INTERNATIONAL JOURNAL FOR INNOVATIVE RESEARCH IN MULTIDISCIPLINARY FIELD
ISSN(O): 2455-0620 [Impact Factor: 7.581]
Monthly, Peer-Reviewed, Refereed, Indexed Journal with IC Value : 86.87
Volume - 9, Issue - 9, September- 2023 Publication Date: 30/09/2023

DOIs:10.2015/IJIRMF/202309037 --:-- Research Paper / Article / Review

Topic Modelling from Repository of Political Speeches

[1]K.Bhuvaneshwari, [2]S.A JyothiRani, [3]V.V. Haragopal
1 Research Scholar, Department of Statistics, Osmania University, India
2 Professor, Department of Statistics, Osmania University,
3 Retd.Professor, Department of Statistics, Osmania University, India
Email – [1]k.bhuvaneshwari7@gmail.com, [2]jyothi2263@gmail.com, [3]haragopalvajjha@gmail.com

Abstract: *To identify Topics in political speeches delivered by politicians among same party and also comparison of topics spoken by other politicians of difference parties in Telangana state. This experiment of applying NLP tools based text analytics proves to be efficient practice for Political Discourse Analysis. For this study, political speeches of certain politicians in the Telangana state were collected from a particular time period. Then the speech transcripts of the political speeches were obtained using Google Cloud. These unstructured textual transcripts were pre-processed to remove the stop words, noise and characters in order to obtain the semantic structures and to make the meaningful corpus. Further, the most powerful NLP tool "Topic Modelling using LDA (Latent Dirichlet Allocation)" has been applied to the resultant corpus to extract the topics spoken by different poli-ticians. Further, Model Perplexity and Coherence Score was computed and trained the resultant model with the LDA parameters. The findings include details on the most likely subjects and ideas that each leader spoke that had an effect on listeners. This is the first study of the use of a popular topic modelling approach to political speeches of various political leaders in Indian scenario. Furthermore, this is the first attempt to use a statistical model to extract the subjects mentioned in a campaign.*

Key Words: *Coherence Score, Latent Dirichlet Allocation, Natural Language Processing (NLP), Political Speeches, Topic Modelling.*

1. INTRODUCTION:

In the present digital era, each day we see a large amount of text data emanate in an unstructured format in the form of news articles, social media posts research pa-pers, transcripts of political speeches etc. Traditional approaches are often used in text mining to find information or terms. Topic modelling varies from a rule-based text mining technique in that it also offers a rapid and concise summary of the essen-tial content in a few words without having to read the entire data. Topic modelling helps to organize and give insights to understand such huge collection of unstructured textual content to make well-versed decisions [1].

Candidates from different political parties often deliver speeches on a range of problems and subjects. There is a sizable reservoir of unstructured data because most powerful or well-known political candidate's speech transcripts are freely accessible online. To analyse such large pool of unstructured data, natural language processing (NLP) is essential for automatically extracting the topics that people are talking about [2]. With the help of multiple topics that best explain the underlying information in a given document, a text document can be represented using the robust topic modelling technique of unsupervised natural language processing [3].

This paper covers research on the topic of "Topic modelling using Latent Dirichlet Allocation (LDA)" in Natural Language Processing (NLP). The proposed framework has been applied to the corpus of political speeches of the influential leaders of Tel-angana state namely, K.ChandraSekhar Rao, K.Taraka Rama Rao, Harish Rao, Bandi Sanjay, Revanth Reddy collected during the months of August to October 2021, for analyzing and extracting the topics from the speeches of these politicians. Speech transcripts in English are considered in the purview of this Textual situation study.

In the recent past, there have been numerous researchers working in the area of textual analytics using topic modelling techniques. Most of the research papers pro-posed on topic modelling, nearly all of the articles have adopted the LDA method because it is the most advanced and popular method for topic modelling [4]. To rapidly and efficiently remove all the extraneous information from high dimensional data in order to extract the most pertinent information and present it concisely. Natu-ral Language Processing (NLP) is an analytical study in computer science that is con-cerned with data management, semantic data mining, and allow computers to derive insights from human-language processing

of textual data [5]. In order to extract the hidden patterns from text data, it is necessary to use particular pre-processing tech-niques and algorithms.

Chauhan U, Shah A [6] stated that Topic modelling methods as robust and intelli-gent techniques that have seen widespread use in NLP to semantic data mining and topic detection from unordered text documents. Jelodar, H.,et al[7] proposed that, LDA is a simple method that may be used to assess how similar the source files are to one another and to determine how each document is distributed over the various subjects. Jacobs T, Tschotschel R [8], the authors proposed a generative framework to identify the hidden (latent) links between opinions and topics that can be helpful for pulling out the political views. A brand-new ”two-layer matrix factorization algo-rithm” has been created by Greene D. and Cross JP [9] for locating subjects in lengthy political speeches corpora gathered from European Parliament speeches and identi-fied topics related to actions at a specific point in time and enduring subjects. In 10] in order to extract the top general topics, authors created a joint Bayesian model that combines topic modelling and event segmentation into a single, integrated frame-work. In the literature on political science, some topic modelling techniques have been used to examine political attention [11]. This study highlights subjects that come up repeatedly and identifies the key viewpoints of political leaders from vari-ous parties. Isoaho K. in [12] proposed a novel unsupervised topic model for self-supervised opinion modelling based on LDA to find the opinions from various points of view.

In the Proposed Work, to understand the insights of the speeches, topic modelling by LDA is used to analyze political speeches of various influential politicians in the Telangana state. It involves gathering of speeches and creating transcripts, processing documents beforehand, document modelling using LDA, and extracting the key words from the topics.

2. METHODOLOGY:

Data collection :

For this study, speeches of the influential politicians in the Telangana state were col-lected. During the Bye Election at Huzurabad constituency, in Telangana state, we collected political speeches of the leaders namely K. Chandrasekhar Rao (KCR), K.Taraka Rama Rao (KTR), Harish Rao from the ruling party i.e, Telangana Rastra samithi , Bandi Sanjay Kumar, state president of the Bharatiya Janata Party and Revanth Reddy from Indian National Congress party. For each leader we gathered 3 speeches during the period of 3 months from August 2021 to October 2021. The speech transcripts were generated over internet using google cloud. Speech transcripts in English were taken into consideration for this study.

Pre-Processing:

Data pre-processing is a primary step for any study in order to transform raw data into useful and efficient format for analysis. The text documents generated from speech transcripts may contain lot of noise stop words, numbers and punctuation. Therefore text data must be cleaned in order to extract hidden information. Pre- processing involves removal of stop words, numeric values, punctuation, stemming and lemmatization.

In this study, Python has been used it is an open-source and interpretive, high- lev-el programming language that offers brilliant functionality for processing semantic data. Additionally, it supports a large number of scientific libraries [13]. The text cor-pus is analysed in Python by importing the re and nltk libraries. Meaningful metadata is extracted from the text documents and gathered in the data frame.

Topic Modelling :

Topic modelling is simply referred as the process of extracting or pulling out topics from the text documents. A topic model is the one which automatically identifies the topics based on the words appearing in a text document. Topic modelling is an unsupervised learning method that enables users to recognise and analyse the themes that are concealed inside text content [14]. It presumes that a corpus has a variety of obscure topics. Although it is unable to comprehend the true meaning and implication of the words and concepts in the textual data, it is able to capture the hidden con-cepts by making use of the context surrounding the words in document collections. It also provides the hidden topics in the documents with a certain probability assigned to terms in each document. Thus, the topic modelling can be thought of as a problem of clustering where the number of clusters is like the topics [15].

Latent Dirichlet Allocation

A generative model called Latent Dirichlet allocation (LDA) was created in 2003 by David Blei, Andrew Ng, and Michael I. Jordan. The shortfalls of TFIDF were high-lighted in their work[6]. The LDA was created as a result of TFIDF’s inability to com-prehend word semantics or word placement in a document. LDA is a generative model, which means it generates every outcome conceivable for a given phenome-non. The fundamental concept is that each document is represented as a random mixture of latent themes, where each subject is represented by a distribution of words. The assumptions of LDA model mathematically can be described as follows:

Select N~Poisson(n (an array of N words within the document that follow Poisson distribution
Select θ~Dir(α (a parameter θ that follow Dirichlet distribution For every Nth word (wn :
Select topic zn ~Multinomial (θ) (Each topic zn follow multinomial probability distribution.)
Select wn from p (wn | zn, β), a multinomial conditional probability on topic zn. (Each topic is viewed as the distribution over the words, where a probability is generated from the probability of the nth word, conditioning upon the topic as well as
β where βij = p(wj = 1|zi = 1) with dimensions k x V)
β = likelihood of a given word,
V = word count in the vocabulary,
k = the dimensionality of the Dirichlet distribution and
θ = the random variable is sampled from the probability simplex. The Frame-work of LDA model is illustrated in Fig.1.

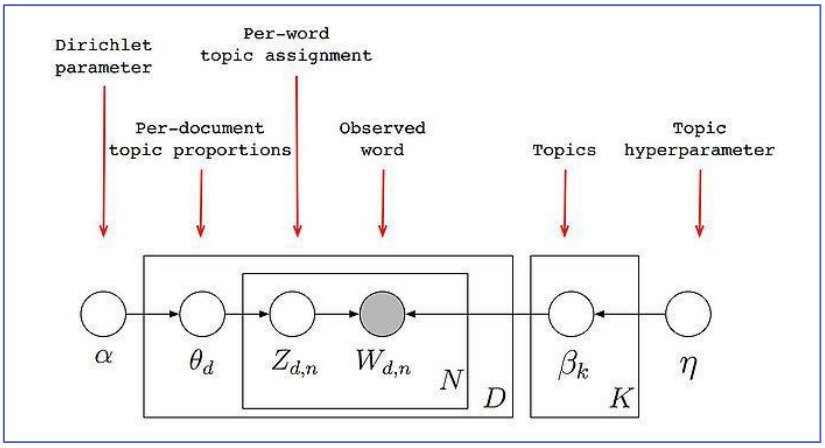

Figure 1. Frame work of LDA model

In this study, after pre-processing the text data, in order to build LDA model, its two primary inputs, namely the dictionary and the corpus were created using genism package in Python. Thereby, fixing the Dirichlet hyper-parameter alpha i.e., Document- Topic Density and Dirichlet hyper-parameter beta i.e., Word-Topic Density, based on the coherence score, the optimal number of topics (K) is determined.

3. RESULTS AND DISCUSSION:

This is a pioneering investigation on the use of the well-known topic modelling ap-proach on political speeches by different politicians in Telangana State. Additionally, this is the first attempt to use a statistical model to extrapolate the concepts stated in a social campaign. This section includes the findings of the study, here we used speeches of the above cited politicians in the Telangana state. We gathered 3 speech-es of each politician i.e., K.Chandra sekhar Rao (KCR), K.Taraka Rama Rao (KTR), Harish Rao, Bandi Sanjay, Revanth Reddy. Initially, a corpus has been built from the transcripts of the speeches collected. Here we used python programming for pre- processing of the text data. Then in order to visualize the corpus, word clouds of speeches by each candidate were created to get insights on the most repeatedly used words shown in Figure 2.

Figure 2. Wordclouds of speeches of politicians

In order to identify the topics by topic modelling with LDA, the coherence scores that have the largest value corresponding to the number of topics (K) by fixing hyper parameters of LDA model were initially used to identify the ideal number of topics spoken. The results are illustrated in Figure 3.

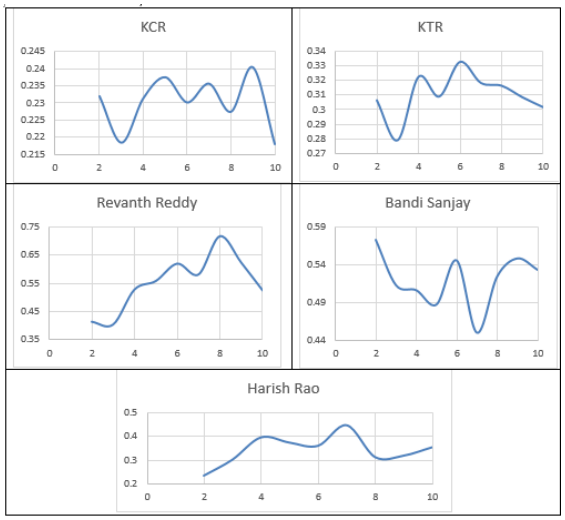

Figure 3. Optimal Number of topics by coherence values

From the above visualization KCR and KTR pattern of topics delivered in their speeches seem to almost similar while Harish Rao's to be entirely different while the opposition party leaders concentrated entirely different and there is no match what-soever. The Optimal number of topics is given in Table 1.

Table 1. Optimal Number of Topics spoken by each leader

S.No	Political Leader	No. of Topics
1	KCR	9
2	KTR	6
3	Harish Rao	7
4	Bandi Sanjay	9
5	Revanth Reddy	8

Next, a matrix of topic probabilities was produced as a result of the LDA model's implementation on a text corpus. The values for document topic likelihood are listed in Tables 2 to 7.

Table 2. Document Probability Values – KCR

Topic	***Speech_1***	***Speech_2***	***Speech_3***
Topic 1	0.0027	0.0476	0.0026
Topic 2	0.0244	0.0489	0.0351
Topic 3	0.0487	0.0578	0.0857
Topic 4	0.1768	0.1877	0.0875
Topic 5	0.0578	**0.3895**	0.4789
Topic 6	0.0499	0.0971	0.0054
Topic 7	0.0473	0.0316	0.004
Topic 8	0.0339	0.0496	0.0129
Topic 9	0.0116	0.016	0.0204

Table 3. Document Probability Values – KTR

Topic	***Speech_1***	***Speech_2***	***Speech_3***
Topic 1	0.0133	0.0587	0.0042
Topic 2	0.0216	**0.5788**	0.0128
Topic 3	0.0166	0.0506	0.0247

Topic 4	0.0099	0.0253	0.0221
Topic 5	0.0247	0.041	0.0255
Topic 6	0.0107	0.03	0.0054

Table 4. Document Probability Values – Harish Rao

Topic	***Speech_1***	***Speech_2***	***Speech_3***
Topic 1	0.0058	0.0107	0.0116
Topic 2	0.0034	0.032	0.0116
Topic 3	0.0081	0.0153	0.0218
Topic 4	0.0118	0.0118	**0.7094**
Topic 5	0.0042	0.0118	0.0148
Topic 6	0.0107	0.017	0.0432
Topic 7	0.0376	0.0234	0.0065

Table 5. Document Probability Values – Bandi Sanjay

Topic	***Speech_1***	***Speech_2***	***Speech_3***
Topic 1	0.0137	0.0097	0.0055
Topic 2	0.0067	0.0063	0.0135
Topic 3	0.0155	0.0052	0.0221
Topic 4	0.0067	0.0075	0.0118
Topic 5	0.006	0.0273	0.0048
Topic 6	**0.6984**	0.0472	0.0066
Topic 7	0.0123	0.0115	0.0128
Topic 8	0.0268	0.0115	0.0247
Topic 9	0.0215	0.0102	0.0107

Table 6. Document Probability Values – Revanth Reddy

Topic	***Speech_1***	***Speech_2***	***Speech_3***
Topic 1	0.0081	0.0113	0.0081
Topic 2	0.7819	0.026	0.0163
Topic 3	0.0163	0.0122	0.017
Topic 4	0.0081	0.0151	0.0156
Topic 5	0.013	**0.8212**	0.0234
Topic 6	0.0081	0.0742	0.0103
Topic 7	0.0081	0.0125	0.0087
Topic 8	0.0368	0.0118	0.0218

Table 7. Topic wise Highest Document probability values of Chosen leaders

Leader	**(topic, speech, probability)**
KCR	(topic 5, speech 2, 0.3895) L
KTR	(topic 2, speech 2, 0.5788)
Harish Rao	(topic 4, speech 3, 0.7094)
Bandi Sanjay	(topic 6, speech 1, 0.6984)
Revanth Reddy	(topic 5, speech 2, 0.8212) H

Table 7, illustrates the topic wise Highest Document probabilities of all the leaders. From this we notice that, among all the leaders. Revanth Reddy is having highest probability value i.e., 0.8212 whereas KCR has least i.e., 03895. This shows that the opposition leaders are in impressing the voters to gain their confidence and concen-trating on more topics and want to be more citizen friendly.

Table 8, illustrates the topics with top (20) words spoken by the leaders that are highest probability values.

Table 8. Topics with highest probability values with top (20) words of Leaders

KCR	KTR	Harish Rao	Bandi Sanjay	Revanth Reddy
Topic 5	***Topic 2***	***Topic 4***	***Topic 6***	***Topic 5***
give	today	dalit	kcr	kcr
government	reform	scheme	trs	make
dalit	land	government	bjp	false
many	government	dalitbandhu	dalit	promise
people	leadership	provide	yatra	village
progress	people	kcr	telangana	family
today	farmer	people	party	fake
telangana	minister	drowntrodden	fake	dalit
welfare	dalit	chief	chief	state
development	state	give	population	government
state	new	Minister	hyderabad	rule
president	give	rupee	demanded	congress
lakh	scheme	bjp	minister	leader
movement	gadwal	make	fund	people
provide	eetela	promise	hatred	politics
scheme	year	welfare	leader	house
come	welfare	development	promise	practically
crore	telangana	huzurabad	unemployment	release
will	chief	constituency	support	funds
constituency	many	state	power	development

Furthers, in order to view the topics at granular level with respect to each speech, world clouds of individuals speeches were prepared which are presented in Figure 4 to Figure 6.

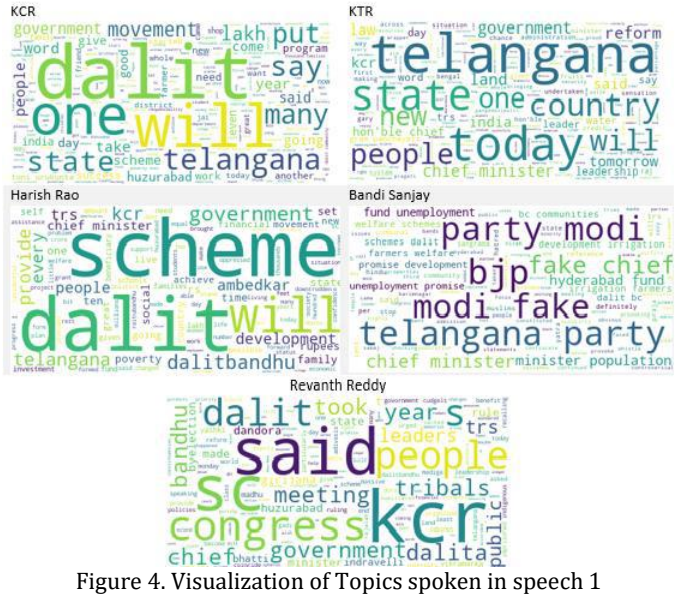

Figure 4. Visualization of Topics spoken in speech 1

INTERNATIONAL JOURNAL FOR INNOVATIVE RESEARCH IN MULTIDISCIPLINARY FIELD
ISSN(O): 2455-0620 [Impact Factor: 7.581]
Monthly, Peer-Reviewed, Refereed, Indexed Journal with IC Value : 86.87
Volume - 9, Issue - 9, September- 2023 Publication Date: 30/09/2023

Figure 5. Visualization of Topics spoken in speech 2

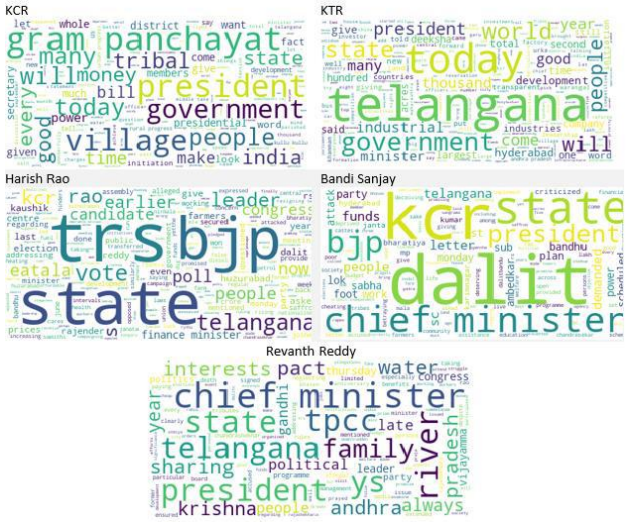

Figure 6. Visualization of Topics spoken in speech 3

Tables 9, 10 and 11 shows the number of common topics of the leaders KCR, KTR, Harish Rao,Bandi Sanjay and Revanth Reddy in Speech 1, Speech 2 and Speech 3 respectively. From these tables we can infer that the ruling party TRS and the opposition leaders are concentrating in most of the common topics to grab the attention of Voters.

Table 9. Common Topics in speech1 of all leaders

Leader	KCR	KTR	Harish Rao	Bandi Sanjay	Revanth Reddy
KCR	-	9	16	5	7
KTR	9	-	7	4	8
Harish Rao	16	7	-	7	4
Bandi Sanjay	5	4	7	-	4
Revanth Reddy	7	8	4	4	-

Table 10. Common Topics in speech2 of all leaders

Leader	KCR	KTR	Harish Rao	Bandi Sanjay	Revanth Reddy
KCR	-	9	7	4	3
KTR	9	-	4	3	4
Harish Rao	7	4	-	4	4
Bandi Sanjay	4	3	4	-	2
Revanth Reddy	3	4	4	2	-

Table 11. Common Topics in speech3 of all leaders

Leader	KCR	KTR	Harish Rao	Bandi Sanjay	Revanth Reddy
KCR	-	16	10	4	5
KTR	16	-	6	3	4
Harish Rao	10	6	-	4	6
Bandi Sanjay	4	3	4	-	3
Revanth Reddy	5	4	6	3	-

4. CONCLUSION:

This study explored the political speeches containing multiple topics with the framework designed. From this analysis, it is obvious that the designed methodology adopted i.e., Topic modelling by Latent Dirichlet Allocation (LDA) has significantly helped in extracting topics from corpus of transcripts of political speeches. This analysis reveals that the various topics or issues being raised by politicians in their speeches.

From table 7, by observing the document probability values of different leaders, it is clear that Revanth Reddy has highest document probability value i.e., 0.8212 where as KCR has lowest document probability value i.e., 0.3895. This implies that the leader Revanth Reddy from Opposition party has covered most of the topics in his speeches which depicts that, being an opponent leader in an effort to acquire power, he is emphasizing on more topics whereas the one who is in power K.Chandrasekhar rao (Chief Minister) from ruling party, Telangana Rastra Samithi has spoken only few number of topics in his speeches.

Further, from table 8, it is clear that there is commonness in topics being spoken by leaders of TRS party i.e., ruling party. The results also demonstrate that, the various topics by opposition party leaders Bandi sanjay and Revanth reddy were mostly about criticizing the ruling party regarding the schemes, programmes and promises made by the chief minister. The overall analysis of the political speeches shows that the leaders were stressing more about dalits, schemes and welfare programs for the downtrodden people of Telangana state.

In addition to that, the most common words in the text are shown as a weighted list of words in figs. 3, 4, and 5. These word clouds provide a comprehensive knowledge of the subjects that are being addressed. In this study it is clear that the leaders Bandi Sanjay and Revanth Reddy had raised more number of topics and addressed the themes such as unemployment, youth, jobs, farmer's welfare, development of state etc.

Further, it can also be noticed that, from tables 9, 10 & 11, the speech wise analysis of each leader depicts the similarity of topics in each speech spoken by the leaders. Here in our analysis, the similarity between topics uttered by KCR and KTR is much more followed by KCR and Harish Rao. In future work, we shall apply this method on more complex text data and try to analyze the semantic textual similarity in order to discover the similarity in meaning between the words that represent the topics.

REFERENCES:

1. Vayansky I. and Kumar S.A., 2020. A review of topic modeling methods. *Information Systems*, *94*, p.101582.
2. Katre P.D., 2019. NLP based text analytics and visualization of political speeches. *International Journal of Recent Technology and Engineering*, *8*(3), pp.8574-8579.
3. Tandel S.S., Jamadar A. and Dudugu S., 2019, March. A survey on text mining techniques. In *2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS)* (pp. 1022-1026). IEEE.
4. Asmussen C.B. and Møller C., 2019. Smart literature review: a practical topic modelling approach to exploratory literature review. *Journal of Big Data*, *6*(1), pp.1-18.
5. Kherwa P. and Bansal P., 2019. Topic modeling: a comprehensive review. *EAI Endorsed transactions on scalable information systems*, *7*(24).
6. Chauhan U. and Shah A., 2021. Topic modeling using latent Dirichlet allocation: A survey. *ACM Computing Surveys (CSUR)*, *54*(7), pp.1-35.
7. Jelodar H., Wang Y., Yuan C., Feng X., Jiang X., Li Y. and Zhao L., 2019. Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey. Multimedia Tools and Applications, 78, pp.15169-15211.
8. Jacobs T. and Tschötschel R., 2019. Topic models meet discourse analysis: a quantitative tool for a qualitative approach. *International Journal of Social Research Methodology*, *22*(5), pp.469-485.
9. Greene D. and Cross J.P., 2015, June. Unveiling the political agenda of the European parliament plenary: A topical analysis. In *Proceedings of the ACM web science conference* (pp. 1-10).

10. Calderón C.A., de la Vega G. and Herrero D.B., 2020. Topic modeling and characterization of hate speech against immigrants on Twitter around the emergence of a far-right party in Spain. *Social Sciences*, *9*(11), p.188.
11. Reber U., 2019. Overcoming language barriers: Assessing the potential of machine translation and topic modeling for the comparative analysis of multilingual text corpora. *Communication methods and measures*, *13*(2), pp.102-125.
12. Isoaho K., Gritsenko D. and Mäkelä E., 2021. Topic modeling and text analysis for qualitative policy research. *Policy Studies Journal*, *49*(1), pp.300-324.
13. Bird S., Klein E. and Loper E., 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. "O'Reilly Media, Inc.".
14. Hager A. and Hilbig H., 2020. Does public opinion affect political speech?. *American Journal of Political Science*, *64*(4), pp.921-937.
15. Guo C., Lu, M. and Wei W., 2021. An improved LDA topic modeling method based on partition for medium and long texts. *Annals of Data Science*, *8*, pp.331-344.