IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

A Survey on Fake News Detection using NLP

Chinmay kumar Rout[1], Padmini Giri[2], Shuchismita Sahu[3], Bibisika Behera[4], Mr.Tapas Das[5]

1.Department of Computer Science & Engeneering ,Balasore College Of Engineering & Technology, Odisha, India
2.Department of Information Technology ,Balasore College Of Engineering & Technology, Odisha, India
3.Department of Information Technology ,Balasore College Of Engineering & Technology, Odisha, India
4.Department of Information Technology ,Balasore College Of Engineering & Technology, Odisha, India
5.Department of Computer Science & Engeneering ,Balasore College Of Engineering & Technology, Odisha, India

Abstract:-
Fake news detection is a major challenging problem in Natural language processing(NLP).The increasing number of social media platform and user .Through these electronic media, The fake news are also increasing exponentially . The increasing reliance of more people on electronic media for their daily news has opened doors for malicious people to spread fake news . So detecting the fake news in this survey paper , Data set has Taken from keggle, Google Dataset search .Then we applied Transformers a Deep Learning approach to sequence and Classification .

Key words :- Fake news detection , Natural language processing , Deep learning , Transformers

1.Introduction :-

Today social media has become popular among people worldwide as their primary source of news. Moreover, online social media is becoming the main platform for people to exchange information, opinions and experiences about different events . These platforms have significantly helped the free speech, political and social awareness , and raising the voice of minorities and oppressed groups . Despite all the benefits, social media can be harmful as well. One of the harms of social media is the spread of misinformation.

Misinformation and its propagation are not limited to the modern information age and social media. In the information age, the lack of proper monitoring and fact-checking and capability of the spread of news by bots [1] have turned social media into a key conduit for fake news.

The spread of misinformation on these platforms can significantly affect different aspects of our lives. Some research also indicates that exposure to fake news can cause attitudes of inefficacy and cynicism toward certain political candidates[3,4]. The influence of misinformation is not limited to Politics. Another study shows that 1225 fake news stories were spread during COVID-19, with half of them coming from social media, putting public health in danger [5]. This significant effect of misinformation on society has motivated extensive research on fake news, especially on Twitter as a major social media. Some of these studies focus on the statistical behavior of fake news. A recent study by Vosoughi et al. shows that diffusion parameters of fake news are different from real news on Twitter. They showed that “falsehood” diffuses deeper and faster . In another study, fake and real news propagate different diffusion topologies on social platforms such as Weibo and Twitter [6].

To address hate speech and disinformation, which have historically troubled the country, Ethiopia enacted the hate speech and disinformation prevention and suppression proclamation in March 2020 [7]. However, while government regulation is necessary to control hate speech, Ethiopia's new law threatens online freedom of expression and access to information. As a result, it seems to be less useful, as fake news and hate speech creators conceal their work, leaving no record for the law. Using various methods, Facebook, Google, Twitter, and YouTube tried to take technological precautions. Linguistic resources are vital in the creation of fake news and hate speech detection approaches. However, "low-resource" languages, primarily African languages, lack such tools and resources [8]. Ethiopia has established a policy to introduce four more working languages in addition to Amharic, which has traditionally served as the country's working language. The government will adopt Afan Oromo, Ethiopia's most frequently spoken language, as well as Afar, Somali, and Tigrigna as official languages in the future [9]. Despite this, Ethiopian languages remain among the world's "low-resource" languages, lacking the tools and resources required for natural language processing applications and other techno-linguistic activities. However, a lack of appropriate datasets and good word embedding have made it difficult to create detection techniques that are reliable enough [11]. Recent improvements in natural language processing and understanding have made it possible to detect and counteract fake news and hate speech in textual streams with greater accuracy by using different approaches. With the growing influence of social media platforms in affecting public opinion and ideas around the world, there has been a greater focus on recognizing and combatting fake news and hate speech on various platforms [12]. Currently, in Ethiopia, hate speech and the spread of fake news have already impacted the lives of millions of people. Some schools, public and private universities, or colleges have recently closed; business activities have been severely hampered due to the closure of major roads in the country; citizen movement has been severely hampered; millions have been displaced, and many thousands have died due to scarcity of food and shelter . Accordingly, all Ethiopians are suffering more from the harmful effects of social media, than those in other developing countries [13]. As described in [14], fighting against fake news and hate information is to save lives. Fake news, misinformation, and hate speech have flourished in Ethiopia's media ecosystem, especially in online systems [15]. This is strongly linked to significant, tragic, real-world consequences, which exacerbated preexisting tensions and contributed to violence and conflict. To date, the Ethiopian government's response to the spread of fake news, misinformation, and hate speech has been heavy-handed, with the go-to response to escalation being to turn off the internet for the entire country. However, as the Internet and social media communications,

such as Twitter, YouTube, and Facebook messages, have evolved, so have the chances and obstacles to developing such solutions. The fake news and hate speech detection method used to detect and counteract fake news and hate speech on social media is far from flawless . For foreign and Ethiopian languages, several studies have been undertaken to detect and counteract fake news and hate speech on various social media platforms. Researchers have been conducted to detect and combat fake news and hate speech from various social media for Ethiopian languages and have advised future researchers to collect more corpora from various sources and use different approaches to improve the performances of the system in detecting and combatting of fake news and hate speech from various social media platforms. This study is planned to review the implemented approaches for fake news and hate speech detection research works in Ethiopian languages and to recommend the best approach regarding the performances of the evaluation metrics for future researchers of the area to minimize the risks that come due to the widespread of fake news and hate speech among the societies. The rest of the paper is organized into different but interrelated sub-sections. The paper begins by discussing the related works in "Related works" section, results and discussions in "Results and discussion" section, and the paper is a conclusion and future works in "Conclusion and recommendation" section. These are the paper we have deeply studied.

Sr. no.	Paper Name	Authors	Year of publication	Merits	Demerits	Methods/tools
1	AutomatedFakeNewsDetectionusingcross-checkingwithreliablesources	Zahra Ghadiri, 1 Milad Ranjbar, 1 Fakhteh Ghanbarnejad,1, ∗ and Sadegh Raeisi1	2022	This approach gives a 70% accuracy which outperforms other generic fake-news classification models.	This is not limited to a specific news story or a category of information.	Trained RandomForest
2	Fake News Detection Using NLP	1 Research Scholar, 2Assistant Professor	2021	They got an accuracy of 81.6%.	Thay have presented six LSTM models and three different methods were used for feature extraction	LSTM(long short term memory), Word2Vec, TF-IDF
3	CODEatCheckThat! 2022: Multi-class fake news detection of news articles with BERT	Olivier Blanc1, Albert Pritzkau 2, Ulrich Schade2 and Michaela Geierhos1	2022	They have got good accuracy	The performance decreases significantly on the test data due to a too large gap between the gold standard and There extended training dataset	Deep Learning, Transformers, BERT
4	NLPIR-UNED at CheckThat! 2022: Ensemble of Classifiers for Fake News Detection	Juan R. Martinez-Rico1, Juan Martinez-Romo1, 2 and Lourdes Araujo1, 2	2022	They have obtained the second best F1 measure (0.3324) among the 25 participating teams with a difference of 0.0066 compared to the team ranked first.		FFNN, Transformers, BERT,LIWC
5	NLyticsatCheckThat! 2022: Hierarchicalmulti-class fake news detection of news articles exploiting the topicstructure	Albert Pritzkau 1, Olivier Blanc2, Michaela Geierhos2 and Ulrich Schade1	2022	Transformer-basedmodelssuchasRoBERTaorLongformerhaveproventobepowerful language representation model for various natural language processing tasks	Thay examine the application of conceptssuchasactivelearning,semi-supervisedlearning aswellasweaksupervision	Transformers, RoBERTa, Longformer, Topic modeling

1.2.Machine Learning :-

In fake news detection, when we use a model to assist a user make predictions on the trustworthiness of a webpage, the trust of the user in the model affects the user's judgment on the webpage. Ultimately, no matter how good the model is at predicting fake news, the user will judge whether to believe or not what is written in the news article. The user's trust is usually low when the user does not understand the model, or the model behaves like a black-box. The undesired black-box model problem has developed into a growing research field in the last few years. Explainable artificial intelligence (XAI) allows a model to output an interpretable result. This result can be explained differently by highlighting parts of its input, contrasting the input with similar ones, or using natural language. Interpretability is essential because many machine learning models are being deployed, and people do not understand how they work or trust them. It also helps designers understand the reasoning process to improve the model. Approaches to explainable machine learning are generally classified into two categories: post hoc explainability and intrinsic explainability. In the post hoc approach, the model is viewed as a black box, and post-processing is necessary following the model's prediction to provide an explanation. One of the most popular approaches is LIME , in which a separate model is trained on the input–output pairs of the original model based on local linear regression to determine which features of the input are important for classification. However, the model is modified to have interpretable parts inside its structure based on the intrinsic explainability approach. One example of such a model contains an attention mechanism that can express the parts of the model's input that are considered when classifying. However, recent published studies have presented an ongoing debate on how useful attention mechanisms are in providing an explanation and whether they capture the decision-making process of the model ,. Although many models and datasets that address the interpretability problem have been developed, there is still much ambiguity in the formal definition of commonly used terms. Therefore, the objective and evaluation methods for these models vary from paper to paper.

1.3.Deep Learning and Pre-trained DeepLanguage Representation :-

Recent work on text classification uses neural networks, particularly Deep Learning (DL). Badjatiya et al. demonstrated that these architectures, including variants of Recurrent Neural Networks (RNN) , Convolution Neural Networks (CNN) , or their combination (Char CNN, Word CNN, and Hybrid CNN), produce state-of-the-art results and outperform baseline methods (character n-grams, TF-IDF or bag-of-words representations). Until recently, the dominant paradigm in approaching NLP tasks has been focused on the design of neural architectures,using only task-specific data and word embeddings such as those mentioned above. This led to the development of models such as Long Short Term Memory (LSTM)networks or Convolution Neural Networks ,which achieve significantly better results in a range of NLP tasks as compared to less complex classifiers such as Support Vector Machines, Logistic Regression or Decision Tree Models. Badjatiya et al. demonstrated that these approaches outperform models based on character and word n-gram representations. In the same paradigm of pre-trained models, methods like BERT and XLNe thave been shown to achieve state-of-the-art performance in a variety of tasks. Indeed, the usage of a pre-trained word embedding layer to convert the text into vectorized input for a

Neural network marked a significant step forward in text classification. The potential of pre-trained language models, e.g. Word2Vec , GloVe , fastText , or ELMo , to capture the local patterns of features to benefit text classification, has been described by Castelle . Modern pre-trained language models use unsupervised learning techniques such as creating RNN embeddings on large text corpora to gain some primal "knowledge" of the language structures before a more specific supervised training steps in. Transformer-based model sare unable to process long sequences due to their self-attention mechanism, which scales quadratically with the sequence length. BERT-based models enforce a hard limit of 512 tokens, which is usually enough to process the majority of sequences in most benchmark datasets.

1.4.BERT,RoBERTa and Longformer :-

BERT stands for Bidirectional Encoder Representations from Transformers. It is based on the Transformer model architectures introduced by Vaswanietal.. The general approach consists of two stages: first,BERTispre-trainedonvastamountsoftext,withanunsupervisedobjective of masked language modeling and next-sentence prediction. Second, thispre-trained network is then fine-tuned on task specific, labeled data. The Transformer architecture is composed of two parts, an Encoder and a Decoder, for each of the two stages. The Encoder used in BERT is an attention-based architecture for NLP. It works by performing a small, constant number of steps. In each step, it applies an attention mechanism to understand relationships between all words in a sentence, regardless of their respective position. By pre-training language representations, theEncoderyieldsmodelsthatcaneitherbeusedtoextracthighqualitylanguagefeaturesfrom text data, or fine-tune these models on specific NLP tasks (classification, entity recognition, question answering, etc.). We rely on RoBERTa , a pre-trained Encoder model which builds on BERT's language masking strategy. However, it modifies key hyperparameters in BERT such as removing BERT's next-sentence pre-training objective, and training with much larger mini-batches and learning rates. Furthermore, in comparison to BERT, the training data set for

Roberta was an order of magnitude larger (160 GB of text) with the maximum sequence length of 512 used for all interations. This allows RoBERTa representations to generalize even better to downstream tasks. To address the limitation of traditional Transformer-based models to 512 tokens, Long former uses an attention pattern that scales linearly with sequence length, making it easy to process documents of thousands of tokens or longer. To this end, the standard self-attention is replaced by an attention mechanism, which combines a local windowed attention with a task motivated global attention, thus allowing up to 4096 position embeddings. Longformer is pre-trained from RoBERTa.

2.Related works :-

Several surveys on fake news have provided overviews of fake news detection methods. They first attempt to define fake news and comprehensively describe the current approaches to the problem. Because the range of is wide, surveys can review explainable models used in different fields or focus on literature with natural language processing (NLP) . In contrast, our survey goes deeper into one of the applications of NLP, fake news detection models, and survey models that have interpretable functions.

3.Datasets for Fake News Detection :-

The creation of datasets is often one of the most overlooked parts of machine learning, yet data are often one of the aspects that determine whether a machine learning model is successful. Numerous fake news and medical misinformation datasets are publicly available to train models. Nonetheless, few have the capability to evaluate interpretability . Every article on the website not only has a label indicating its truthfulness, but a reason for the label is also provided . The dataset not only contains common aspects, such as news content, social engagement, and user information, but it also includes an explanation of the ground truth. there are problems with free form explanations, such as annotator bias and the individual definition of explanation, which may cause the data quality to decrease.

4.Conclusion :-

Fake news detection is a research area in which interpretability is important for both users and designers to make models more trustworthy and clearer. This paper presented publications that show how interpretability can be implemented and even presented to users. It also examined other aspects such as explainable datasets, explanation ML ,DL ,BERT ,RoBERTa ,Longformer and model performance. Finally, we recommend future directions where efforts are needed to advance explainable fake news detection, including 1) preparing explainable datasets with labeled attributes; 2) establishing the evaluation metrics on the quality of the created explanations; 3) researching on the visualization scheme for end-users' easy understanding of explanations; and 4) additional types of explainability, e.g., website-level trustworthy detection, to enhance the fake news detection.

5.References

1. Buzea MC, Trausan-Matu S, Rebedea T. Automatic fake news detection for romanian online news. Information. 2022;13(3):1–13. https:// doi. org/ 10. 3390/ info1 30301 51.

2. Shu K, Sliva A, Wang S, Tang J, Liu H. Fake news detection on social media. ACM SIGKDD Explore News. 2017;19(1):22–36. https:// doi. org/ 10. 1145/ 31375 97. 31376 00.

3. Zhou X, Zafarani R. A survey of fake news: fundamental theories, detection methods, and opportunities. ACM Comput Surv. 2020;53(5):1–37. https:// doi. org/ 10. 1145/ 33950 46.

4. Chakraborty T, Masud S. Nipping in the Bud: Detection, Diffusion, and Mitigation of Hate Speech on Social Media. 2022: 1–9.

5. Arega KL. Classification and detection of amharic language fake news on social media using machine learning approach. Electr Sci Eng. 2022; 4: 1–6.

6. Hadj Ameur MS, Aliane H. "AraCOVID19-MFH: Arabic COVID-19 Multi-label Fake News & Hate Speech Detection Dataset. Procedia CIRP. 2021;189:232–41. https:// doi. org/ 10. 1016/j. procs. 2021. 05. 086.

7. Chekol MA, Moges MA, Nigatu BA. Social media hate speech in the walk of Ethiopian political reform: analysis of hate speech prevalence, severity, and natures. Inf Commun Soc. 2021;0(0):1–20. https:// doi. org/ 10. 1080/ 13691 18X. 2021. 19429 55.

8. HaqCheck, Annual Report on Disinformation in Ethiopia _ Addis Zeybe - Digital Newspaper. 2021.

9. WHO. Director-General ' s remarks at the media briefing on 2019 novel coronavirus on 8th of February 2020," Who, no., 2020; 2019–2021

10. Alsenoy B. General data protection regulation. Data protection law in the EU: roles, responsibilities, and liability. Proce Comput Sci. 2019. https:// doi. org/ 10. 1017/ 97817 80688 459. 021.

11. Gereme F, Zhu W, Ayall T, Alemu D. Combating fake news in 'low-resource' languages: Amharic fake news detection accompanied by resource crafting. Inf. 2021;12(1):1–9. https:// doi. org/ 10. 3390/ info1 20100 20.

12. Kovács G, Alonso P, Saini R. "Challenges of Hate Speech Detection in Social Media. SN Comput Sci. 2021;2(2):1–15. https:// doi. org/ 10. 1007/ s42979- 021- 00457-3.

13. Gazette FN. Federal Negarit Gazette of the Federal Democratic Republic of Ethiopia, Content. 2020; 2–7.

14. Shaban ARA. Ethiopia cabinet approves bill to combat fake news, hate speech | Africanews. 2019; 1–2.

15. Admin. Facebook expands third-party fact-checking to Ethiopia, more African countries -. 2019; 1–2.

16. AutomatedFakeNewsDetectionusingcross-checkingwithreliablesources Zahra Ghadiri,1 Milad Ranjbar,1 Fakhteh Ghanbarnejad,1,∗ and Sadegh Raeisi1

17. Fake News Detection Using NLP Samrudhi Naik1, Amit Patil2

18. CODEatCheckThat! 2022: Multi-classfakenews detectionofnewsarticleswithBERT Olivier Blanc1, Albert Pritzkau2, Ulrich Schade2 and Michaela Geierhos1

19. NLPIR-UNEDatCheckThat! 2022: Ensembleof ClassifiersforFakeNewsDetection Juan R. Martinez-Rico1, Juan Martinez-Romo1,2 and Lourdes Araujo1,2

20.NLyticsatCheckThat!2022:Hierarchicalmulti-class fake news detection of news articles exploiting the topic structure Albert Pritzkau1, Olivier Blanc2, Michaela Geierhos2 and Ulrich Schade1

21. AIT_FHSTPatCheckThat! 2022: Cross-LingualFake News Detection with a LargePre-Trained Transformer Mina Schütz, Jaqueline Böck, Medina Andresel, Armin Kirchknopf, Daria Liakhovets, Djordje Slijepčević and Alexander Schindler

22. AENeT: an attention-enabled neural architecture for fake news detection using contextual features, Vidit Jain , Rohit Kumar Kaliyar,Anurag Goswami1, Pratik Narang,Yashvardhan Sharma

23. Arabic Fake News Detection Based on Textual Analysis, Hanen Himdi · George Weir · Fatmah Assiri · Hassanin Al-Barhamtoshy