A COMPARISON OF TOPIC MODELING AND CLASSIFICATION MACHINE LEARNING ALGORITHMS ON LUGANDA DATA

Anonymous authors
Paper under double-blind review

ABSTRACT

Extracting functional themes and topics from a large text corpus manually is usually infeasible. There is a need to build text mining techniques such as topic modeling, which provide a mechanism to infer topics from a corpus of text automatically. This paper discusses topic modeling and topic classification models on Luganda text data. For topic modeling, we considered a Non-negative matrix factorization (NMF) algorithm, and with topic classification, we considered classic approaches, neural networks and pretrained algorithms. BERT and Support vector Machine (SVM) algorithms provide the best results for text classification. Our results indicate that topic modeling and topic classification algorithms produce relatively similar results when topic classification algorithms are trained on a balanced dataset.

Topic modelling,Topic classification,Word embeddings,Luganda

1 INTRODUCTION

In this information age, vast amounts of text data are continuously generated from social media platforms and the world wide web. The extraction of meaningful information from text data requires that data be categorized into various themes and topics. The topics are essential to understand subject domains, generate secondary data for casual discoveries, and in the end, design tools that facilitate decision-making processes.

Social media platforms like Twitter, a microblogging platform, provide a way of topic mining using its 'Hashtag' and 'mention' feature. Although these features also facilitate the discovery of most trending topics on the platform in real-time, this is not sufficient as most trending topics are represented using either a name of an individual, hashtags, or words in other languages. In most cases, it is not easy to understand what the trending topics are about (Lee et al., 2011). Modeling or classifying these topics into general categories with high accuracy and precision remains an important research area in Natural Language Processing (NLP) as it facilitates better information retrieval.

Topic modeling in NLP is an unsupervised machine learning technique that involves detecting word and phrase patterns within a set of sentences or documents and automatically clustering these word groups and similar expressions into topics (Zheng et al., 2021). On the other hand, text classification is a supervised machine learning technique that involves the extraction of features from the data and the use of the extracted features to assign a set of pre-defined labels to open-ended text(Hamed et al., 2020). Previous research in Hamed et al. (2020); Lee et al. (2011) has shown that various efforts have been made in building topic classification and topic modeling models for the English language. However, there has been less focus on low-resourced languages like Luganda, a Bantu language spoken in the African Great Lakes region by more than fifteen million people (of Statistics, 2016).

This research compares topic modeling and topic classification machine learning algorithms on our unique Luganda dataset. Our contribution is threefold: (a) We develop FastText, Glove, Paragram, and word2vec Luganda word embeddings of 300 dimensions, which other NLP researchers can use to train models to understand the characteristics of the Luganda Language. (b)We develop a Luganda topic model using the NMF algorithm to uncover latent topics under study within the Luganda dataset. Topics like Education, Land, COVID-19 and Security emerge with the highest weights and

thus can be easily discovered by our Luganda topic model. (c) We develop topic classification models using classic approaches like neural and pre-trained algorithms. BERT, a pre-trained algorithm, performs better than the other classification algorithms with a precision and recall of 0.97 on the Luganda text data.

The rest of the paper is organised as follows: Section 2 discusses related work; Section 3 discusses the methodology which contains a description of the data, steps taken to preprocess the data, topic modeling and classification approaches used in this study and the evaluation metrics used to assess the performance of the models. Section 4 discusses the results from both the topic modelling and topic classification algorithms. Section 5 concludes the paper.

2 Related Work

Machine Learning approaches have been used in various Natural Language Processing (NLP) tasks including sentiment classification Yoon (2014),statistical machine translation Jacob et al. (2014), text classification Rubungo Andre et al. (2020) and topic modeling(Hamed et al., 2020). Hamed et al. (2020) present an automated extraction of COVID-19 related discussions from social media which leverages topic modeling approaches to uncover various issues related to COVID-19 from public opinions. Lee et al. (2011) present a two scheme based approach: text-based and network-based classification for classifying trending topics in twitter data. The experimental results suggest that Naive Bayes Multinomial classifier using text from trend definition, 100 tweets, and a maximum of 1000 word tokens per category gives the best accuracy of 65.36%. They also discover that some topics could fall under more than one category for example actor's biography would fall under TV, movies and books categories.

Recently, there has been research carried out in NLP for African languages. For example the creating of open source data sets like the: Agriculture keyword dataset for building speech keyword spotter models for Luganda (Mukiibi et al., 2020). The authors in Vukosi & Tshephisho (2020) created a news classification dataset that was limited to headlines. The researchers created a benchmark for research in Setswana and Sepidi, two Bantu South African languages. The authors in Rubungo Andre et al. (2020) created the KINNEWS and KIRNEWS dataset in Kinyarwanda and Kirundi respectively. The dataset was created for multi-class classification on the news articles. These datasets have been used to be used to create benchmark experiments for topic classification (Vukosi & Tshephisho, 2020; Rubungo Andre et al., 2020).

2.1 Topic modeling and Topic classification on low-resourced languages

Vukosi & Tshephisho (2020) discuss work where four machine learning models (Logistic Regression, Support Vector Classification, XGBoost and MLP Neural network) were used to classify news in Setswana and Sepidi. The XGBoost model provided the best results for Sepedi news headline classification. Rubungo Andre et al. (2020) present benchmark experiments on Kinyarwanda and Kirundi news articles using different classic and neural approaches. The classic approaches experimented on the Multinomial Naive Bayes, Logistic Regression and Support Vector Machine with SGD. The neural model approaches that were used in the study included the Character-level Convolutional Neural Networks, Convolutional Neural Network, and Bidirectional Gated Recurrent Unit. In this paper, SVM and BiGRU yielded the best results in the classical and neural approaches respectively.

Zahara (2020) focuses on constructing a viable focused-analysis topic model which learns topics from Twitter data written in a closely related group of non-standardized varieties in Arabic widely spoken in the Levant region.Results of the baseline model and a targeted topic model are compared and designed to serve the purpose of focused analysis. Nasim (2021) combines deep-learning-based language modeling with classic topic modeling techniques to produce interpretable topics for a given set of documents in Urdu, a low resource language. They propose building an accurate Part of Speech tagger for the Urdu language, using semantically rich feature extraction approaches like word2Vec and BERT, they experiment with different clustering and topic modeling techniques to produce a list of potential topics from a given set of documents.

Although the focus on previous work in topic classification and modelling is on low-resourced languages, these languages have different dialects, syntax and lexical semantics. Therefore, it is im-

portant to build a machine learning model that can be trained on specific natural language corpora to capture specific language features. Currently, there is hardly any NLP research involving the Luganda language in classifying and modeling topics from a dataset or a set of documents. In an effort to close this gap, this research explores the application of various machine learning approaches classic, neural and pretrained models for the Luganda dataset.

3 Methodology

3.1 Data

The initial seed of the corpus was retrieved from websites that publish Luganda newspapers. However, since this initial seed corpus was limited in size and topic representation, more data was scraped from English-language websites and newspapers. . This English source text data was then translated to Luganda by a team of Linguists from the Department of African Languages, Makerere University. The data was preprocessed to focus with content on local events in Uganda, and after this process, we remained with 15,000 sentences. While exploring the data, a word cloud was generated to provide a visual representation of the data. District names like Nebbi, Yumbe, Moyo, Koboko were identified as some of the most common words in the data as depicted in Figure 1, which shows that the data was about the local regions in Uganda.

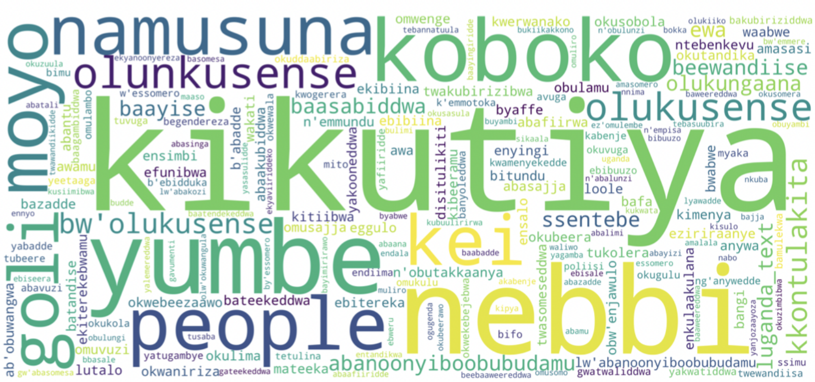

Figure 1: Luganda Wordcloud.

3.2 Data Annotation

We developed annotation guidelines which were followed by annotators who are Luganda Linguists to review and label each sentence with the corresponding topic code based on the sentence subject. Annotators categorized sentences into 14 topics that have been captured in both English and Luganda as depicted in Table1.

The distribution of the Luganda corpora used in this study across the topics is as depicted in Figure 2.

3.3 Data Preprocessing

Luganda is a low-resourced language which implies that the available resources for NLP researchers such as lemmatizers, stemmers, a rich set of stop-words are limited and not available. Data preprocessing involved identifying and removing stop words from the corpora and generating word embeddings.

Table 1: English/Lugand topic names in the dataset and their respective topic codes.

Topic Name	Topic code
Covid (“Kolona”)	Covid
Security (“Ebyokwerinda”)	SE
Agriculture (“Ebyobulimi”)	Agri
Culture (“Ebyobuwangwa”)	C
Transport (“Ebyentambula”)	T
Environment (“Ebyobutonde”)	Env
Politics (“Ebyobufuzi”)	P
Health (“Ebyobulamu”)	H
Religion (“Eddiini”)	R
Sports (“Ebyemizannyo”)	S
Business (“Ebyenfuna”)	B
Land (“Ebyettaka”)	Land
Legal (“Amateeka”)	L
Education (“Ebyenjigiriza”)	Educ

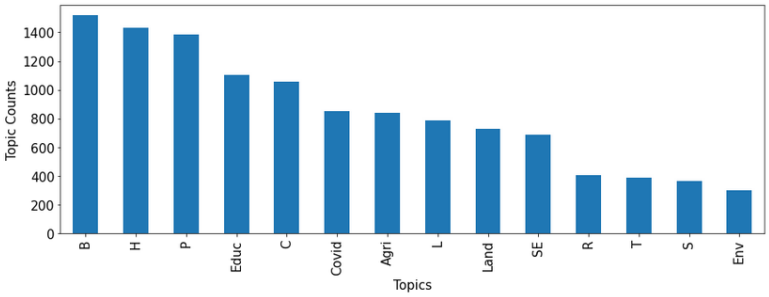

Figure 2: Topic distribution in the dataset.

3.3.1 LUGANDA STOP WORDS

Together with the linguists, we created a list of Luganda stop words as depicted in Table 2. The list was compiled from the Luganda corpora, and this is a resource that other NLP researchers can use as a benchmark while dealing with Luganda data and/or code-mixed.

Table 2: A list of Luganda stop words.

Luganda Stop Words.
a,singa,neera,yenna,nze,ne,ba,nga,wansi
ku,naye,byonna,zonna,bonna,bombi yaffe,liryo
kyaffe,kuwa,nabo, ebyo,yina,ziba,tewali,byabwe,kino
erimu, ye,kyennyini,bya,atya,bokka, mu,nnyigi
ga,bibye,ayinza,nedda,kiki, bo,ekyo,abava,
gumu,gujja,edda,nedda,nze,bingi,nnina,ajja

3.4 TOPIC MODELING

We used the Non-negative Matrix Factorization(NMF) topic modeling technique to detect and discover meaningful topics from the unlabeled monolingual Luganda corpora. In the NMF model, each sentence was considered as a data point. The NMF model operates by decomposing high dimensional vectors into lower-dimensional representations, and these lower-dimensional vectors are non-negative, which means that their coefficients are non-negative as well (Hyun Ah & Soo-Young, 2013). Using the original matrix (A), NMF gives two matrices (W and H) where W is the topics the model found, H is the coefficient(weight) of a topic. Equivalent to this is; A matrix holds records by words, H matrix holds records by topics and W is a representation of topics by words.

The NMF model was trained on features extracted from the TF-IDF vectorizer. The TF-IDF vectors were considered as high dimensional vectors to enable the model to modify the initial values of W and H so that the product approaches A until either approximation error converges or the maximum iterations are reached. Most of the hyper-parameters in the NMF model were set to their default values. However, some parameters were changed such as solver which was set to 'mu' value; max-interations which was set to 1000, alpha set to 0.01, and l1-ratio set to 0.5.

3.5 TOPIC CLASSIFICATION

After annotating the data, the data was unevenly distributed across the different topics as depicted in Figure 1. The "business" topic was the majority class with 1423 data points whereas the "environment" was the minority class with 350 sentences.

Standard classifiers usually get biased towards the minority class (Cristian & Mihaela Elena, 2019). Random oversampling was applied to the training dataset to create a balanced set. With random oversampling, samples of the data are taken from the minority classes randomly and duplicate instances are created so that the minority class reaches a size comparable with the majority class (Cristian & Mihaela Elena, 2019). After resampling the size of the corpus increased to 20922 sentences. However, experiments were done on both the imbalanced set and re-sampled set and the results are presented in the subsequent sections.

3.5.1 WORD EMBEDDINGS

Word embeddings are commonly used in NLP research as real-valued representations because of their ability to capture lexical semantics from the natural language corpora they are trained on (Amir, 2018). According to recent studies, the transfer learning approach has been used for text classification in low-resource languages. This approach involves using the features of high-resource languages, which are learned by pre-trained word embeddings to train models on low-resource languages. However, this technique might not be practical or even applicable in a low-resource setting due to insufficient parallel corpus for both the low-resource and high-resource languages Rubungo Andre et al. (2020). On the other hand, these existing pre-trained models such as Glove, Fast text, Paragram, and word2vec are not applicable for Luganda since they were exclusively trained on high resource languages.

We have developed Luganda word embeddings in this research, i.e., Fasttext, Glove, Paragram, and Word2vec Embeddings on a 300-dimensional input feature from the 15,000 monolingual Luganda dataset. These were trained using the Gensim [1] and Glove [2] frameworks.

3.5.2 TOPIC CLASSIFICATION METHODS

We utilized different classification methods to learn topic classification on the Luganda dataset. These included: baseline models Logistic Regression, Support vector machine, multi-layer perceptron, XGBoost model, neural network models, and pre-trained models.

Baseline Models For the baseline models, the data were transformed into tokens using a count vectorizer. A count vectorizer provides a simple way to tokenize a collection of sentences, build a vocabulary of known words, and also encode new documents using that vocabulary.

Neural Models For all the proposed neural models, As input into the neural network, every sentence was tokenized to produce a vector and then the input sequence was truncated and padded. The padded sentences generated for every sentence in the data were then taken as input into the model to output the different class topics.

BIDIRECTIONAL LSTM WITH 2D MAX POOLING We used Bidirectional LSTM for text classification as proposed by Peng et al. (2016) with default parameters. However, we changed the original feature map to 250 and a min-batch to 64. This model was trained with and without word

[1] https://radimrehurek.com/gensim/models/word2vec.html

[2] https://nlp.stanford.edu/projects/glove/

embeddings in the embedding layer. While training with word embeddings, we used three different word embeddings i.e., Glove, fasttext and word2vec embeddings of the same dimensions (300 dimensions).

GATED RECURRENT UNIT (GRU) We explored a GRU for text classification proposed by Junyoung et al. (2014) with default parameters. An embedding layer was used to input word-embeddings and in another experiment word-embeddings were not used in the first layer, the second layer was the SpatialDropout1D layer with a dropout of 0.2. The GRU layer was used as the fourth layer followed by the fifth layer as the output dense layer.

Pre-trained Models A transformer model architecture was used to implement the pre-trained classification models. The architecture uses a stacked self-attention and point-wise, fully connected layers for both the encoder and decoder Ashish et al. (2017), shown in the left and right parts of Figure 4 respectively. We implemented both the BERT and RoBERTa models using hugging face sentence transformers.

We used an encoder with 6 identical layers in which each layer consists of a multi-head self-attention mechanism and a fully connected feed-forward network. The multi-head attention system and feed-forward network both had a residual connection and a normalization layer.

The decoder consisted of 6 identical layers with an additional sub-layer in each of the 6 layers. The additional sub-layer helps to perform multi-head attention over the output of the encoder stack. An attention mechanism for mapping of a query and a set of key-value pairs to an output, where the query, keys, values, and output are all vectors. BERT breaks the familiar left-to-right setting to use only the encoder part of the transformer architecture.

3.6 EVALUATION METRICS

3.6.1 COHERENCE SCORE

The performance of the NMF model was measured using coherence score. Coherence score measures the relative distance between words in topics. Mainly there two kinds of coherence score: c_v which is the most popular and u_mass. c_v is known for accuracy whereas u_mass is faster. For this study, we used c_v which ranges between 0 and 1 with 1 being the best value where the topics are perfectly coherent.

3.6.2 ACCURACY SCORE

We measured and assessed the performance of the topic classification models based on different evaluation metrics such as Accuracy score, Precision, and Recall These were computed using the values in the confusion matrix that is: True Positives (TP), True Negatives (TN), False Positives (FP), False Negatives (FN), as is depicted in Figures 3 and 4. We evaluated the model using the accuracy, precision and recall and F_scores.

4 RESULTS

Figure 5 show the results from the topic modelling techniques. The results show the meaningful topics generated from the NMF model. Additionally, the top 10 words under each topic are captured in Figure 5. The results show a coherence score for the NMF model is 0.56.

We also present results using the accuracy score evaluation metric for all the classifiers on both the unsampled dataset and resampled dataset as depicted in Table 3 and 4. In the groups of classic models, SVM merges the best performer on both experiments whereas for the pre-trained models BERT outperformed the RoBERTa model.

4.1 DISCUSSION OF RESULTS

As depicted in Figure 7, the model clusters topics: COVID ("Ssenyinga omukambwe"), Land ("Ebyettaka"), Security ("Ebyokwerinda"), and Education ("Ebyenjigirza") under more than one

Table 3: Model Performance before Resampling

Classifier	Accuracy	Precision	Recall
XGB Classifier	0.69	0.76	0.64
MLP classifier	0.75	0.78	0.71
Logistic Regression	0.72	0.79	0.66
SVM	0.76	0.75	0.67
GRU	0.64	0.58	0.57
Bi-LSTM	0.57	0.51	0.47
RoBerta	0.73	0.75	0.72
Bert	0.77	0.75	0.77

Table 4: Model performance after resampling.

Classifier	Accuracy	Precision	Recall
XGB Classifier	0.89	0.89	0.89
MLP classifier	0.90	0.90	0.90
Logistic Regression	0.96	0.96	0.96
SVM	0.97	0.97	0.97
GRU	0.96	0.96	0.96
Bi-LSTM	0.96	0.96	0.96
RoBerta	0.96	0.96	0.96
Bert	0.97	0.97	0.97

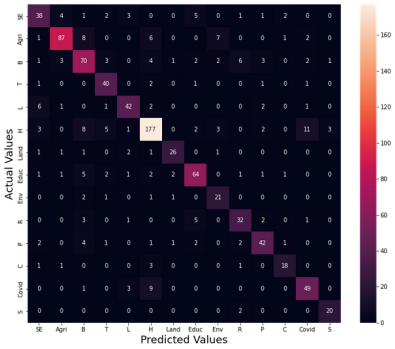

Figure 3: BERT classifier

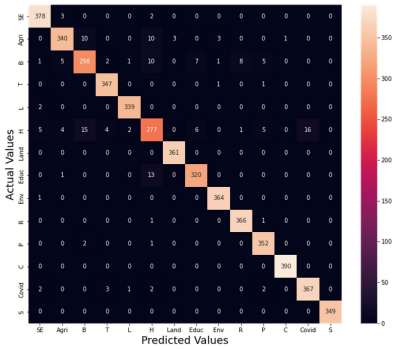

Figure 4: SVM classifier

cluster. This shows the weights of these topics over the other topics but it also shows us more words that are associated with that topic as stored in the W matrix.

The NMF was not able to linearly separate the Health topic and this could be attributed to the fact that some of the keywords generated under the COVID topic such as “ekirwadde”, “obulwadde”, “abalwadde”, “obulamu”, “akawuka” can be used while discussing or referring to either topics but also given the fact that this data was collected during the COVID-19 pandemic, there were more co-occurrences of these words with words like “senyiga”, “okusasana” which were commonly used to refer to COVID-19.

About resampling the training data for topic classification models so as to balance the dataset across all the topics, the performance increases across all models as depicted in Table 4. We observed a significant difference between the two different datasets and thus the sampling technique worked well for topic classification of the Luganda dataset.

The results of the neural models with the Luganda word embeddings that were developed from the dataset using word2vec models are not presented in this paper as they are slightly lower than the results obtained from the default layers of the neural network. This could be attributed to the size of the data on which the embeddings were developed.

Topic 2	Topic 5	Topic 6	Topic 8	Topic 9	Topic 10	Topic 13
akawuka	ettaka	poliisi	ttaka	abayizi	ttiimu	abalimi
kolona	enkaayana	emisango	obukuubagano	ssomero	omupiira	ebirime
abalwadde	ensonga	omusango	enkaayana	abasomesa	abawagizi	beetaaga
abalina	ssentebe	abateeberezebwa	nkaayana	akamalirizo	abazannyi	abalunzi
abalala	lyabwe	okunoonyereza	bwannannyini	bubi	baayo	ebibiina
okusaasaana	zirina	okwekalakaasa	obwannannyini	masomero	mpaka	okulima
kawuka	ekyalo	kaduukulu	bungi	ensoma	ebyemizannyo	balima
omuwendo	lisobola	ekitundu	lyabwe	bajja	ebigere	katale
abasawo	okuwandiisa	okukuuma	okugonjoola	omusomesa	amaanyi	abasinga
obulwadde	mingi	ekwata	mateeka	abali	omutendesi	obwegassi

Topic 14	Topic 15	Topic 16	Topic 17	Topic 18	Topic 19	Topic 20
kkooti	okulonda	kitundu	enjawulo	enguudo	emiti	katonda
emisango	omwaka	obutebenkevu	obuwangwa	embi	obutonde	kkanisa
enkulu	ogujja	obutali	ebifo	obubenje	ensi	omuntu
mingi	pulezidenti	mugaso	amawanga	zirina	okutema	obulamu
musango	bwenkanya	emirembe	bulina	nguudo	okusimba	alina
yasindikiddwa	ennaku	obuwangwa	enzikiriza	ennungi	okukuuma	abakrisito
guno	omuntu	amagye	byobuwangwa	nnyingi	ensigo	okubeera
ensala	kwaliwo	butebenkevu	ngeri	mbeera	kutema	abasumba
obujulizi	okuba	entambula	ebintu	ensimbi	kyonoona	kisa
mateeka	kujja	ebyokwerinda	bingi	entambula	batema	

Topic 21	Topic 22	Topic 24	Topic 25
abaana	essomero	bangi	ssente
okusoma	omukulu	ekirwadde	okukola
abazadde	ebibiina	ssenyiga	bizinensi
ssomero	bbaasi	omukambwe	okufuna
masomero	abasomesa	abavubuka	pulojekiti
bateekeddwa	abalala	obulamu	nnyingi
beetaaga	okuzimba	bafudde	oluguudo
amasomero	omukulembeze	ggwanga	emirimu
engeri	ebisale	bwabwe	ekitundu
abato	amazzi	ababbi	kkampuni

Figure 5: Latent topics and the top 10 words derived by the topic modeling model.

To better assess the models both the classifiers and the NMF model but also to understand the error set(points which are off the main diagonal in the confusion matrix) we downloaded more Luganda data from a website and used this to test the best classifiers and the NMF model, as presented in Figure 5. From that set we observed that for some instances the NMF model prediction deviates from that of the classifiers and also the actual label. And this is because of the duality issue, where the sentences can belong to more than one topic. Figure 6 shows some sample test results from the models. A sentence such as *"Omusirikale wa Poliisi asindikiddwa mu kkomera okumala emyezi mukaaga"*. translated as: *A police officer has been sentenced to jail for six months* in English can belong to either the *Legal topic* or *Security topic*.

5 Conclusion and Future Work

This research presents the first topic classification and modeling benchmark for Luganda, a low resource language and the most common native language in Uganda. The performance of the models builds a confidence that we can develop downstream models in low-resourced languages which can be used in different applications. Considering the gap between high resource languages and low resource languages, our future work will look at enriching the dataset to address the issue of duality in topic classification and also use the data to develop better word embeddings for the language.

References

Bakarov Amir. A survey of word embeddings evaluation methods. 2018.

Vaswani Ashish, Shazeer Noam, Parmar Niki, Uszkoreit Jakob, Jones Llion, Gomez Aidan N, Kaiser Łukasz, and Polosukhin Illia. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Padurariu Cristian and Breaban Mihaela Elena. Dealing with data imbalance in text classification. *Procedia Computer Science*, 159:736–745, 2019.

English Sentence	Luganda Sentence	Actual Label	BERT	SVM	NMF
Youth are urged to have medical check-ups early enough.	Abavubuka bakubirizibwa okwekebeza nga bukyali.	H	H	H	Other
A police officer has been sentenced to jail for six months.	Omusirikale wa Poliisi asindikiddwa mu kkomera okumala emyezi mukaaga.	L	L	L	SE
A lawyer has been sued by the court for three cases.	Kkooti eggudde emisango esatu ku munnamateeka.	L	L	L	L
The police has warned people who abuse other people rights while pretending to exercise there rights.	Poliisi erabudde abantu abatyoboola eddembe ly'abantu nga abeerimbika mu kutuukiriza obuvunaanyizibwa bwabwe.	SE	SE	SE	SE
Youth are urged not to sell off land.	Abavubuka bakubirizibwa obuteetundako ttaka lyabwe.	Land	Land	Land	Land
What is in the deal of chasing bodaboda riders out of the city?	Biki ebiri mu ddiiru y'okugoba abagoba ba boodabooda mu kibuga?	T	T	T	B
Let's go and play football.	Tugende tuzannye omupiira.	S	S	S	S
Those who teach students illegally will be punished.	Aba abasomesa abayizi mu bumenyi bw'amateeka baakukangavvulwa.	Educ	Educ	Educ	Educ
He urged the public to go for corona vaccination.	Yakubirizza abantu okugenda bagemebwe akawuka ka Kkolona.	Covid	Covid	Covid	Covid
He went to his garden to dig.	Yagenze mu nnimiro ye okulima.	Agric	Agric	Agric	Agric

Figure 6: Sample test set results from SVM and BERT model versus the NMF model.

Jelodar Hamed, Wang Yongli, Orji Rita, and Huang Shucheng. Deep sentiment classification and topic discovery on novel coronavirus or covid-19 online discussions: Nlp using lstm recurrent neural network approach. *IEEE Journal of Biomedical and Health Informatics*, 24(10):2733–2742, 2020.

Song Hyun Ah and Lee Soo-Young. Hierarchical representation using nmf. *In International conference on neural information processing Springer, Berlin, Heidelberg,*, 159:466–473, 2013.

Devlin Jacob, Zhongqiang Huang Rabih Zbib, Lamar Thomas, Schwartz Richard, and Makhoul John. Fast and robust neural network joint models for statistical machine translation. *In proceedings of the 52nd annual meeting of the Association for Computational Linguistics*, 1: Long Papers:1370–1380, 2014.

Chung Junyoung, Gulcehre Caglar, Cho KyungHyun, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. 2014.

Kathy Lee, Diana Palsetia, Ramanathan Narayanan, Md. Mostofa Ali Patwary, Ankit Agrawal, and Alok Choudhary. Twitter trending topic classification. In *2011 IEEE 11th International Conference on Data Mining Workshops*, pp. 251–258, 2011. doi: 10.1109/ICDMW.2011.171.

Jonathan Mukiibi, Claire Babirye, Andrew Katumba, and Joyce. Nakatumba. Agriculture keywords dataset (version one) [data set]. zenodo. `https://doi.org/10.5281/zenodo.4347308`, 2020.

Zarmeen Nasim. On building an interpretable topic modeling approach for the urdu language. 2021.

Uganda Bureau of Statistics. The national population and housing census 2014 - main report. 2016.

Zhou Peng, Qi Zhenyu, Zheng Suncong, Xu Jiaming, Bao Hongyun, and Bo. Xu. Text classification improved by integrating bidirectional lstm with two-dimensional max pooling. 2016.

Niyongabo Rubungo Andre, Qu Hong, Kreutzer Julia, and Huang Li. Kinnews and kirnews: Benchmarking cross-lingual text classification for kinyarwanda and kirundi. 2020.

Marivate Vukosi and Sefara Tshephisho. Improving short text classification through global augmentation method. *In International Cross-Domain Conference for Machine Learning and Knowledge Extraction*, 18:385–399, August 2020.

Kim Yoon. Convolutional neural networks for sentence classification. *In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1746–1751, 2014.

Shorouq Zahara. Targeted topic modeling for levantine arabic. 2020.

Fang Zheng, He Yulan, and Procter Rob. A query-driven topic model. 2021.

6 APPENDIX

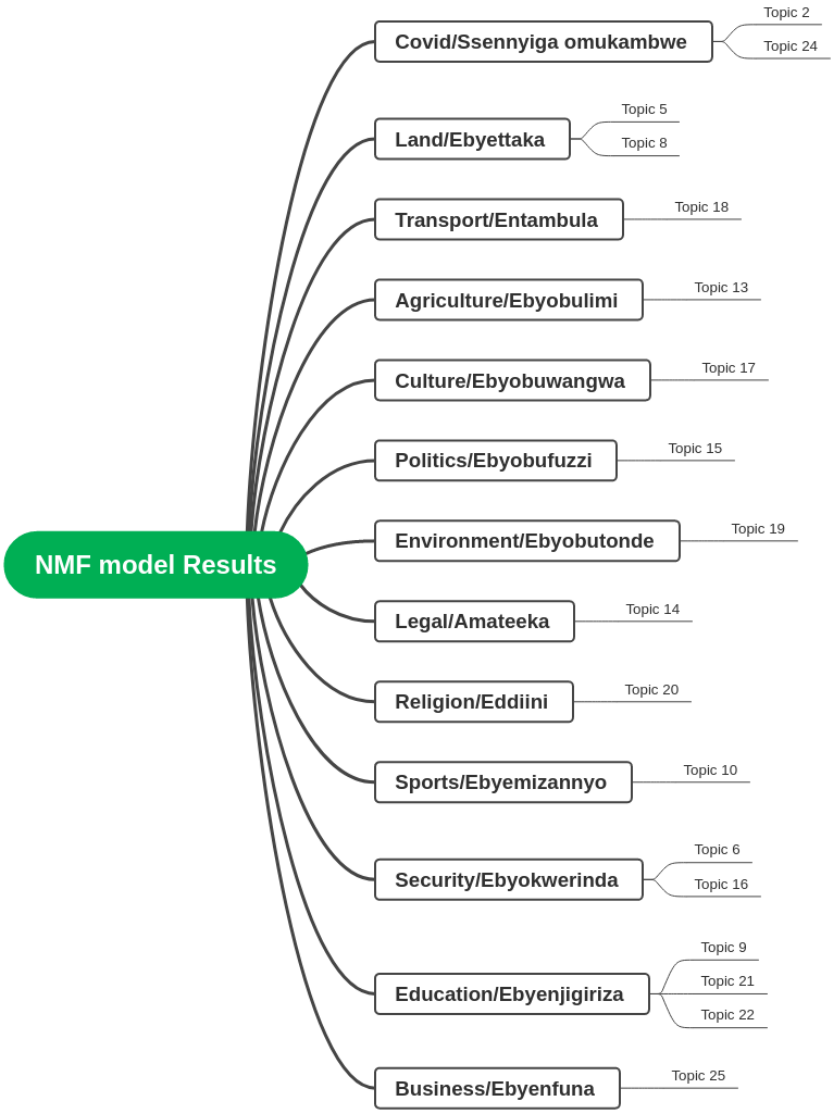

Figure 7: A dendrogram depicting latent topics generated from the NMF model.