ALT-TECH TOPIC MODELING INCORPORATING A TOPIC MODEL SELECTION STRATEGY

Sanaz Rasti[1] and Sarah Anne Dunne[2] and Eugenia Siapera[2]

[1]School of Computer Science, University College Dublin, Dublin, Ireland
[2]School of Information and Communication Studies, University College Dublin, Dublin, Ireland

ABSTRACT

The rapid growth of Alt-tech platforms and concerns over their less stringent content moderation policies, make them a good case for opinion mining. This study aims at investigating the topic models that exist in specific Alt-tech channel on Telegram, using data collected in two time points of 2021 and 2023. Three different topic models of LDA, NMF and Contextualized NTM were explored and a model selection procedure was proposed to choose the best performing model among all. To validate the model selection algorithm quantitatively and qualitatively, the approach was tested on publicly available labelled datasets. For all the experiments, data was pre-processed employing an effective NLP pre-processing procedure along with an Alt-tech customised list of stop-words. Using the validated topic model selection algorithm, LDA topics with Ngram range = (4, 4) were extracted from the targeted Alt-tech dataset. The findings from topic models were qualitatively evaluated by a social scientist and are further discussed. The conclusion of the work suggests that the proposed model selection procedure is effective for corresponding corpus length and context. Future work avenues are suggested to improve the outcome of Alt-tech topic modeling.

KEYWORDS

Topic Modeling, Topic Model Selection, LDA, NMF, Contextualized NTM, Alt-tech

1. INTRODUCTION

Topic modeling refers to a family of text mining tools for analysing text corpora. It offers a high-level summary of document collections, providing an insight into their overall structure and subject matter. When it comes to the unsupervised nature of topic modeling, quantitative and qualitative analysis of topics offer model reliability and stability, topic coherency, semantic meaningfulness, topic model latent variable discrepancy, etc.

Social media is a collective term encompassing online websites and applications focusing on communication, content sharing, online social interaction and collaboration. It has great value for data mining as an up-to-date resource of public opinion. Social media platforms apply different content moderation regimes and some of the platforms are more tolerant of extreme speech than others. In 2016, the European Commission sought to harmonise approaches, a voluntary agreement was established with social media platforms to monitor and control illegal hate speech and misinformation through the Code of Conduct and Code of Practice. While all the main social media platforms signed the Code, there are still platforms that do not adhere to it and maintain indifferent regulation practices. At the same time, the application of the Code in the European

David C. Wyld et al. (Eds): MLNLP, NWCOM, DTMN, ASOFT, SIGPRO, AIFZ, CSITY, CLSB - 2023
pp. 01-13, 2023. CS & IT - CSCP 2023 DOI: 10.5121/csit.2023.131601

context led to the deplatforming of accounts and pages deemed harmful by mainstream social media platforms. Deplatforming has thus led to the rise of alternative social media platforms (Alt-tech),which generally operate with less stringent content moderation policies.

Hence, the enactment of more stringent content moderation policies has not eradicated misinformation from the internet; rather these policies have displaced it[1]. Conspiracy theorists and far-right activists migrate onto Alt-tech platforms, transitioning into spaces where misinformation and hate speech can coalesce. Alt-tech platforms offer content creators and their followers a place to regroup and strengthen their collective identity in the event of deplatforming. As regulation and moderation is limited on these platforms, disinformation continues unhindered which facilitates the radicalization of users who visit these websites, exposing them to an unprecedented volume of hate speech and misinformation.

Opinion mining is a Natural Language Processing toolkit offering emotional tone explanation of Alt-tech textcorpus. Topic modeling is a popular technique for analysing Alt-tech data[2]. A recent study conducted by Curley et.al. [6], presented a promising result for analyzing far-right terms and discourses in Alt-tech Telegram channels using topic modeling.

In this paper, a topic model selection methodology was developed and proposed for producing and validating topic models. In previous work [3] robustness of the proposed methodology was extensively evaluated using qualitative rating of the best and worst performing chosen models on several unlabelled dataset and presented as a poster [3].In this work we present two systematic quantitative measures for evaluating the proposed methodology, backed by qualitative rating on labelled dataset. The proposed topic model selection algorithm measures the similarity between topics and their corresponding documents. Our main corpus of interest is a corpus of posts collected from the account of an influencer on Telegram.

The main contributions of this work are listed as following:

- Alt-tech data was collected from a single Telegram channel.
- The data was pre-processed
- Three different topic models of LDA, NMF and NTM were implemented
- DTS topic model selection algorithm was proposed and tested on labelled dataset
- DTS topic model selection was evaluated quantitatively
- DTS topic model selection was evaluated qualitatively
- DTS topic model selection was employed on Alt-tech corpus
- Qualitative analysis of topics extracted from Alt-tech corpus by a social scientist

The remainder of this paper is structured as follows. Section 2 covers details of the dataset. The methodology and evaluation procedure is explained in Section 3; covering data pre-processing in subsection 3.1, topic modeling and evaluation procedure explained in subsections 3.2 and 3.3, respectively. The experiments and results are presented in Section 4.The paper closes by conclusions and proposes future research avenues in Section 5.

2. DATASET

The performance of our proposed topic model selection algorithm is investigated employing labelled datasets. Details of these dataset are presented in subsection 2.1. The targeted Alt-tech corpus is an unlabelled dataset, which is detailed in subsection 2.2.

2.1. Labelled Datasets

Amazon Reviews corpus is a well-known and popular corpus for data analysis purposes. Our study benefits from the labelled Amazon Cell Phone Reviews dataset which was originally collected for the problem of product rating prediction by [4], from the official Amazon website (Amazon.com) in the cell phone and accessories category. This dataset with two specific label ratings of 0 and 1 respectively for the negative and positive reviews was employed in this study. The corpus includes 973 documents and the count of tokens in the original corpus is 10,246 terms.

Another two labelled dataset which were experimented in this study are Yelp Reviews and Computer Science Abstracts corpuses, with the latter collected during this project by the corresponding author and can be found in our GitHub repository[5]. Details of the labelled datasets are listed in Table 1.

2.2. Unlabelled Dataset

Computing Forever is one of the most successful Irish Alt-tech channels, which primarily posts content relating to technology and social commentaries; discussions about politics, future technology avenues, gaming and film reviews and vlogs, and criticisms of social justice, political correctness and hyper-consumerism. The content creator of this channel claims to have a critical eye and a sceptical mind. The number of subscribers, activity and mentions of this channel grew from the year 2021 to 2023, possibly owing to the Covid-sceptic content which has became a focus for the content creator and his followers. Currently over 10,000 and 113,700 subscribers follow the channel on Telegram and BitChutevlog, respectively. The data was collected at two time points (year 2021 and 2023); details of which can be found in Table 1 by Computing Forever (I) and (II) annotations, respectively.

3. METHODS AND EVALUATION

Every successful NLP project requires data pre-processing and our work is not exempt. The procedure of data pre-processing which was employed in this work, is detailed in subsection 3.1, subsection 3.2 goes over three methods of topic modeling explored for Alt-tech data and our proposed model selection algorithm is presented in subsections 3.3. The approach for quantitatively and qualitatively evaluating the model selection algorithm is explained in subsections 3.4 and 3.5, respectively.

3.1. Data Pre-processing

Unicode is a standard coding system which assigns characters of the language with a unique integer code between 0 and 0x10FFFF. Text content of original corpus was converted into lower case and subsequently Unicode format; this allows adopting a standard encoding for language processing.

To help treating each text equally, punctuation characters, emoji patterns and links were removed from the corpus. A list of English stop-words was appended with our customized list of stop-words and the updated material were removed from the remainder of the corpus to adopt more focus on important information. Finally, the empty document strings were removed.

The pre-processed data was stored and herein is referenced as stripped text data. The count of corpus tokens for each channel before and after pre-processing can be found in Table 1. On average, data pre-processing reduced the count of corpus' tokens by 56.4%.

Table 1. Datasets

	Dataset	Length of Corpus	count of tokens: original corpus	count of tokens cleaned corpus
Labelled Dataset	Amazon Reviews	973	10,246	3,688
	Yelp Reviews	976	10,894	3,859
	Computer Science Abstracts	100	19,217	9,960
Unlabelled Dataset	Computing Forever I	994	31,743	14,995
	Computing Forever II	11,273	481,884	229,556

3.2. Topic Modeling

Topic modeling is an unsupervised machine learning technique for text analysis, which potentially extracts abstract topics from a collection of documents and helps to rapidly identify patterns in the corpus. This work includes two popular conventional topic modeling approaches of (a) Latent Dirichlet Allocation (LDA) and (b) Non-Negative Matrix Factorization (NMF), and state of the art Neural Topic Modeling with contextualization:(c) Contextualized Neural Topic Modeling (NTM).

LDA is an unsupervised clustering technique which presents the documents as a collection of word-topics and topics as a collection of words. LDA is considered a "bag-of-words" model, in which the order of words does not matter, and the documents are generated word-by-word by choosing a topic mixture.

Non-NegativeMatrix factorization is a statistical method employed in Natural Language Processing (NLP) to approximate a breakdown of the original corpus' term-document matrix into a word-topic and topic-document matrices. This automatic extraction of sparse matrices is easily interpretable in text clustering allocations and provides a semantically coherent set of topics [6].

Neural Topic Models (NTMs) emerged while presenting a better topic coherence compared to traditional topic models. Further studies searched for a more coherent topic model while extending state of the art Neural Topic Model ProdLDA (Product of Experts LDA [7]) with the contextualized representation of corpus[8]. The authors presented a significant improvement in topic model coherence while employing contextualized Neural Topic model on four corpuses of Wiki-20K, Stack-Overflow and Google-News[9], and Tweets. However, the method failed to provide better coherence on 20-News-Group dataset compared to MetaLDA topic modeling[10] on the same corpus.

In this work, two Contextualized-NTM models of I) CombinedTM[8] including two main components of ProdLDA and SBERT embedding representations [11], and II) ZeroShotTM[12], were employed. For evaluation, the performance of aforementioned topic models were tested using publicly available labelled datasets and an Alt-tech unlabelled dataset (see Table 1).

3.3. Topic Model Selection

A study conducted by [13], suggested that investigating the correlation between topic model coherence and classifier accuracy leads to a promising topic model selection procedure. Selecting a topic model which has the greatest document topic coherence is the outcome of topic model quantitative analysis. In a study conducted by [14], the authors introduced TopDocs as documents referring to a topic with highest probability. Targeted coherence score was calculated from vectorized selected documents and their related topics (TopDocs).

Algorithm 1 Model Selection

```
procedure ModelSelection(CleanedCorpus,MODELS)
    nlp_pipeline = spacy.load('en_core_web_sm')
    DTS = [ ]
    Consider model_ij as the model with Ngram range = (i, j):
    for model_ij in MODELS do
        sim = 0
        Produce the TOPICS from CleanedCorpus
        for each topic in TOPICS do
            Produce doc_topic employing nlp_pipeline
            Produce N-TopDocs (N-top related documents to topic)
            for each document of N-Docs do
                Produce doc_document employing nlp_pipeline
                similarity = doc_topic.similarity(doc_document)
                sim = sim + similarity
            end for
        end for
        Append DTS with average(sim) for model_ij
    end for
    Find model_ij where DTS has it's maximum
    Return model_ij as the best performing model
end procedure
```

Algorithm 1. Pseudocode of proposed model selection algorithm.

Our study developed a topic model selection algorithm based on Document Topic Similarity (DTS) of TopDocs, in order to choose the right topic model which yields the best performance. The topic Ngram range defines how many tokens\ words per term of a topic are allowed to be extracted. It is beneficial to consider different Ngram ranges while analyzing social media text data as the online conversations are often unstructured or in the form of short sentences and not consistently adhering to grammatical rules. To cover this inconsistency, LDA and NMF topic models were produced for various Ngram ranges.

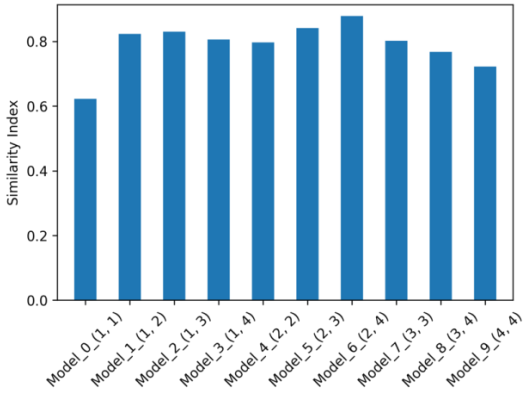

Figure 1. DTS for NMF topic modeling of Amazon Reviews corpus with n = 4 topics.
The best performing model is Model_6 with Ngram ranges = (2, 4).

For each produced topic model with specific Ngram = (i , j), (*model_ij*), five top relevant documents to topics were produced (d = 5 TopDocs) and the DTS were measured using spaCy pipeline [15]. The similarity index was averaged over all different Ngram ranges of topic models and *model_ij* with maximum similarity index was selected for topic modeling of unlabelled dataset, see Algorithm 1 for pseudocode of model selection scheme. For contextualized NTM topic models, DTS was calculated and averaged over the experimented Combined TM and ZeroShotTM models.

In order to quantitatively validate the performance of proposed topic model selection algorithm, the following procedure and metrics are calculated:

1. Calculating topic model's rand-score for labelled-datasets.
2. DTS score between the topics and their five-top corresponding documents.
3. Plotting DTS-RI for various models of LDA, NMF and Contextualized-NTM.
4. Calculating Spearman correlation Coefficient (SpC) for the plotted DTS-RI.
5. Choosing the best performing model.
6. Asserting if the chosen model is the same as model selected via our proposed model selection algorithm.

Employing LDA and NMF model, topics with Ngram ranges of (1, 1), (1, 2), (1, 3), (1, 4), (2, 2), (2, 3), (2, 4), (3, 3), (3, 4), (4, 4) were produced from Amazon Reviews dataset. Three different number of topics were tested as follows: n = 2, n = 4 and n = 8. For each topic t = 10 and d = 5 number of terms and TopDocs were respectively extracted.

Table 2. Result of topic modeling DTS, RI and Spearman Correlation between DTS-RI, on Amazon Reviews Corpus for targeted n=2topics.

Model	N-gram Ranges	Average DTS	Average RI	SpC
LDA	(1, 1)	0.7	0.47	0.4
	(1, 2)	0.7	0.47	
	(1, 3)	0.68	0.64	
	(1, 4)	0.71	0.47	
	(2, 2)	0.75	0.47	
	(2, 3)	0.74	0.53	
	(2, 4)	0.67	0.47	
	(3, 3)	0.7	0.47	
	(3, 4)	0.66	0.44	
	(4, 4)	0.66	0.53	
NMF	(1, 1)	0.62	0.44	
	(1, 2)	0.76	0.47	
	(1, 3)	0.83	0.47	
	(1, 4)	0.76	0.47	
	(2, 2)	0.72	0.53	
	(2, 3)	0.85	0.64	
	(2, 4)	0.88	0.64	
	(3, 3)	0.76	0.64	
	(3, 4)	0.78	0.53	
	(4, 4)	0.73	0.53	
NTM	CombinedTM	0.58	0.53	
	ZeroShotTM	0.54	0.47	

An example of DTS plotted figure for different Ngram ranges of a model can be found in Figure 1. The figure shows a sample of DTS results, for extracted NMF topic models on Amazon Reviews corpus, with n = 4 topics. The stats depict that among different NMF models employed on this corpus, Model_6 with Ngram range = (2, 4) yielded the best DTS with similarity index = 0.88. Hence for n = 4 number of topics, Model_6 has the potential to be selected for further investigation of the corpus.

Table 2 presents the result of average DTS, average RI and SpC for Amazon Reviews labelled corpus and n = 2 number of topics. The aforementioned setting yields the best result of DTS and RI while employing NMF models.

DTS-RI is plotted in Figure 2 for n=2 number of topics indicating LDA models with [L0, L1, …, L9], NMF models with [N0, N1, …, N9] and the two NTM models. The Spearman correlation applied on DTS-RI sets of data achieved a positive correlation coefficient of SpC = 0.40.

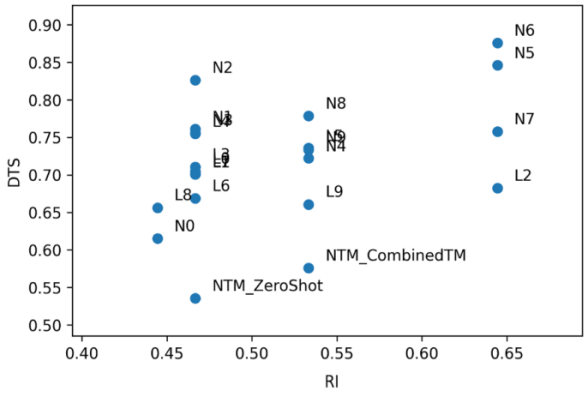

Figure 2. Plotted DTS-RI for various topic models on Amazon Reviews corpus with n = 2 number of topics.

The result of averaged DTS and RI as well as achieved SpC between DTS-RI for n = 2, 4 and 8 number of topics which was implemented on labelled Amazon Reviews corpus, is presented in Table 3.

Table 3. Average DTS, RI and SpC for n = 2, 4 and 8 topics on Amazon Reviews corpus.

Corpus	Number of Topics	Model	Average DTS	Average RI	SpC
Amazon	2	LDA	0.70	0.5	0.4
		NMF	0.77	0.54	
		NTM	0.56	0.5	
	4	LDA	0.68	0.49	0.266
		NMF	0.79	0.5	
		NTM	0.55	0.5	
	8	LDA	0.68	0.5	0.082
		NMF	0.78	0.51	
		NTM	0.59	0.52	

The quantitative evaluation (steps 1 to 6) was also tested on Yelp Reviews and Computer Science Abstract corpuses. However, the results showed a poor connection between average DTS, average RI and the greatest SpC which was achieved. The reason behind this mis-alignment for Yelp reviews corpus might be diverse topics exist in Yelp Reviews context, whereas for Computer Science Abstract corpus a possible justification could be the very short length of corpus (100 documents) with long abstract documents. The detailed experiments results using these corpuses can be found in our GitHub repository Appendix[5].

3.4. Qualitative Topic Model Analysis

NLP is intrinsically an ambiguous task and is often interpreted subjectively; which makes the evaluation of text processing a challenging task. A valid approach toward topic model analysis is Qualitative Analysis. In this work, qualitative evaluation of topics was conducted by rating (QR) the TopDocs which were assigned in quantitative analysis.

Three discrete score ratings were employed for QR. These scores are -1, 0 and 1 respectively for (a) non-topic and/or vague document, (b) irrelevant pair of TopDoc, and (c) relevant pair of TopDoc.

Table 4. Qualitative analysis of selected topic model on Amazon Reviews corpus with n=2 number of topics.The selected model via DTS model selection algorithm is NMF Topic model with Ngram range = (2, 4).

Topic number	Topic Terms	TopDocs	DTS	QR
0	['waste money', 'please waste money', 'please waste', 'terrible product waste money', 'product waste money', 'product waste', 'terrible product', 'terrible product waste', 'customer service', 'highly recommend']	Doc_9: What a waste of money and time!.	0.87	1
		Doc_837: I don't like this Nokia either.	0.87	1
		Doc_356: The only thing that I think could improve is the sound leaks out from the headset.	0.87	1
		Doc_93: The ear buds only play music in one ear.	0.87	1
		Doc_394: I highly recommend these and encourage people to give them a try.	0.87	-1
1	['battery life', 'motorola razor', 'love battery life', 'motorola razor vi', 'range battery life', 'range battery', 'love battery', 'phone battery life', 'phone battery', 'phone battery life operates']	Doc_362: It is light, easy to use, and has very clear reception and transmission.	0.91	1
		Doc_759: Its the best headset I have used.	0.88	1
		Doc_564: Yet Plantronincs continues to use the same flawed charger design.	0.83	1
		Doc_368: Love This Phone.	0.87	1
		Doc_260: Good product - incredible value.	0.91	1
			Average QR	**0.8**

To validate topic model selection procedure and justify the performance of quantitative analysis, the chosen topic model on labelled Amazon Reviews corpus with n = 2 number of topics were qualitatively evaluated and rated according to our proposed QR values, see Table 4. The model achieved average QR = 0.8; presenting that topics clustered correctly for 9 out of 10 documents.

Although identifying a topic model with coherent TopDocs is a challenging task, the model selection algorithm was proven successful quantitatively and qualitatively to overcome the task on Amazon Cell Phone Reviews corpus.

4. EXPERIMENTS AND RESULTS

Topic model selection algorithm was evaluated and approved effective on Amazon Reviews labelled dataset in section 3.In this part of study, the targeted unlabelled Alt-tech corpus which was collected from Telegram channel is investigated. The corpus includes the Computing Forever channel of data which was crawled with two time domains in 2021 and 2023 and herein are referenced as Computing Forever (I) and (II), respectively. Topic models were produced with n = 10 number of topics and 10 terms followed by a qualitative evaluation which was performed by a social scientist. Table 5 includes average DTS and QR for n=10 topics of the Computing Forever channel.

Table 5. Average DTS, QR for n = 10 topics on Computing Forever (I) and (II).

	Computing Forever (I)			Computing Forever (II)		
Topic Number	**General Topic Terms**	**Average DTS**	**Average QR**	**General Topic Terms**	**Average DTS**	**Average QR**
0	populist conspiracies, globalism, transhumanism	0.89	1	LGBTQ, health concerns	0.94	0.8
1	vaccines, vaccine passports, lockdown	0.88	0.8	health, EMF devices, surveillance and controls	0.93	1
2	vaccine, lockdowns	0.78	0.8	health, pharmaceutical industry, vaccines transhuman	0.94	0.6
3	transhumanism, media dominance	0.74	0.2	health, covid	0.81	0.2
4	Climate change, transhumanism	0.85	0.8	globalisation, food supply, Irish water	0.89	0.6
5	risks of vaccine	0.84	0.2	globalisation, concerns over protests	0.86	0.2
6	masks and vaccination	0.84	0.8	vaccine	0.81	0.4
7	vaccine, covid	0.85	1	PCR, covid, vaccine	0.91	0.4
8	online harms	0.85	0.8	not a clear topic	0.73	-1
9	conspiracy theory	0.88	0.8	health	0.97	0.4

The following subsections of 4.1 and 4.2 present comprehensive result for Computing Forever (I) and (II), respectively.

4.1. Computing Forever (I)

DTS model selection on Computing Forever (I) corpus, proposed LDA model with Ngram range = (4, 4) as the best performing model. The qualitative rating of d = 5 TopDocs presented excellent and good coherence between topics and TopDocs respectively for Topic_0, and Topic_7 with QR=1 and Topic_1, Topic_2, Topic_4, Topic_6, Topic_8 and Topic_9 with QR= 0.8.

The key terms in these topics cover populist conspiracies, globalism, transhumanism, the Covid-19 pandemic, particularly relating to vaccines and lockdown restrictions. The corresponding documents to these topics (TopDocs) contain some of the following contexts:

- Various concerns and conspiracies related to the vaccine
- Anti-lockdown discussions which often paint the restrictions as unnecessary forms of control
- Conspiracies relating to the vaccine and vaccine passport as producing new forms of “human augmentation” permitting surveillance and controls of the population
- Beliefs that Covid and its variants are a manufactured means of enforcing further controls over the population by a frequently unnamed societal elite
- Information about euthanization and a new "suicide capsule"
- Unease with the vaccine passports and the “medical apartheid” and “medical tyranny” they would impart on the masses.

- Using vaccines to inject people with 5G
- Masks becoming a norm and how society has been "socially reengieered"

On the other hand, the TopDocs assigned to Topic_3 and Topic_5 are poorly coherent in terms of context and offer no clear linkages. For these topics, the terms include transhumanism, media dominance, Ireland energy crisis, finance airports and climate. The TopDocs context is circulating around the following:

- Social compliance and brainwashing
- Cashless society
- Discusses propaganda and people who are not critical of liberalism
- Popular show on TV
- Vaccine
- Covid certification

4.2. Computing Forever (II)

DTS model selection on Computing Forever (II) corpus, proposed LDA model with Ngram range = (4, 4) as the best performing model. The qualitative rating of d = 5 TopDocs presented excellent and good coherence between topics and TopDocs respectively for Topic_1, with QR = 1 and Topic_0 with QR = 0.8.

The key terms in these topics coverhealth, EMF devices, surveillance, controls and LGBTQ. The corresponding documents to these topics (TopDocs) contain coherent context with a few are listed in the following:

- Protection from Electronic and Magnetic Fields (EMFs) with a neutraliser
- Trillionaires/elites controlling the world
- Surveillance/spying on citizens
- Link to a new "virus" being used to control people
- Inclusion of non-useful materials in schools, asks for holistic healing
- LGBTQ+ "social credit score" coming into effect
- Issue with transgender people
- Health food related - more an advert/promotion
- Pharmaceutical products identified as poisonous

TopDocs assigned to Topic_3 and Topic_5 are poorly coherent in terms of context. For these topics, the topic terms are about health, globalisation and concerns over protests. The context of irrelevant TopDocs for these topics includes the following terms:

- Swiss neutrality and financial capitalism
- Fianna Fail boosting votes via false harassment claims
- Concerns over China-Russia meetings
- Airplane mode allows for better sleep
- Link to nature apothecary
- Spam messages

This topic model produced a non-topic (Topic_8) with unclear terms. The TopDocs assigned to this topic are not clear either. This resulted in average QR = -1 for this topic and a negative effect on overall average QR.

The prominent difference between the topic model theme extracted for Computing Forever (I) and Computing Forever (II) is related to Covid-19 topics. For Computing Forever (I) a majority of the topics were related explicitly to the pandemic, with notable emphasis on lockdown restrictions, the vaccine and populist Far-Right conspiracies. In Computing Forever (II) corpus, Covid-19 topics are present, though less explicitly. Concerns with health and well-being are expressed as a reflection on Covid-19 concerns, often vilifying Big Pharma companies and advocating for more holistic methods of care. The data further depicts a growth in anti-LGBTQ+ narratives and climate denial to a greater extent than Computing Forever (I) corpus.

A similar topic appearing in both corpuses regarded concerns relating to transhumanism and human body augmentation. These topics appeared to be expressed in relation to vaccines, the digitization of society and the potential of a "cashless society becoming reality". Each of these concerns relates to fears of surveillance by an often unnamed global elite seeking to control the greater population, a fear which is expressed in both Computing Forever (I) and Computing Forever (II).

5. CONCLUSIONS

Due to a less stringent content moderation of Alt-tech channels, opinion mining of these platforms is imperative. This study employs topic modeling as a successful tool for analysing social media data and Alt-tech platforms. An algorithm was proposed for topic model selection using Document Topic Similarity (DTS) index. The method was validated quantitatively by calculating Rand Index (RI) and reporting the Spearman correlation Coefficient between DTS and RI findings. The proposed model selection algorithm was investigated further qualitatively and scored with 80% accuracy on Amazon Cell Reviews labelled dataset.

Applying topic model selection algorithm to choose the best performing topic model on Alt-tech corpus, proposed to choose LDA topic model with Ngram range of (4, 4). The achieved qualitative analysis of this topic model on unlabelled Alt-tech corpus was asserted by a social scientist and scored with average QR = 0.72 and QR = 0.36 for Computing Forever (I) and Computing Forever (II) corpuses.

The proposed procedure of topic model selection was evaluated on Amazon Reviews Corpus with 973 documents and proved to be effective on Computing Forever (I) corpus with 994 documents. However, the qualitative rating of the selected topic model on Computing Forever (II) corpus with 11,273 documents was poor. The low-quality topics extracted from the Computing Forever (II) corpus could be the result of large corpus length, occurrence of unimportant terms in both topics and documents, stated general context in TopDocs, advertisement, documents including spam or no context, etc. Strategic future work toward addressing and resolving these issues for topic modeling of Alt-tech data is encouraged.

ACKNOWLEDGEMENTS

This research was funded by the Irish Research Council under grant number COALESCE/2021/39. The authors would like to thank Prof. Padraig Cunningham and Dr.Mansurah Afifa Khan for their helpful advice during this research project.

REFERENCES

[1] R. Rogers, 'Deplatforming: Following extreme Internet celebrities to Telegram and alternative social media', *Eur. J. Commun.*, vol. 35, no. 3, pp. 213–229, Jun. 2020, doi: 10.1177/0267323120922066.

[2] 'Peering into parler: topic modeling and data exploration on a dataset of parler posts'.

[3] S Rasti, MA Khan, Brendan Scally, Eugenia Siapera, Padraig Cunningham, 'Alt-tech topic modeling with coherence similarity measure for model selection', presented at the AICS 2022 Conference, MTU Cork, MTU Cork, Dec. 2022.

[4] J. McAuley and J. Leskovec, 'Hidden factors and hidden topics: understanding rating dimensions with review text', in *Proceedings of the 7th ACM conference on Recommender systems*, in RecSys '13. New York, NY, USA: Association for Computing Machinery, Oct. 2013, pp. 165–172. doi: 10.1145/2507157.2507163.

[5] S Rasti, 'Alt-tech Topic Model Selection'. University College Dublin, Apr. 2023. [Online]. Available: https://github.com/sanaz-rasti/Alt-tech-Topic-Model-Selection-

[6] S. Latif, F. Shafait, and R. Latif, 'Analyzing LDA and NMF Topic Models for Urdu Tweets via Automatic Labeling', *IEEE Access*, vol. 9, pp. 127531–127547, 2021.

[7] A. Srivastava and C. Sutton, 'Autoencoding variational inference for topic models', *ArXiv Prepr. ArXiv170301488*, 2017.

[8] F. Bianchi, S. Terragni, and D. Hovy, 'Pre-training is a hot topic: Contextualized document embeddings improve topic coherence', *ArXiv Prepr. ArXiv200403974*, 2020.

[9] J. Qiang, Z. Qian, Y. Li, Y. Yuan, and X. Wu, 'Short text topic modeling techniques, applications, and performance: a survey', *IEEE Trans. Knowl. Data Eng.*, vol. 34, no. 3, pp. 1427–1445, 2020.

[10] H. Zhao, L. Du, W. Buntine, and G. Liu, 'MetaLDA: A topic model that efficiently incorporates meta information', in *2017 IEEE International Conference on Data Mining (ICDM)*, IEEE, 2017, pp. 635–644.

[11] N. Reimers and I. Gurevych, 'Sentence-bert: Sentence embeddings using siamese bert-networks', *ArXiv Prepr. ArXiv190810084*, 2019.

[12] F. Bianchi, S. Terragni, D. Hovy, D. Nozza, and E. Fersini, 'Cross-lingual contextualized topic models with zero-shot learning', *ArXiv Prepr. ArXiv200407737*, 2020.

[13] A. R. Hadiat, 'Topic Modeling Evaluations: The Relationship Between Coherency and Accuracy', PhD Thesis, 2022.

[14] D. Korenčić, S. Ristov, and J. Šnajder, 'Document-based topic coherence measures for news media text', *Expert Syst. Appl.*, vol. 114, pp. 357–373, 2018.

[15] 'spaCy: open-source software library for advanced natural language processing'. https://spacy.io/ (accessed Mar. 28, 2023).

AUTHOR

Sanaz Rasti, PhD, Research Software Engineer, School of Computer Science, University College Dublin, Dublin, Ireland