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

Tobius Saul Bateesa, Claire Babirye, Joyce Nakatumba-Nabende
Department of Computer Science
Makerere University
Uganda, Kampala
tobiusaolo21@gmail.com, clarybits68@gmail.com, joyce.nabende@mak.ac.ug

Andrew Katumba
Department of Electrical and Computer Engineering
Makerere University
Uganda, Kampala
andrew.katumba@mak.ac.ug

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) which is an unsupervised machine learning algorithm that extracts hidden patterns from unlabeled text data to create latent topics, and for topic classification, we considered classic approaches, neural networks, and pretrained algorithms. The Bidirectional Encoder Representations from Transformers(BERT), a pretrained model that uses an attention mechanism that learns contextual relations between words (or sub-words) in a text, and a Support Vector Machine (SVM) algorithm, a classic model which analyzes particular properties of learning within text data, record the best results for topic 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 (UBOS, 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 50 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 Luganda 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 of 0.9958 and a recall of 0.996 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(Bi-GRU). Rubungo Andre et al. (2020), shows that Support Vector Machine(SVM) and Bidirectional Gated Recurrent Unit(Bi-GRU) 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 important 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 both supervised (classic, neural and pretrained models) and unsupervised(NMF) on the Luganda dataset.

3 Methodology

3.1 Data

A Corpus of 15000 English sentences was retrieved from websites and newspapers that publish articles in Uganda. This English text data was sub divided into sets of 100 sentences and distributed for translation among a team of Luganda Linguists from the Department of African Languages Makerere University. 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.

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

Topic Name	Topic code
COVID-19 (“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

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

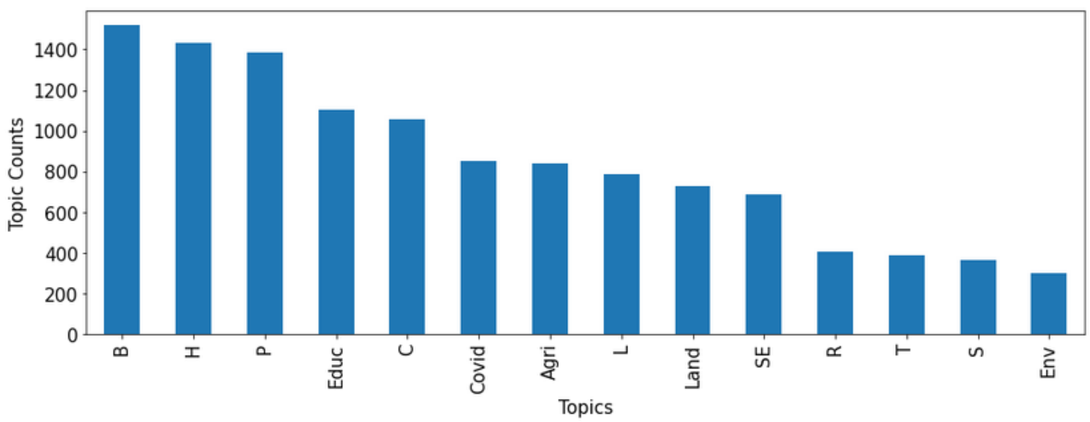

Figure 2: Topic distribution in the dataset.

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.

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.

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,

Table 2: A list of 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

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 2. 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 majority class (Cristian & Mihaela Elena, 2019). Random oversampling was applied to the training dataset to create a balanced set across all the topics. With random oversampling, samples of the data are taken from the minority classes(business class) 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 by 30% to make a dataset of 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 50-dimensional input feature from the 15,000 monolingual Luganda dataset. These were trained using the Gensim [1] and the Glove framework[2].

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

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

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 was 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. All the baseline models above were implemented with the help of the scikit-learn framework [3]. The dataset was split into: training set and testing set, with a ratio of 9:1.

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. The dataset was split into: training set, validation set and testing set, with a ratio of 8:1:1.

BIDIRECTIONAL LSTM WITH 2D MAX POOLING We used Bidirectional LSTM for text classification as proposed by Peng et al. (2016) with default hyperparameters. However, we changed the original feature map to 250 and a min-batch to 64 since we were training the model on a small dataset. This model was trained with and without word 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 (50 dimensions).

GATED RECURRENT UNIT (GRU) We explored a GRU for text classification proposed by Junyoung et al. (2014) with default hyperparameters. 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 Spatial 1 Dimension of Dropout(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 We used the BERT and RoBERTa transformer models which are pretrained on English data for text classification from the Hugging Face platform [4]. As proposed by Ashish et al. (2017) we retrained both the BERT and RoBERTa models using a Luganda dataset to implement Luganda text classification using the above mentioned models. The dataset was split into: training set, validation set and testing set, with a ratio of 8:1:1.

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 are two kinds of coherence scores: CV Coherence and UMass. CV Coherence, creates content vectors of words using their co-occurrences and, after that, calculates the score using normalized pointwise mutual information (NPMI) and the cosine similarity. UMass calculates how often two words appear together in the corpus. For this study, we used CV Coherence which ranges between 0 and 1 with 1 being the best value where the topics are perfectly coherent.

3.6.2 MODEL RESULTS

We measured and assessed the performance of the topic classification models based on different evaluation metrics such as F1 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 depicted in Figures 3 and 4.

[3]https://scikit-learn.org/stable/

[4]https://huggingface.co/docs/transformers

4 RESULTS

Figure 5 in the Appendix section show the results from the NMF model. The results include the meaningful topics generated from the model. Additionally, the top 10 words under each topic are captured in Table 3. We achieved a coherence score of 0.56 on evaluating the NMF model.

Topic 2	**Topic 5**	**Topic 6**	**Topic 8**	**Topic 9**	**Topic 10**
akawuka	ettaka	poliisi	ttaka	abayizi	ttiimu
kolona	enkaayana	emisango	obukuubagano	ssomero	omupiira
abalwadde	ensonga	omusango	enkaayana	abasomesa	abawagizi
abalina	ssentebe	abateeberezebwa	nkaayana	akamalirizo	abazannyi
abalala	lyabwe	okunoonyereza	bwannannyini	bubi	baayo
okusaasaana	zirina	okwekalakaasa	obwannannyini	masomero	mpaka
kawuka	ekyalo	kaduukulu	bungi	ensoma	ebyemizannyo
omuwendo	lisobola	ekitundu	lyabwe	bajja	ebigere
abasawo	okuwandiisa	okukuuma	okugonjoola	omusomesa	amaanyi
obulwadde	mingi	ekwata	mateeka	abali	omutendesi
Topic 13	**Topic 14**	**Topic 15**	**Topic 16**	**Topic 17**	**Topic 18**
abalimi	kkooti	okulonda	kitundu	enjawulo	enguudo
ebirime	emisango	omwaka	obutebenkevu	obuwangwa	embi
beetaaga	enkulu	ogujja	obutali	ebifo	obubenje
abalunzi	mingi	pulezidenti	mugaso	amawanga	zirina
ebibiina	musango	bwenkanya	emirembe	bulina	nguudo
okulima	yasindikiddwa	ennaku	obuwangwa	enzikiriza	ennungi
balima	guno	omuntu	amagye	byobuwangwa	nnyingi
katale	ensala	kwaliwo	butebenkevu	ngeri	mbeera
abasinga	obujulizi	okuba	entambula	ebintu	ensimbi
obwegassi	mateeka	kujja	ebyokwerinda	bingi	entambula
Topic 19	**Topic 20**	**Topic 21**	**Topic 22**	**Topic 24**	**Topic 25**
emiti	katonda	abaana	essomero	bangi	ssente
obutonde	kkanisa	okusoma	omukulu	ekirwadde	okukola
ensi	omuntu	abazadde	ebibiina	ssenyiga	bizinensi
okutema	obulamu	ssomero	bbaasi	omukambwe	okufuna
okusimba	alina	masomero	abasomesa	abavubuka	pulojekiti
okukuuma	abakrisito	bateekeddwa	abalala	obulamu	nnyingi
ensigo	okubeera	beetaaga	okuzimba	bafudde	oluguudo
butebenkevu	ngeri	amasomero	omukulembeze	ggwanga	emirimu
kyonoona	kisa	engeri	ebisale	bwabwe	ekitundu
batema	okuweereza	abato	amazzi	ababbi	kkampuni

Table 3: Latent topics and the top 10 words derived by the topic modeling model.

We also present results from the topic classification experiments using the evaluation metrics i.e., F1 score, precision and recall for all the classifiers used in this study on both the unsampled dataset and resampled dataset as depicted in Tables 4 and 5 respectively. In the groups of classic models, SVM emerges the best performer on both the unsampled and resampled datasets whereas for the pre-trained models BERT performed better than the RoBERTa model.

4.1 DISCUSSION OF RESULTS

As depicted in Figure 5, the model clusters topics: COVID ("Ssenyinga omukambwe"), Land ("Ebyettaka"), Security ("Ebyokwerinda"), and Education ("Ebyenjigirza") under more than one 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 Luganda topic model 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" are also used while discussing or referring to the Health topic but also given the fact that this data was collected during the COVID-19 pandemic,

Table 4: Model Performance before data resampling.

Classifier	f1 Score	Precision	Recall
XGB Classifier	0.584	0.617	0.59
MLP classifier	0.610	0.676	0.0.61
Logistic Regression	0.570	0.662	0.561
SVM	0.59	0.752	0.752
GRU	0.645	0.643	0.654
Bi-LSTM	0.574	0.511	0.575
RoBERTa	0.731	0.759	0.726
BERT	0.77	0.747	0.766

Table 5: Model performance after data resampling.

Classifier	f1 score	Precision	Recall
XGB Classifier	0.908	0.915	0.906
MLP classifier	0.975	0.9759	0.975
Logistic Regression	0.9708	0.972	0.971
SVM	0.979	0.979	0.979
GRU	0.963	0.964	0.965
Bi-LSTM	0.965	0.964	0.965
RoBERTa	0.958	0.959	0.958
BERT	**0.996**	0.9958	0.996

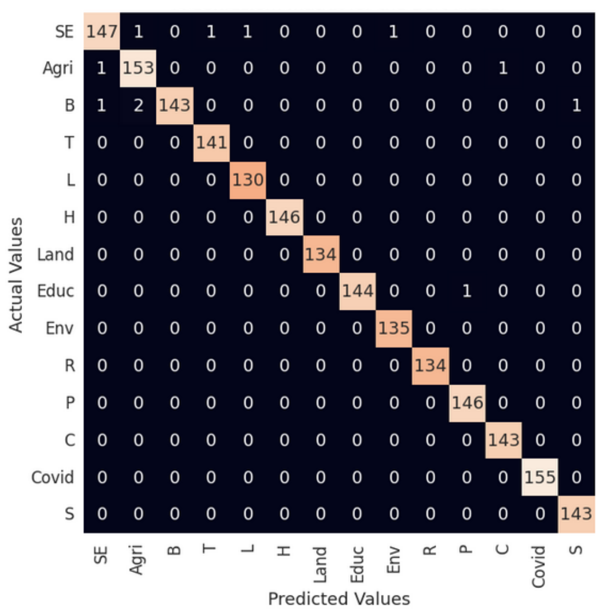

Figure 3: BERT classifier

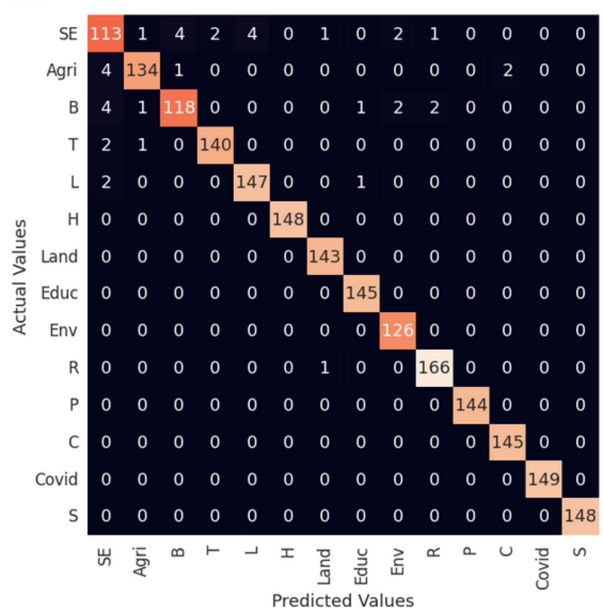

Figure 4: SVM classifier

there were more co-occurrences of these words with words like “senyiga”, “okusasana” which were commonly used to refer to COVID-19.

On resampling the training data for topic classification, the performance for all the models increased as depicted in Table 5. There was a significant difference in the performance of the models between the two datasets(unsampled and resampled) 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.

ENGLISH SENTENCE	**LUGANDA SENTENCE**	**ACTUAL LABEL**	**PREDICTED LABEL**		
			BERT	SVM	NMF
Youth are urged to have medical check-ups early enough	Abavubuka bakubiribiwa 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 omukaaga	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's rights while pretending to exercise there rights	Poliisi erabudde abeerimbika mu ddembe lyabwe okuvvoola erya abalala	SE	SE	SE	SE
Youth are urged not to sell off land	Abavubuka basabiddwa obuteetundako ttaka	Land	Land	Land	Land
What comprise of the deal to send bodaboda riders out of the city?	Ebiri mu ddiiru egoba takisi ne boda boda mu Kibuga	T	T	T	B
Lets go and play football	Ka tugende tuzannye omupiira	S	S	S	S
The teacher reached earlier in class than us	Omusomesa yetusoose mukibiina	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 kulima	Agric	Agric	Agric	Agric

Table 6: Sample test set results from the BERT, SVM and NMF models.

To better assess the models both the Luganda topic classification models and the Luganda topic model, but also to understand the error set (points that are off the main diagonal in the confusion matrices shown in figure 3 and figure 4) we used a sample set of ten new Luganda sentences to test and compare the results of the deployed best Luganda topic classification models and the Luganda topic model, as presented in Table 6. From that comparison, we observed that for some instances the Luganda topic model prediction deviates from that of the Luganda topic classification models and also the actual label. This was attributed to the issue of duality, a scenario where a sentence can belong to more than one topic. 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*. Since the models were trained on a single topic classification problem, in a scenario where the provided sentence belongs to more than one topic, the models can only provide one of the topics as the result.

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.

ACKNOWLEDGMENTS

This work was carried out with support from Lacuna Fund, an initiative cofounded by The Rockefeller Foundation, Google.org, and Canada's International Development Research Centre.

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.

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.

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.

UBOS. The national population and housing census 2014 - main report. 2016.

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

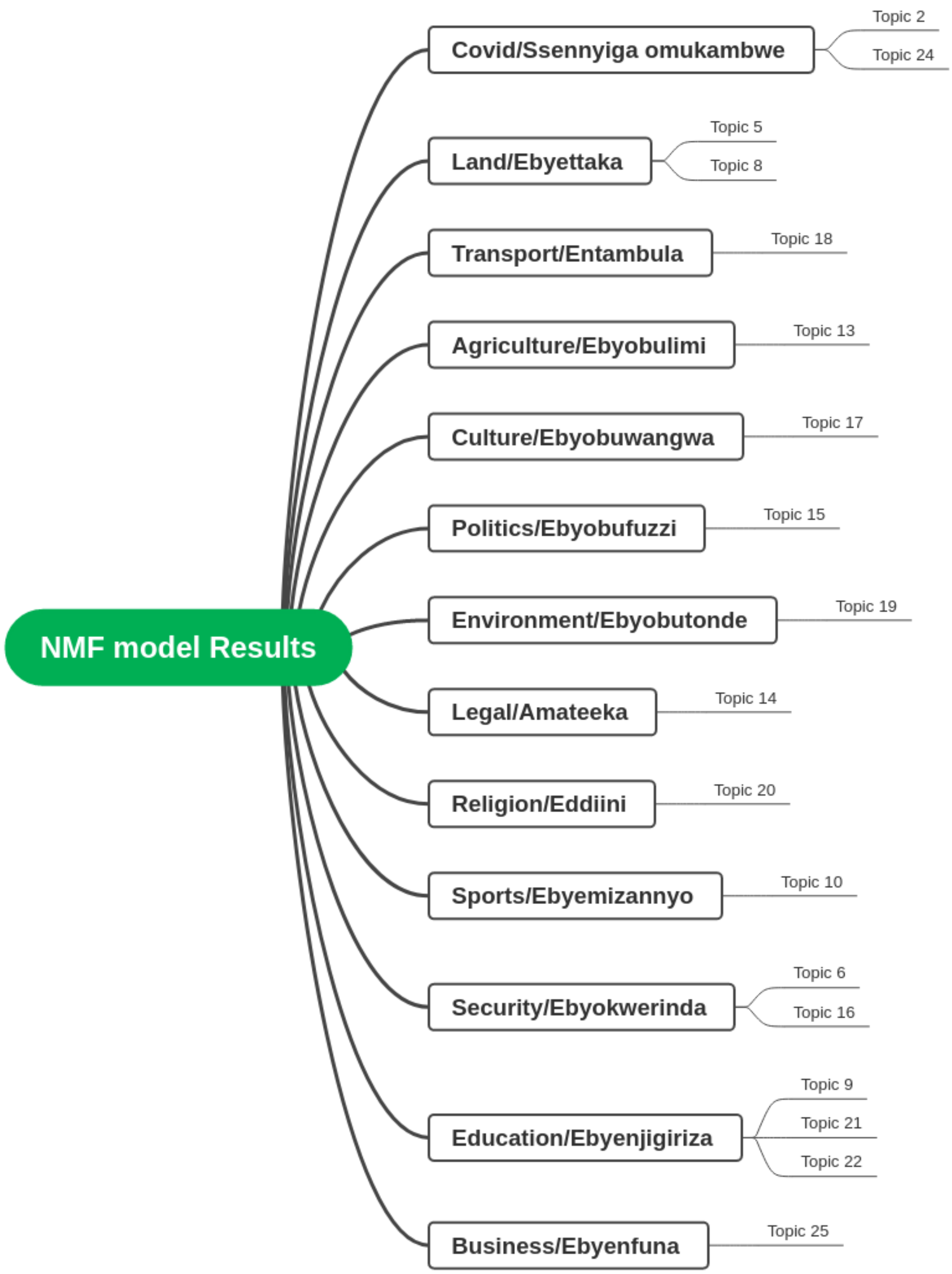

Figure 5: A dendrogram depicting latent topics generated from the NMF model.