Enhancing Civic Engagement and Accountability of Kenyan Parliament Using Topic Modeling

NG'ETICH STEPHEN KIPCHIRCHIR, Strathmore University, KENYA

This paper proposes a method for enhancing civic engagement and accountability of the Legislative arm of the Kenyan government using topic modelling. This provides insights into key issues being debated and the position of different stakeholders. This study analysed the Kenyan parliament Hansard from 2006 to 2023 using the Latent Dirichlet Allocation model to identify key topics. The coherence score was used to select the best model based on the optimal number of topics. The best model has 72 topics with a coherence score of -4.874. Parliament procedural terms, personal identification and droughts-related topics are among the top discussed topics ranked by coherence score. From the findings, this paper provides a valuable contribution to efforts to improve civic engagement and accountability in Kenya and highlights the potential of topic modelling as a tool for analysing political discourse in the context of parliamentary proceedings.

Additional Key Words and Phrases: Political Science, Topic modeling, Civic Engagement, Hansard, Parliament

ACM Reference Format:
Ng'etich Stephen Kipchirchir. 2023. Enhancing Civic Engagement and Accountability of Kenyan Parliament Using Topic Modeling. 1, 1 (February 2023), 5 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Since gaining independence in 1963, Kenya has had presidential, parliamentary, and local government elections every five years in accordance with the nation's constitution. Kenyans exercise their democratic right by voting for the candidates who represent their views. Voting gives a sense of involvement by offering citizens an opportunity to express discontent and enthusiasm by selecting the right leaders with the sound manifestos [Makarenko 2015]. However, patriotic citizens have a doubting task to ensure their elected leaders are fulfilling the manifestos during their 5-year tenure. This solely relies on access to information which is key in liberal democracy [Nyabuga 2023].

Citizens are actively involved in the government when they are allowed to seek and receive public documents. This helps to fight corruption, enabling citizens to participate in public life, and making government efficient. The Hansard is the official public printed document that captures the opinions and attitudes of every member -question, response and debates national and county assemblies on a variety of crucial societal issues in Kenya. However, manual interpretation of information from the Hansard is a difficult effort for the average citizen due to the sheer volume of recorded material, esoteric speaking style, and complex procedural language employed in Parliament.

Topic modelling is a popular statistical tool for analysing large text data and can be used to uncover hidden patterns and trends in parliamentary and county assembly debates which could allow the public to more easily assess and aggregate the contributions that their elected representatives make in Parliament.

In this paper, we propose to use topic modelling to analyse legislative debates and identify the main issues and themes discussed. This approach will provide an adequate and comprehensive way to analyse parliamentary debates and will promote civic engagement, scrutiny and accountability of parliaments.

1.1 Problem Statement

Understanding the key issues and themes discussed in parliament debates is crucial for the average citizen to track political decision-making and legislative processes. However, manually analysing the large volume of parliamentary debates can be time-consuming and may not provide a comprehensive overview of the topics discussed.

1.2 Objectives

The objective of this research is to use topic modelling to analyse parliamentary debates and identify the main issues and themes discussed in order to gain a better understanding of the legislative process and decision making of elected officials. Specifically, this research aims to:

- Utilise topic modelling techniques to automatically identify key themes and issues discussed in parliamentary debates.
- Compare and contrast the topics discussed in different parliamentary sessions to identify changes and trends in the legislative agenda.

This research will provide a more efficient and comprehensive way to analyse parliamentary debates and will have implications for scholars of political science, policymakers and other stakeholders interested in understanding the legislative process and decision making.

1.3 Assumptions and Scope

The paper will review literature on topic modelling and its application in political science with a focus on parliamentary debates and political science, describe the data and methods used, present and discuss the findings, and conclude with future research directions.

The analysis is limited to english text.

2 LITERATURE REVIEWS

In [Quinn et al. 2006] background of the study, the authors brought out the importance of understanding how political debates evolve over time and it helps in decision-making. The problem is that traditional manual methods of topic modelling are time-consuming and

Author's address: Ng'etich Stephen Kipchirchir, stephen.ng'etich@strathmore.edu, Strathmore University, P.O. Box 59857, Nairobi, KENYA, 59857-00200.

XXXX-XXXX/2023/2-ART $15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

labour-intensive, making it difficult to conduct large scale-analysis of legislative speech. The study proposed dynamic hierarchical prior distribution for the documents' latent topic indicators to identify and categorise the topic discussed in the 105th-108th US senate proceedings. Their dataset, the United States Senate agenda from 1997 to 2004, had over 70,000 documents and over 70,000,000 words. The results show the potential of this method for large-scale analysis of political discourse and for understanding how political discourse evolves over time.

[Guldi 2019] employs topic modelling on British parliamentary speeches(Hansard) to match the known tuning points, key players and technologies in the history of British infrastructure. This study supports dynamic topic models as a resource for periodizing technological advancements. In dynamic topic modelling(DTM), each document in a collection of documents is represented as a combination of topics in a generative model and the topics change over time. The study demonstrated that topic modelling using digital tools complemented the secondary sources in some cases it showed new patterns.

[DiPierro 2018] noted the rapid decrease of media companies has led to fewer journalists covering local government meetings which in turn has negatively impacted civic engagement. In light of this problem, the researchers used topic modelling on 3000 meeting-related documents such as minutes provided by 16 local government agencies in the Bay Area, Carlifornia to generate noteworthy news tips. The model applied is Latent Dirichlet Allocation (LDA) and TopicVec.The authors improved the quality of the topics generated in the model by removing the names of a government agency from the documents. The study found that online LDA and TopicVec techniques can effectively identify some breaking news situations.

[Kleynhans 2014] study used unsupervised topic modelling on South African parliament audio data. Before analysing the audio, the author converted audio into text which used topic modelling to summarise the debate. The study also used written Hansard documents to train the model. Before training, they converted word documents to UTF-8, removed UTF-8 characters, extracted spoken text only, parsed the text and merked the entities, normalised and finally removed non-English words.The model identified the topics with an accuracy of 92.3

[Altaweel et al. 2019] study analysed 3,344 e-petitions filed by US citizens on an electronic petition platform created by the Obama administration dubbed "we the people". The authors used LDA to generate topic models from the e-petitions. Their aim was to identify the main topics and issues addressed by petitioners. Before training the model, they preprocess the data by stemming the words, filtering out less frequent words and common terms and removing stopwords. The authors used coherence to evaluate the optimal number of topics by comparing the coherence values generated by the model and human coding. The study showed that computer-assisted content analysis complements manual content analysis extracting topics with significantly lower levels of human bias.

[Quinn et al. 2010] study aims to develop a method of analysing political attention which is a critical component in moulding the public opinion and public outcomes. The authors used a dynamic multi topic model to analyse US senate legislative speeches from 1997 to 2004. Their dataset has over 118,000 speeches from the US senate debates. The paper presented methods of evaluating the reliability of the topics using the concept of semantic validity. Semantic validity is the degree to which each category or document is consistent in meaning(intra-topic validity) and the degree to which the categories are meaningfully connected to one another(inter-topic validity). According to the authors, their approach offers a valid and transparent indication of political interest that is not influenced by the same presumptions and biases as previous measures. They recommend using their topic models as an exploratory tool applied to very large corpora.

[Greene and Cross 2015] the study aims to analyse the political agenda of the European Parliamentary speeches. The authors suggest that whereas past research has concentrated on the voting behaviour of Members of the European Parliament(MEPs), the topics and issues that are argued in the plenary have received less attention. This article seeks to fill this void by revealing the political agendas and interests of the European Parliament. The authors use topic modelling, specifically a non-negative matrix factorization framework, to analyse the content of MEPs debates from the 5th-7th term of the European Parliament. The dataset has 269,696 unique speeches in 24 languages. The researchers opted for English language speeches, which represents 77.95% since multilingual topic modelling or automated translation of documents reduced model accuracy. They divided the corpus into 60 quarterly time windows from 1993-Q3 to 2014-Q2. Standard preprocessing was done as cited in the previous works. The researchers identified 20 main topics and issues that were discussed in the plenary such as economic governance and climate change.

There are numerous topic models that have been developed to aid in content analysis.[Vayansky and Kumar 2020] provides a comprehensive overview of various topic modelling techniques apart from the most common Latent Dirichlet Allocation(LDA) model to optimise and infer parameters from text-based data and their applications, highlighting their strengths and limitations. These techniques include Latent Dirichlet Allocation (LDA), Non-Negative Matrix Factorization (NMF), Latent Semantic Analysis (LSA), Probabilistic Latent Semantic Analysis (pLSA), Correlated Topic Model (CTM), Hierarchical Dirichlet Process (HDP), Time-based models such as Dynamic topic models and continuous-time topic modelling.

LDA is a generative model that assumes each document is a collection of topics, and each topic is a word distribution. The non-negative matrix factorization (NMF) approach divides the document-term matrix into two non-negative matrices. LSA is a dimensionality reduction technique that finds a low-rank approximation to lower the dimensionality of the document-term matrix. CTM models the cross-document correlations and incorporates them into the topic modelling process. HDP is a non-parametric Bayesian model that can infer the number of topics from grouped data.

[Vayansky and Kumar 2020] came with decision tree charts for researchers looking to select a suitable model depending on the available dataset as shown below in figure 1

One critical component of developing topic modelling is evaluation. [Trenquier 2018] addresses the problem of improving the semantic quality of LDA models used in forensic investigations. They sought to answer what is the optimal number of topics for an LDA model, the number of iterations influencing the quality of

Fig. 1. Word cloud of the most popular words

the model and ways to improve the evaluation of semantic quality. The study used a collection of emails. They evaluated their models using 3 coherence measures: U_{MASS}, C_v and $C_{word2vec}$. $C_{word2vec}$ was used as an external model to avoid the bias of the other two coherence measures. They use pre-processing methods to remove mail signatures, dates, address names and HTML-encoded emails. They used the gensim python model to train their LDA model. The study showed that the 3 coherence behaves differently when the number of topics increases and the number of iterations has little significance in the quality of the mode.

3 METHODOLOGY

Development of the NLP solution involved the following steps: web scraping of the dataset, data cleaning, model training and model evaluation.

3.1 Description of the Dataset

This study uses the Hansard of the Kenyan parliament and the senate from 2006 to 2023. The data was scraped from the Mzalendo.com website and saved on a CSV file. The goal of Mzalendo, a non-partisan group that monitors Parliament in Kenya, is to "support the realisation of open, inclusive, and responsible Parliaments across Kenya and Africa.". The dataset has 4 columns as shown in table 1

Table 1. Description of the dataset

Columns	Type of Data	Description
chamber	categorical	represent the houses the Kenyan parliament - the senate and the national assembly
session_date	date	the date of the separate sittings within a calendar year
politician	text	the member of parliament (MP)
speech	text	the respective MP address in the sitting

3.2 Data pre-processing

Data preprocessing is a crucial step in data mining to increase data effectiveness [Verma 2019]. The LDA models are served with a single concatenated string of the parliamentary speeches as the models' only input. The first process in the preprocessing pipeline was removing symbols, numbers and URL links. In the second step, non-English words and stopwords were removed. Stopwords are the most widespread words in any language (such as articles, prepositions, pronouns, conjunctions, etc.), which do not significantly add value to the text's content. The speech text was also converted to lowercase. Stemming and lemmatization is the third step. Lemmatization is a text normalisation technique used to group different forms of the same words[noa 2009]. Lemmatization helps to improve the accuracy of the model.

The fourth step is generating bi-grams and tri-grams. These are two or three phrases that when used together have different meanings [Tan et al. 2002]. The final step is to filter significant words using TF-IDF with a set threshold value.

78 models were generated from a number of topics ranging from 2 to 80 with an interval of 2.For each model, coherence was computed using U_{mass} measure.C_{umass} is based on document co-occurrence counts, a one-preceding segmentation and a logarithmic conditional probability as confirmation measure. The best model has the lowest coherence values based on the local maxima. This also provides the model with the optimal number of topics [Trenquier 2018].

4 RESULTS AND ANALYSIS

Speaker, follow, excellency, and proceed were some of the popular words mentioned in the plenary speeches as shown in figure 2.These are words that are common in plenary debates.

Fig. 2. Word cloud of the most popular words

78 models were trained with a different number of topics ranging from 2 to 80 with an interval of 2. For every model , the coherence scores using C_{mass} were calculated and plotted against the number of topics as shown in Figure 2. The best model with 72 topics had a coherence score of -4.874 which was the lowest of the local minima generated as shown in figure 3

Fig. 3. Plot of coherence scores against the number of topics

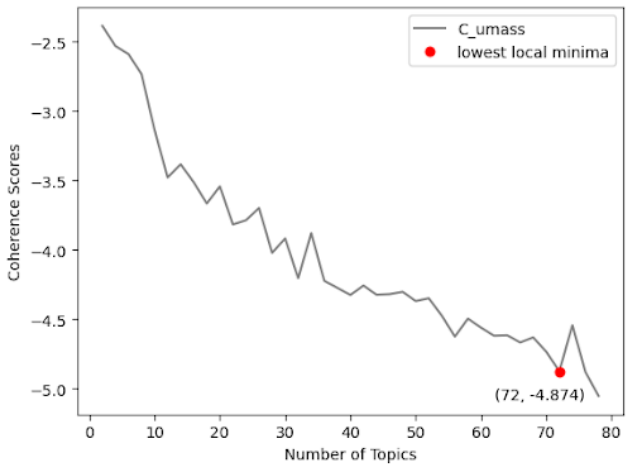

Figure 4 shows the top 4 models racked by coherence score. Topic 0,1,2 and 4 is related to action on legislative clauses, renewable energy, cash crop farming and drought respectively.

5 CONCLUSION

The paper has shown the use of topic modelling to analyse a large corpus of the Kenyan parliamentary debates and identify the main issues and themes discussed to gain a better understanding of the planery process and decision-making of elected officials. The work can be further extended to understand how the topics evolve over time using dynamic topic modelling.

REFERENCES

2009. Introduction to Information Retrieval. https://nlp.stanford.edu/IR-book/html/htmledition/irbook.html

Mark Altaweel, Christopher Bone, and Jesse Abrams. 2019. Documents as data: A content analysis and topic modeling approach for analyzing responses to ecological disturbances. *Ecological Informatics* 51 (May 2019), 82–95. https://doi.org/10.1016/j.ecoinf.2019.02.014

Amy DiPierro. 2018. Generating news tips for local journalism with topic modeling. (2018).

Derek Greene and James P. Cross. 2015. Unveiling the Political Agenda of the European Parliament Plenary: A Topical Analysis. In *Proceedings of the ACM Web Science Conference*. ACM, Oxford United Kingdom, 1–10. https://doi.org/10.1145/2786451.2786464

Jo Guldi. 2019. Parliament's Debates about Infrastructure: An Exercise in Using Dynamic Topic Models to Synthesize Historical Change. *Technology and Culture* 60, 1 (2019), 1–33. https://doi.org/10.1353/tech.2019.0000

N. Kleynhans. 2014. Unsupervised topic modelling on South African parliament audio data. (Nov. 2014). https://researchspace.csir.co.za/dspace/handle/10204/7947 Accepted: 2015-03-12T10:20:02Z Publisher: Pattern Recognition Association of South Africa.

Boris Makarenko. 2015. The Role of Elections in Democracy. In *Democracy in a Russian Mirror*, Adam Przeworski (Ed.). Cambridge University Press, Cambridge, 130–146. https://doi.org/10.1017/CBO9781107282070.008

George Nyabuga. 2023. Devolved Power: A Critical Interrogation of the Place, Roles and Obligations of the Media at the Grassroots in Kenya. *Africa Development* 4 (2023).

Kevin M. Quinn, Burt L. Monroe, Michael Colaresi, Michael H. Crespin, and Dragomir R. Radev. 2010. How to Analyze Political Attention with Minimal Assumptions and Costs. *American Journal of Political Science* 54, 1 (Jan. 2010), 209–228. https://doi.org/10.1111/j.1540-5907.2009.00427.x

Kevin M Quinn, Burt L Monroe, Michael Colaresi, Dragomir R Radev, and Michael H Crespin. 2006. An Automated Method of Topic-Coding Legislative Speech Over Time with Application to the 105th-108th U.S. Senate. (May 2006).

Chade-Meng Tan, Yuan-Fang Wang, and Chan-Do Lee. 2002. The use of bigrams to enhance text categorization. *Information Processing & Management* 38, 4 (July 2002), 529–546. https://doi.org/10.1016/S0306-4573(01)00045-0

Henri Trenquier. 2018. Improving Semantic Quality of Topic Models for Forensic Investigations. (2018).

Ike Vayansky and Sathish A.P. Kumar. 2020. A review of topic modeling methods. *Information Systems* 94 (Dec. 2020), 101582. https://doi.org/10.1016/j.is.2020.101582

Ujjawal Verma. 2019. Data Cleaning and Preprocessing. https://medium.com/analytics-vidhya/data-cleaning-and-preprocessing-a4b751f4066f

Received 24 February 2023

Fig. 4. Top 4 topic for the best model