A Systematic Literature Review on Topic Modelling and Sentiment Analysis

Josip Kunsabo, Jasminka Dobša
Faculty of Organization and Informatics
University of Zagreb
Pavlinska 2, Varaždin, Croatia
{jkunsabo, jasminka.dobsa}@foi.hr

Abstract. *The human need for organization has always existed, but due to the ever-increasing generation of digital records, it is becoming almost necessary. With the development of information technologies, the use and application of various machine learning methods in the processing of natural language is becoming more widespread. In this paper we will give an overview of methods of topic modelling and sentiment analysis as well as approaches in processing of documents needed for their application. These two tasks of natural language processing are naturally related in the problem of identification of sentiments in the topics automatically extracted from the collection of observed documents.*

Keywords. topic modelling, sentiment analysis, machine learning, deep learning, sentiment lexicon

1 Introduction

With the development and availability of the Internet, large amounts of information, i.e. digital records, are being created. Such records include various online publications, portals, blogs, forums, social networks, streaming sites, public administration data, company data, various databases, and the like. Timeliness and easy availability with the possibility of open and free use provide readers with access to a large and diverse amount of information. Metadata processing can help structure available records, but their large number and unstructuredness, often make this challenge impossible.

The availability of computer systems and information technologies with the application of natural language processing methods enable a significant shift in the speed and quality of processing text sets of data by various statistical methods, as well as machine and deep learning methods. Thus, methods of topic modelling and sentiment analysis can be used to identify and structure textual content.

Topic modelling is a data mining task of discovering topics that appear in the observed set of documents (Blei et al., 2003). The basic approach involves machine learning to group similar topics, often through various lexicons that define affiliation to a particular topic. Probabilistic methods of unsupervised machine learning have also been widely implemented to find hidden topics in electronic publications.

Computer sentiment analysis, also known as “opinion mining”, is an area of text mining that analyzes human opinions, sentiments, attitudes, and emotions toward entities such as products, services, organizations, individuals, events, topics, and their attributes (Pang and Lee, 2008). There are several approaches to sentiment analysis but the most common is the approach in which the text is labeled as positive or negative. Sentiment classification can be either a binary or a multi-class problem. Binary sentiment analysis is the classification of texts into positive and negative classes, while multi-class sentiment analysis focuses on classifying data into fine-grained labels or multi-level intensities. (Minaee et al., 2021).

Topic modelling, as well as sentiment analysis, attracted researchers from academia and industry, and this paper will provide an overview of the methods used in this area.

2 Literature review

Topic modelling as well as sentiment analysis are well-studied methods in the last decade, as evidenced by the large increase in results in the past period. Based on searches of Ebsco, Scopus and Web of Science databases by keywords “topic modelling”, “sentiment analysis” and combinations of words “topic modelling” and “sentiment analysis”, a large increase in publication of scientific papers related to these topics can be observed in the period from 2012 and later compared to the period before 2012 (Table 1.). Particularly, huge increase can be observed for the topic of sentiment analysis and application of topic modelling methods for it. The reasons for this can be found in the greater availability of computer systems and information technologies, as well as in the fact that technological development has enabled complex or extensive computer tasks to be processed in a shorter time. Furthermore, increasingly diverse areas of application and promising results are of growing interest to researchers and to the industry.

With the aim of analyzing methods for topic modelling and sentiment analysis of electronic publications, the results of the last six years of articles covering the above keywords were isolated and 313 articles in English were filtered for further analysis. A review of the results shows domination of health-related topics, mostly related to COVID-19, while topics before COVID-19 were mostly related to the oil and financial sector. Considering that COVID-19 was an integral part of the daily life of a large part of the world's population in the past period, it is understandable to observe its impact from all aspects (from health to economy).

Table 1. Statistics for database search

Query	Period	Database		
		Ebsco	Scopus	Web of Science
Topic modelling	<2012	8,091	27,891	192,212
	2012-2022	12,575	52,192	384,245
Sentiment analysis	<2012	2,239	2,186	1,175
	2012-2022	15,800	26,850	23,898
Topic modelling and sentiment analysis	<2012	183	38	28
	2012-2022	986	1,454	2,976

Looking at data sources, most of them are data from the social network Twitter and specialized databases or other electronic publication. The reasons for using Twitter as a source of data is in the fact that it is one of the most used social networks, which also allows easy and open access to data.

For further analysis of articles relevant to topic modelling and sentiment analysis, based on the reading of abstracts, 60 papers were selected, i.e. 45 of them remained after reading the whole paper. An analysis of the selected papers with an overview of type of textual documents and their amount used, text preprocessing methods and the methods used for topic modelling and sentiment analysis are shown in Table 2. The column that describes the data sets used contains two values, “article” and “post”. “Article” includes text documents with longer content such as publicly available articles, blogs, financial or media reports and reviews, while “post” means shorter messages such as posts on social networks or comments on articles.

3 Discussion

Topic modelling as well as sentiment analysis are confirmed as very current topics. Research shows that it attracts researchers from all over the world to understand the events around them. Research related to COVID-19 dominates thematically, but the observed aspects of the impact of COVID-19 are very diverse. Thus, topics related to crisis and health communication during the pandemic, as well as topics related to fears and views on vaccination are addressed. Also, comparison of communication in the news in relation to social networks and analysis of online education during a pandemic and the like is being made.

In order to select the appropriate research sample, when processing “post” data it is noticeable the dominance of using Twitter as a source of data with the use of filtering by keywords related to the research topic. In such processing, samples range from several thousand to several million records. When processing “article” data, only a few researchers included a very large number of publicly available news articles, which were not previously filtered on the basis of keywords selected according to the area of interest. The remaining research used specific reports and records of targeted exploration areas such as healthcare or oil.

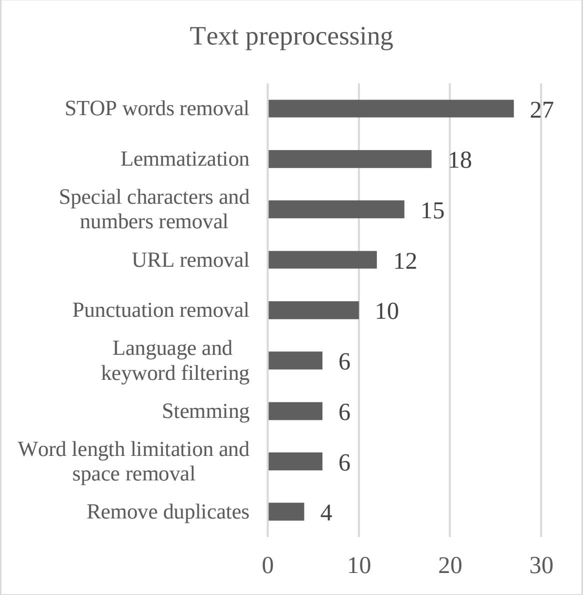

Figure 1. Distribution of text preprocesing methods

Textual records in most research go through a preprocessing phase of data. As shown on Figure 1, preprocessing often includes removing of stop words, URL links, punctuation and special characters, duplicate results, and changing uppercase letters to lowercase letters.

Lemmatization and tokenization are also often applied in the preprocessing phase. Lemmatization is the operation that maps words into their basic form and tokenization is a step that involves recognizing and

Table 2. Overview of the analyzed papers

Paper	Dataset	Preprocessing	Methods of topic modelling	Methods/lexicons for sentiment analysis
Kim et al. (2015)	16,189 articles 7,106,297 posts	stop words removal, Lemmatization	n-gram LDA	PKDE4J SO-CAL
Bandhakavi et al. (2016)	13,750 articles 280,000 posts	-	PMI, sLDA	DSEL
Liu et al. (2017)	1,455,947 posts	stop words removal, Stemming	DTSA	
Morimoto et al. (2017)	298,205 articles	stop words removal	MDTM	-
Cerchiello et al. (2018)	553,666 articles	stop words removal, Word length limitation	STM	-
Huang et al. (2018)	21,984 articles	stop words removal, Lemmatization	LDA	VADER
Li et al. (2018)	6,756 articles	stop words removal,	LDA DTM	Textblob
Carrera, Berny & Jung, Jae-Yoon. (2018)	143,876 posts	Language filtering, Punctuation and stop words removal, Lemmatization	LDA	Naive Bayes
Gambino et al. (2018)	17,743 articles 4,412 posts	Lemmatization, URL and numbers removal,	Binary Relevance (BR)	Naive Bayes, Random Forest, SVM
Tahmassebi et al. (2018)	Article stream	stop words removal, Lemmatization	LDA	SentiWordNet
Zolnoori et al. (2019)	3,763,737 articles	Special characters removal, Lemmatization, Stemming	TKM	VADER
Xu et al. (2020)	3,715 articles	Space removal, Punctuation, stop words removal, Lemmatization	MLbased	
Daou (2020)	1,216,343 posts	-	LDA, NMF, LSA	VADER SentiStrength
Liu et al. (2020)	7,791 articles	Remove duplicates, stop words removal, Stemming	LDA	-
Valdez et al. (2020)	86,581,237 posts	stop words removal	LDA	VADER
Abd-alrazaq et al. (2020)	167,073 posts	Language filtering, Punctuation and stop words removal, Lemmatization	LDA	Textbloob
Li et al. (2020)	9,715 articles	-	DP-Sent-LDA	Bi-LSTM
Muhamedyev, et al. (2020)	804,829 articles	-	BigARTM	BERT

Zhou et al. (2020)	2,211 articles	Numbers removal, Punctuation and stop words removal, Word length limitation	STM	DUST
Chu et al. (2020)	1,040 articles	Punctuation and stop words removal	LDA	NRC, Bing, AFINN,
Rudra et al. (2020)	600 articles	Special characters removal	LDA	Naive Bayes
Shah et al. (2021)	96,234 posts	URL removal, Special characters and stop words removal, Lemmatization	DTM	CrystalFeel
Yadav et al. (2021)	12,451,298 posts	Special characters and URL removal, stop words removal, Stemming, Lemmatization	LDA	SentiWordNet AFINN
Melo et al. (2021)	18,413 articles 1,597,934 posts	URL removal, Special characters, Punctuation and stop words removal, Lemmatization, Stemming	LDA	VADER
Monselise et al. (2021)	7,948,886 posts	Remove duplicates, URL removal and stop words removal, Lemmatization	NMF	VADER BERT
Lyu et al. (2021)	227,840 posts	Remove duplicates, URL and stop words removal, Lemmatization	LDA	NRC
Zhang (2021)	404 articles	stop words removal	LDA	LSD
Cheng et al. (2021)	800,000 posts	-	LDA	VADER, NER, RF
Febro and Catindig (2021)	9,842 articles	Punctuation, Special characters and stop words removal	TKM	-
Ghanem et al. (2021)	747,018 posts	Keyword filtering, Language filtering	MD-ULM	
Liew and Lee (2021)	672,133 posts	Space removal, Punctuation and stop words removal, Lemmatization	STM	VADER
Kuo et al. (2021)	125,570 articles	-	LDA	ANTUSD
Gerts D et al. (2021)	1,800,000 posts	URL removal, stop words removal	DTM	AFIN
Yakunin et al. (2021)	1,142,735 articles	-	LDA	Word-Net, ConceptNet,
Liu and Huang (2021)	30,789 articles	-	ODEE	VADER

Smith and Cipolli, (2021).	8,013 posts	Space removal, URL removal, Punctuation, Numbers and stop words removal	LDA	NRC
Kregel et al. (2021)	95,000 articles	URL removal, Lemmatization	LDA	SentiWordNet
Ghasiya et al. (2021)	102,278 articles	URL removal	top2vec	RoBERTa
van der Veen and Bleich (2021)	15,121 articles	-	NMF	lexicon
Dornick et al. (2021)	57,921 articles	Language filtering, Special characters and numbers removal, stop words removal, Lemmatization	BERT	TextBlob
Zheng et al. (2022)	315,136 posts	Remove duplicates, Special characters and URL removal, stop words removal, Lemmatization	LDA	LIWC
Oliveira et al. (2022).	1,700,000 posts	Special characters and numbers removal, Word length limitation, Lemmatization	LDA	RoBERTa
Waheeb et al. (2022)	853,000 posts	-	LDA	VADER, ULMS, TextBloob, SentiWordNet, TSA
Idler et al. (2022)	704 articles	Keyword filtering	STM	Sentimentr
Yin et al. (2022).	75,665 posts	URL removal, Special characters and stop words removal, Punctuation, Lemmatization, Stemming	LDA	VADER

separating individual phrases. Stemming is the process of reducing words to their word stem or root form. Already created packages in R or Python are most often used when performing preprocessing.

The Latent Dirichlet Allocation (LDA) method (Blei et al., 2003) was used in topic modelling in 23/45 (51%) of the research papers. LDA is a probabilistic model with a hierarchical structure for its components, which include documents, topics, and words. It assumes that a given document is generated from a mixture of topics and these topics produce the words in the documents according to their probabilistic distributions. LDA backtracks and derives the hidden topics that create those documents on the basis of the statistics of the included words.

In addition to the LDA, research also uses methods based on LDA, such as Dynamic Topic Modelling (DTM) (Blei and Lafferty, 2007), supervised Latent Dirichlet Allocation (sLDA) (Blei and McAuliffe, 2007), Dependency Parsing Sentence Latent Dirichlet Allocation (DP-Sent-LDA) (Li, et al., 2018) and Big Additive Regularization of Topic Models (BigARTM) (Muhamedyev, et al., 2020). DTM divides data into time slices based on the time they were generated. The set of topics at each time slice is then assumed to evolve from the set of topics at the previous time slice using a state space model. The result is an evolving probability distribution of words for each topic that shows how certain words become more or less important over time for the same topic (Gerts D et al., 2021). In Multiscale Dynamic Topic Model (MDTM), current topic-specific distributions over words are assumed to be generated based on the multiscale word distributions of the previous epoch. (Iwata T et al., 2010). In sLDA, a response variable associated with each document is added to LDA. sLDA jointly models the documents and the responses, in order to find latent topics that will best predict the response variables for future unlabeled documents (Blei and McAuliffe, 2007). The LDA is a kind of bag-of-words model, which assumes that the order of words within a document does not matter, and different words in one sentence are likely to come from different topics. On the other hand, Sent-LDA model relaxes the assumption of bag-of-words and makes the assumption that all words in one sentence are sampled from the same topic (Bao and Datta 2014). Redundant words in sentences can disturb the effectiveness of discovering when using Sent-LDA, so Dependency Parsing (DP) is introduced (Li, et al., 2018). DP is a natural language processing technology that can reveal the syntactical structure of sentences and improves the effectiveness of Sent-LDA by removing the redundant words within the short text. BigARTM is an open-source library for parallel construction of thematic models on large text cases whose implementation is based on the additive regularization approach (ARTM) (Muhamedyev, et al. 2020).

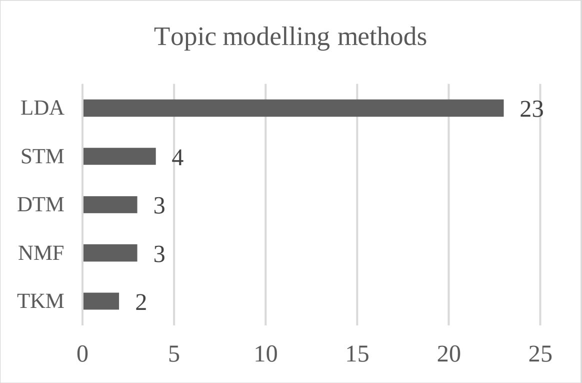

Figure 2. Frequency of usage of topic modelling methods used in at least two analysed papers

As shown on Figure 2., the usage of Structural Topic Modelling (STM) (Roberts et al. 2016), Topic Keyword Model (TKM) (Schneider and Vlachos, 2018). and Non-negative Matrix Factorization (NMF) (Lee and Seung, 1999) is noticeable. STM uses document-level metadata to estimate how the prevalence of topics might vary (Roberts et al. 2016). TKM considers the word order in a text document for topic modelling. TKM links a word to a topic if it or its adjacent words have a high association score with the topic (Schneider and Vlachos, 2018). NMF is a partitioning method that uses nonnegative factorization of the term-document matrix of a given corpus (Lee and Seung, 1999).

When analyzing sentiment, special lexicons are often used. All the lexicons and data sets in analyzed papers are in English except a few data sets in Portuguese (Melo et al., 2021), Arabic (Ghanem et al., 2021) or Chinese (Liu et al., 2020). Beside unsupervised approach of sentiment analysis using lexicons, a supervised approach using standard classification algorithms such as Naive Bayes and Random Forest classifiers is also noticeable.

The Valence Aware Dictionary and Sentiment Reasoner (VADER) lexicon was used in 11/45 (24%) sentiment analyses (Liew, Lee, 2021). To identify key sentiment (positive, neutral and negative), VADER uses a human-curated lexicon of 7500 emotion-related words as well as five human-interpretable rules that identify sentiment intensity.

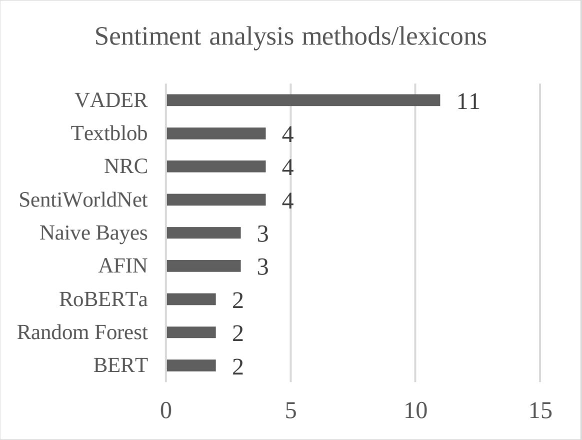

Figure 3. Frequency of usage of sentiment analysis methods/lexicons used in at least two analysed papers

In addition to VADER, National Research Council of Canada Emotion (NRC) (Mohammad and Turney, 2013), TextBlob, and SentiWordNet lexicons as well as BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2019) and RoBERTa (Robustly Optimized BERT Pretraining Approach) (Y. Liu et al., 2019) models were used for sentiment analysis (Figure 3). NRC is a sentiment and emotion lexicon that contains over 14,000 English words and their associations with two sentiments (positive, negative) and eight basic emotions: anger, fear, anticipation, trust, surprise, sadness, joy, and disgust (Mohammad and Turney, 2013). Textblob is a Python library which provides a simple API for diving into common natural language processing tasks, such as sentiment analysis, classification, translation, etc (Li et al., 2018), while SentiWordNet was designed to support sentiment analysis applications by providing an annotation based on three numerical sentiment values of positivity, negativity, and neutrality (Kregel et al., 2021). BERT is a word representation model that uses unannotated text to perform various natural language processing tasks such as classification and question answering (Devlin et al., 2019). RoBERTa is an extension of the original BERT model. Compared to BERT, RoBERTa is trained on more data, with longer sequences, bigger batches, for a longer time and does not include the next sentence prediction objective during pretraining (Y. Liu et al., 2019).

Some research that have chosen to approach a common model for topic modelling and sentiment analysis such as Moroccan Dialect Universal Language Model (MD-ULM) (Ghanem et al., 2021), Manifold Learning-based model (ML-based) (Oikawa et al. 2016) and Dynamic Topic-based Sentiment Analysis model (DTSA) (Liu et al., 2017) models can be highlighted. MD-ULM is pre-trained model on large unlabeled corpora, which is fine-tuned using COVID-19 dataset (Ghanem et al., 2021). Manifold learning (ML) is an effective method for nonlinear dimensionality reduction (NLDR), whose main objective is to discover meaningful low dimensional structures hidden in high dimensional data (Oikawa et al. 2016). DTSA assumes that topic-based sentimental evolution of the current time epoch is influenced by its historical sentiment of multiple time slides (Liu et al., 2017).

When evaluating results in the case of topic modelling, the most common is the application of coherence score (Syed and Spruit, 2017) or cosine similarity (Rahutomo et al., 2012). When observing the evaluation of sentiment analysis results, it is common to use metrics of precision, recall, and F1-score measures. Data visualization in most research includes tabular and temporal representations. Some research uses a “text cloud” to describe topics, which clearly shows what the topic covers, while some researchs stand out using Social Network Analysis (SNA) methods and present their results via a network graph.

4 Conclusion

Topic modelling is a challenging area that provides insight into structure of data in electronic media and enables identification of sentiments present in the data. Although models such as the LDA have long been present in the scientific community and have been well studied, their dominant use confirms that they serve their purpose in determining the topics of text documents. Other models, although not necessarily new, are used to a lesser extent.

While the topic modelling task is independent of the language, language is important when analyzing sentiment. Most of the observed research determined sentiment with the help of one of the lexicons, i.e. words associated to the sentiment. As the words themselves have a different spelling in different languages, to determine sentiment, it is necessary to have a lexicon. This is the reason why English, as the most common language, was covered in most of the research. The frequency of using VADER lexicons is noticeable, but the diversity of lexicons tells us that there are still many different options that researchers are experimenting with. In addition to the lexicon approach, there are also models based on machine learning with the “gold standard” as well as pre-trained BERT models, which will be proven in future research. The most often used methods of machine and deep learning, as well as the amount and types of dictionaries that exist on different natural languages, will be part of future research.

References

Abd-alrazaq, A., Alhuwail, D., Househ, M., Hamdi, M., & Shah, Zubair. (2020). Top Concerns of Tweeters During the COVID-19 Pandemic: Infoveillance Study. Journal of Medical Internet Research. 22. 10.2196/19016.

Bandhakavi, A. S., Wiratunga, N., Deepak, P., & Massie, S. (2016). Lexicon based Feature Extraction for Emotion Text Classification. Pattern Recognition Letters. 93. 10.1016/j.patrec.2016.12.009.

Bao, Y., & Datta, A. (2014). Simultaneously discovering and quantifying risk types from textual risk disclosures. Management Science, 60, 1371–1391.

Blei, D., Ng, A., & Jordan, M. (2003). Latent Dirichlet Allocation. Journal of Machine Learning Research. 3. 993-1022. 10.1162/jmlr.2003.3.4-5.993.

Blei D, & Lafferty J. Dynamic topic models. In: ICML '06: Proceedings of the 23rd international Conference on Machine learning. 2006 Jun Presented at: 23rd International Conference on Machine Learning; June 25-29, 2006; Pittsburgh, PA p. 113-120.

Blei, D., & McAuliffe, J. (2007), Supervised topic models, in: Advances in Neural Information Processing Systems 20 (NIPS 2007).

Carrera, B., & Jung, J.Y. (2018). SentiFlow: An Information Diffusion Process Discovery Based on Topic and Sentiment from Online Social Networks. Sustainability. 10. 2731. 10.3390/su10082731.

Cerchiello, P., & Nicola, G. (2018). Assessing News Contagion in Finance. Econometrics. 6. 5. 10.3390/econometrics6010005.

Cheng, I., Heyl, J., Lad, N., Facini, G., & Grout, Z. (2021). Evaluation of Twitter data for an emerging crisis: an application to the first wave of COVID-19 in the UK. Scientific Reports. 11. 10.1038/s41598-021-98396-9.

Chu, C.Y., Park, K., & Kremer, G. (2020). A global supply chain risk management framework: An application of text-mining to identify region-specific supply chain risks. Advanced Engineering Informatics. 45. 10.1016/j.aei.2020.101053.

Daou, H. (2020). Sentiment of the public: the role of social media in revealing important events. Online Information Review. ahead-of-print. 10.1108/OIR-12-2019-0373.

Devlin J, Chang M, Lee K, & Toutanova K. BERT: pretraining of deep bidirectional transformers for language understanding. 2019 Presented at: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics; June 2019; Minneapolis, Minnesota. [doi: 10.18653/v1/n18-2]

Dornick, C., Kumar, A., Seidenberger, S., Seidle, E., & Mukherjee, P. (2021). Analysis of Patterns and Trends in COVID-19 Research. Procedia Computer Science. 185. 302-310. 10.1016/j.procs.2021.05.032.

Febro, J., & Catindig, M. (2021). Exploring cyber violence against women and girls in the Philippines through Mining Online News [Explorando la ciberviolencia contra mujeres y niñas en Filipinas a través de Mining Online News]. Comunicar. 30. 10.3916/C70-2022-10.

Gambino, O., & Calvo, H. (2018). Modelling distribution of emotional reactions in social media using a multi-target strategy. Journal of Intelligent & Fuzzy Systems. 34. 2837-2847. 10.3233/JIFS-169471.

Gerts, D., Shelley, C., Parikh, N., Pitts, T., Watson, R.C., Fairchild, G., Vaquera, C. N., & Daughton, A. “Thought I’d Share First” and Other Conspiracy Theory Tweets from the COVID-19 Infodemic: Exploratory Study JMIR Public Health Surveill 2021;7(4):e26527 URL: https://publichealth.jmir.org/2021/4/e26527 DOI: 10.2196/26527

Ghanem, A., Asaad, C., Hafidi, H., Moukafih, Y., Guermah, B., Sbihi, N., Zakroum, M., Ghogho, M., Dairi, M., Cherqaoui, M., & Baïna, K. (2021). Real-Time Infoveillance of Moroccan Social Media Users’ Sentiments towards the COVID-19 Pandemic and Its Management. International Journal of Environmental Research and Public Health. 18. 12172. 10.3390/ijerph182212172.

Ghasiya, P., & Okamura, K. (2021). Investigating COVID-19 News across Four Nations: A Topic Modelling and Sentiment Analysis Approach. IEEE Access. 9. 36645-36656. 10.1109/ACCESS.2021.3062875.

Huang M., ElTayeby O., Zolnoori M & Yao L. Public Opinions Toward Diseases: Infodemiological Study on News Media Data. J Med Internet Res. 2018 May 8;20(5):e10047. doi: 10.2196/10047. PMID: 29739741; PMCID: PMC5964307.

Idler, E., Bernau, J., & Zaras, D, (2022). Narratives and counter-narratives in religious responses to COVID-19: A computational text analysis. PloS one. 17. e0262905.; 0.1371/journal.pone.0262905.

Iwata, T., Yamada, T., Sakurai, Y., & Ueda, N. (2010). Online Multiscale Dynamic Topic Models. 663-672. 10.1145/1835804.1835889.

Kim, H.J., Jeong, Y.K., Kim, Y., Kang, K., & Song, M. (2015). Topic-based content and sentiment analysis of Ebola virus on Twitter and in the news. Journal of Information Science. 42. 10.1177/0165551515608733.

Kregel, I., Koch, J., & Plattfaut, R. (2021). Beyond the Hype: Robotic Process Automation's Public Perception Over Time. Journal of Organizational Computing and Electronic Commerce. 31. 1-21. 10.1080/10919392.2021.1911586.

Kuo, H.Y., Chen, S.Y., & Lai, Y.T. (2021). Investigating COVID-19 News before and after the Soft Lockdown: An Example from Taiwan. Sustainability. 13. 11474. 10.3390/su132011474.

Lee DD, & Seung HS. Learning the parts of objects by non-negative matrix factorization. Nature 1999 Oct 21;401(6755):788-791. [doi: 10.1038/44565] [Medline: 10548103]

Li, J., Li, G., Yao, Y., & Zhu, X. (2018). Identifying the influence factors of commodity futures market through a new text mining approach. In The 7th international conference on futures and other derivatives. Shanghai, China.

Li, X., Shang, W., & Wang, S. (2018). Text-based crude oil price forecasting: A deep learning approach. International Journal of Forecasting. 35. 10.1016/j.ijforecast.2018.07.006.

Li, J., Li, G., Liu, M., Zhu, X., & Wei, L. (2020). A novel text-based framework for forecasting agricultural futures using massive online news headlines. International Journal of Forecasting. 10.1016/j.ijforecast.2020.02.002.

Liew, T., & Lee, C. (2021). Examining the Utility of Social Media in COVID-19 Vaccination: Unsupervised Learning of 672,133 Twitter Posts (Preprint). JMIR Public Health and Surveillance. 7. 10.2196/29789.

Liu, J. & Huang, X. (2021). Forecasting Crude Oil Price Using Event Extraction. Papers 2111.09111, arXiv.org.

Liu, P., Gulla, J. & Zhang, L. (2017). A joint model for analyzing topic and sentiment dynamics from large-scale online news. World Wide Web. 21. 1-23. 10.1007/s11280-017-0474-9.

Liu, Q., Zheng, Z., Zheng, J., Chen, Q., Liu, G., Chen, S., Chu, B., Zhu, H., Akinwunmi, B., Huang, J., Zhang, C., & Ming, W.K. (2020). Health Communication Through News Media During the Early Stage of the COVID-19 Outbreak in China:

A Digital Topic Modelling Approach. 10.1101/2020.03.29.20043547.

Lyu, J., Han, E., & Luli, G. (2021). COVID-19 Vaccine–Related Discussion on Twitter: Topic Modelling and Sentiment Analysis (Preprint). 10.2196/preprints.24435.

Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad K.N., Asgari-Chenaghlu, M., & Gao, J. (2021). Deep Learning--based Text Classification: A Comprehensive Review. ACM Computing Surveys. 54. 1-40. 10.1145/3439726.

Melo, T., & Figueiredo, C. (2021). Comparing News Articles and Tweets About COVID-19 in Brazil: Sentiment Analysis and Topic Modelling Approach. JMIR Public Health and Surveillance. 7. 10.2196/24585.

Mohammad, S. & Turney, P. (2013). Crowdsourcing a Word-Emotion Association Lexicon. Computational Intelligence. 29. 10.1111/j.1467-8640.2012.00460.x.

Monselise, M., Chang, C.H., Ferreira, G., Yang, R., & Yang, C. (2021). Topics and Sentiments of Public Concerns Regarding COVID-19 Vaccines: Social Media Trend Analysis. Journal of Medical Internet Research. 23. 10.2196/30765.

Morimoto, T., & Kawasaki, Y. (2017). Forecasting Financial Market Volatility Using a Dynamic Topic Model. Asia-Pacific Financial Markets. 24. 10.1007/s10690-017-9228-z.

Muhamedyev, R., Yakunin, K., Mussabayev, R., Buldybayev, T., Kuchin, Y., Murzakhmetov, S., & Yelis, M. (2020). Classification of Negative Information on Socially Significant Topics in Mass Media. Symmetry. 12. 1945. 10.3390/sym12121945.

Oikawa, M.A., Dias, Z., de Rezende Rocha, A. (2016). Manifold learning and spectral clustering for image phylogeny forests. IEEE Transactions on Information Forenstcs and Security, 11(1), 5–19.

Oliveira, F., Haque, A., Mougouei, D., Evans, S., Sichman, J., & Singh, M. (2022). Investigating the Emotional Response to COVID-19 News on Twitter: A Topic Modelling and Emotion Classification Approach. IEEE Access. 10. 16883-16897. 10.1109/ACCESS.2022.3150329.

Pang, B. & Lee, L. (2008). Opinion Mining and Sentiment Analysis. Foundations and Trends in Information Retrieval. 2. 1-135. 10.1561/1500000011.

Rahutomo, F. & Kitasuka, T. & Aritsugi, M. (2012). Semantic Cosine Similarity.

Roberts, M.E., Brandon M. S., & Dustin T. 2016a. Navigating the Local Modes of Big Data: The Case of Topic Models. In Data Analytics in Social Science, Government, and Industry. New York: Cambridge University Press.

Rudra B.P., Kokatnoor, S., Yadav, C., & Mounika, M. (2020). Polarity Detection on Real-Time News Data Using Opinion Mining. 10.3233/APC200124.

Schneider J & Vlachos M. Topic modelling based on keywords and context. In: Proceedings of the 2018 SIAM International Conference on Data Mining. SIAM; 2018.

Shah, A., Yan, X., Tariq, S., & Shah, A. (2021). Tracking patients healthcare experiences during the COVID-19 outbreak: Topic modelling and sentiment analysis of doctor reviews. Journal of Engineering Research. 9. 10.36909/jer.v9i3A.8703.

Smith, H., & Cipolli, W. (2021). The Instagram/Facebook ban on graphic self-harm imagery: A sentiment analysis and topic modelling approach. Policy & Internet. 10.1002/poi3.272.

Syed, S. & Spruit, M. (2017). Full-Text or Abstract? Examining Topic Coherence Scores Using Latent Dirichlet Allocation. 165-174. 10.1109/DSAA.2017.61.

Tahmassebi, A., McCann, I., Meyer-Base, A., Erlebacher, G. & Foo, S. (2018). NewsAnalyticalToolkit: an online natural language processing platform to analyze news. 23. 10.1117/12.2304646.

Valdez D., ten Thij M., Bathina K., Rutter L., & Bollen J., Social Media Insights Into US Mental Health During the COVID-19 Pandemic: Longitudinal Analysis of Twitter Data J Med Internet Res 2020;22(12):e21418 URL: https://www.jmir.org/2020/12/e21418 DOI: 10.2196/21418

van der Veen, M., & Bleich, E. (2021). Atheism in US and UK Newspapers: Negativity about Non-Belief and Non-Believers. Religions. 12. 291. 10.3390/rel12050291.

Waheeb, S., Khan, N., & Shang, X. (2022). Topic Modelling and Sentiment Analysis of Online Education in the COVID-19 Era Using Social Networks Based Datasets. Electronics. 11. 715. 10.3390/electronics11050715.

Xu, Y., Li, Y., Liang, Y., & Cai, L. (2020). Topic-sentiment evolution over time: a manifold learning-based model for online news. Journal of Intelligent Information Systems. 55. 10.1007/s10844-019-00586-5.

Yadav, A., & Vishwakarma, D. (2021). A Language-independent Network to Analyze the Impact of COVID-19 on the World via Sentiment Analysis.

ACM Transactions on Internet Technology. 22. 1–30. 10.1145/3475867.

Yakunin, K., Kalimoldayev, M., Muhamedyev, R., Mussabayev, R., Barakhnin, V., Kuchin, Y. Murzakhmetov, S., Buldybayev, T., Ospanova, U., Yelis, M., Zhumabayev, A., Gopejenko, V., Meirambekkyzy, Z., & Abdurazakov, A. (2021). KazNewsDataset: Single Country Overall Digital Mass Media Publication Corpus. Data. 6. 31. 10.3390/data6030031.

Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, & V. Stoyanov, ``RoBERTa: A robustly optimized BERT pretraining approach," 2019, arXiv:1907.11692.

Yin, H., Song, X., Yang, S., Li, J. (2022). Sentiment analysis and topic modelling for COVID-19 vaccine discussions. World Wide Web. 25. 1-17. 10.1007/s11280-022-01029-y.

Zhang, C. (2021). Media Framing of Color-Blind Racism: A Content Analysis of the Charlottesville Rally*. Race and Social Problems. 13. 10.1007/s12552-021-09321-8.

Zheng, H., Goh, D., Lee, E., & Lee, C. S., & Theng, Y. (2022). Understanding the effects of message cues on COVID-19 information sharing on Twitter. Journal of the Association for Information Science and Technology. 73. 847-862. 10.1002/asi.24587.

Zhou, Pi., He, Y., Lyu, C., & Yang, X. (2020). Characterizing News Report of the Substandard Vaccine Case of Changchun Changsheng in China: A Text Mining Approach. Vaccines. 8. 691. 10.3390/vaccines8040691.

Zolnoori, M., Huang, M., Patten, C., Balls-Berry, J., Goudarzvand, S., Brockman, T., Pour, E., & Yao, L. (2019). Mining News Media for Understanding Public Health Concerns. Journal of Clinical and Translational Science. 5. 1-29. 10.1017/cts.2019.434.