IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

CNN BASED CLASSIFICATION OF NEWSTEXTS USING DOC2VEC MODEL

GONTHINA KAVYA [*1], MR. PRASANNA KUMAR LAKINENI [*2]

[*1]M. Tech Scholar, Department of Computer Science & Engineering,

[*2]Associate Professor, Department of Computer Science & Engineering,

DADI Institute of Engineering and Technology, (Affiliated to Jawaharlal Nehru Technological University, Kakinada), NH-16,Anakapalle,531002.

ABSTRACT

The rapid increment in internet usage has also resulted in bulk generation of text data . Therefore, investigation of new techniques for automatic classification of textual content is needed as manually managing unstructured text is challenging. The main objective of text classification is to train a model such that it should place an unseen text into correct category. In this study, text classification was performed using the Doc2vec word embedding method on the Turkish Text Classification 3600 (TTC-3600) dataset consisting of Turkish news texts and the BBC-News dataset consisting of English news texts. As the classification method, deep learning- based CNN and traditional machine learning classification methods Gauss Naive Bayes (GNB), Random Forest (RF), Naive Bayes (NB) and Support Vector Machine (SVM) are used. In the proposed model, the highest result was obtained as 94.17% in the Turkish dataset and 96.41% in the English dataset in the classification made with CNN.

KEY WORDS:

Turkish Text Classification, Support Vector Machine, Naive Bayes, Random Forest, Gauss Naive Bayes.

1. INTRODUCTION

Technology has a significant impact on society and has significantly changed the way people access information. News is a well-known and standard service. Recent technological advancements have considerably changed the way news is produced, consumed, and disseminated. It has enabled more frequent and on-spot news reporting that smartphones can access anywhere and anytime. Therefore, people now expect to receive news of their interest in real-time. The news sources are already flooded with colossal information. Therefore, it is essential to automatically classify the news in specific categories based on the information content to allow timely and efficient information dissemination. Automatic Document Classification (ATC) can be used to efficiently manage text based information (i.e. news) [1-3]. Itallows timely and efficient information retrieval in the search phase.

Automatic Document Classification can assign a relevant category to a news from a predefined set of reference categories based on the text feature extraction by correctly understanding the meaning and context of words. The time required to categorize the news correctly is directly proportional to the quantity of the text. In the newspaper's archive, the comprehensive range of articles varies from business to technology, so it is inconceivable that humans could manage this abundant content of information in a reasonable time frame. Manual document classification is cumbersome and resource-exhaustive. The news category predictor aims to recognize and categorize different articles based on content/information type. The automatic news classification plays a vital role in processing a massive amount of articles. It can classify and label the news articles by analyzing the content (i.e. extracting feature values) to quickly access where they are focused in, allowing efficient and speedy news dissemination. Additionally, news websites can also increase their visibility by developing a recommendation system that suggests/recommends relevant news to attract more attention[12].

2. LITERATURE SURVEY

In this section we will mainly discuss about the background work that is carried out in order to prove the performance of our proposed Method. Literature survey is the most important step in software development process. For any software or application development, this step plays a very crucial role by determining the several factors like time, money, effort, lines of code and company strength. Once all these several factors are satisfied, then we need to determine which operating system and language used for developing the application. Once the programmers start building the application, they will first observe what are the pre-defined inventions that are done on same concept and then they will try to design the task in some innovated manner.

MOTIVATION

The main knowledge which is required to design this current application is text mining which is nothing but the process of dividing the text into multiple parts and where each and every individual part need to be saved in a separate array items.The term text mining, sometimes also referred to as text data mining, roughly equivalent to text analytics, refers to the process of deriving high-quality information from text data se[5]-[8]t. Normally from statistical pattern learning we mainly try to derive the patterns and trends that are calculated from the high quality of information. As we all know that the process of extracting or structuring the input text by identifying the main features and removing the unrelated data from that main document and finally convert the document in a structured way is known as text mining [10]. In this paper we use a word like high quality information, which is typically derived through a set of patterns and trends that are used in pattern learning applications. Also the term high quality in this paper clearly states that combination of some relevance and interestingness for the topic that was available in that conversation file. In the primitive text mining ,there are a possibilities like to scan a set of documents written in a natural language and either model the document set for predictive classification purposes or populate a database or search index with the information extracted

3. EXISTING METHODOLOGY

In the existing system there was no proper method to identify the text news prediction using any well known algorithms. All the news classification is done under manual approach and hence it is very tough for one to classify the news under manual approach.The following are the main limitations in the existing system.

LIMITATION OF EXISTING SYSTEM

1. More Time Delay in finding the news information from text message
2. All the primitive methods failed to predict the accurate news
3. All the primate methods failed to achieve high accuracy.
4. The primitive methods unable to classify for bulk amount of data.

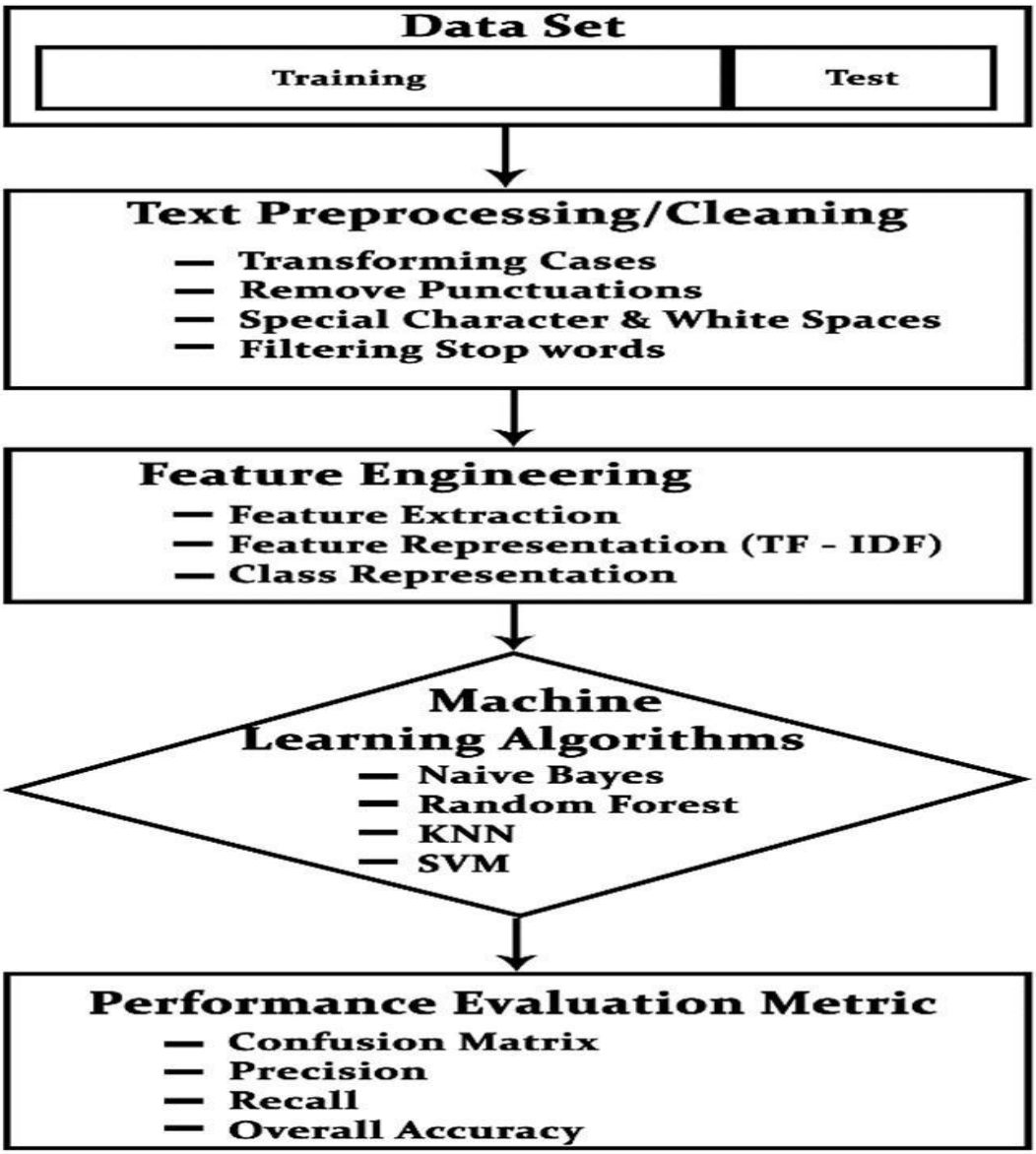
4. PROPOSED SYSTEM

In proposed system we are applying Doc2Vec model for predicting the news from text messages. This is one of the best model in which text classification is easily done.

ADVANTAGES OF PROPOSED SYSTEM

1. It is the simplest process to know the testing results in a few minutes.
2. It is very accurate in generating the result.
3. The proposed model can train the data under more accuracy.

4. The proposed model can train for large amount of data rather than small amount of data



5. PROPOSED METHODOLOGY

Implementation is the stage where the theoretical design is converted into programmatically manner. In this stage we will divide the application into a number of modules and then coded for deployment..The application is divided mainly into following 5 modules. They are as follows:

1. Load Dataset Module
2. Visualize the Data
3. Data Pre-Processing
4. Train the Model Using Several ML Algorithms
5. Find the Performance of ML Algorithms

1. LOAD DATASET MODULE

In this module we will try to load the dataset which is collected from KAGGLE website and once the dataset is downloaded from kaggle, we will give that dataset as input for the system.

```
[ ] from google.colab import files
    files.upload()

    Choose Files  No file chosen    Upload widget is only available when the cell has been executed in the current browser session. Please rerun this cell to enable.
    Saving kaggle.json to kaggle.json
    {'kaggle.json': b'{"username":"b131258","key":"5cc292bc58798bdbf958636f8f9ed5bd"}'}

[ ] ! pip install -q kaggle

[ ] !mkdir ~/.kaggle

[ ] !cp kaggle.json ~/.kaggle

[ ] !chmod 600 ~/.kaggle/kaggle.json

    ! kaggle datasets download -d rmisra/news-category-dataset

    Downloading news-category-dataset.zip to /content
     35% 9.00M/25.4M [00:00<00:00, 37.9MB/s]
    100% 25.4M/25.4M [00:00<00:00, 83.6MB/s]
```

2) VISUALIZE THE DATA MODULE

In this module we try to load the dataset which is downloaded from a pre-defined location:

```
[ ] import pandas as pd
    import numpy as np
    import matplotlib.pyplot as plt
    import seaborn as sns
```

```
[ ] df=pd.read_json('News_Category_Dataset_v2.json',lines=True)
    df.head()
```

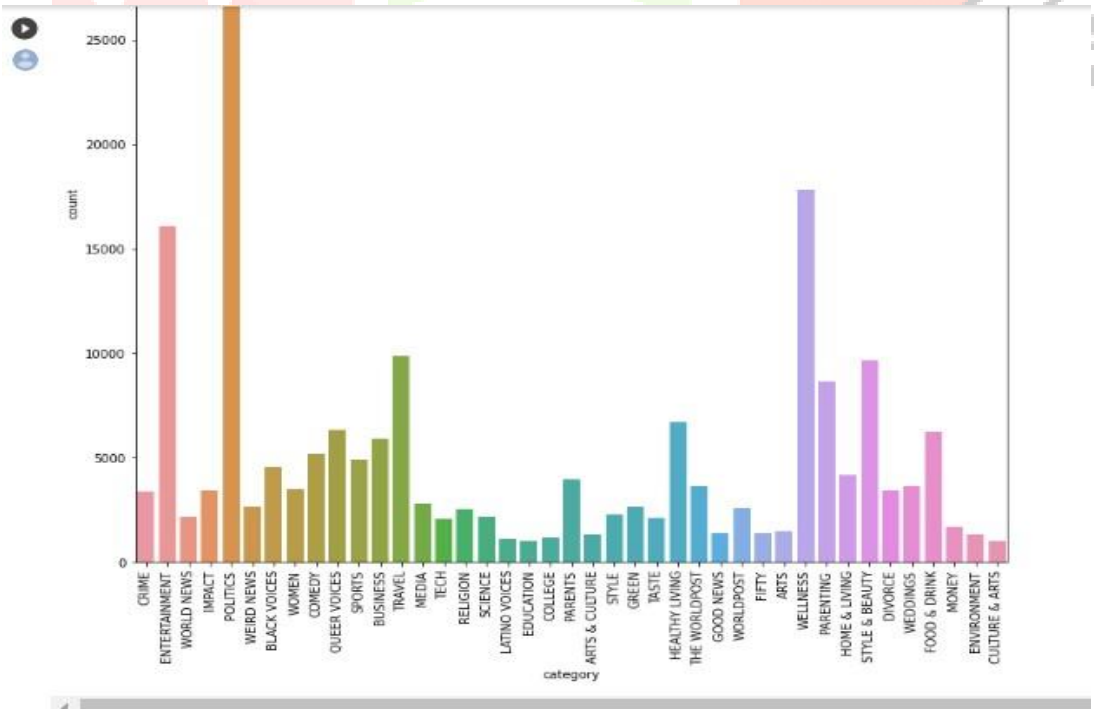
	category	headline	authors	link	short_description	date
0	CRIME	There Were 2 Mass Shootings In Texas Last Week...	Melissa Jeltsen	https://www.huffingtonpost.com/entry/texas-ama...	She left her husband. He killed their children...	2018-05-26
1	ENTERTAINMENT	Will Smith Joins Diplo And Nicky Jam For The 2...	Andy McDonald	https://www.huffingtonpost.com/entry/will-smit...	Of course it has a song.	2018-05-26
2	ENTERTAINMENT	Hugh Grant Marries For The First Time At Age 57	Ron Dicker	https://www.huffingtonpost.com/entry/hugh-gran...	The actor and his longtime girlfriend Anna Ebe...	2018-05-26
3	ENTERTAINMENT	Jim Carrey Blasts 'Castrato' Adam Schiff And D...	Ron Dicker	https://www.huffingtonpost.com/entry/jim-carre...	The actor gives Dems an ass-kicking for not fi...	2018-05-26
4	ENTERTAINMENT	Julianna Margulies Uses Donald Trump Poop Bags...	Ron Dicker	https://www.huffingtonpost.com/entry/julianna-...	The "Dietland" actress said using the bags is ...	2018-05-26

```
df.category.value_counts()
```

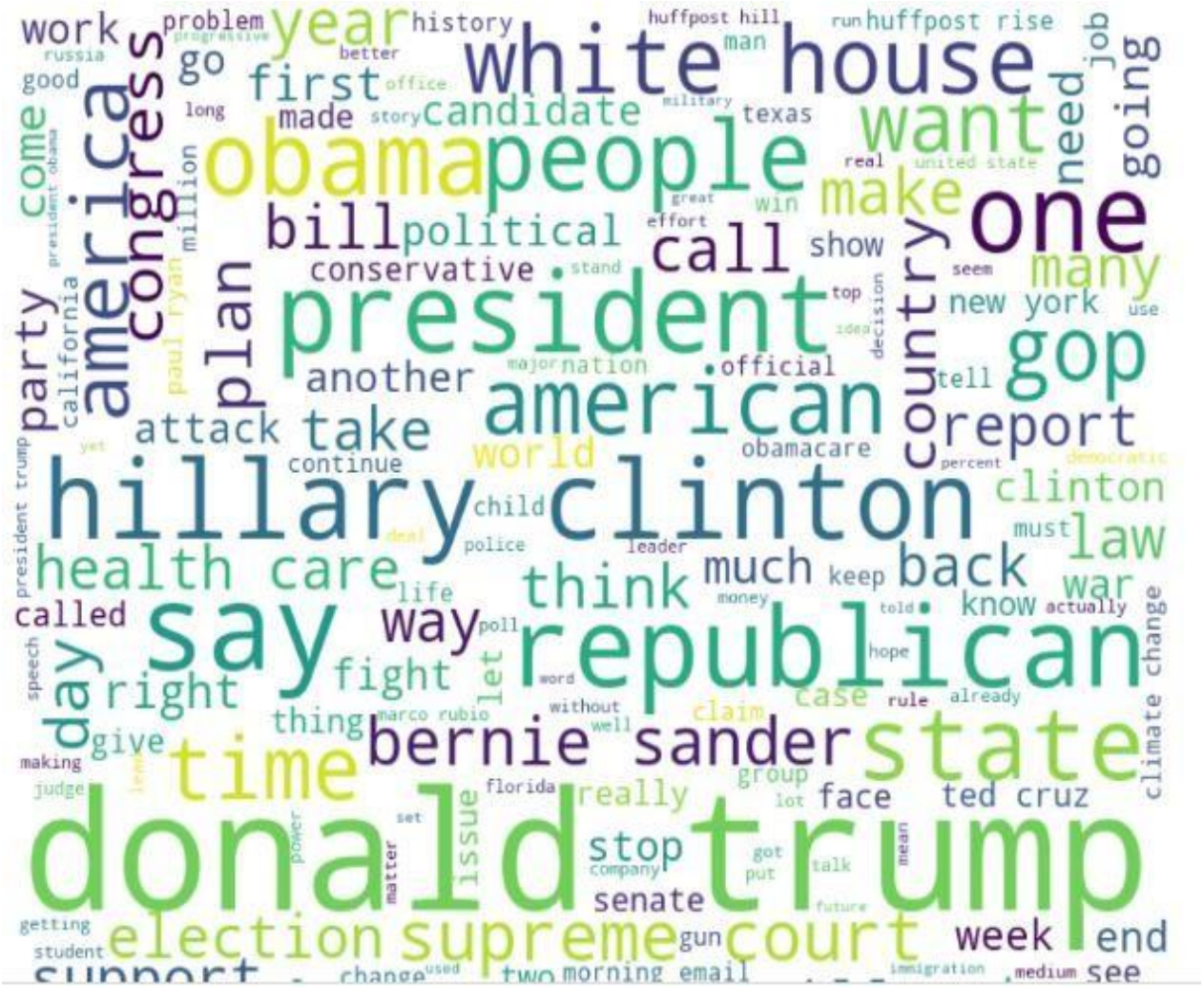
Here we can see the data in tabular manner which contains the following attributes such as category,headline,authors,link,description and date.

3) DATA PRE-PROCESSING MODULE

Here the data pre-processing is performed and if there is any in complete data present in the dataset those incomplete records need to be removed from the dataset and only which are complete and accurate must only be kept in the input dataset.



WORD CLOUD



4) TRAIN THE MODEL USING SEVERAL ML ALGORITHMS

Here we try to apply classification algorithms like :

1. LSTM
2. Naïve Bayes,

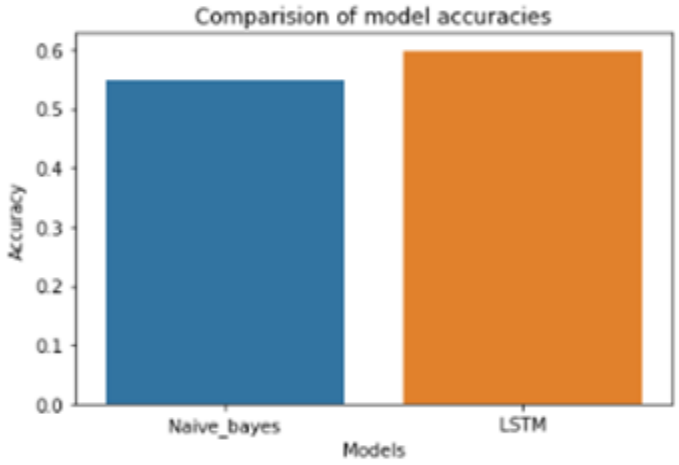
And try to check the accuracy of two classification algorithms separately and then conclude which one is best in order to classify the news.

5) PERFORMANCE ANALYSIS MODULE

In this module we try to classify each and every algorithm and then find out the accurate algorithm from a set of classification algorithms. Finally we try to find out the best algorithm based on accuracy.Here LSTM gives more accuracy compared with naïve bayes and we finally conclude that DOCV2MODEL is used for data pre-processing and the processed data is send as input for the LSTM to classify the news based on categories.

6. RESULT AND DISCUSSION

PERFORMANCE ANALYSIS



From the above window we can clearly see comparision of two algorithms for news classification based on category and we can finally tell that LSTM is more accurate compared with other algorithms.

7. CONCLUSION

This paper presents a comparative analysis of the multiclass category predictor's prediction performance. News category predictors were developed by deploying/training well- known machine learning algorithms (Naïve Bayes, Random Forest, K-Nearest Neighbor, and Support Vector Machine) on a BBC news dataset having five categories (business, sports, technology, politics, and entertainment). Later, using performance evaluation metrics, we analyzed the Confusion Matrix and quantified the test dataset's Precision, Recall, and overall Accuracy. As a result, the SVM model was proven to be the best among the four supervised learning models in correctly categorizing news articles with 98.3% accuracy. The lowest accuracy was obtained by the KNN model with K=5.However, the KNN model's performance can be enhanced by investigating the value of the optimal number of neighbors K.As future work, deep learning schemes will be introduced to further improve the classifier performance.

8. REFERENCES

[1] A. A. Hakim, A. Erwin, K. I. Eng, M. Galinium, and W. Muliady, "Automated document classification for news article in Yogyakarta,Indonesia, Oct. 2014, doi: 10.1109/ICITEED.2014.7007894.

[2] G. Mujtaba, L. Shuib, R. G. Raj, R. Rajandram, and K. Shaikh, "Prediction of cause of death from forensic autopsy reports using text classification techniques: A comparative study," Journal of Forensic and Legal Medicine, vol. 57, pp. 41–50, Jul. 2018, doi: 10.1016/j.jflm.2017.07.001.

[3] V. S. Padala, K. Gandhi, and D. V. Pushpalatha, "Machine learning: the new language for applications," IAES International Journal of Artificial Intelligence (IJ-AI), vol. 8, no. 4, pp. 411–421, Dec. 2019, doi:10.11591/ijai.v8.i4.pp411-421.

[4] F. Miao, P. Zhang, L. Jin, and H. Wu, "Chinese News Text Classification Based on Machine Learning Algorithm," in 2018 10th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC), Hangzhou, China, Aug. 2018, vol. 02, pp. 48–51,doi: 10.1109/IHMSC.2018.10117.

[5] S. M. H. Dadgar, M. S. Araghi, and M. M. Farahani, "A novel text mining approach based on TF-IDF and Support Vector Machine for news classification," in 2016 IEEE International Conference on Engineering and Technology (ICETECH), Coimbatore, India, Mar.2016, pp. 112–116, doi: 10.1109/ICETECH.2016.7569223.

[6] G. L. Yovellia Londo, D. H. Kartawijaya, H. T. Ivariyani, Y. S. Purnomo W. P., A. P.Muhammad Rafi, and D. Ariyandi, "A Study of Text Classification for Indonesian NewsArticle," in 2019 International Conference of Artificial Intelligence and Information Technology (ICAIIT), Yogyakarta, Indonesia, Mar. 2019, pp. 205–208, doi: 10.1109/ICAIIT.2019.8834611.

[7] R. Wongso, F. A. Luwinda, B. C. Trisnajaya, and O. R. Rusli, "News Article Text Classification in Indonesian Language," Procedia Computer Science, vol. 116, pp. 137–143, 2017, doi: 10.1016/j.procs.2017.10.039.]

[8] A. N. Chy, Md. H. Seddiqui, and S. Das, "Bangla news classification using Naive Bayes classifier," in 16th Int'l Conf. Computer and Information Technology, Khulna, Bangladesh, Mar. 2014, pp. 366–371,doi: 10.1109/ICCITechn.2014.6997369.

[9] I. Dilrukshi, K. De Zoysa, and A. Caldera, "Twitter news classification using SVM," in 2013 8th International Conference on Computer Science Education, Colombo, Sri Lanka, Apr. 2013, pp. 287–291, doi:10.1109/ICCSE.2013.6553926.

[10] H. Sawaf, J. Zaplo, and H. Ney, "Statistical classification methods for arabic news articles," presented at the Natural Language Processing in ACL2001, Toulouse, France, 2001.

[11] I. J. Mrema and M. A. Dida, "A Survey of Road Accident Reporting and Driver's Behavior Awareness Systems: The Case of Tanzania," Engineering, Technology & Applied Science Research, vol. 10, no. 4,pp. 6009–6015, Aug. 2020.