web browser regardless of the device type (i.e., smartphone, laptops, PC, etc.). Annotators were presented with instructions explaining the task's requirements and then with the platform's interface (see Figure 3. The first pane contained

a randomly selected topic modelling output to be annotated, as well as the remaining number of annotations left to complete. For the LDA outputs, to increase interpretability, annotators were presented with non-stemmed tokens. The subsequent panes contained the annotation choices for the aforementioned metrics. For each metric, annotators were required to label each output on a five-point Likert scale, where a score of zero signified poor quality, no usefulness, and a task that is not efficient, and a score of 4 signified high quality descriptions that contain clear and effective text that is easy to extract meaning from, that are useful in describing a topic, and that perform the annotation task in the best possible manner with the least amount of effort.

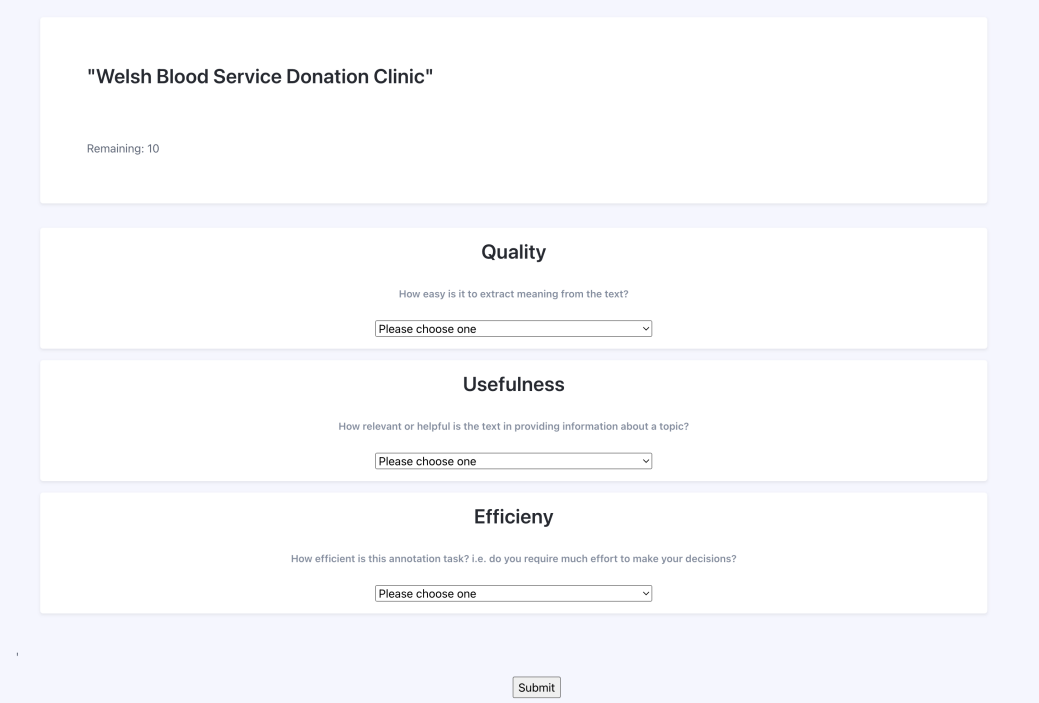

Figure 3. Bespoke annotation platform.

The crowdsourcing of labelling natural language often uses a limited number of annotators with the expectation that they are perceived to be experts [50]. However, the task of annotating text is considered as being highly subjective and varies with the annotator's age, gender, experience, cultural location, and individual psychological differences [51]. For instance, Snow et al. [52] investigate collecting annotations from a wide audience of non-expert annotators over the Web. They show high agreement between the 10 annotations provided by non-experts and those provided by experts.

In this case, to crowdsource annotations in this study and to reach a diverse range of annotators, Twitter was leveraged as a distribution channel. The study was able to attract a broad audience interested in contributing to the annotation process. Annotators were not selected selectively; instead, participation was open to anyone who expressed interest in contributing their insights. To maintain anonymity and avoid duplication of annotations, annotators were distinguished solely by their IP addresses. No personal information was collected from annotators, as explained in the privacy policy, ensuring the confidentiality of their identities. All annotation results were securely stored in a relational database.

A total of 1600 annotations were collected for all 80 outputs described in Section 4 and 80 extended outputs described in Section 5, with 10 annotations per output. A total of 63 annotators participated in the study.

The distributions of annotations across quality, usefulness, and efficiency scales for extended topic modelling outputs across each dataset described in Table 1 are shown in Figures 4–6, respectively, whilst Figures 7–9 report the distributions of annotations across

LDA outputs. For both quality and usefulness, the extended outputs incurred by far a highest usage of 4 on the Likert scale in comparison to the LDA outputs (quality = 302 out of 800 (37.8%), usefulness = 286 out of 800 (35.8%)). An interesting observation is that dataset ID = 8 did not receive any annotation scores of 0 or 1 for both metrics. Conversely, the LDA outputs incurred by far the highest usage of 0 (quality = 425 out of 800 (53.1%), usefulness = 474 out of 800 (59.3%)). Further analysis demonstrates that datasets ID = 2 and 6 did not receive any annotations of scores of 4 for both metrics.

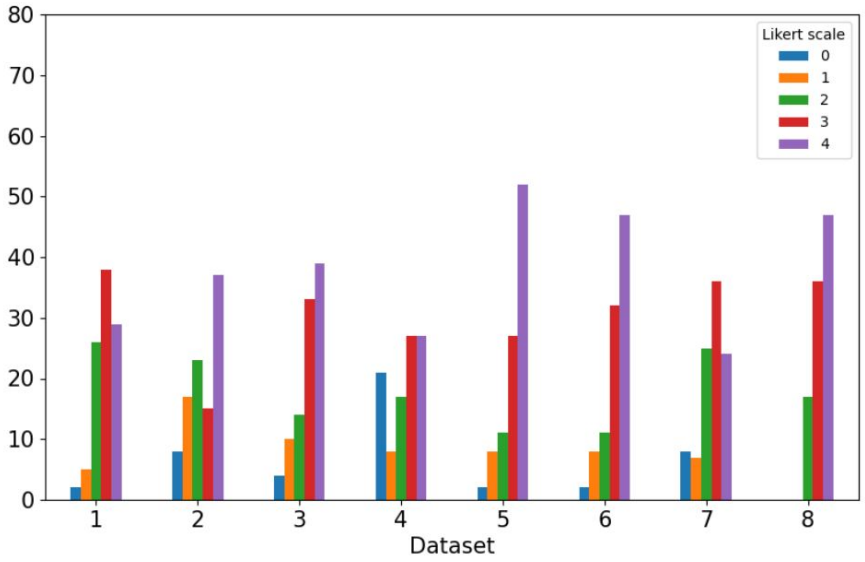

Figure 4. Distribution of annotations across the quality of extended outputs.

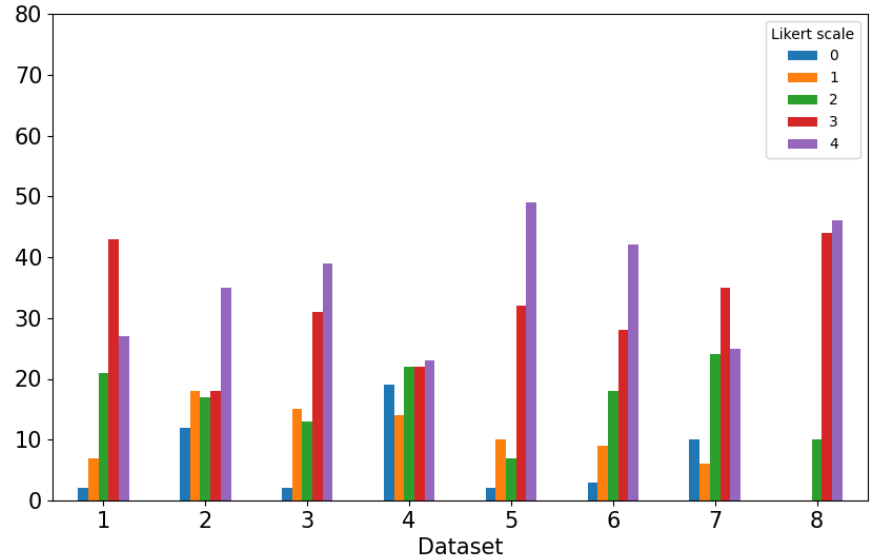

Figure 5. Distribution of annotations across the usefulness of extended outputs.

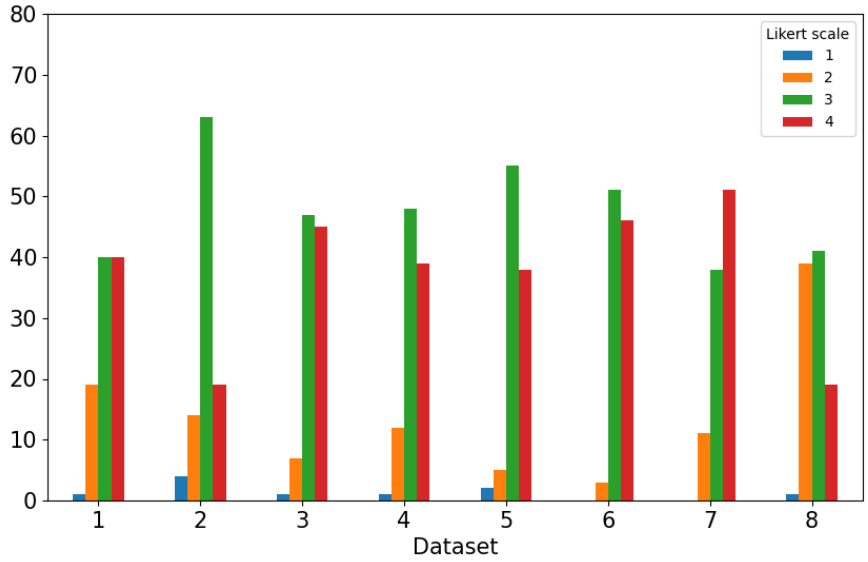

Figure 6. Distribution of annotations across the efficiency of extended outputs.

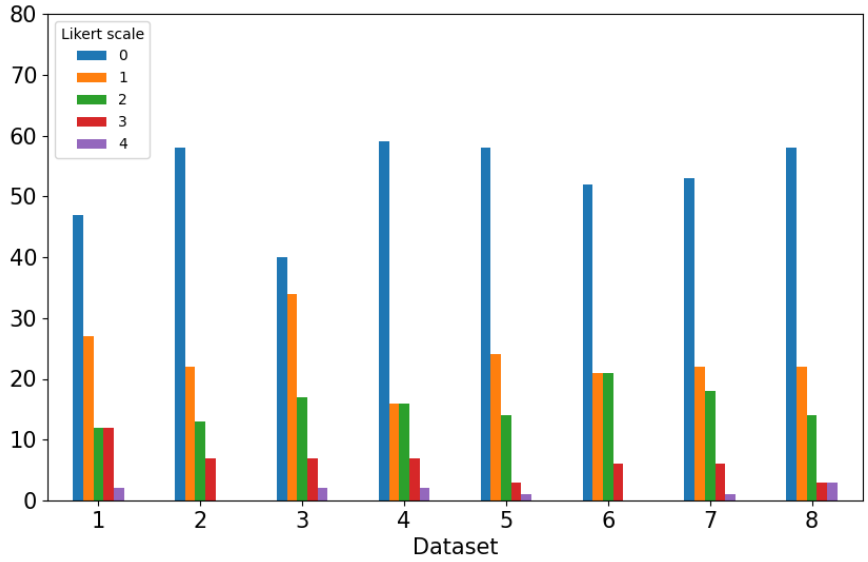

Figure 7. Distribution of annotations across the quality of LDA outputs.

In terms of task efficiency, similar results can be observed. For the extended outputs, large distributions of 3 and 4 on the Likert scale were used by annotators, denoting that the task of interpreting the extended topic descriptions did not require a significant amount of effort to complete. On the other hand, the LDA outputs report to have a high distribution of 0 on the Likert scale, indicating that annotators required much more effort in defining whether the quality and usefulness of the topic descriptions.

We hypothesise that when an output is unambiguous and helpful in providing information about a topic, then the likelihood of independent annotators selecting higher scores for the metrics increases. This leads to a higher inter-annotator agreement, which indicates that a topic output is more interpretable. Krippendorff's alpha coefficient [53] was used to measure the inter-annotator agreement. As a generalisation of known reliability indices, it was used as it (1) applies to any number of annotators, not just two, (2) applies to any number of categories, and (3) corrects for chance expected agreement. Krippendorff's alpha coefficient of 1 represents full agreement, 0 represents no agreement beyond chance, and -1 represents disagreement.

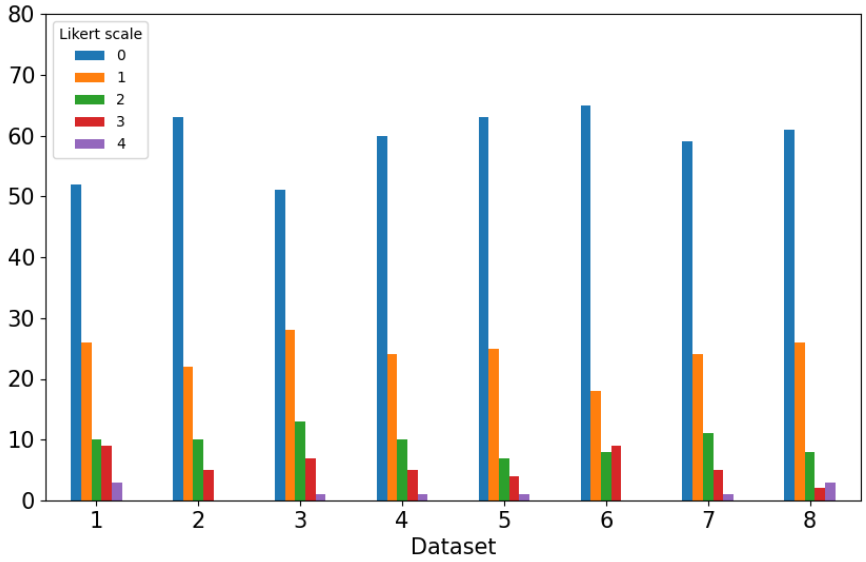

Figure 8. Distribution of annotations across the usefulness of LDA outputs.

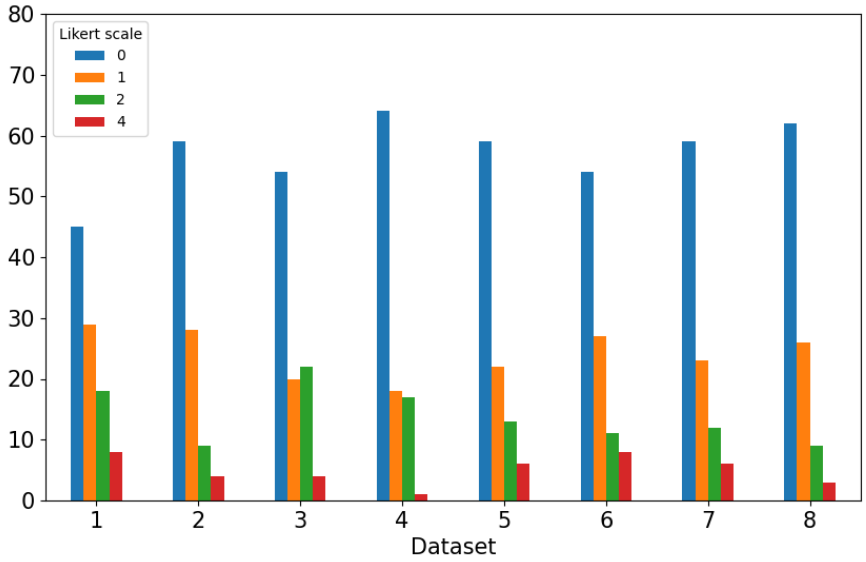

Figure 9. Distribution of annotations across the efficiency of LDA outputs.

Figures 10–12 report the inter-annotator agreement for the quality and usefulness of both the topic modelling outputs across each dataset as well as the efficiency of the annotation task. For quality, the extended outputs reported the highest agreement of $\alpha = 0.522$ (dataset ID = 4) and the lowest agreement of $\alpha = 0.216$ (dataset ID = 8). Likewise, for usefulness, the extended outputs reported the highest agreement of $\alpha = 0.575$ (dataset ID = 1) and the lowest agreement of $\alpha = 0.220$ (dataset ID = 5 and 6). In terms of the LDA outputs, for quality, the highest agreement reported was $\alpha = 0.075$ (dataset ID = 3) and the lowest agreement was $\alpha = -0.036$ (dataset ID = 7). Likewise, for usefulness, the LDA outputs reported the highest agreement of $\alpha = 0.090$ (dataset ID = 3) and the lowest agreement of $\alpha = -0.040$ (dataset ID = 1).

For extended outputs, the relatively high agreement for both quality and usefulness may be explained by the fact that they received a higher distribution of higher scoring annotations. For instance, referring back to the ongoing example of dataset ID = 5 (see Table 2), the extended output Group Email Address received a unanimous agreement of a score of 4 across both quality and usefulness, whereas due to the ambiguity of the topic description given by the LDA output ("address, request, item, power, group, receive, start, environment, real, differ"), it received the following distribution of annotations: quality (0 = 8 annotations, 1 = 1 annotation, 2 = 1 annotation) and usefulness (0 = 8 annotations, 1 = 2 annotations).

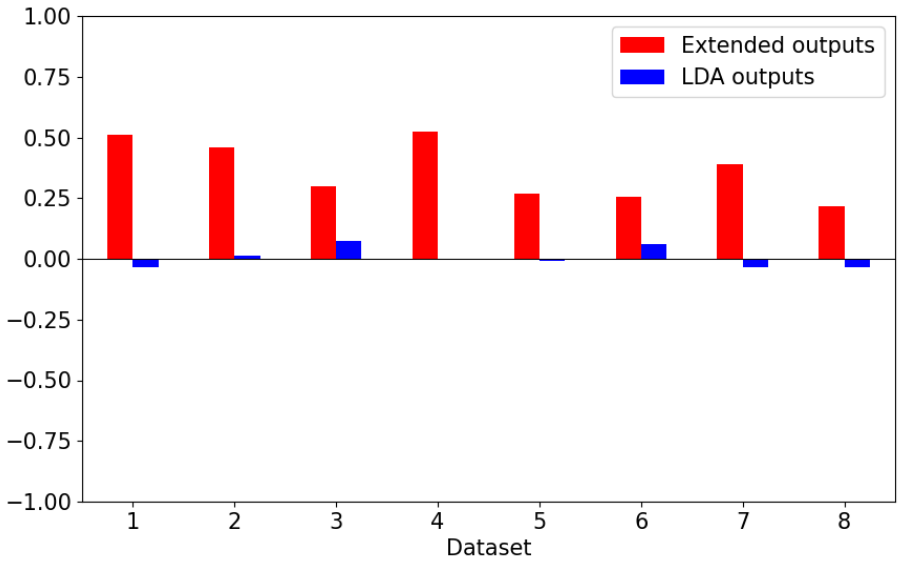

Figure 10. Inter-annotator agreement across quality.

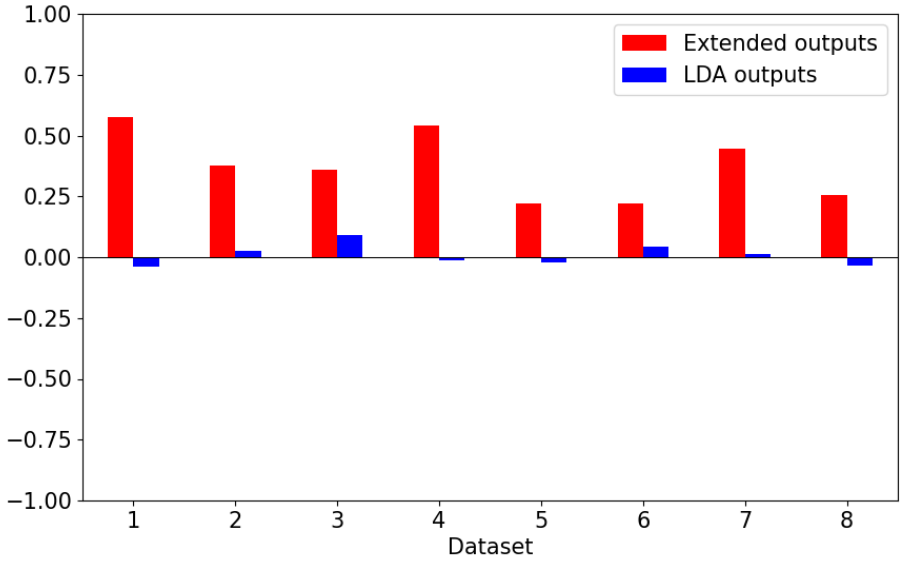

Figure 11. Inter-annotator agreement across usefulness.

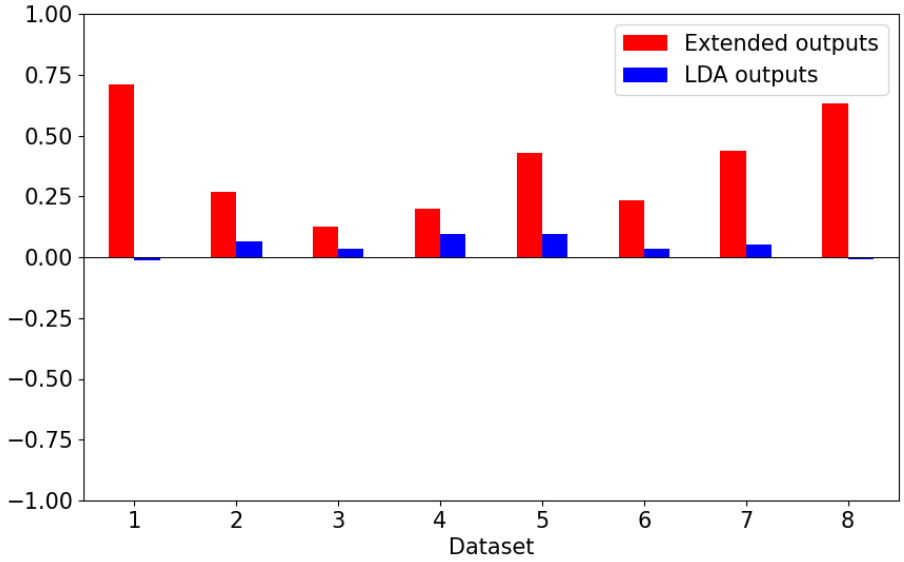

Figure 12. Inter-annotator agreement across efficiency.

In terms of task efficiency, similar agreements can be found. It was reported, for extended outputs, the highest agreement of $\alpha = 0.709$ (dataset ID = 1) and the lowest agreement of $\alpha = 0.126$ (dataset ID = 3). For the LDA outputs, the highest agreement of $\alpha = 0.094$ (dataset ID = 4) and the lowest agreement of $\alpha = -0.012$ (dataset ID = 1) were reported. Overall, the results demonstrate that the extended outputs are reliably more interpretable to human readers compared to LDA outputs, and the task of annotating their quality and usefulness is particularly less challenging for annotators when observing the extended outputs in comparison to LDA outputs.

7. Generalisation of the Proposed Method Using Other Topic Modelling Approaches

The initial experiments conducted with LDA provided a solid foundation, demonstrating the effectiveness of the approach in generating intepretable topic descriptors. Building upon this, to demonstrate the generalisability and robustness of the proposed approach, the methodology is extended to include BERTopic and Top2Vec, two advanced state-of-the-art topic modelling techniques.

Figures 13 and 14 report the extended outputs across each of the topics extracted for all datasets when BERTopic and Top2Vec were used as the underlying topic modelling method, respectively. Both methods were applied without any constraint on the number of topics to be generated.

The implementation of BERTopic and Top2Vec present a broader spectrum of insights. BERTopic, with its deep learning foundation, offered a more nuanced understanding of the text data, generating topic descriptions that were not only coherent but also contextually rich. This method was particularly effective in capturing the subtleties and complexities within the datasets, which LDA might have overlooked due to its probabilistic nature. For instance, in dataset ID = 4, BERTopic provided a multi-dimensional view of topics, reflecting a deeper layer of thematic understanding.

Top2Vec, on the other hand, presented a unique perspective by generating an extensive array of topics, exceeding 100 in dataset ID = 4. This method's ability to produce such a wide range of topics highlights its utility in exploring large and diverse datasets, offering a comprehensive landscape of themes and ideas present in the data. However, it is important to note the presence of single-word topic descriptions in some cases. While these may seem less informative at first glance, they play a crucial role in offering a focus on specific themes, as seen with terms like Insurance and Docbot.

Topic	Dataset ID = 1
1	Good Practice Portal
2	Text Push Messaging / Digital Info Screens / Deliver Key Messages
3	Operational News Relevant

Topic	Dataset ID = 2
1	Police Staff Safety Training
2	Promoting Officer Fitness Improving Officers Mental Health
3	DMI Software Update Superusers
4	Builders Van Props Clothing Signs
5	Investigate Cyber Crime OSINT Tools
6	Dedicated Digital Investigators
7	CHIs Referrals Process / CHIs Tasking Process

Topic	Dataset ID = 3
1	Emergency Service Personnel / Emergency Service Staff
2	Additional Roof Coat Bag
3	Oil Rig Business / Construction Based Companies
4	Colourful Coats / Homeless People / Safe Space
5	Waterproof Jacket / Keep Police / Waterproof Coat / Security Dry
6	Shopping Bag / Reducing Plastic

Topic	Dataset ID = 4
1	Investors Marketers UXI Designers Software Developers Cardiff Council Main Customer
2	Updated Website Consistent Social Media Posting Good Customer Satisfaction
3	Fashion Business Making Unique Handmade Artisanal
4	10 Month Subscription
5	Leeds 60 Study Abroad Offices
6	Flyingint Education
7	Prospective Language Service Providers Translation Agencies
8	Virtual Law Firms Attract Customers
9	Played Murder Mystery Type Games
10	Providing High Quality Magazines
11	Targeted Marketing Site / Brands University Students / Target Young People
12	Mobile Money Users
13	8 UK Medical Schools Offering Medicine
14	Help Reducing Food Waste
15	Interactive Employee Training Video
16	Wedding Photo Video Shooting Package Services
17	Suite Business Workers
18	Rusty Design Limited
19	Accountant Lawyer Web App Developer
20	Enactus Candle / Selling B2C / Selling Directly / Internet Access
21	Gym Owners Dashboard Saas Website Gym Clients
22	Student Safety App
23	Blackonblack Business Canvas
24	Uncommon Ground Coffee Shop
25	Second Shooter Accountant Lawyer
26	Illegally Parked Car
27	Sport Rehabilitation Clinics Reach
28	Customer Pays Lipa Baadaye

Topic	Dataset ID = 5
1	Trip Catcher App
2	Aid Mental Health
3	Team Email Addresses
4	Welsh Blood Service Donation Clinic

Topic	Dataset ID = 6
1	Wales Branded Umbrella
2	Full DBW Staff Directory Transferred
3	Add Details / Exchange Idea

Topic	Dataset ID = 7
1	Specially Trained Officers Conducting Joint Investigations
2	Plain Clothes Allowance CID Specialist Crime Officers

Topic	Dataset ID = 8
1	Daily 5 Minute Stretch Exercise
2	Hays Support Remote Working
3	Imaginative Best Dressed Team

Figure 13. Extended topic outputs for all datasets using BERTopic as the topic model.

Topic	Dataset ID = 1
1	Public / Staff
2	Weekly Podcast Ted Talks

Topic	Dataset ID = 2
1	Supported Appropriately
2	CHIs Tasking Process / CHIs Referrals Process
3	Middle Management / Easier Tasking
4	DSU Staff
5	Relevant Form
6	Car Parking Issues
7	Chief Constables Delivery Plan
8	Data Extractions / Software Updates
9	SMS Messages
10	Attention
11	Covert Communication Equipment
12	Introduction / Function
13	Cyber Crime Printed
14	Full Time Quality Manager
15	Conducting Counter Surveillance / Inclement Weather Clothing
16	Example Spin Bike
17	User Friendly / NCALT Package
18	Staff Engagement
19	Prevent Long Term Sickness
20	Full Time Quality Manager
21	Implement Rag System / Agreed Timeline Set / 3 Month Review
22	Covert Communication Equipment
23	Child Sex Offender
24	Profound Impact
25	Exhibits Seized
26	Meaning Greater Efficiency
27	Cyber Crime Printed
28	Embed Career Development
29	Greatly Assist
30	Screen
31	Fully Knowing Current Structure
32	Better Structure
33	Admin Record Receipt
...	

Topic	Dataset ID = 3
1	Waterproof Jacket / Security Dry / Keep Police / Safe Space / Homeless People / Reasonable Idea / Umbrella Man
2	Easy Resources / Cold Tour

Topic	Dataset ID = 4
1	Boar Sports Media
2	1320
3	Enactus Candle / Soy Wax
4	Society Web
5	Mobile Money / Digital Bank
6	Student Safety App
7	Interactive Employee Training Video
8	Blackonblack Business Canvas
9	Caesar
10	Farming Unions
11	Docbot
12	Cruelty Free False Eyelashes Brand Based
13	Medium Size Organisation
14	Anaesthetic Training Modules
15	Students Save
16	Mobile Money / Digital Bank
17	Accredited Landlords
18	Parents Sellers
19	Helping Businesses Find Development Properties
20	Manage Workloads Effectively
21	Insurance
22	Purchasing Products Worth
23	Vale Uhb Annual Report
24	Web App Developer Investors
25	Biggest Logistics Service Platform
26	Kashmir Untold
27	Rusty Design Limited
29	Boar Sports Media
30	International English Language School
31	Farming Unions
32	Interactive Employee Training Video
33	Buy Products
...	

Topic	Dataset ID = 5
1	Offer Stress Relieving Massages
2	People Accidently Clicking
3	Welsh Blood Service Donation Clinic
4	Mail Distribution Lists / Mileage Claim Revamp / Trip Catcher App

Topic	Dataset ID = 6
1	Defibrillators Save Lives / Customer Referral Programme / Save Paper Plastic / Incentivised Referral Programme
2	Save Paper Plastic
3	Meeting Scheduling Software
4	Offer Letter Storage

Topic	Dataset ID = 7
1	Special Payment
2	Help Eliminate Bad Practice
3	Current 9-5 Time Frame
4	CID Awareness Days
5	Day Basis
6	Bank Holiday / Physical Locations
7	Fear Interviewing Suspects
8	Ring Fenced Admin Day
9	Local CPS Workers

Topic	Dataset ID = 8
1	MS Teams Drop
2	Daily 5 Minute Stretch Exercise

Figure 14. Extended topic outputs for all datasets using Top2Vec as the topic model.

The comparative analysis of topic descriptions across LDA, BERTopic, and Top2Vec revealed interesting patterns. The consistency in topic descriptions across different methods, as shown in dataset ID = 8 with descriptions such as Daily 5 Minute Stretch Exercise, underscores the reliability of the approach. This consistency is also an affirmation of the underlying thematic structures within the datasets. Furthermore, the alignment of topic descriptions across various models indicates a convergence of thematic interpretation, despite the inherent differences in the methodologies the topic models. This congruence suggests that this approach is capable of capturing the core essence of topics regardless of the underlying topic modelling technique used.

8. Comparing the Proposed Method with Large Language Models (LLMs)

Sections 4 and 7 present initial experiments in generating interpretable topic descriptors when using traditional and advanced topic modelling techniques. To further build upon this, to enhance the evaluations presented herein and to provide a more comprehensive analysis, the results reported in Figures 2–14 were compared against two current cutting-edge tools in the field of natural language understanding and generation, OpenAI's GPT-3.5 [16] and Meta's Llama 2 [54].

For each dataset, the original text responses in Section 3 were given as input into both LLMs. To ensure a balanced comparison and that any differences in the topic generation were attributable solely to the models' interpretative algorithms and not to variations in input instructions, both LLMs were directed to generate topics from the raw text provided using the following prompt: 'Given this text, perform topic modelling and generate 10 topics'.

Figures 15 and 16 report the topics generated across each of the ten topics extracted for all datasets by Llama and ChatGPT, respectively. When comparing such results with those generated by the proposed method in Figure 2, each topic modelling method has consistently generated themes with a focus on workplace-related terms such as well-being, training and development, and engagement (dataset IDs = 1, 2, 5, and 6). Customer engagement and business operations (dataset IDs = 2, 7), as well as community engagement and corporate social responsibility (dataset IDs = 2, 4, 6, 7, 8), are prevalent topics in all outputs. Both ChatGPT and Llama occasionally generate specialised topics, such as Cybercrime and Digital Forensics (dataset ID = 6) and Ethical Considerations and Social Responsibility (dataset ID 10), as well as offering a broader perspective that applies to wider contexts, with topics such as Role Awareness and Understanding (dataset ID = 1) and Team Building and Engagement (dataset ID = 5).

As similarly discussed in Section 7, when observing the topics generated by each method, a high degree of correspondence across each dataset may be found, indicating a strong thematic alignment. The proposed method not only matched but occasionally enhanced the topic coverage offered by the aforementioned state-of-the-art language models. For example, in dataset ID = 2, whereas ChatGPT and Llama generated the topic descriptors Employee Well-being and Satisfaction and Staff Wellbeing, the proposed method generated Dedicated Mental Wellbeing Support Group. Another example is dataset ID = 1, where ChatGPT and Llama generated the topic descriptor Training and Development, and the proposed method generated Force Training Days Comms Opportunity. This shows that not only does the proposed method resonate with the thematic coverage found in ChatGPT and Llama, but it often surpasses such models by bridging the gap between broad thematic elements and granular details.

Topic	Dataset ID = 1
1	Detective Workload and Stress
2	Financial Incentives
3	Training and Development
4	Career Progression
5	Uniform Staff Perceptions
6	Response Officer Challenges
7	Case Submission Delays
8	Clothing Allowance
9	Public Interaction
10	Valuing Staff Appreciation

Topic	Dataset ID = 2
1	Office Environment Improvements
2	Staff Wellbeing
3	Development and Training
4	Communication and Collaboration
5	Health and Fitness
6	Company Culture
7	Office Storage and Organization
8	Legal and Compliance
9	Morale Boosting
10	Innovation and Creativity

Topic	Dataset ID = 3
1	Communication Channels
2	Productivity and Wellbeing
3	Training and Development
4	News and Updates
5	Community Engagement
6	Technology and Innovation
7	Work-Life Balance
8	Leadership and Management
9	Performance Management
10	Employee Wellbeing

Topic	Dataset ID = 4
1	Office Efficiency
2	Employee Welcome
3	Flexible Work Arrangements
4	Charity Events
5	Information Security
6	Team Bonding
7	Environmental Impact
8	Communication Improvements
9	Employee Benefits
10	Workplace Culture

Topic	Dataset ID = 5
1	Virtual Meetings
2	Mental Health Support
3	Remote Working
4	Communication
5	Socializing
6	Wellbeing
7	Support Networks
8	Professional Development
9	Technology
10	Fun Activities

Topic	Dataset ID = 6
1	Staffing and Resources
2	Training and Development
3	Equipment and Software
4	Collaboration and Communication
5	Mental Health and Wellbeing
6	Working Hours and Arrangements
7	Office Space and Facilities
8	Technical Support and Resources
9	Case Management Systems
10	Partnership and Collaboration Opportunities

Topic	Dataset ID = 7
1	Education
2	Travel
3	Career Development
4	Technology
5	Health and Wellness
6	Social Media Marketing
7	Inheritance Rights
8	Gender and Identity
9	Cultural Exchange
10	Social Impact

Topic	Dataset ID = 8
1	Comfort and Convenience
2	Versatility and Multipurpose Use
3	Sustainability and Environmental Benefits
4	Community and Social Impact
5	Innovation and Technology
6	Fashion and Style
7	Emergency Response and Military Use
8	Outdoor Activities and Recreation
9	Marketing and Sales Strategies
10	Ethical Considerations and Social Responsibility

Figure 15. Topic modelling results generated by Llama.

Topic	Dataset ID = 1
1	Financial Support and Incentives for Detectives
2	Recruitment and Retention Strategies
3	Training and Development
4	Streamlining Processes and Resources
5	Role Awareness and Understanding
6	Recognition and Reward
7	Workload Management
8	Professionalism and Image
9	Collaboration and Communication
10	Career Pathways and Progression

Topic	Dataset ID = 2
1	Business Efficiency and Productivity
2	Employee Well-being and Satisfaction
3	Brand Awareness and Promotion
4	Office Environment Improvements
5	Employee Development and Education
6	Customer Engagement and Referral Programs
7	Health and Safety Measures
8	Sustainable Practices
9	Networking and Communication
10	Corporate Social Responsibility

Topic	Dataset ID = 3
1	Internal Communication Channels and Tools
2	Information Dissemination and Sharing
3	Staff Engagement and Well-being Initiatives
4	Corporate Messaging and Strategy Communication
5	Operational Updates and Policy Changes
6	Productivity Improvement and Time Management
7	External Communication and Engagement
8	Training and Development Opportunities
9	Employee Feedback and Recognition
10	Continuous Improvement and Innovation

Topic	Dataset ID = 4
1	Workplace Well-being and Health Initiatives
2	Environmental Sustainability and Recycling
3	Office Efficiency and Process Improvement
4	Staff Engagement and Recognition
5	Flexibility and Work-life Balance
6	Communication and Information Sharing
7	Community Engagement and Corporate Social Responsibility
8	Training and Development Opportunities
9	Cost-saving Measures and Budget Allocation
10	Office Amenities and Facilities

Topic	Dataset ID = 5
1	Employee Well-being and Engagement Activities
2	Remote Work Tools and Practices
3	Recognition and Rewards
4	Fun and Social Activities
5	Training and Development Opportunities
6	Environmental and Health Initiatives
7	Communication and Collaboration
8	Team Building and Engagement
9	Creativity and Personal Development
10	Work-Life Balance and Flexibility

Topic	Dataset ID = 6
1	Law Enforcement Operations and Procedures
2	Workplace Environment and Well-being
3	Cybercrime and Digital Forensics
4	Career Development and Training
5	Data Management and Systems Improvement
6	Specialist Crime Investigation Support
7	Interdepartmental Collaboration
8	Employee Engagement and Recognition
9	Resource Management and Procurement:
10	Community Engagement and Crime Prevention

Topic	Dataset ID = 7
1	Customer Engagement and Retention Strategies
2	Business Operations and Costs
3	Target Customer Demographics and Characteristics
4	Marketing and Advertising Channels
5	Product and Service Offerings
6	Financial Management and Revenue Generation
7	Customer Content Consumption Habits
8	Customer Needs and Expectations
9	Market Research and Targeting
10	Business Partnerships and Collaborations

Topic	Dataset ID = 8
1	Military and Emergency Services Gear
2	Homelessness and Survival
3	Outdoor Activities and Camping
4	Fashion and Design
5	Environmental Impact and Sustainability
6	Market Opportunities and Sales Strategies
7	Accessibility and Ease of Implementation
8	Versatility and Multi-functionality
9	Community Engagement and Social Impact
10	Comfort and Safety

Figure 16. Topic modelling results generated by ChatGPT.

9. Generalisation of the Proposed Method Using New Data

The initial experiments conducted herein were with a private dataset provided by our industrial research collaborator. To demonstrate the generalisability of the proposed approach of extending topic model descriptors, the method is applied when using an unseen dataset.

The 20 Newsgroups dataset [55] is commonly used for text mining applications. It is a collection of approximately 20,000 news group documents, distributed (nearly) evenly across 20 different news groups. The documents within the 20 Newsgroups dataset exhibit a range of lengths and cover a spectrum of topics, from religion to politics to automotive and technological discussions, providing a comprehensive basis for testing the robustness of the proposed approach in a more heterogeneous textual environment, enhancing the external validity of this research. Table 5 shows how the 20 news groups are partitioned according to subject matter.

Table 5. 20 Newsgroup dataset categories.

comp.graphics comp.os.ms-windows.misc comp.sys.ibm.pc.hardware comp.sys.mac.hardware comp.windows.x	rec.autos rec.motorcycles rec.sport.baseball rec.sport.hockey	sci.crypt sci.electronics sci.med sci.space
misc.forsale	talk.politics.misc talk.politics.guns talk.politics.mideast	talk.religion.misc alt.atheism soc.religion.christian

Figures 17–19 report an excerpt of categories from the 20 Newsgroup dataset, where the extended outputs across each of the topics extracted when the LDA, BERTopic, and Top2Vec methods were used as the underlying topic modelling method, respectively. For LDA, the model was requested to distribute texts from each category into 1 of 10 topics. The BERTopic and Top2Vec were applied without any constraint on the number of topics to be generated.

talk.politics.mideast	talk.religion.misc	soc.religion.christian	rec.autos
Danny Keren	Brian Kendig	Existing Religious Newsgroups	Engines Eliot
Serdar Argic Armenians	Lord Jesus Christ	Inerrant Bible Community Views Homosexual Acts	High Speed Road Holding Ability
People Started Making Announcements	Long Judas Remained Hanging	Answering Questions	Kids Drop 20 Lbs Rocks
Jewish National Fund Bought	Secular Humanist Kent Sandvik	Good Life	Best Radar Detector Keywords
Osmanli Devleti	Malcolm Lee Steve Bittrolff	Illicitly Consecrated Chinese Bishops	Handy Automated Mailing List Package Named Listserv
Iraqi Death Toll Numbers	Physical Universe Larson Predicted	Boswell Defines Arsenokoitai	5 Year Ownership Costs
ADI Paid Roy Bullock	Objective Values Exist	Lord Jesus Christ	Oklahoma Law Center Callison
Onur Yalcin	LDS Church Claims Devine Authority	Catholic Church Walter Hooper Mass Bishop Bruskewitz	Honda Clutch Chatter Organization
Fact Israel Didnot Attack Jordan Till Jordan Attacked Israel	Homosexual Nazis	Reason Hell	Legal Car Buying Problems Organization
Muslim Bosnians Organization / Muslim Bosnians Lines	School Prayer Types	Holy Spirit Proceeds Principally / Holy Spirit Proceeds Jointly	Changing Brake Fluid
alt.atheism	**rec.motorcycles**	**talk.politics.misc**	**comp.graphics**
Atomic Model	Place Tom Coradeschi Tcora	Job Work Experience Combined	3D Computer Graphics Software
God Exists Typical Posting	Speedy Mercer	Russian Officer Resettlement Senior Administration Official	Mode Capabilities
Argumentum Ad Hominem Occurs	Org Austin Area Ride Mailing List Ride	Court Granted Atlantic Cement	24 Bits Viewer Organization
Case Western Reserve University Lines / Case Western Reserve University Nntp	Bmw Bikes	Government Politics	Jpeg Gif Converter Based
Marital Sex Helps Break	Harley Riders Seldom Wave	American Private Insurance Plans Cover Travel Expenses	DXF IFF Distribution
Religious People Chose Religion	Passenger Helmet Sizing Organization	Urgent Ted Frank Wanted	Characterizing Cubic Bezier Curves
Objective Values	Flow Air Cleaner Dark Candy Red	Clayton Cramer Writes	5 FPS Frame Rate
Supporting Rushdie	Bourbon Country	Wiretapping Initiative Organization	Fast Polygon Routine Needed Organization
Jon Livesey Writes	Shaft Drive Behavior	Police State USA Message	Gopher Client Point
Innocent German Civilians Killed	Generic Bike Riders	Secret American Bolshevik Naval Fleets	Format Supports 24 Bit Color Images

Figure 17. An excerpt of the extended topic outputs for the 20 Newsgroups dataset using LDA as the topic model.

Upon the reapplication of the proposed method to the 20 Newsgroups dataset, the extended descriptors illustrate coherent and distinctive topics that align with known categories within the dataset. For instance, when using all three models, it is reported that topics related to 'talk.politics.mideast' are accurately associated with region-specific political discussions (e.g., Jewish National Liberation Movement, Iraqi Death Toll Numbers), while categories like 'soc.religion.christian' and 'rec.autos' have keywords strongly related to religious discourse (e.g., Lord Jesus Christ, Jehovah Thy Redeemer) and automotive subjects (e.g., Changing Brake Fluid, BMW Motorcycles), respectively. These outputs also demonstrate a nuanced understanding of the dataset, capturing the finer subtleties and variations within the broader topics. For example, they distinguish between different facets of religious discussion, separating general Christian talk from more specific debates around biblical interpretation. Similarly, in the automotive category, a separation between general automotive discussions and more technical conversations about car maintenance and mechanical issues may be observed.

talk.politics.mideast	talk.religion.misc	soc.religion.christian	rec.autos
Armenians Demand Justice	Jehovah Thy Redeemer	God Loves / Knowing God	BMW Motorcycles
Typical Primitive Muslim Psychopathological Psychotic Behavior	Objective Moral System Exists	Presbyterian Church Split	Extremely Long Oil Change Intervals Claimed
Arab Liberation Army Attacks	Late Iranian Tradition Linking Zarathushtra	Christian Religion	80 Mph / Speed Collisions
Intrepid Israeli Soldiers	Health Insurance Includes Coverage	Marriage Commitment / Marriage Ceremony / Wedding Ceremony	Approaching Highway Speeds
West Bank Car Bomb Explosion Israel Defense Forces Radio	Biblical Knowledge Commentary	Particle Man	Diesel Emissions
Gaza Ghetto	LDS Church Claims Devine Authority	Eucharistic Prayer Blessed	Automatic Transmissions Allow Drivers
Newly Captured East Jerusalem	Historical Orthodox Christian Beliefs	Biblically Sound Basic Treatment	Hard Driven Car
Peace Talks Resume Today	Physical Universe Larson Predicted	Serious Struggle Going	93 Ford Probe GT Engine Problems
Israeli Solution Wouldnot Preserve Human Rights	Obsessed Christians Resorting / Interpret Biblical Scriptures / Christian Faith Stands / Lot Jesus Christ / Accepted Jesus Christ	Defending Infant Baptism / Fold Baptismal Formula / Baptismal Formulae Occurs / Catholics Baptize Babies	Trouble Shifting Gears Smoothly
Jewish National Liberation Movement	Holy Prophet Muhammad	Holy Spirit Proceeds Principally / Holy Spirit Proceeds Jointly	35000 Mile Difference Comparing
…	…	…	…
alt.atheism	**rec.motorcycles**	**talk.politics.misc**	**comp.graphics**
Human Behavior Mimics Animal Behavior	Helmet Manufacturers Provide Inspections Services	Limited Governments Versus Failed Governments News	Handmade Software Offers Free Jpeg Gif Conversion Tools
Believes God Exists	BMW Vintage Bulletin Tech Editor Dod	President Appointed Commerce Secretary Ron Brown	Vesa Local Bus Graphic Cards
Spectrum Theism Agnosticism Weak Atheism Strong Atheism	Idiot Response Dogs	Rodney King Trial	3Do Arm Qt Compact Video Lines
Knowingly Transgresses Islamic Teachings	Motorcycle Safety Foundation Riding Course	Health Insurance Company Offering Basic Care	Display 24 Bits Images
People Choose Religion	Ducati 400 Opinions Wanted	Batf Threw Concussion Grenades	Spinning Earth Organization
Contact American Atheist Veterans	Renounced Drunk Driving	Drugs Legalization Organization	Donate Fully Configured Vertigo 3D Graphics Software Worth
Biblical Contradictions Wanted Typical Posting	Bosch Air Horns Ordered	Second Fire Starts	Nok 895 Post Conference Tour
Fulfilling Earlier Jewish Prophecy	Gentler BMW Mailing List Reply	San Francisco Authorities Simultaneously Released Voluminous Documents	Adobe Photo Shop Type Software
Disciples Stole Jesus	CNN California MC Helmet Law Article Article	Atomic Energy Commision Chairman	Newsgroup Split Disclaimer / Newsgroup Split Organization / Proposed Newsgroup Split
Jesus Christ / Jesus Existed / Associate Christmas	Bikes Sold Long Distances	Times Staff Writer San Francisco Police	Recommend 3D Graphics Library
…	…	…	…

Figure 18. An excerpt of the extended topic outputs for the 20 Newsgroups dataset using BERTopic as the topic model.

talk.politics.mideast	talk.religion.misc	soc.religion.christian	rec.autos
Risk Small Arms Fire	Mormon Organization	Catholic Church Walter Hooper Mass Bishop Bruskewitz	Honda Accord Brake Problem Organization
Violate Federal Education Records Privacy Laws	Christian Bible Including Matthew 17	Real Interesting Question	Car 2 Years Ago / Car Ten Years Gk / Car Ten Years Ahead / Exotic Foreign Sports Cars
Reported Civilian Justifiable Homicides Involving Firearms	Christian Faith Stands	Greek Septuagint Versus Hebrew Scripture	Fleet MGR
Firearms Incendiary Ammunition	Objective Moral System Exists	Eternal Death Exactly	Better Buy / Consider Buying
NRA Gun Safety Course	Sensing Reality / Proved Conceptually Wrong / Dial Read 1 / Purely Philosophical Arguments	Roman Society Treated Male Sexuality	Shifting Accord Automatic Transmissions
Trade Center CDT	Christians Spouting Bible Verse	Christian Religion / Christian Theology	Ford Explorer Toyota 4Runner Nissan Pathfinder Currently
Modern Revolver Designs Incorporating Hammer Blocks	Historical Orthodox Christan Beliefs / Highly Christian Religious Order	Bible Teaches	Motor Oil Designation
Zionists Deny Gazans Equal Rights	Fake Objective Morality Exists / Objective Moral System Exists	Rich Picture Word Describing	Car Accidents
Constantly Spouting Baseless Lies	Lord God / Jehovah Elohim	Catholic Church Poland Organization	Insurance People Claim Car
Israel Rescued Jews Ranging / Jews Dismiss Palestinian Nationalism	Jewish Scripture	Believed God / Hope Good / Hoping Futilely	Average Luxury Car Dealer
…	…	…	…
alt.atheism	**rec.motorcycles**	**talk.politics.misc**	**comp.graphics**
Jews John Jesus	Started Riding Street Bikes	Profoundly Promiscuous Homosexual Men	Good Concave Convex Polygon Algorithm Organization
Spectrum Theism Agnosticism Weak Atheism Strong Atheism	Motorcycle Riders	Waco Survivors 1715 19 April Summary / Waco Survivors 1715 19 April Organization	Machines Support Vr
Presuming Rushdie Did Violate Islamic	Ryan Cousinetc 1982 Yamaha Vision Xz550	President Yeltsin Personally Assured President Clinton	Image Processing Institute University
Human Behavior Mimics Animal Behavior	Shaft Drive Bike	Standard Civil Rights Discussion Classes Based	Existing Image Processing Programs
Religious People Chose Religion	Motorcycle Helmet Organization	Limited Governments Versus Failed Governments News	Send Email Address
Hgod Exists	Bike Riding Dogs	Percent Tax / Deficit Spending / Economic Stimulus	Longer Worth Reading
Logical Argument Apart	Chain Drive	American Private Insurance Plans Cover Travel Expenses	Shareware Program Called Graphics Workshop / Pc Based Shareware Paint Programs
Clearly Defined Parties Wage War	35 Mph Speed Limit	Drugs Dea Wod Legalization Organization	256 Color Vga Rainbow Organization
Death Penalty Constituted Cruel Punishment	Late Lamented Rochester BMW Motorcycles	Bill Clinton Cares	Vesa Local Bus Graphic Cards
Absolutely Unalterably True	Passenger Seat	Republican Government	Format Conversion Reply / Pd Format Converters / Image Format Conversions
…	…	…	…

Figure 19. An excerpt of the extended topic outputs for the 20 Newsgroups dataset using Top2Vec as the topic model.

10. Limitations

The proposed approach utilises a safeguard against the scenario where a text document may not contain a keyword that accurately encapsulates its topic. As keyword extraction is employed on a collective set of texts associated with a topic rather than on individual documents, the likelihood of not finding a representative keyword is significantly reduced. The aggregation of texts for each topic increases the probability that relevant keywords will emerge from the collective context, rather than relying on a singular document to provide them. Moreover, the intersectionality between tokens generated by topic modelling methods and extracted keywords further refines the relevance as only those keywords

sharing common tokens with the topic modelling output are considered. This filtration ensures that the final topic descriptions are not only grounded in the statistically derived patterns of the topic model but are also contextualised by the actual language usage within the texts. However, despite the robustness of this method and its effectiveness in providing topic descriptions for all topics, there remains a possibility that the keywords may not always perfectly match the topic modelling output. In instances where a clear and representative keyword is not identified, or if the extracted keywords do not align well with the modeled topics, such topics can be labeled as 'miscellaneous'. This categorisation serves as an acknowledgment of the method's limitations and provides an opportunity for further exploration and analysis, possibly involving manual review or alternative methods, to better understand and describe these less straightforward topics.

In addition, while the proposed method of expanding topic descriptors in this paper has shown to be effective when using different underlying topic modelling approaches, as well as when it is applied on unseen data, the limitation of this study is the comparison of the proposed approach against relevant baselines discussed in Section 2. As noted in Section 2, other proposed methods are not generalised and are instead tailored to specific datasets and resources, supporting topic extraction from particular genres of text, such as legal documents. These methods also often aim to provide generalised terms for topics based on the specific nature and source of the texts, rather than offering a broad application across various domains. This specialisation limits the comparability of these methods with the more generalised approach proposed in this study. Thus, while a comparison with these methods would be informative, the distinct methodologies and domain-specific applications make direct comparisons complex and perhaps not entirely equitable.

11. Conclusions

This paper presents an approach towards extending the output of traditional topic modelling methods beyond a list of isolated tokens. The proposed approach removes the dependence on external sources by using the textual data itself by extracting high-scoring keywords and mapping them to the topic modelling method's token outputs. This approach, in turn, increases the interpretability of topic descriptions for human readers by allowing more context and information to be identified, which is an important factor to consider in use cases such as decision making and information retrieval.

To support the experiments presented herein, eight datasets containing short textual responses were used. Once each dataset was pre-processed, the LDA algorithm was applied and set to distribute each text response across one of ten topics. For each topic, the keyword extraction algorithm, RAKE, was applied to the original text responses to extract important keywords and phrases. The token outputs generated by the LDA were subsequently mapped to outputs extracted by RAKE. The keywords with the highest number of intersecting tokens with those produced by the LDA model, as well as those that achieved the highest score from RAKE, were assigned as the topic's main description.

Under the hypothesis that the extended topic modelling outputs are more interpretable than the LDA algorithm's output, such results were expected to be associated with high inter-annotator agreement across interpretability scores. Krippendorff's alpha coefficient was used to measure inter–annotator agreement according to which the extended outputs achieved higher agreement for quality ($\alpha = 0.522$) and usefulness ($\alpha = 0.575$), whereas the LDA outputs achieved lower agreement for quality ($\alpha = 0.075$) and usefulness ($\alpha = 0.090$). In terms of the efficiency of the annotation task, the extended outputs achieved higher agreement ($\alpha = 0.709$) in comparison to the LDA outputs ($\alpha = -0.012$). Subsequently, the analysis highlights that proposed approach increases the interpretability of the topic model results from a human perspective.

To investigate how the proposed method benchmarks against the performance of the state of the art, the approach was further evaluated by comparing its results with topics generated by two LLMs. The proposed method consistently yielded rich, contextually grounded topic descriptors that resonate with the thematic intricacies of the data, aligning

closely with, and at times surpassing, the outputs from the LLMs. In addition, to demonstrate and reinforce the generalisation of the proposed method, the approach was further evaluated using two other topic modelling methods as the underlying models, BERTopic and Top2Vec, and when using a heterogeneous unseen dataset.

12. Future Work

While the current method focuses on making topic descriptors more interpretable to human readers, it is crucial to determine whether these descriptors accurately reflect the underlying content of the topics. To assess this, as part of future work, a combination of qualitative and quantitative methods can be employed. Qualitatively, user studies can be conducted where domain experts and potential users evaluate the relevance and clarity of the topic descriptors. To elevate this approach beyond its current comparative standing with state-of-the-art LLMs, there is room to investigate the semantic layers of text. Statistical methods such as coherence measures, which assess the degree of semantic similarity within the topics, can be applied to evaluate the precision of the descriptors. Furthermore, computational experiments comparing the descriptors with unaltered topic outputs could provide quantitative insights into their accuracy and utility. Conducting this further work may offer valuable insights into optimising or fine-tuning the current methodology. This comparative analysis could highlight specific areas where the current approach may be adapted or improved to better suit different text genres or data sources. Additionally, such a study might also pave the way for developing a new, more robust method for expanding topic modeling descriptors. This potentially new method could incorporate the strengths of both generalised and specialised approaches, offering a more versatile solution that can be effectively applied across all domains and text types. By exploring these possibilities, we aim not only to refine our current approach but also to contribute to the broader field of topic modeling, enhancing its applicability and effectiveness in various contexts.

While the experiments herein demonstrated that the computational resources and processing times remained within acceptable bounds for datasets of up to 20,000 entries, the challenge of scaling to larger datasets remains. Real-world applications often involve rapidly expanding datasets that can scale to millions of entries, with the added expectation of real-time processing. Consequently, it is important for future research to focus on optimising the proposed method to better handle large-scale and dynamic datasets. This involves investigating more efficient data structures and algorithms and possibly leveraging advancements in parallel and distributed computing. The aim would be to reduce the computational footprint and improve the speed of processing without compromising the method's performance. Additionally, it would be advantageous to explore machine learning techniques that facilitate incremental learning, allowing the model to adapt and learn from new data as it becomes available. This can significantly enhance the method's applicability in real-time scenarios where data streams continuously and decisions need to be made promptly.

Author Contributions: Conceptualization, L.W.; methodology, L.W. and L.A.; validation, L.W. and E.A.; writing—original draft preparation, L.W. and E.A.; writing—review and editing, L.W., E.A. and P.B.; funding acquisition, P.B. All authors have read and agreed to the published version of the manuscript.

Funding: For the purpose of open access, the author has applied a CC BY public copyright licence (where permitted by UKRI, 'Open Government Licence' or 'CC BY-ND public copyright licence' may be stated instead) to any Author Accepted Manuscript version arising. This work was funded by the Economic and Social Research Council (ESRC), grant 'Discribe—Digital Security by Design (DSbD) Programme'. REF ES/V003666/1.

Data Availability Statement: The dataset provided by our industrial research collaborator cannot be made available. The scripting produced for the paper is available on Github (https://github.com/LowriWilliams/Topic_Modelling_Beyond_Tokens (20 April 2024)). The 20 Newsgroups dataset [55] used to further evaluate the approach presented in this paper is accessible using Python's Scikit-Learn library (https://scikit-learn.org/0.19/datasets/twenty_newsgroups.html (12 December 2023)).

Acknowledgments: The authors are grateful to Will Webberly and John Barker, our industry collaborators, from SimplyDo Ideas, Cardiff, for providing initial data to facilitate the experiments herein as well as providing informal feedback on the interpretability of the proposed approach towards extending topic model descriptors.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Bakshy, E.; Rosenn, I.; Marlow, C.; Adamic, L. The role of social networks in information diffusion. In Proceedings of the 21st International Conference on World Wide Web, Lyon, France, 16–20 April 2012; pp. 519–528.
2. Kang, H.J.; Kim, C.; Kang, K. Analysis of the trends in biochemical research using Latent Dirichlet Allocation (LDA). *Processes* **2019**, *7*, 379. [CrossRef]
3. Curiskis, S.A.; Drake, B.; Osborn, T.R.; Kennedy, P.J. An evaluation of document clustering and topic modelling in two online social networks: Twitter and Reddit. *Inf. Process. Manag.* **2020**, *57*, 102034. [CrossRef]
4. Chinnov, A.; Kerschke, P.; Meske, C.; Stieglitz, S.; Trautmann, H. An overview of topic discovery in Twitter communication through social media analytics. In Proceedings of the 21st Americas Conference on Information Systems (AMCIS), Fajardo, Puerto Rico, 13–15 August 2015; pp. 1–10.
5. Weng, J.; Lim, E.P.; Jiang, J.; He, Q. Twitterrank: Finding topic-sensitive influential twitterers. In Proceedings of the Third ACM International Conference on Web Search and Data Mining, New York, NY, USA, 3–6 February 2010; pp. 261–270.
6. Resnik, P.; Armstrong, W.; Claudino, L.; Nguyen, T.; Nguyen, V.A.; Boyd-Graber, J. Beyond LDA: Exploring supervised topic modeling for depression-related language in Twitter. In Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, Denver, CO, USA, 5 June 2015; pp. 99–107.
7. Spasic, I.; Button, K. Patient Triage by Topic Modeling of Referral Letters: Feasibility Study. *JMIR Med. Inform.* **2020**, *8*, e21252. [CrossRef] [PubMed]
8. Griffiths, T.L.; Steyvers, M. Finding scientific topics. *Proc. Natl. Acad. Sci. USA* **2004**, *101*, 5228–5235. [CrossRef] [PubMed]
9. Wright, L.; Burton, A.; McKinlay, A.; Steptoe, A.; Fancourt, D. Public Opinion about the UK Government during COVID-19 and Implications for Public Health: A Topic Modelling Analysis of Open-Ended Survey Response Data. *medRxiv* **2021**, *17*, e0264134.
10. Jacobi, C.; Van Atteveldt, W.; Welbers, K. Quantitative analysis of large amounts of journalistic texts using topic modelling. *Digit. J.* **2016**, *4*, 89–106. [CrossRef]
11. Greene, D.; O'Callaghan, D.; Cunningham, P. How many topics? stability analysis for topic models. In Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Ghent, Belgium, 14–18 September 2014; Springer: Berlin/Heidelberg, Germany, 2014; pp. 498–513.
12. Morstatter, F.; Liu, H. In search of coherence and consensus: Measuring the interpretability of statistical topics. *J. Mach. Learn. Res.* **2018**, *18*, 1–32.
13. Boyd-Graber, J.; Mimno, D.; Newman, D. Care and feeding of topic models: Problems, diagnostics, and improvements. In *Handbook of Mixed Membership Models and Their Applications*; Taylor & Francis Group Ltd.: Oxfordshire, UK, 2014.
14. Mimno, D.; Wallach, H.; Talley, E.; Leenders, M.; McCallum, A. Optimizing semantic coherence in topic models. In Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, Edinburgh, UK, 27–29 July 2011; pp. 262–272.
15. Chuang, J.; Gupta, S.; Manning, C.; Heer, J. Topic model diagnostics: Assessing domain relevance via topical alignment. In Proceedings of the International conference on machine learning, PMLR, Atlanta, GA, USA, 17–19 June 2013; pp. 612–620.
16. OpenAI. ChatGPT-3.5 Version. 2023. Available online: https://chat.openai.com/ (accessed on 16 April 2024).
17. Yu, J.; Egger, R. Tourist experiences at overcrowded attractions: A text analytics approach. In *Information and Communication Technologies in Tourism 2021*; Springer International Publishing: Berlin/Heidelberg, Germany, 2021; pp. 231–243.
18. Kumari, R.; Jeong, J.Y.; Lee, B.H.; Choi, K.N.; Choi, K. Topic modelling and social network analysis of publications and patents in humanoid robot technology. *J. Inf. Sci.* **2021**, *47*, 658–676. [CrossRef]
19. Mei, Q.; Shen, X.; Zhai, C. Automatic labeling of multinomial topic models. In Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Jose, CA, USA, 12–15 August 2007; pp. 490–499.
20. Hindle, A.; Ernst, N.A.; Godfrey, M.W.; Mylopoulos, J. Automated topic naming. *Empir. Softw. Eng.* **2013**, *18*, 1125–1155. [CrossRef]
21. Lee, T.Y.; Smith, A.; Seppi, K.; Elmqvist, N.; Boyd-Graber, J.; Findlater, L. The human touch: How non-expert users perceive, interpret, and fix topic models. *Int. J. Hum.-Comput. Stud.* **2017**, *105*, 28–42. [CrossRef]
22. Lau, J.H.; Grieser, K.; Newman, D.; Baldwin, T. Automatic labelling of topic models. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Stroudsburg, PA, USA, 21 June 2011; pp. 1536–1545.

23. Magatti, D.; Calegari, S.; Ciucci, D.; Stella, F. Automatic labeling of topics. In Proceedings of the 2009 Ninth International Conference on Intelligent Systems Design and Applications, Pisa, Italy, 2 December 2009; IEEE: Piscataway, NJ, USA, 2009; pp. 1227–1232.
24. Bhatia, S.; Lau, J.H.; Baldwin, T. Automatic Labelling of Topics with Neural Embeddings. In Proceedings of the COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, Osaka, Japan, 11–16 December 2016; pp. 953–963.
25. Basave, A.E.C.; He, Y.; Xu, R. Automatic labelling of topic models learned from twitter by summarisation. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Baltimore, MD, USA, 22–27 June 2014; pp. 618–624.
26. Aletras, N.; Stevenson, M. Labelling topics using unsupervised graph-based methods. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Baltimore, MD, USA, 22–27 June 2014; pp. 631–636.
27. Hulpus, I.; Hayes, C.; Karnstedt, M.; Greene, D. Unsupervised graph-based topic labelling using dbpedia. In Proceedings of the sixth ACM International Conference on Web Search and Data Mining, Rome, Italy, 4–8 February 2013; pp. 465–474.
28. Allahyari, M.; Pouriyeh, S.; Kochut, K.; Arabnia, H.R. A knowledge-based topic modeling approach for automatic topic labeling. *Int. J. Adv. Comput. Sci. Appl.* **2017**, *8*, 335. [CrossRef]
29. Kinariwala, S.A.; Deshmukh, S. Onto_TML: Auto-labeling of topic models. *J. Integr. Sci. Technol.* **2021**, *9*, 85–91.
30. Béchara, H.; Herzog, A.; Jankin, S.; John, P. Transfer learning for topic labeling: Analysis of the UK House of Commons speeches 1935–2014. *Res. Politics* **2021**, *8*, 20531680211022206. [CrossRef]
31. Wan, X.; Wang, T. Automatic labeling of topic models using text summaries. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Berlin, Germany, 7–12 August 2016; pp. 2297–2305.
32. Aletras, N.; Baldwin, T.; Lau, J.H.; Stevenson, M. Evaluating topic representations for exploring document collections. *J. Assoc. Inf. Sci. Technol.* **2017**, *68*, 154–167. [CrossRef]
33. Kou, W.; Li, F.; Baldwin, T. Automatic labelling of topic models using word vectors and letter trigram vectors. In Proceedings of the AIRS, Brisbane, Australia, 2–4 December 2015; Springer: Berlin/Heidelberg, Germany, 2015; pp. 253–264.
34. Chang, J.; Gerrish, S.; Wang, C.; Boyd-Graber, J.L.; Blei, D.M. Reading tea leaves: How humans interpret topic models. In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 7–10 December 2009; pp. 288–296.
35. Hofmann, T. Probabilistic latent semantic indexing. In Proceedings of the 22nd annual international ACM SIGIR Conference on Research and Development in Information Retrieval, Berkeley, CA, USA, 15–19 August 1999; pp. 50–57.
36. Dumais, S.T. Latent semantic analysis. *Annu. Rev. Inf. Sci. Technol.* **2004**, *38*, 188–230. [CrossRef]
37. Blei, D.M.; Ng, A.Y.; Jordan, M.I. Latent dirichlet allocation. *J. Mach. Learn. Res.* **2003**, *3*, 993–1022.
38. Grootendorst, M. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv* **2022**, arXiv:2203.05794.
39. Angelov, D. Top2vec: Distributed representations of topics. *arXiv* **2020**, arXiv:2008.09470.
40. Blei, D.; Carin, L.; Dunson, D. Probabilistic topic models. *IEEE Signal Process. Mag.* **2010**, *27*, 55–65. [CrossRef]
41. Hendry, D.; Darari, F.; Nurfadillah, R.; Khanna, G.; Sun, M.; Condylis, P.C.; Taufik, N. Topic modeling for customer service chats. In Proceedings of the 2021 International Conference on Advanced Computer Science and Information Systems (ICACSIS), Virtual, 23–26 October 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 1–6.
42. Egger, R.; Yu, J. A topic modeling comparison between lda, nmf, top2vec, and bertopic to demystify twitter posts. *Front. Sociol.* **2022**, *7*, 886498. [CrossRef] [PubMed]
43. Scikit-Learn. 0.24.1 Linear Discriminant Analysis. Available online: https://scikit-learn.org/stable/modules/generated/sklearn.\discriminantanalysis.LinearDiscriminantAnalysis.html (accessed on 5 May 2023).
44. Rose, S.; Engel, D.; Cramer, N.; Cowley, W. Automatic keyword extraction from individual documents. *Text Mining Appl. Theory* **2010**, *1*, 1–20.
45. Řehůřek, R. Gensim: Topic Modelling for Humans. Available online: https://radimrehurek.com/gensim_3.8.3/summarization/keywords.html (accessed on 3 April 2024).
46. Mihalcea, R.; Tarau, P. Textrank: Bringing order into text. In Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, Barcelona, Spain, 25–26 July 2004; pp. 404–411.
47. Campos, R.; Mangaravite, V.; Pasquali, A.; Jorge, A.; Nunes, C.; Jatowt, A. YAKE! Keyword extraction from single documents using multiple local features. *Inf. Sci.* **2020**, *509*, 257–289. [CrossRef]
48. Ramanujam, N.; Kaliappan, M. An automatic multidocument text summarization approach based on Naive Bayesian classifier using timestamp strategy. *Sci. World J.* **2016**, *2016*, 1784827. [CrossRef]
49. Davare, S.; Sindwani, D.R.; Castelino, P.; George, A. Text Mining Scientific Data to Extract Relevant Documents and Auto-Summarization. *IJSTE-Int. J. Sci. Technol. Eng.* **2017**, *4*, 109–114.
50. Tarasov, A.; Delany, S.J.; Cullen, C. Using crowdsourcing for labelling emotional speech assets. In Proceedings of the W3C Workshop on Emotion ML, Paris, France, 1–5 October 2010.
51. Passonneau, R.J.; Yano, T.; Lippincott, T.; Klavans, J. Relation between agreement measures on human labeling and machine learning performance: Results from an art history image indexing domain. In *Computational Linguistics for Metadata Building*; European Language Resources Association (ELRA): Paris, France, 2008; p. 49.

52. Snow, R.; O'connor, B.; Jurafsky, D.; Ng, A.Y. Cheap and fast–but is it good? Evaluating non-expert annotations for natural language tasks. In Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, Honolulu, HI, USA, 25–27 October 2008; pp. 254–263.
53. Krippendorff, K. *Content Analysis: An Introduction to Its Methodology*; Sage Publications: Sauzendoaks, CA, USA, 2018.
54. Meta. Llama-2 Version. 2023. Available online: https://llama.meta.com/ (accessed on 16 April 2024).
55. Lang, K. Newsgroups Data Set. Available online: https://scikit-learn.org/0.19/datasets/twenty_newsgroups.html (accessed on 3 November 2023).