Indonesia's News Topic Discussion about Covid-19 Outbreak using Latent Dirichlet Allocation

1st Razief Perucha Fauzie Afidh
Faculty of Mathematics and Natural Science
Universitas Syiah Kuala
Banda Aceh, Indonesia
razief@unsyiah.ac.id

2nd Zainal A. Hasibuan
Faculty of Computer Science
Universitas Dian Nuswantoro
Semarang, Indonesia
zhasibua@gmail.com

***Abstract*— News related to Covid-19 or Corona dominates the news in several media. From various perspectives, a number of news have been produced regarding the corona outbreak. The purpose of this research is to look at various news topics related to corona, both in national online media and in local online media. This study uses the Latent Dirichlet Allocation (LDA) algorithm to find out news topics related to corona. Preprocessing was carried out on existing articles such as removing punctuation marks, numbers, and removing stopwords. This process is preceded by lowering the text to get unique words. The number of articles collected was 12.883 news titles from the national online media RMOL and the Aceh Tribunnews regional online media in the period January to May 2020. The news items came from various news categories such as politics, law, health, economy, sports and others. Based on this research, by using the LDA algorithm, evaluate using coherence value and visualize the topics, will have the best perspective of how many topics can be created. RMOL news articles have 12 topics discussion with the coherence value is 0.538795. Aceh Tribunnews has 8 topics with the coherence value of 0.522946.**

Keywords— Covid-19, News, Topic modeling, LDA

I. INTRODUCTION

Since January 2020 the Covid-19 outbreak, known as the Coronavirus, has spread throughout the world. Indonesia is one of the countries that has experienced the spread of this virus outbreak. Various media, both online digital media and social media, have reported massively about the Coronavirus. The number of media coverage with various news viewpoints provides various information to the public.

In this paper, we implemented one of topic modeling algorithm Latent Dirichlet Allocation is known as LDA. LDA as a topic modeling used to monitoring the discussion topic in the online digital news during the Covid-19 outbreak. In order to get an overview of corona news, we collected booths from national and regional online news. This paper is divided into several sections. Part 1, the introduction, is an introductory part. In part 2, the related works with the topic modeling for the online digital news will be discussed briefly. In part 3, we will explain the experiment process. Section 4 will explain the result and analysis of the research experiment. Conclusions and future work will be presented in section 5.

II. RELATED WORKS

Research conducted [1] uses Latent Dirichlet Allocation (LDA) to conduct topic modeling on user comments on social media. This study uses data users' comments on the Tokyo Electric Power plant accident interview comments. In addition, this study also uses news data that is most read and commented on Yahoo. A total of 14450 comments and 1005 news stories along with their comments.

News data as research data related to topic modeling was also carried out in this study[2]. This study uses 1000 news data to classify news in e-news in Punjabi. This study uses LDA to obtain words from topic modeling for further use in news classification. Another study that also uses LDA to build a news classification model is research conducted by [3]. This study used a 20 newsgroup dataset consisting of 20 different news groups.

Research conducted by [4] uses LDA to classify Indonesian language news taken from Radar Semarang news data. A total of 100 articles from 5 newsgroups were used in this study. This study conducted a comparison of LDA parameters with several scenarios, namely for 5, 10 and 15 topics. The accuracy of the model produced in this study reaches 70% for 5 newsgroups; economy, tourism, crime, sport and politics. Research [5] uses LDA on Spark MLlib to model large amounts of news data. This research uses 50 thousand news data taken randomly on the news portals Sohu news, Xinhua and NetEase news. This study uses Loglikelihood and logPerplexity to measure the quality of the model obtained.

[6] In his research used 645.000 news data from detik.com to do topic modeling to map locations with case studies on Jakarta floods and diphtheria diseases in Indonesia. This study uses LDA and word2vec to see the suggestion of positive keywords and negative keywords which will be aggregated with georeferenced data. In the health sector, the use of news articles as the main data to extract information is also widely used. Research related to the disease by extracting data from the news was also conducted by [7]. This study looked at public sentiment and topics discussed related to various diseases and health conditions. This study uses 3.5 million archive news articles sourced from Reuters news data. LDA is used to model the topics discussed in these articles. The Study [8] using 7710 news data from Chosunilbo analyzed popular topics related to women's health conditions and diseases that changed over time. [9] In his research, he collected media reports related to Covid-19. By using LDA and several other python libraries, perform the process of extracting information from Chinese news and other articles contained in the WiseSearch database. 7791 articles were used in this study out of a total of 11200 news articles available.

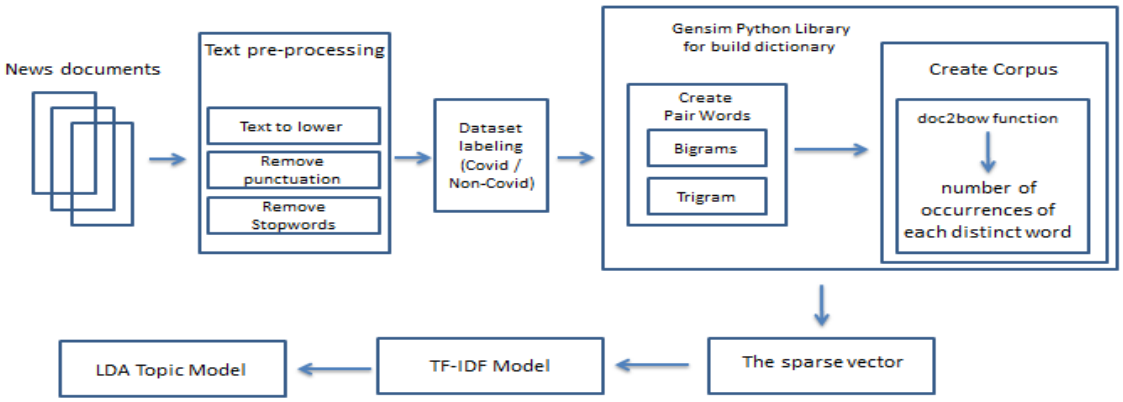


Fig. 1. Topic Modeling Process

[10] In his research, he analyzes the topic of news in Korea through articles on the internet related to the fourth industrial revolution in various fields such as business, government, and industry. 8428 news items obtained from January to September 2018 were used to analyze articles using LDA and centrality analysis.

In the political field, research conducted by [11] analyzes how media coverage of hydraulic fracturing and political conflict is. 7616 The data is used to perform modeling using LDA. Research conducted by [12] using a dataset from the News24 coverage, identified discussion topics in the media related to elections. 53897 data were used in this study which generated 20 latent topics.

III. METHODOLOGY

The main objective of this research is to monitor the topic discussed in the Indonesian news using the news title dataset. In this part, we will discuss more the detail about the data collecting process, including the data pre-processing process. LDA as an algorithm that is used will be explained briefly in this section.

A. Dataset

In this study, using a dataset extracted from the online news portal RMOL.id and Aceh Tribunnews for the period January 2020 to May 2020. The data crawling process uses the Scrapy library in Python. The existing data is referred to as a stored dataset in the csv format. There are 43.254 news titles collected. In this research, we focused on the specific news topic which is Corona/Covid-19. We used 12.883 news titles for the topic modeling process. The distribution number of the news by categories are following in table 1:

TABLE I. DATASET

News Portal	Document Count
RMOL	6543
Aceh Tribunnews	6340
Total 9 Categories	12883

This data is further cleaned by converting the whole data into the lowercase by eliminating punctuation and stopwords. In this topic modeling process, no difference between words using uppercase letters and words with lowercase letters. Therefore, making words in the same format is important. The calculation process of the number of words can be done properly.

B. Preprocessing

Punctuation elimination aims to get words without punctuation. So that punctuation is not part of the word. In some cases, punctuation is considered a separate character, so it is counted as words.

In a text, some words that often appear and do not have a significant influence on the topic modeling process. These words can be removed. This study uses self-defined stopwords based on their appearance in the dataset. The implementation of stopwords in this dataset can be seen in the following figure 2, where the dominance of the appearance of stopwords that was defined earlier can be eliminated.

Word Frequency before remove stopword
All frequencies, including STOPWORDS:

Word	Frequency
di	3950
corona	2597
dan	2560
covid19	2475
yang	1784
tidak	1515
jokowi	1377
untuk	1308
pemerintah	1149
dari	1085

Word Frequency before cleaning
(text-to-lower,remove punctuation,
remove stopword)
All frequencies, not including STOPWORD

Word	Frequency
corona	2596
jokowi	1377
pemerintah	1149
virus	929
china	820
indonesia	797
kasus	747
kpk	702
warga	675
anies	655

Fig. 2. Example of Words Frequency before and after data cleansing

An example of the results of text pre-processing is as shown in figure 3. These clean data will use for topic modeling as an input in the LDA model.

Before text preprocessing

```
0    New York Times: Indonesia Terlambat, Virus Cor...
1    Rusia Mulai Uji Klinis Vaksin Covid-19 Pada Ma...
2    Lima Bulan, Kasus Covid-19 Di Dunia Sudah Temb...
3    India Perpanjang Lockdown Di Zona-zona Berisik...
4    Usul Berlakukan New Normal Di Tempat Kerja, Pe...
```

After text preprocessing

```
0    new york times indonesia terlambat virus coron...
1              rusia uji klinis vaksin pada manusia
2        lima bulan kasus dunia tembus enam juta jiwa
3          india perpanjang lockdown berisiko tinggi
4    usul berlakukan new normal tempat kerja pelaku...
```

Fig. 3. Example of Data Pre-processing result

C. LDA Algorithm

Dictionary and corpus are the main inputs for building the LDA topic model. For this reason, in this study a dictionary was built which contains a list of combinations of 2 words (bi-gram) and 3 words (tri-gram) which are more dominant along with their frequency of occurrence. The sparse vector is the output of this dictionary which provides *word_id* information and the number of occurrences (frequent) in a document. An example of the resulting sparse vector is as shown in the following figure 4:

```
[(3, 1), (4, 1), (5, 3), (6, 1), (85, 1), (164, 1), (446, 1)]
```

Fig. 4. Example of sparse vector

The number of words with id 3 appears once in the first document. Next, the words id 4, 6, 85, 164 and 446 also appear once in this first document. Meanwhile, the word with id 5 appears 3 times in this first document.

The objective of topic models is to extract the underlying topics from a given collection of text documents. Each document in the text is considered as a combination of topics, and each topic is considered as a combination of related words. We use TF-IDF to count the occurrences of words in each document.

The formula to find the weight of words with TF-IDF is (1):

$$w_{ij} = tf_{ij}\, idf$$
$$idf = log \frac{N}{df_j} \quad (1)$$

Where w_{ij} is the weight of word i at the document of j, N the number of documents, and the term frequency tf_{ij} is the sum of the presence of word i in document j, df_j (document frequency) is the number of document j which contains word i.

In addition, to determine the number of topics to be formed, this study uses a coherence score. Coherence score is a measure used to evaluate the topic modeling in which a good model will produce topics with high topic coherence scores.

In order to implement the LDA model, we use LDA Model library from Gensim in python. The scheme of the LDA model, as shown in figure 5.

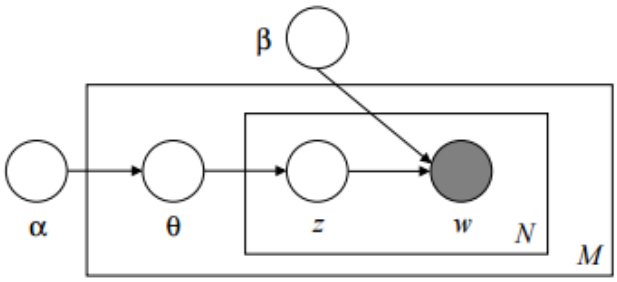


Fig. 5. Graphical model representation of LDA[13]

Where:

α, β: a variable of the Corpus label

- α: Dirichlet distribution on the subject, showing the pattern of the subject
- β: Dirichlet distribution for words, showing the pattern of words
- θ: a variable of the document level showing of the topic word distribution of one document
- z, w: a variable of the word level
- z: an assigned topic in that word
- w: a word of actual observed for the topic

In order to visualize the topic that extracted from the dataset, pyLDAvis python library is used. pyLDAvis imported from the Gensim python library.

IV. Result and Analysis

Based on the model and the dataset used, the following results are explained:

A. Word cloud topic visualization

In our research, we found that the use of stopword has a significant impact on the number of the topic that will create. As shown in figure 6, if we use all of the words including the stopwords, the stopwords as the most frequent words will appear as a keyword. Furthermore, this word will influence the number of topics that will construct.

(a) (b)

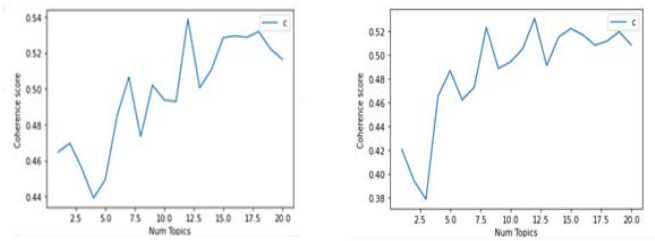
Fig. 6. Word cloud visualization with stopwords RMOL (a) and Aceh Tribunnews (b)

On the other hand, as shown in figure 7, stopwords removal will give a significant impact. The model counts the frequent word in order to build the topic cluster based on the relevance words. In our case, there are 12 topics for RMOL news article and 8 topics for Aceh Tribunnews articles.

(a) (b)

Fig. 7. Word cloud visualization without stopword RMOL (a) and Aceh Tribunnews (b)

B. Coherence Value



(a) (b)

Fig. 8. Coherence value for RMOL (a) and Aceh Tribunnews (b)

According to figure 8, for RMOL topic modeling, we take 12 with the coherence value is 0.538795 and The Aceh Tribunnews is 8 with the coherence value 0.522946. This number used to visualize the topic cluster using LDA algorithm with the implementation of pyLDAvis python library.

C. LDA analysis

Each topic consists of words that relevance to that topic with some specific value. As shown in Table II, there are top 10 words for each topic that construct the topic on Aceh Tribunnews. Table III shows the top 10 words that construct the topic in RMOL portal news.

TABLE II. THE RESULT OF LDA ANALYSIS OF ACEH TRIBUNNEWS (TOP 10)

Topic	Word	Value
1	rumah_sakit	0.054664
	virus_corona	0.025917
	sakit	0.020673
	plt_gubernur	0.014181
	aceh_tamiang	0.014091
	rumah	0.013284
	pakai_masker	0.012897
	ditangkap	0.010828
	cerita	0.010163
	aceh_barat	0.010065
2	virus_corona	0.059727
	virus	0.020807
	korban	0.015973
	meninggal	0.015738
	penerima_bantuan	0.013424
	bertambah	0.012276
	meninggal_dunia	0.012159
	wni	0.01125
	kasus	0.00976
	indonesia	0.00955
3	virus_corona	0.11363
	virus	0.040064
	china	0.035013
	terinfeksi_virus	0.028301
	wuhan	0.018189
	wabah	0.010427
	terinfeksi	0.009605
	singapura	0.009195
	pkh	0.008712
	hari	0.008557
4	nagan_raya	0.045666
	serahkan_bantuan	0.024565
	bantuan	0.023523
	virus_corona	0.023435
	korban_kebakaran	0.022585
	nagan	0.015205
	raya	0.01482
	salurkan_bantuan	0.014431
	banda_aceh	0.012278
	masa_panik	0.011765
5	virus_corona	0.051773
	wuhan	0.021379
	virus	0.016701
	tewas	0.013883
	doa	0.011957
	kota_langsa	0.011139
	china	0.010598
	lansia	0.009517
	terpapar	0.009087
	wabah	0.008934
6	terima_bantuan	0.031298
	virus_corona	0.028246
	bandara_sim	0.027969
	bandara	0.021318
	aceh_selatan	0.019531
	wali_kota	0.017517
	banda_aceh	0.016991
	terjangkit	0.01688
	bantuan	0.015645
	kelelawar	0.014934
7	virus_corona	0.042419
	penyebaran_virus	0.024074
	aceh_singkil	0.022868
	penyakit	0.022433
	virus	0.018305
	kapal_pesiar	0.014927
	taiwan	0.011046
	muncul	0.010687
	video	0.010013
	pencegahan	0.009938
8	virus_corona	0.039333
	mahasiswa_aceh	0.029566
	mahasiswa	0.027
	cegah_penyebaran	0.015421
	virus	0.014955
	aceh_utara	0.014178
	china	0.011693
	senjata	0.011198
	tangani_pasien	0.011121
	makanan	0.011045

TABLE III. THE RESULT OF LDA ANALYSIS RMOL (TOP 10)

Topic	Word	Value
1	virus_corona	0.093243
	virus	0.03311
	wuhan	0.031867
	korban_meninggal	0.026533
	china	0.019395
	korban	0.014684
	asal	0.01385
	mahasiswa	0.013397
	karyawan	0.01199
	bandara	0.011875

2	observasi	0.029384
	virus_corona	0.024338
	wni	0.021186
	umumkan	0.020775
	ruang_isolasi	0.018867
	dikarantina	0.017209
	karantina	0.014611
	terduga	0.011734
	saleh_daulay	0.011356
	pantau	0.011119
2	meninggal_dunia	0.03145
	panik	0.024561
	presiden_jokowi	0.015086
	jokowi	0.014053
	meninggal	0.013621
	pencegahan	0.012476
	besar	0.01138
	keluar_rumah	0.010272
	presiden	0.010193
	dunia	0.009762
3	pemprov_dki	0.027612
	rapid_test	0.025694
	korea_selatan	0.022631
	virus_corona	0.018314
	dievakuasi	0.013892
	sri_mulyani	0.01321
	jalani	0.012476
	selatan	0.011834
	bentuk	0.010038
	abk	0.009863
4	rumah_sakit	0.032071
	kemenkes	0.019272
	virus_corona	0.016339
	pulau	0.016067
	diamond_princess	0.014314
	penumpang	0.01341
	anies_baswedan	0.013282
	wuhan	0.011908
	sakit	0.011437
	rumah	0.01073
5	penyebaran_virus	0.052557
	air	0.023264
	pasien_positif	0.023153
	tutup	0.019539
	tim_medis	0.018508
	pasien	0.016316
	penyebaran	0.013383
	virus_corona	0.011813
	positif	0.011732
	puan_maharani	0.011502
6	evakuasi_wni	0.042815
	evakuasi	0.038516
	wuhan	0.026529
	virus_corona	0.024419
	wni	0.023169
	rp_miliar	0.020528
	diduga	0.015859
	menular	0.013837
	sehat	0.013276
	langsung	0.010351
7	virus_corona	0.032653
	terjangkit	0.030106
	hong_kong	0.021841
	penerbangan	0.01788
	arab_saudi	0.016619
	maskapai	0.014953
	china	0.014335
	dinyatakan_positif	0.013892
	dki_jakarta	0.013199
	hoax	0.012936
8	virus_corona	0.134312
	virus	0.045406
	kapal	0.020277
	kapal_pesiar	0.018937
	china	0.014104
	ekonomi	0.014041
	cegah_penyebaran	0.012552
	xi_jinping	0.01248
	taiwan	0.012121
	pesiar	0.011997
9	komisi_ix	0.020922
	petugas_medis	0.019578
	komisi	0.015874
	pesawat	0.015724
	tenaga_medis	0.014468
	rsud	0.013141
	pulangkan	0.011936
	menhub	0.01151
	ix	0.010498
	meninggal	0.009487
10	menkes_terawan	0.03475
	menkes	0.021013
	virus_corona	0.019146
	terawan	0.014536
	tempat	0.014059
	lockdown	0.013235
	karantina	0.011764
	anggota_dpr	0.011218
	tangkal	0.010782
	kasus_positif	0.010323
11	virus_corona	0.098003
	virus	0.037429
	china	0.030518
	natuna	0.028256
	wabah	0.019418
	wuhan	0.015973
	kematian_akibat	0.014651
	menyebar	0.012113
	perangi	0.010458
	korea_utara	0.009858

LDA Topic Visualization

Topic visualization create using pyLDAvis python library as shown in figure 9 and figure 10. The topic position describes the coherency between topic that are constructed. The number of words and the values are shows on the right side of the dashboard. We can adjust the relevance metric by slide the value of the λ in order to know which words and the frequency of words that contribute to construct the topic cluster.

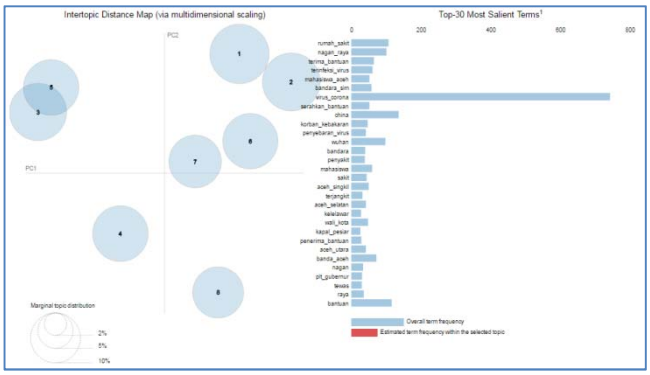


Fig. 9. Topic Visualization for Aceh Tribunnews

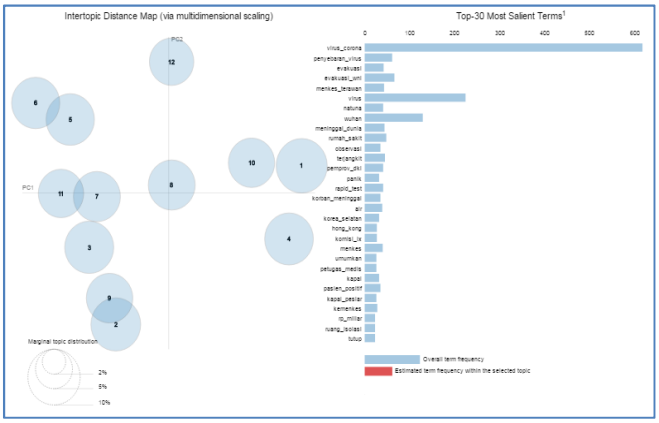


Fig. 10. Topic Visualization for RMOL

V. Conclusion And Future Research

According to our experiment, by improving the use of stopwords in the preprocessing phase will give us the proper number of words that later on will be used as an input for the topic modeling process. Topic coherence score can be used as an option to choose the best number of a topic that can be constructed from our dataset. Based on our experiment, 8 topics are constructed from our dataset during January until May 2020 for Aceh Tribunnews and 12 topics for RMOL portal online. As future research, we can trace and predict the evolution of the topic in order to prepare the solution if there is a problem or positive impact on the activities.

References

[1] D. Ramamonjisoa, "Topic Modeling on Users ' s Comments," *Third ICT Int. Student Proj. Conf.*, pp. 177–180, 2014.

[2] A. Verma and A. K. Gahier, "Topic modeling of E-news in Punjabi," *Indian J. Sci. Technol.*, vol. 8, no. 27, pp. 1–10, 2015.

[3] Z. Li, W. Shang, and M. Yan, "News text classification model based on topic model," *2016 IEEE/ACIS 15th Int. Conf. Comput. Inf. Sci. ICIS 2016 - Proc.*, pp. 1–5, 2016.

[4] R. Kusumaningrum, M. I. A. Wiedjayanto, S. Adhy, and Suryono, "Classification of Indonesian news articles based on Latent Dirichlet Allocation," *Proc. 2016 Int. Conf. Data Softw. Eng. ICoDSE 2016*, pp. 1–5, 2016.

[5] J. Gui and Q. Wang, "Topic modeling of news based on spark Mllib," *2016 13th Int. Comput. Conf. Wavelet Act. Media Technol. Inf. Process. ICCWAMTIP 2017*, vol. 2018-Febru, pp. 224–228, 2017.

[6] A. Dewandaru, S. I. Supriana, and S. Akbar, "Event-Oriented Map Extraction From Web News Portal : Binary Map Case Study on Diphteria Outbreak and Flood in Jakarta," *ICAICTA 2018 - 5th Int. Conf. Adv. Informatics Concepts Theory Appl.*, no. October, pp. 72–77, 2018.

[7] M. Huang, O. ElTayeby, M. Zolnoori, and L. Yao, "Public opinions toward diseases: Infodemiological study on news media data," *J. Med. Internet Res.*, vol. 20, no. 5, pp. 1–12, 2018.

[8] T. Modeling, K. Won, S. Young, and Y. Woon, "Osong Public Health and Research Perspectives Analysis of Women ' s Health Online News Articles Using," vol. 10, no. 3, pp. 158–169, 2019.

[9] Q. Liu *et al.*, "Health communication through news media during the early stage of the covid-19 outbreak in China: Digital topic modeling approach," *J. Med. Internet Res.*, vol. 22, no. 4, 2020.

[10] K. A. Jang, M. Park, and W. J. Kim, "Topic analysis for the fourth industrial revolution using LDA and centrality analysis in Korea," *Proc. - Int. Conf. Mach. Learn. Data Eng. iCMLDE 2018*, pp. 108–116, 2019.

[11] R. Berardo *et al.*, "Hydraulic fracturing and political conflict: News media coverage of topics and themes across nine states," *Energy Res. Soc. Sci.*, vol. 70, no. October 2019, p. 101660, 2020.

[12] A. Moodley and V. Marivate, "Topic modelling of news articles for two consecutive elections in South Africa," *2019 6th Int. Conf. Soft Comput. Mach. Intell. ISCMI 2019*, pp. 131–136, 2019.

[13] J. C. Campbell, A. Hindle, and E. Stroulia, "Latent Dirichlet Allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, 2003.