ISSN: 2454-9940

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

E-Mail :
editor.ijasem@gmail.com
editor@ijasem.org

www.ijasem.org

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

ISSN 2454-9940

www.ijasem.org

Vol 19, Issue 2, 2025

Enhancement of textual data for word embeddings in the classification of fraudulent news

KASANI DURGA SAI SRI

Student,Computer Science and Engineering Department

Sasi Institute of Technology Engineering

Tadepalligudem, INDIA

durgasri.kasani@sasi.ac.in

PERUMALLA NAGARAJU

Student,Computer Science and Engineering Department

sasi Institute of Technology Engineering

Tadepalligudem,INDIA

nagaraju.perumalla@sasi.ac.in

ANDE BENJAMIN TYCHICUS

Student,Computer Science and Engineering Department

sasi Institute of Technology Engineering

Tadepalligudem,INDIA

benjamin.ande@sasi.ac.in

ALLURI ABHINAY

Student,Computer Science and Engineering Department

Sasi Institute of Technology Engineering

Tadepalligusem, INDIA

abhinay.alluri@sasi.ac.in

Dr.Kalli Srinivasa Nageswara Prasad

Professor,Computer Science and Engineering Department

Sasi Institute of Technology Engineering

Tadepalligudem,INDIA

ksnprasad@sasi.ac.in

***Abstract*— Data amount and quality affect categorization algorithm efficiency. Algorithm accuracy is great when the size of the data corpus is enormous. A low level of algorithm accuracy is indicative of a little data set. In order to better classify fake news, this article use Text Data Augmentation. In this study, the author evaluated methods for synonym replacement, back translation, and function word reduction. Word2Vec Skip-gram models will transform text enriched using the aforementioned approach into a numerical vector. Bernoulli NB, RF, SVM, and LR are some of the classifiers that will be used to train Word2Vec. Random Forest on original text, improved Logistic Regression with 'Reduction of Function Words,' and SVM and Bernoulli naïve Bayes on Back Translation upgraded text all lead to high accuracy.**

Keywords*— *Original Corpus, Synonym replacement (SR), Back translation (BT), Reduction of function words (FWD)

I. INTRODUCTION

Data augmentation eliminates the need to provide data while increasing the diversity of training examples [1]. By subjecting machine

learning models to a variety of scenarios, data augmentation aims to boost their performance and resilience. Augmenting data is useful in computer vision, NLP, and speech recognition. By expanding the variety of the data, data augmentation decreases overfitting. Putting an end to the process of the machine learning using example data used for training. Regardless of the apparent success, NLP studies including computer vision issues have not reaped significant advantage from DA systems thus far [2]. The most common methods for improving text data are paraphrasing, random insertion, swapping, and deletion; back translation; and synonym substitution. Outlined in papers [1], [2] are data augmentation campaigns. Many modern apps rely on text vectorization Manage natural language processing tasks involving classification. Word2Vec, Doc2Vec, and Glove are just a few of the popular word embedding methods that continue to thrive by capitalizing on words' semantic similarities.

Practically speaking, word vectors have a wide range of uses. In this part, the focus is on classification jobs. Improving classification performance metrics through data augmentation strategies is the focus of this study. To identify which data augmentation strategy offers the best assistance with classification difficulties, this post will take a look at a few popular options. From the various alternatives, the following techniques were chosen:

- Substituting a term with one of its synonyms is known as synonym replacement (SR). For this purpose, we use WordNet, an extensive database of words and their meanings.

- When it comes to improving written content, back translation (BT) is both easy and effective. It involves translating the source text from one language to another and vice versa. The real text is often slightly altered using this procedure yet important details are preserved.

- Random Deletion is similar to Function Word Reduction (FWD). Using a probability parameter, the approach extracts words from a given text. The likelihood parameter was restricted to words pertaining to function and content due to our context.

- The original corpus is consulted for reference purposes only; no data augmentation methods are employed. We used the Original as a benchmark and tested various methods against it [1, 2].

Time and computing power are needed to apply these strategies to classification tasks. The relevance of these ideas is up for debate. We set out to find out in this post whether these tactics really do boost classification model performance metrics. We also want to find out which approach improves categorization model results the most.

The purpose of this essay is to compare and contrast different classification systems and their relative merits. By determining their relevance, researchers will be better able to choose approaches for future categorization issues.

In order to get the corpus ready, we follow the supplied procedures. Train the Word2Vec Skip-gram model using the prepared corpus for classification tasks that involve word embeddings. For WELFakedataset, the task was to categorize false news. The classification's performance metrics will be evaluated next. So, we won't be evaluating Data.
We will attack the classification issue by utilizing augmentation tactics and then evaluate their efficiency.
classification.

There are a number of options for evaluating different corpus preparation methods for huge language models. In their performance evaluation, Nazir et al. [3] used WordSim-353 [4] and SimLex-999 [5], two sets of records that contain the similarity between words. Word vectors are frequently used in educational examples to show how to compute word similarities.

II. LITERATURE REVIEW

1. *A survey of data augmentation approaches for NLP*

A Survey of Data Augmentation Approaches for NLP

Abstract: Interest in DA for natural language processing has grown in recent years, particularly in research pertaining to new problems, low-resource domains, and large-scale neural networks that require a substantial amount of training data. The discontinuity of the linguistic evidence may

explain why this subject has remained unexplored despite increasing studies in the field. This research compiles and organizes the literature on data augmentation in NLP. Next, we'll go over some important tactics for data augmentation in NLP. Neural Network Processing (NLP) methods for typical jobs and uses are outlined below. Finally, we cover current difficulties and potential directions for future study. Finally, we hope that this effort will either shed light on data augmentation for NLP literature or inspire other studies in this area.

2. *Data augmentation techniques in natural language processing.*

https://arxiv.org/pdf/2110.01852

Abstract: In cases when deep learning is unsuccessful, DA can be used to address data shortages. The extensive usage of it in computer vision and NLP has improved several issues. A primary goal of DA methods is to increase the model's generalizability to novel testing data by diversifying the training data. We classify DA methods as either paraphrasing, noising, or sampling in this study according to the different types of enhanced data. The aforementioned DA methods are the focus of our investigation. Aside from that, we go over the issues and uses of their natural language processing capabilities. Key resources are in Appendix A.

3. *Toward the development of large-scale word embedding for low resourced language*

Toward the Development of Large-Scale Word Embedding for Low-Resourced Language | IEEE Journals & Magazine | IEEE Xplore

Abstract: To syntactically and semantically alter unlabeled data, natural language processing employs word embedding. With the help of vector space representations of the recovered corpus features, It is feasible to do things like summarize, simplify, and forecast the following line, among other things. By considering the frequency and co-occurrence of words, the matrix Noisey contrastive estimation, factorization, skip-gram, and hierarchical-structure regularizer embed them. Researchers have concentrated on Urdu because of its large speaker population (231.3 million) and the fact that these methods have yielded fully developed word vectors for the majority of spoken languages. In this study, we look at word embedding in Urdu. Among the categories represented in our dataset were: Industry, athletics, medicine, government, show business, scientific discoveries, and international news. This dataset yielded 288 million tokens. We used the skip-gram (word2vec) model for word vector creation. No more than 100, 200, 300, 400, 500, 128, 256, or 512 vector dimensions could be used for embedding. Assessments were conducted using Lexsim-999 and Wordsim-353 annotated datasets. The suggested study indicated that wordsim-353 and Lexsim-999 had Spearman correlation coefficients of 0.66 and 0.439, respectively. Compared to state-of-the-art, the results were superior.

4. *A study on similarity and relatedness using distributional and WordNet based approaches*

A Study on Similarity and Relatedness Using Distributional and WordNet-based Approaches

Abstract: Methods for similarity based on WordNet and distributional analysis are compared in this paper. Each relatedness and similarity method is discussed, along with its advantages and disadvantages, and a hybrid approach is proposed. Each of our methods outperforms the competition on the RG and WordSim353 datasets, and when combined, they provide the top results across the board. Our methods can be easily modified with little loss, and we pioneer cross-lingual similarity.

5. *SimLex-999: Evaluating seman tic models with (Genuine) similarity estimation*

[1408.3456v1] SimLex-999: Evaluating Semantic Models with (Genuine) Similarity Estimation

Abstract: To evaluate distributional semantic models, we present SimLex-999, a benchmark that outperforms existing tools in multiple respects. In contrast to industry leaders like WordSim-353 and MEN, which place a broader emphasis on relatedness and association, it performs poorly categorizing things that are similar but different [Freud, psychology]. Compared to conceptual association models, SimLex-999 models are more versatile because they prioritize similarity. Along with concreteness and (free) association strength grades, SimLex-999 provides abstract and concrete noun, adjective, and verb pairs. The ability to assess model performance on various ideas with granularity is made possible by this diversity, which in turn reveals opportunities to enhance

structures. When tested on SimLex-999, modern models consistently fall short of the inter-annotator agreement ceiling. In this way, SimLex-999 can gauge progress in distributional semantic models, which will power future representation-learning systems.

6. *EDA: Easy data augmentation techniques for boost ing performance on text classification tasks*

[1901.11196] EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks

Abstract: Through the use of data augmentation, EDA simplifies text classification. Simple but successful methods used by EDA include swapping, deletion, random insertion, and synonym substitution. RNNs and CNNs both benefit from EDA's enhanced performance on five different text classification tasks. Training EDA with only half of the training set yields results that are on par with full training, according to five different datasets. Our reasonable numbers were the result of thorough ablation studies.

7. *Improving short text classification through global augmentation methods*

(PDF) Improving Short Text Classification Through Global Augmentation Methods

Abstract: Various techniques for improving text are investigated. To do this, we scour three databases of news articles and social media posts. Our goal is to enlighten researchers and practitioners on how to select augmentation for classification applications. We discover that augmentation based on Word2Vec can function even in the absence of a formal synonym model such as WordNet. By reducing overfitting in tested deep learning models, Mixup improves all text-based augmentations. For common and low-resource use cases, round-trip translation through a translation service is too costly and unavailable.

8. *Contextual augmentation: Data augmentation by words with paradigmatic relations*

Contextual Augmentation: Data Augmentation by Words with Paradigmatic Relations - ACL Anthology

Abstract: Our latest data augmentation technique is labeled sentence contextual augmentation. Even when paradigmatic relations stand in for actual words, we still consider sentences to be naturally occurring. Predicting arbitrary word substitutions is the job of a bidirectional language model. Word improvement is a good fit for the many context-predicted words. In addition, we extend sentences without impacting label compatibility by modifying a language model with a label-conditional design. We found that the proposed method improves the performance of classifiers built using convolutional or recurrent neural networks across six different text categorization tasks.

9. *SynoExtractor: A novel pipeline for Arabic synonymextraction using Word2 Vecwordembeddings*

SynoExtractor: A Novel Pipeline for Arabic Synonym Extraction Using Word2Vec Word Embeddings - Al-Matham - 2021 - Complexity - Wiley Online Library

Abstract: Automated synonym extraction is a feature of many NLP systems, such as those that retrieve information or answer questions. Recent work has used word embeddings to extract semantic linkages by capturing word relatedness and similarity. Synonym extraction is challenging since word embeddings do not have the ability to determine if two words are synonyms or have any other kind of semantic link. This research presents the SynoExtractor pipeline, a tool for preserving synonyms by filtering related word embeddings according to linguistic principles. For our experiments, we utilized Gigaword and KSUCCA embeddings, as well as CBOW and SG models. Using Alma'any Arabic synonym thesaurus, we compared automatically extracted synonyms. We had two native Arabic speakers review it by hand. When compared to cosine similarity, our results demonstrate that the SynoExtractor pipeline significantly enhances the accuracy of synonym extraction. Gigaword and the King Saud University Corpus of Classical Arabic were both enhanced using SynoExtractor; the former by 21% and the latter by 0.605. Comparing Sketch Engine to SynoExtractor, the latter performed 32% better when it came to MAP synonym extraction. The ability to extract synonyms demonstrates the method's generalizability to different languages.

10. *On the use of text augmentation for stance and fake news detection*

(PDF) On the use of text augmentation for stance and fake news detection

ISSN 2454-9940
www.ijasem.org
Vol 19, Issue 2, 2025

Abstract: DA(DATA Augmentation)can be used to generate latest training examples by updating old ones. Notable benefits of DA include alleviating class imbalance, avoiding data scarcity, and enhancing generalization. This study delves at the use of DA to identify bogus news and take a stance. The impact of DA methods on conventional classification algorithms is investigated in the first part of our study. The flawed "the more, the better" approach to text augmentation is exposed by our study, which demonstrates that no one-size-fits-all strategy exists. Part two of our research presents an ensemble learning method that uses augmentations. To enhance ensemble prediction performance, the proposed method augments base learners with text to increase their diversity and accuracy. Class imbalance with DA is the focus of our final experiment. As a result of social stratification, algorithms for detecting bias and fake news are skewed. Both moderate and severe imbalances can be remedied through text augmentation, according to this study.

III. PROPOSED SYSTEM

To make the suggested method better at identifying fake news, increase the amount of text data in the training dataset. To transform text into numerical vectors, the Word2Vec Skip-gram model employs SR, BT, and FWD reduction.

The impact of vector augmentation on classification accuracy will be assessed by a number of ML classifiers, including RF, SVM, Logistic Regression, and Bernoulli Naïve Bayes.

Performance will be enhanced with the addition of XGBoost and other modern ensemble techniques. In order to improve the system's ability to identify false news and overcome dataset restrictions, this strategy is employed. Measures of performance will include F1-score, recall, precision, and accuracy.

Extension: We enhanced the suggested system with deep learning models and powerful ensemble algorithms including LightGBM, CatBoost, and XGBoost. Training and categorization of features are both improved by this extension. A news classification user interface was also developed by us using Flask.

Advantages:

- The proposed method makes advantage of text data augmentation to incorporate more diverse and variable samples into the training dataset, which in turn improves the model's classification abilities.
- To increase the accuracy and resilience of detecting faked news, the proposed method employs ensemble learning with advanced techniques like XGBoost. This combines the skills of several classifiers.
- The suggested method use the Word2Vec Skip-gram model to transform textual information into comprehensive numerical vectors that encompass semantic relationships and context. This enables machine learning classifiers to identify minute differences in false news.
- Look at performance measures like F1-score, accuracy, precision, and recall to see how well the system is doing and how far the model has come.
- Using complex algorithms such as XGBoost can enhance the accuracy and dependability of classifying bogus news.
- Flexibility to meet future demands is guaranteed by the ease of scaling to incorporate new algorithms or datasets.
- It is easy to upload news articles for classification using the Flask interface.
- Instantaneous feedback on the veracity of user-generated news is available.

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT
ISSN 2454-9940
www.ijasem.org
Vol 19, Issue 2, 2025

II.SYSTEM ARCHITECTURE

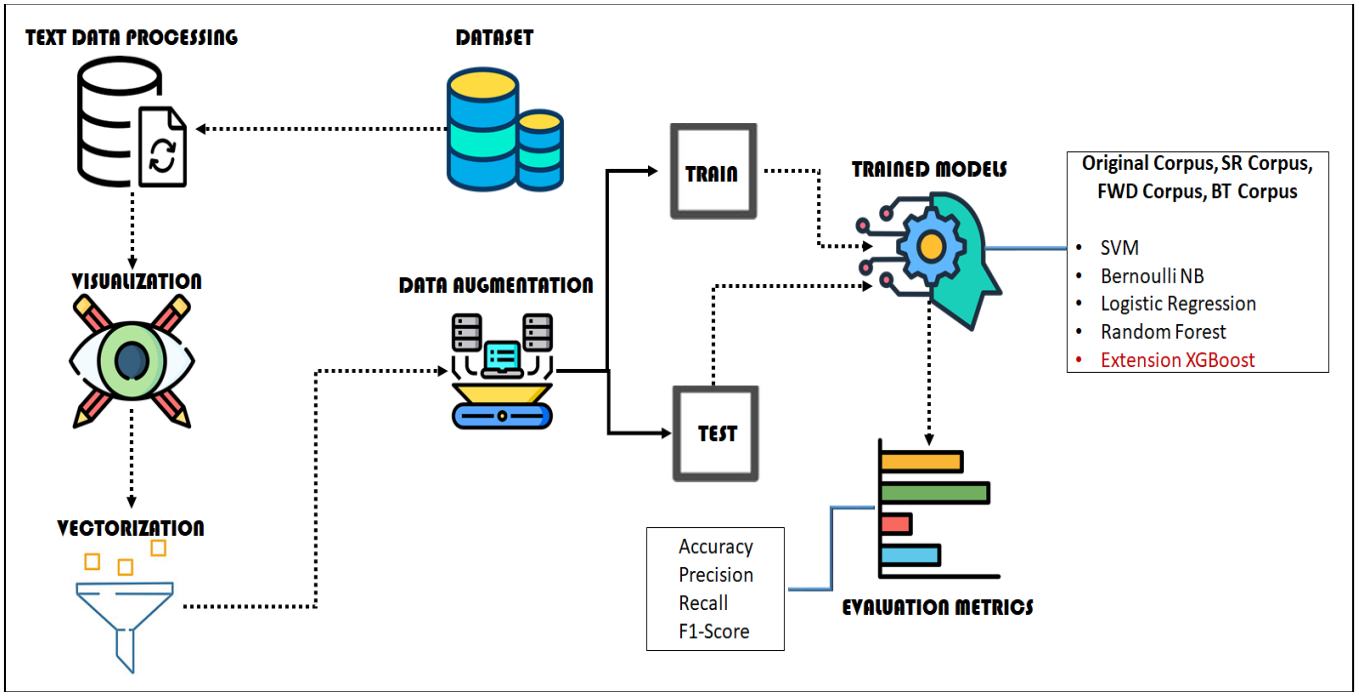


FIG 1.system architecture

The proposed system for fake news classification integrates text data processing, augmentation, and machine learning techniques to enhance classification accuracy. Initially, the raw text data undergoes preprocessing steps such as tokenization, stop word removal, and normalization to prepare it for analysis. Following this, visualizations such as word clouds and frequency plots are generated to gain insights into the dataset. The cleaned data is then vectorized using techniques like TF-IDF or word embeddings to convert it into numerical format. To increase the diversity and generalization ability of the model, data augmentation methods—such as synonym replacement (SR), forward translation (FWD), and back translation (BT)—are applied to generate new samples. The dataset is split into training and testing sets, where multiple machine learning models including SVM, Bernoulli Naïve Bayes, Logistic Regression, Random Forest, and an extended version of XGBoost are trained. These models are evaluated using performance metrics like accuracy, precision, recall, and F1-score to determine their effectiveness. The architecture is designed to handle a variety of textual inputs and is optimized for improved detection of fake news content through robust augmentation and modeling strategies.

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

ISSN 2454-9940

www.ijasem.org

Vol 19, Issue 2, 2025

IV.EXPERIMENETAL RESULT AND DISUSSION

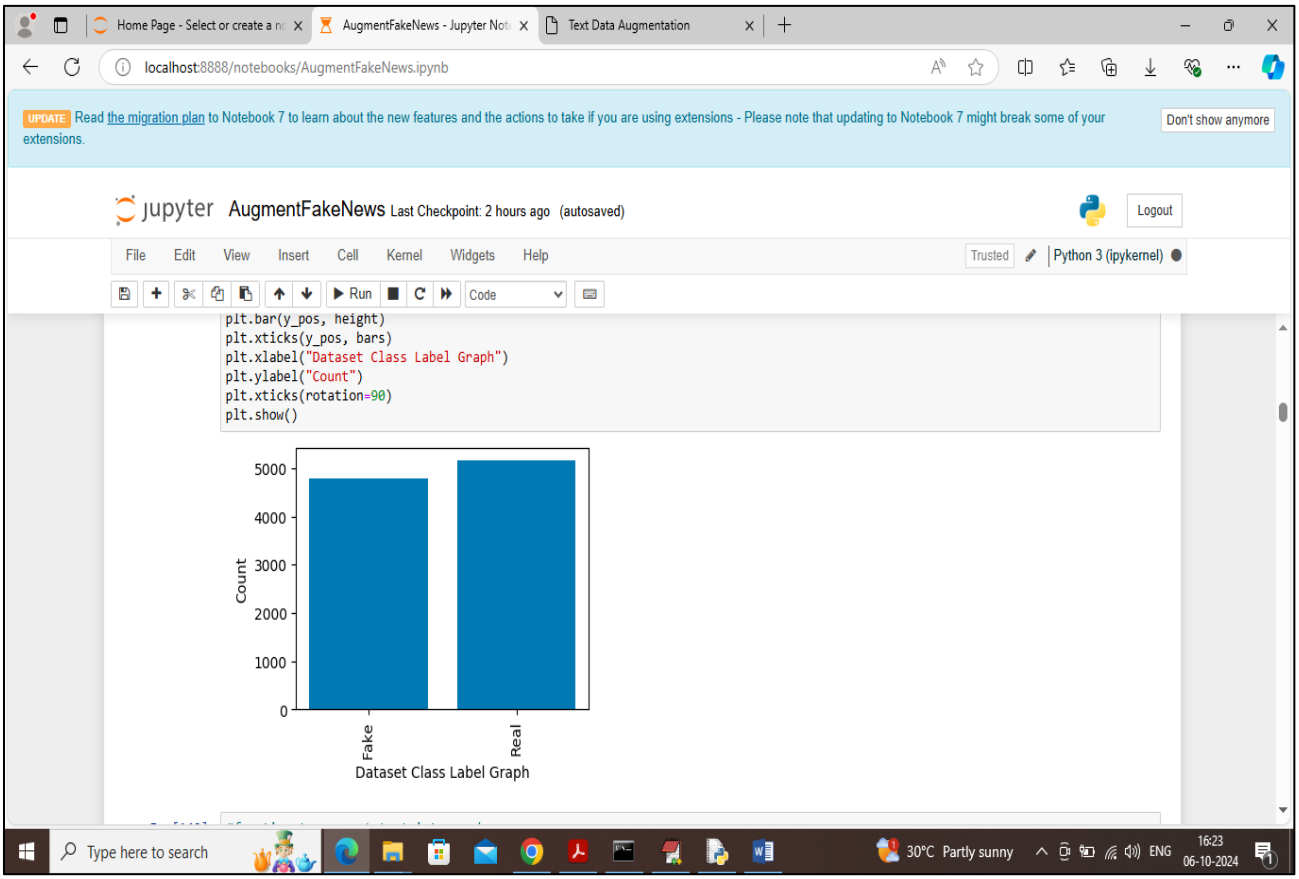


FIG 2 . detect fake or real

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

ISSN 2454-9940

www.ijasem.org

Vol 19, Issue 2, 2025

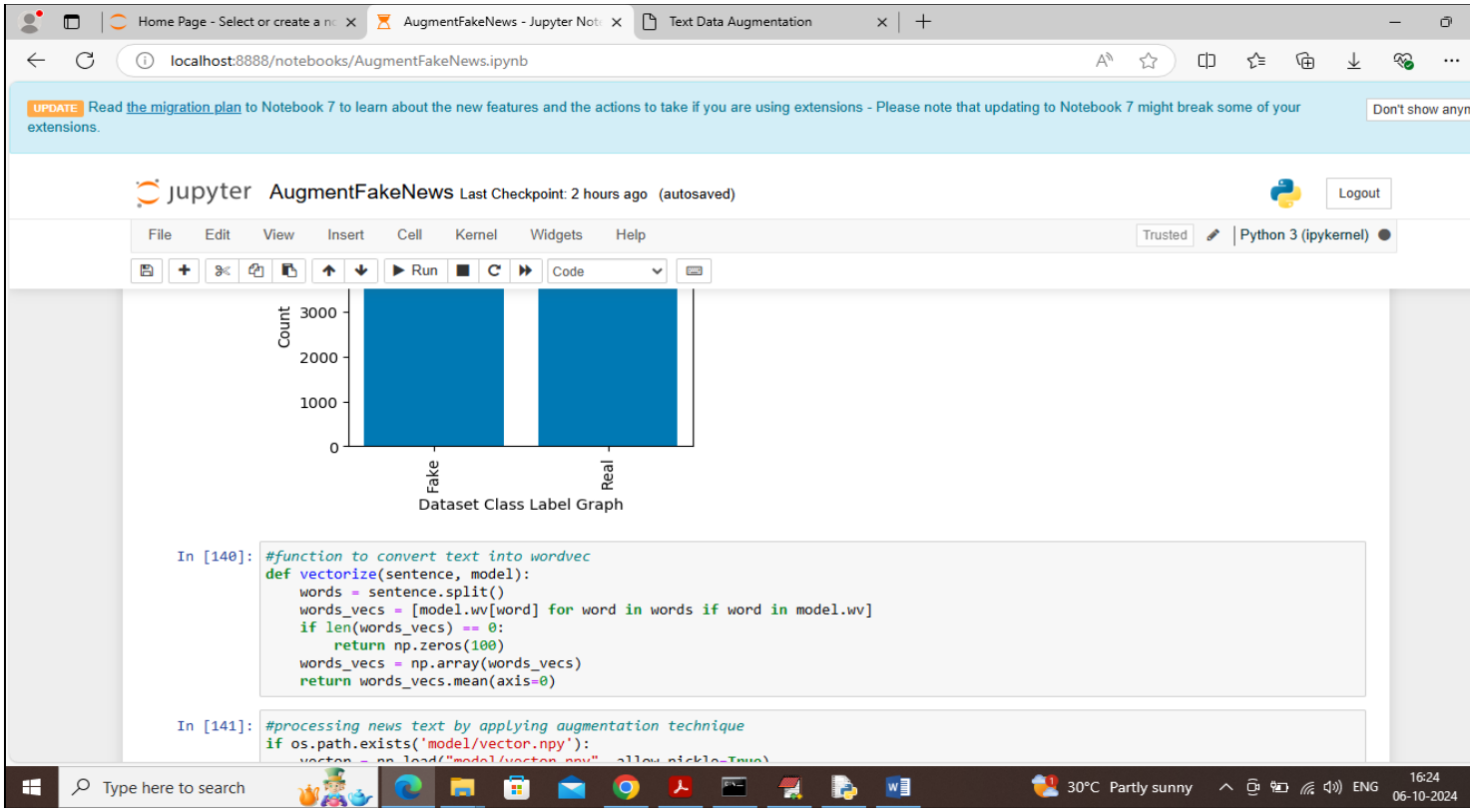
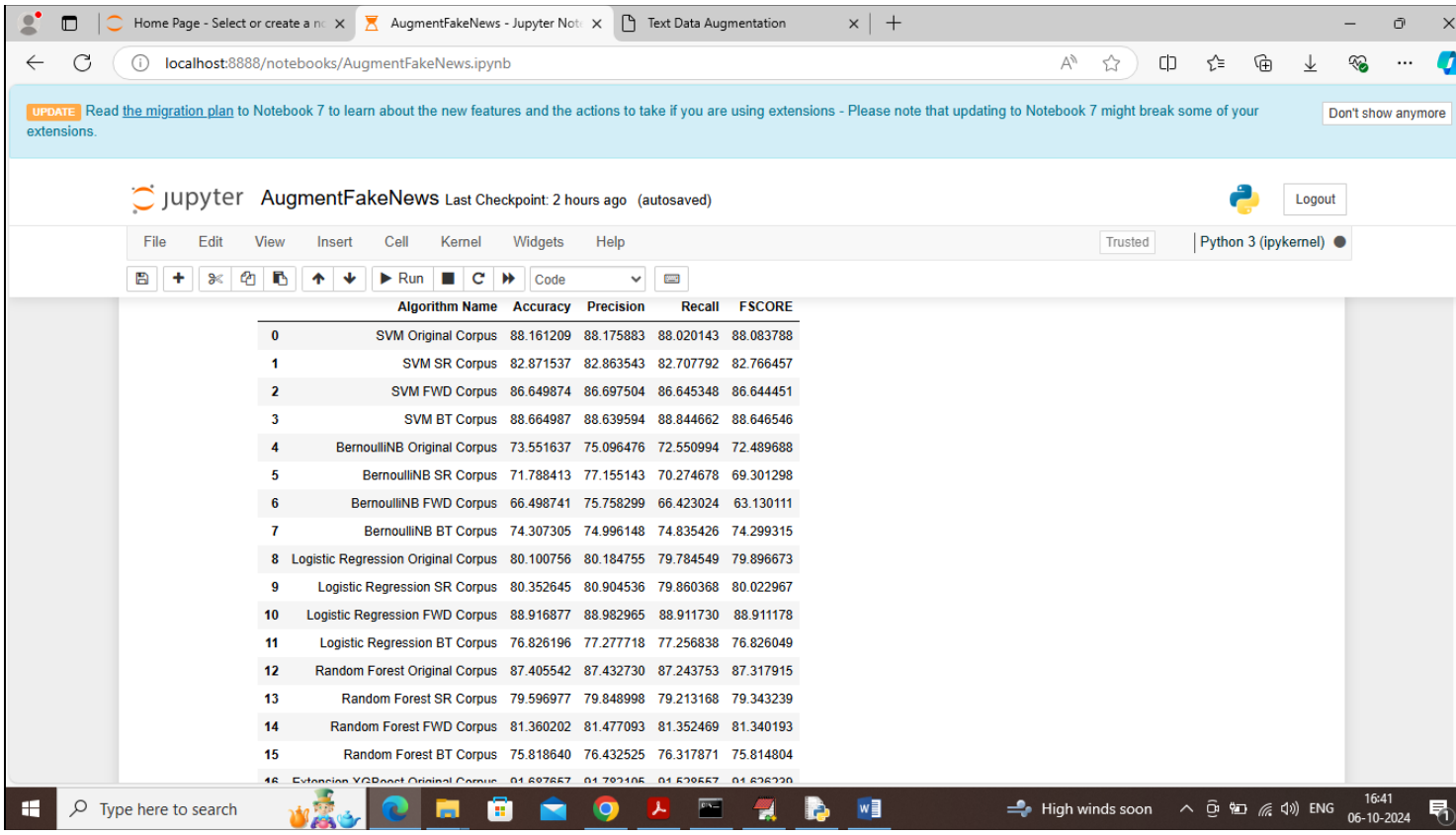


FIG 3. to convert words into vector



	Algorithm Name	Accuracy	Precision	Recall	FSCORE
0	SVM Original Corpus	88.161209	88.175883	88.020143	88.083788
1	SVM SR Corpus	82.871537	82.863543	82.707792	82.766457
2	SVM FWD Corpus	86.649874	86.697504	86.645348	86.644451
3	SVM BT Corpus	88.664987	88.639594	88.844662	88.646546
4	BernoulliNB Original Corpus	73.551637	75.096476	72.550994	72.489688
5	BernoulliNB SR Corpus	71.788413	77.155143	70.274678	69.301298
6	BernoulliNB FWD Corpus	66.498741	75.758299	66.423024	63.130111
7	BernoulliNB BT Corpus	74.307305	74.996148	74.835426	74.299315
8	Logistic Regression Original Corpus	80.100756	80.184755	79.784549	79.896673
9	Logistic Regression SR Corpus	80.352645	80.904536	79.860368	80.022967
10	Logistic Regression FWD Corpus	88.916877	88.982965	88.911730	88.911178
11	Logistic Regression BT Corpus	76.826196	77.277718	77.256838	76.826049
12	Random Forest Original Corpus	87.405542	87.432730	87.243753	87.317915
13	Random Forest SR Corpus	79.596977	79.848998	79.213168	79.343239
14	Random Forest FWD Corpus	81.360202	81.477093	81.352469	81.340193
15	Random Forest BT Corpus	75.818640	76.432525	76.317871	75.814804

FIG 4. tabular format

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

ISSN 2454-9940

www.ijasem.org

Vol 19, Issue 2, 2025

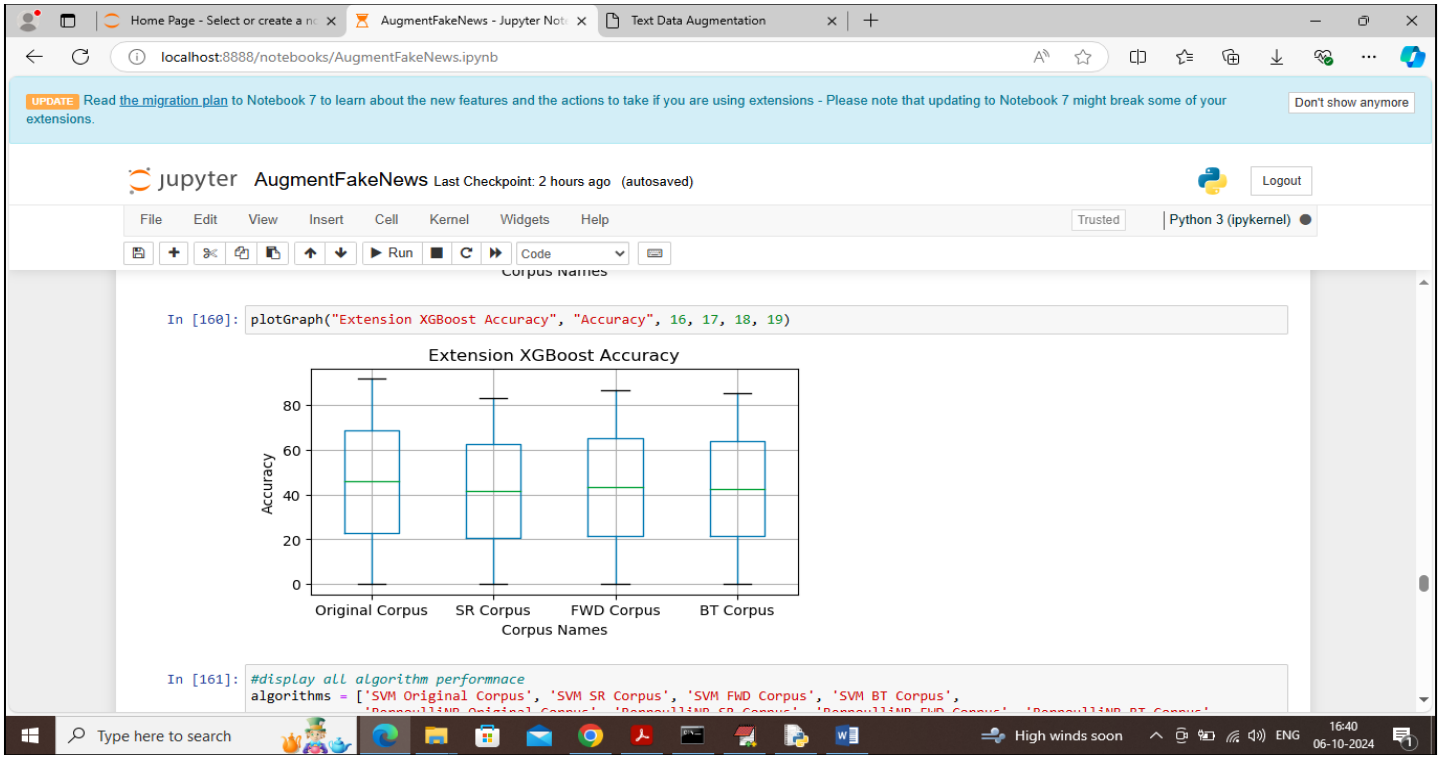


FIG 5. extension XGBOOST

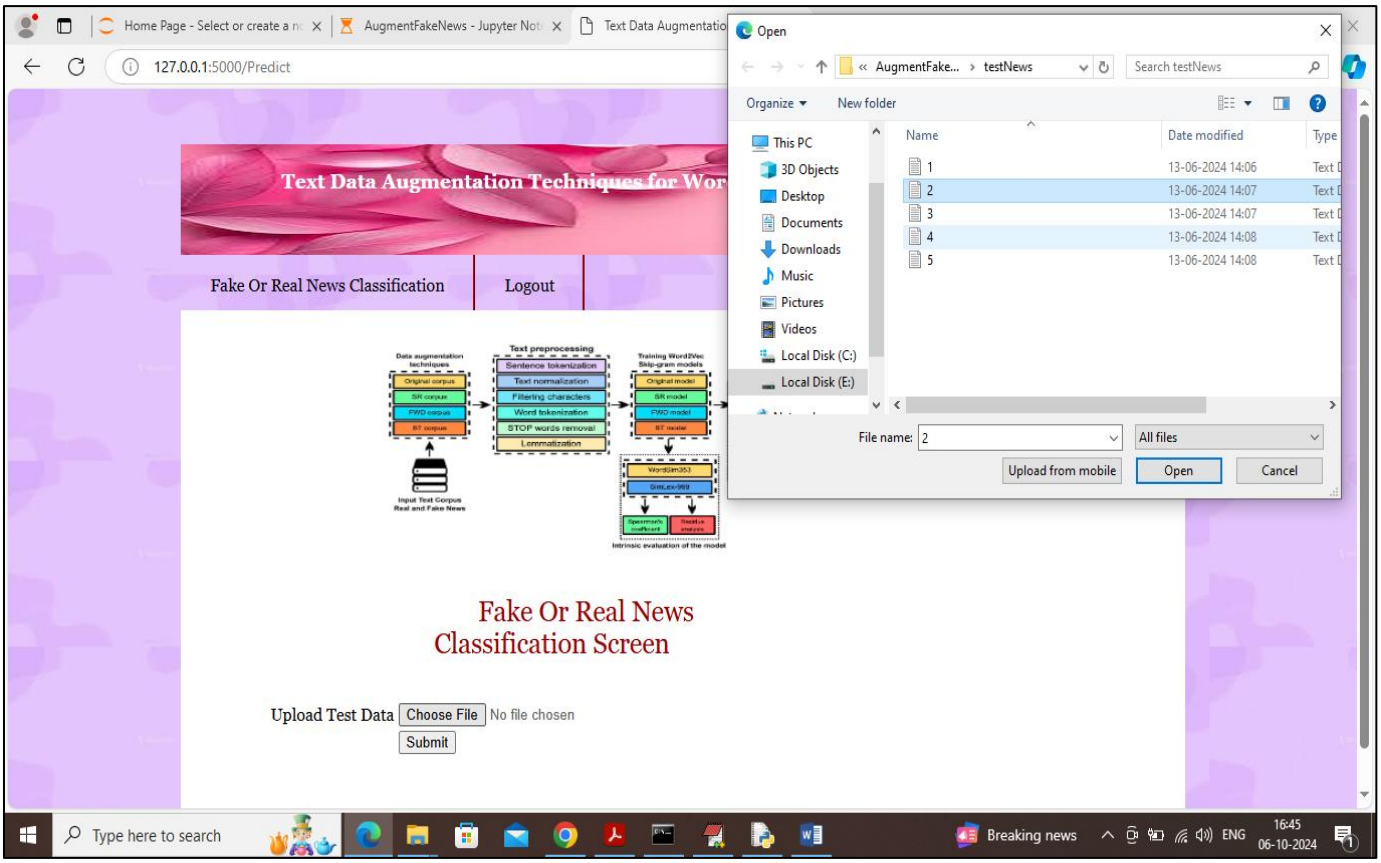


FIG .6 'Open and Submit'

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT
ISSN 2454-9940
www.ijasem.org
Vol 19, Issue 2, 2025

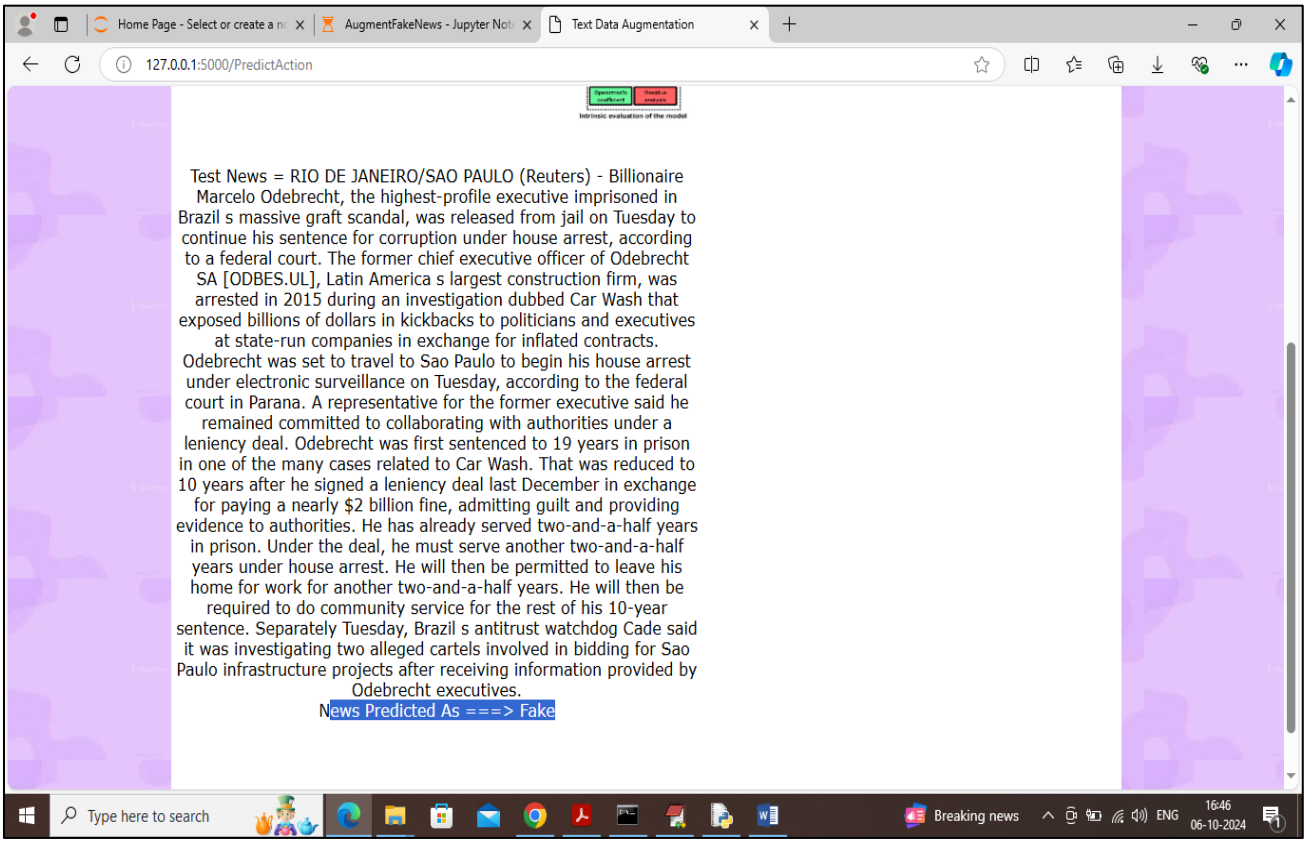


FIG.7 Results obtained by Extention XGboost

IV. CONCLUSION

Results show that text data augmentation helps classify fake news better. With the use of Function Word Reduction, Synonym Replacement, and Back Translation, the classification accuracy and datasets were greatly enhanced.

The best algorithms for Back Translation-augmented text were SVM and Bernoulli Naïve Bayes, among all those that were tested. In contrast, Logistic Regression demonstrated the impact of augmentation strategies on classifier performance by outperforming other methods while using Reduction of Function Words. When testing on the original corpus without any augmentation, Random Forest yielded the best results. The most accurate, though, was XGBoost, an ensemble algorithm that boosts predictive strength by combining decision trees.

This approach demonstrates that textual data can improve classification accuracy in limited datasets. The accuracy of the system's fake news identification is enhanced by advanced technologies such as XGBoost.

FUTURE SCOPE

Utilizing Contextualized Word Embeddings and Word Embedding Averaging can enhance this project.
Transformers and BERT are two examples of advanced deep learning models that could improve classification accuracy.

Potentially more effective methods include hybrid models that use machine learning techniques and ensemble procedures.

To further aid the algorithm in detecting fake news in many languages, multi-lingual data augmentation can be considered.

REFERENCES

[1] S. Y. Feng, V. Gangal, J. Wei, S. Chandar, S. Vosoughi, T. Mitamura, and E. Hovy, ‘‘A survey of data augmentation approaches for NLP,’’ 2021, arXiv:2105.03075.

[2] L.F.A.O.Pellicer,T.M.Ferreira, andA.H.R.Costa,‘‘Dataaugmentation techniques in natural language processing,’’ Appl. Soft Comput., vol. 132, Jan. 2023, Art. no. 109803, doi: 10.1016/j.asoc.2022.109803.

[3] S. Nazir, M. Asif, S. A. Sahi, S. Ahmad, Y. Y. Ghadi, and M. H. Aziz, ‘‘Toward the development of large-scale word embedding for low resourced language,’’ IEEE Access, vol. 10, pp. 54091–54097, 2022, doi: 10.1109/ACCESS.2022.3173259.

[4] E. Agirre, E. Alfonseca, K. Hall, J. Kravalova, M. Pasca, and A. Soroa, ‘‘A study on similarity and relatedness using distributional and WordNet based approaches,’’ in Proc. Human Lang. Technologies, Annu. Conf. North Amer. Chapter Assoc. Comput. Linguistics, 2009, pp. 19–27.

[5] F. Hill, R. Reichart, and A. Korhonen, ‘‘SimLex-999: Evaluating semantic models with (Genuine) similarity estimation,’’ Comput. Linguistics, vol. 41, no. 4, pp. 665–695, Dec. 2015, doi: 10.1162/coli_a_00237.

[6] J. Wei and K. Zou, ‘‘EDA: Easy data augmentation techniques for boosting performance on text classification tasks,’’ in Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process., 2019, pp. 6381–6387, doi: 10.18653/v1/d19-1670.

[7] G. A. Miller, ‘‘WordNet,’’ Commun. ACM, vol. 38, no. 11, pp. 39–41, Nov. 1995, doi: 10.1145/219717.219748.

[8] V. Marivate and T. Sefara, ‘‘Improving short text classification through global augmentation methods,’’ in Proc. Int. Cross-Domain Conf. Mach. Learn. Knowl. Extraction, 2020, pp. 385–399, doi: 10.1007/978-3030-57321-8_21.

[9] S. Kobayashi, ‘‘Contextual augmentation: Data augmentation by words with paradigmatic relations,’’ in Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol., 2018, pp. 452–457, doi: 10.18653/v1/n18-2072.

[10] R. N. Al-Matham and H. S. Al-Khalifa, ‘‘SynoExtractor: A novel pipeline for Arabic synonymextraction using Word2 Vecwordembeddings,’’Complexity, vol. 2021, pp. 1–13, Feb. 2021, doi: 10.1155/2021/6627434.

[11] I. Salah, K. Jouini, and O. Korbaa, ‘‘On the use of text augmentation for stance and fake news detection,’’ J. Inf. Telecommun., vol. 7, no. 3, pp. 359–375, Jul. 2023, doi: 10.1080/24751839.2023.2198820.

[12] I. Salah, K. Jouini, and O. Korbaa, ‘‘Augmentation-based ensemble learning for stance and fake news detection,’’ in Proc. Int. Conf. Comput. Collective Intell., 2022, pp. 29–41, doi: 10.1007/978-3-031-16210-7_3.

[13] M. Bucos and G. Ţucudean, ‘‘Text data augmentation techniques for fake news detection in the Romanian language,’’ Appl. Sci., vol. 13, no. 13, p. 7389, Jun. 2023, doi: 10.3390/app13137389.

[14] A.J.Keya, M.A.H.Wadud, M.F.Mridha, M.Alatiyyah, and M.A.Hamid, ‘‘AugFake-BERT:Handling imbalance through augmentation of fake news using BERT to enhance the performance of fake news classification,’’Appl. Sci., vol. 12, no. 17, p. 8398, Aug. 2022, doi: 10.3390/app12178398.

[15] G. Haralabopoulos, M. T. Torres, I. Anagnostopoulos, and D. McAuley, ‘‘Text data augmentations: Permutation, antonyms and negation,’’ Expert Syst. Appl., vol. 177, Sep. 2021, Art. no. 114769, doi: 10.1016/j.eswa.2021.114769.

[16] A. Dahou, A. A. Ewees, F. A. Hashim, M. A. A. Al-Qaness, D. A. Orabi, E. M. Soliman, E. M. Tag-Eldin, A. O. Aseeri, and M. A. Elaziz, ‘‘Optimizing fake news detection for Arabic context: A multitask learning approach with transformers and an enhanced nutcracker optimization algorithm,’’ Knowl.-Based Syst., vol. 280, Nov. 2023, Art. no. 111023, doi: 10.1016/j.knosys.2023.111023.

[17] J. Hua, X. Cui, X. Li, K. Tang, and P. Zhu, ‘‘Multimodal fake news detection through data augmentation-based contrastive learning,’’ Appl. Soft Comput., vol. 136, Mar. 2023, Art. no. 110125, doi: 10.1016/j.asoc.2023.110125.
[18] M. I. Marwat, J. A. Khan, M. D. Alshehri, ‘‘Sentiment analysis of product reviews to identify deceptive rating information in social media: ASentiDeceptive approach,’’ KSII Trans. Internet Inf. Syst., vol. 16, no. 3, pp. 830–860, Dec. 2022, doi: 10.3837/tiis.2022.03.005.
[19] J. A. Khan, A. Yasin, R. Fatima, D. Vasan, A. A. Khan, and A. W. Khan, ‘‘Valuating requirements arguments in the online user’s forum for requirements decision-making: The CrowdRE-VArg framework,’’ Softw., Pract. Exper., vol. 52, no. 12, pp. 2537–2573, Dec. 2022, doi: 10.1002/spe.3137.
[20] M. Risdal. (2016). Getting Real About Fake News. Kaggle. Accessed: Dec. 28, 2023. [Online]. Available: https://www.kaggle.com/code/ anthonyc1/gathering-real-news-for-oct-dec-2016/output 31550
[21] S.Bird,E.Klein,andE.Loper,NaturalLanguageProcessingWithPython: Analyzing Text With the Natural Language Toolkit, 1st ed. Sebastopol, CA, USA: O’Reilly Media, 2009.
[22] P. K. Verma, P. Agrawal, I. Amorim, and R. Prodan, ‘‘WELFake: Word embedding over linguistic features for fake news detection,’’ IEEE Trans. Computat. Social Syst., vol. 8, no. 4, pp. 881–893, Aug. 2021. [Online].
Available: https://www.kaggle.com/datasets/saurabhshahane/fake-newsclassification/data
[23] J. Tiedemann and S. Thottingal, ‘‘OPUS-MT—Building open translation services for the world,’’ in Proc. 22nd Annu. Conf. Eur. Assoc. Mach. Transl., A. Martins, H. Moniz, S. Fumega, B. Martins, F. Batista, L. Coheur, C. Parra, I. Trancoso, M. Turchi, A. Bisazza, J. Moorkens, A. Guerberof, M. Nurminen, L. Marg, M. L. Forcada, Eds., 2020, pp. 479–480.
[24] R. Řehuřek and P. Sojka, ‘‘Software framework for topic modelling with large corpora,’’ in Proc. LREC Workshop New Challenges for NLP Frameworks, ELRA, 2010, pp. 45–50.
[25] T. Mikolov, K. Chen, G. Corrado, and J. Dean, ‘‘Efficient estimation of word representations in vector space,’’ 2013, arXiv:1301.3781.
[26] M. Zhai, J. Tan, and J. Choi, ‘‘Intrinsic and extrinsic evaluations of word embeddings,’’ in Proc. AAAI Conf. Artif. Intell., Nov. 2016, vol. 30, no. 1, pp. 4282–4283, doi: 10.1609/aaai.v30i1.9959.
[27] Y. Shi, Y. Zheng, K. Guo, L. Zhu, and Y. Qu, ‘‘Intrinsic or extrinsic evaluation: An overview of word embedding evaluation,’’ in Proc. IEEE Int. Conf. Data Mining Workshops (ICDMW), Nov. 2018, pp. 1255–1262,
doi: 10.1109/ICDMW.2018.00179.