

# Text Data Augmentation Techniques for Word Embedding's in Fake News Classification

Dr. N. Sai Sindhuri <sup>[1]</sup>, Thummalapenta Bhargav <sup>[2]</sup>, Thupakula Chandu <sup>[3]</sup>, Shaik Taju <sup>[4]</sup>, Padavala Sreyanth <sup>[5]</sup>,  
Sada Vinay <sup>[6]</sup>,

Associate Professor <sup>[1]</sup>, Student <sup>[2][3][4][5][6]</sup>, Department of Computer Science and Engineering, Geethanjali  
Institute of Science and Technology, Nellore, Andhra Pradesh-524137

**Abstract:** Contemporary language models rely heavily on large training corpora to enhance their ability to capture semantic relationships. However, limited corpus size can negatively impact classification accuracy. To mitigate this, various text data augmentation techniques were applied to improve fake news classification. Techniques such as Synonym Replacement (SR), Back Translation (BT), and Reduction of Function Words (FWD) were utilized, with text representations converted into numeric vectors using the Word2Vec Skip-gram model. Classifiers, including Random Forest, Support Vector Machine (SVM), Logistic Regression, Bernoulli Naïve Bayes, and XGBoost, were evaluated on the WEL Fake News dataset. The experiments revealed that SVM and Naïve Bayes performed best on BT-augmented text, Logistic Regression achieved the highest accuracy with FWD, and Random Forest excelled with the original text. The highest overall accuracy of 91% was achieved using XGBoost on the original corpus. Performance metrics such as accuracy, precision, recall, and F1-score were used for evaluation, highlighting the significant potential of data augmentation in enhancing classification performance in scenarios with limited data.

**Index Terms** - Code Smells, Bug Prediction, Machine Learning, Random Forest, PROMISE Dataset, Code Metrics, Voting Classifier, Software Maintainability, Accuracy, Anti-Patterns.

## 1. INTRODUCTION

Data Augmentation (DA) refers to any technique that increases the diversity of training examples without the need to collect new data explicitly. The primary objective of DA is to enhance the performance and robustness of machine learning models by exposing them to a broader range of variations and scenarios [11], [12]. Additionally, DA mitigates overfitting by introducing variation into the training data, thereby reducing the likelihood of the model memorizing specific

examples rather than generalizing effectively [6], [12].

While data augmentation has been extensively successful in domains such as computer vision and speech recognition, its application in natural language processing (NLP) has been comparatively limited [11]. Popular techniques for text data augmentation include Back Translation, Synonym Replacement, Paraphrasing, Random Insertion, Random Swap, and Random Deletion [6], [8], [9]. A comprehensive overview of these techniques is provided in recent surveys and studies [11], [12].

In NLP, the necessity of text vectorization techniques for classification tasks has grown significantly. Word embedding models such as Word2Vec, Doc2Vec, and Glove rely on capturing semantic similarities among words and remain widely used in this domain [13], [7]. These embedding models enable the effective transformation of textual data into numeric vectors, forming the foundation for many machine learning-based classification tasks. Despite the progress, there is a growing need for innovative DA methods tailored to NLP challenges to achieve the same level of success observed in computer vision applications [9], [11].

## 2. RELATED WORK

Text data augmentation has been widely studied in recent years, particularly as a solution to improve model performance in scenarios with limited training data. Early efforts like Easy Data Augmentation (EDA) introduced techniques such as synonym replacement, random insertion, and random deletion to increase textual diversity [6]. These methods are simple yet effective and continue to be a baseline for augmentation in text classification tasks.

Back Translation, a technique where a sentence is translated into another language and then back to the original language, has emerged as one of the most robust approaches to text augmentation. Studies have demonstrated its effectiveness in improving classification performance, particularly in multilingual contexts [8]. Similarly, contextual augmentation, which replaces words with contextually appropriate alternatives, has also shown promise [9].

More recent advancements have focused on integrating deep learning-based methods for augmentation. Models like BERT and GPT have

been employed to generate paraphrased or augmented text that preserves semantic consistency [14]. In addition, multitask learning frameworks combined with optimization techniques, such as the Nutcracker Optimization Algorithm, have demonstrated significant improvements in fake news detection tasks within specific languages and domains [3].

Multimodal approaches have also gained traction. For example, data augmentation-based contrastive learning methods have been proposed to enhance multimodal fake news detection, achieving state-of-the-art results by leveraging both textual and visual data [4]. Additionally, ensemble methods utilizing multiple augmentation techniques have shown effectiveness in improving stance detection and fake news classification performance [5].

In the context of word embeddings, models like Word2Vec and Glove have been instrumental in converting text into numeric representations. These embeddings, when combined with augmentation techniques, further improve downstream classification tasks [7], [13]. For low-resource languages, augmentation techniques have been particularly critical. Synonym extraction using word embeddings has shown success in enhancing text diversity for underrepresented languages like Arabic [10].

Surveys and comparative studies highlight the breadth of data augmentation techniques available and their applications across various NLP tasks. These studies underscore the need for tailored augmentation strategies to address task-specific challenges, particularly in domains like fake news detection [11], [12].

## 3. MATERIALS AND METHODS

In this proposed system, we aim to enhance fake news classification by applying text data augmentation techniques to expand the training dataset. Techniques such as Synonym Replacement (SR) [15], Back Translation (BT) [8], and Reduction of Function Words (FWD) [9] will be employed to create varied versions of the original text, which will then be transformed into numeric vectors using the Word2Vec Skip-gram model [13]. These vectors will be used as inputs for multiple machine learning classifiers, including Random Forest, Support Vector Machine (SVM), Logistic Regression, and Bernoulli Naïve Bayes, to assess the impact of data augmentation on classification accuracy [6], [7]. Additionally, to improve performance further, advanced ensemble algorithms such as XGBoost will be integrated into the system [14]. This system is designed to address the challenges posed by limited datasets and improve the overall effectiveness of fake news detection. Performance evaluation will be conducted using metrics such as accuracy, precision, recall, and F1-score, with the goal of demonstrating how augmentation techniques can enhance the classification performance in scenarios with limited data [11], [12].

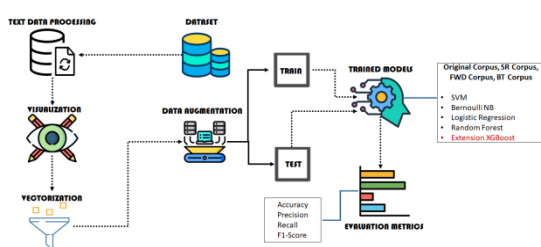


Fig.1 Proposed Architecture

The system architecture (fig. 1) depicts a machine learning pipeline for text classification. It begins with text data processing, where raw text is prepared and stored in a dataset. Data is visualized and vectorized to convert text into numerical representations. Data augmentation is applied to enhance the dataset's diversity. The data is then split

into training and testing sets. Various models, including SVM, Bernoulli Naive Bayes, Logistic Regression, Random Forest, and XGBoost, are trained and tested. The trained models' performance is evaluated using metrics such as accuracy, precision, recall, and F1-score. The workflow ensures robust and reliable text classification outcomes.

### i) Dataset Collection:

(WELFake) is a dataset of 72,134 news articles with 35,028 real and 37,106 fake news. For this, authors merged four popular news datasets (i.e. Kaggle, McIntire, Reuters, BuzzFeed Political) to prevent over-fitting of classifiers and to provide more text data for better ML training.

Dataset contains four columns: Serial number (starting from 0); Title (about the text news heading); Text (about the news content); and Label (0 = fake and 1 = real).

There are 78098 data entries in csv file out of which only 72134 entries are accessed as per the data frame.

Unnamed: 0	0	title	text	label
0	0	LAW ENFORCEMENT ON HIGH ALERT Following Threat...	No comment is expected from Barack Obama Membe...	1
2	2	UNBELIEVABLE! OBAMA'S ATTORNEY GENERAL SAYS MO...	Now, most of the demonstrators gathered last ...	1
3	3	Bobby Jindal, raised Hindu, uses story of Chri...	A dozen politically active pastors came here f...	0
4	4	SATAN 2: Russia unveils an image of its terrif...	The RS-28 Sarmat missile, dubbed Satan 2, will...	1
5	5	About Time! Christian Group Sues Amazon and SP...	All we can say on this one is it's about time ...	1

Fig 2 Dataset Collections

### ii) Pre-Processing:

Preprocessing is a crucial step in preparing the dataset for building predictive models, ensuring the data is suitable for analysis and machine learning.

**a) Text Data Processing:** Text data processing involves cleaning the dataset by removing stop words, stemming, and lemmatization, which standardizes the text and improves the quality and

relevance of the data for classification [12], [6]. This preprocessing step ensures that noise in the data is reduced, allowing machine learning algorithms to focus on meaningful patterns.

**b) Visualization:** Visualization techniques help in understanding the distribution of real and fake news within the dataset. By plotting graphs that represent the counts of different news types, insights into the composition of the data can be derived, which aids in the analysis and helps inform model selection [14].

**c) Vectorization (Word to Vector):** Vectorization transforms text into numerical representations using models like Word2Vec [13]. The Skip-gram model is employed to generate word vectors that capture semantic relationships, enabling algorithms to process textual data more effectively by leveraging the semantic similarity between words [7], [13].

**d) Data Augmentation:** Data augmentation techniques, including Synonym Replacement (SR) [15], Back Translation (BT) [8], and Reduction of Function Words (FWD) [9], are applied to enhance the dataset. These methods introduce diversity into the training data, increasing its robustness and improving the accuracy of classification models [6], [12]. By augmenting the dataset, the system can overcome challenges associated with limited data, ultimately leading to better performance in fake news detection tasks.

### iii) Training & Testing:

The training process involves using the preprocessed and augmented dataset to train multiple machine learning classifiers, including Random Forest, SVM, Logistic Regression, and Bernoulli Naïve Bayes. The models are trained on the augmented data, enabling them to learn from the variations introduced by techniques like Synonym

Replacement, Back Translation, and Reduction of Function Words. For testing, the models are evaluated on a separate test set, and performance metrics such as accuracy, precision, recall, and F1-score are computed. The system's robustness is further evaluated using advanced ensemble algorithms like XGBoost to assess improvements in fake news detection [6], [12], [14].

### iv) Algorithms:

**SVM (Support Vector Machine):** The Support Vector Machine (SVM) classifies text data by identifying the optimal hyperplane that separates different classes. It uses various corpora, including the Original Corpus, Synonym Replacement (SR) Corpus [15], Reduction of Function Words (FWD) Corpus [9], and Back Translation (BT) Corpus [8], to improve classification accuracy, ensuring robust differentiation between fake and real news.

**Bernoulli Naïve Bayes:** Bernoulli Naïve Bayes employs a probabilistic approach, classifying text based on the presence or absence of words. It effectively utilizes the Original Corpus, SR Corpus, FWD Corpus, and BT Corpus [6], [8] to enhance performance, providing reliable results in distinguishing fake news from real news.

**Logistic Regression:** Logistic Regression is suited for binary classification, predicting outcomes like fake or real news. By incorporating the Original Corpus, SR Corpus [15], FWD Corpus [9], and BT Corpus [8], it improves accuracy and assesses the impact of various text augmentation techniques on model performance.

**Random Forest:** Random Forest, an ensemble learning method, constructs multiple decision trees for classification. By using the Original Corpus, SR Corpus [15], FWD Corpus [9], and BT Corpus [8], it enhances classification accuracy and robustness

against overfitting, distinguishing effectively between fake and real news [6].

**XGBoost:** XGBoost is an advanced ensemble algorithm utilizing gradient boosting to improve prediction accuracy. Leveraging the Original Corpus, SR Corpus [15], FWD Corpus [9], and BT Corpus [8], it significantly enhances performance metrics and is highly effective in classifying news articles as either fake or real [14].

#### 4. RESULTS & DISCUSSION

**Accuracy:** The accuracy of a test is its ability to differentiate the patient and healthy cases correctly. To estimate the accuracy of a test, we should calculate the proportion of true positive and true negative in all evaluated cases. Mathematically, this can be stated as:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (1)$$

**Precision:** Precision evaluates the fraction of correctly classified instances or samples among the ones classified as positives. Thus, the formula to calculate the precision is given by:

$$Precision = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}} \quad (2)$$

**Recall:** Recall is a metric in machine learning that measures the ability of a model to identify all relevant instances of a particular class. It is the ratio of correctly predicted positive observations to the total actual positives, providing insights into a model's completeness in capturing instances of a given class.

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

**F1-Score:** F1 score is a machine learning evaluation metric that measures a model's accuracy. It combines the precision and recall scores of a model. The accuracy metric computes how many times a model made a correct prediction across the entire dataset.

$$F1 \text{ Score} = 2 * \frac{Recall \times Precision}{Recall + Precision} * 100 \quad (1)$$

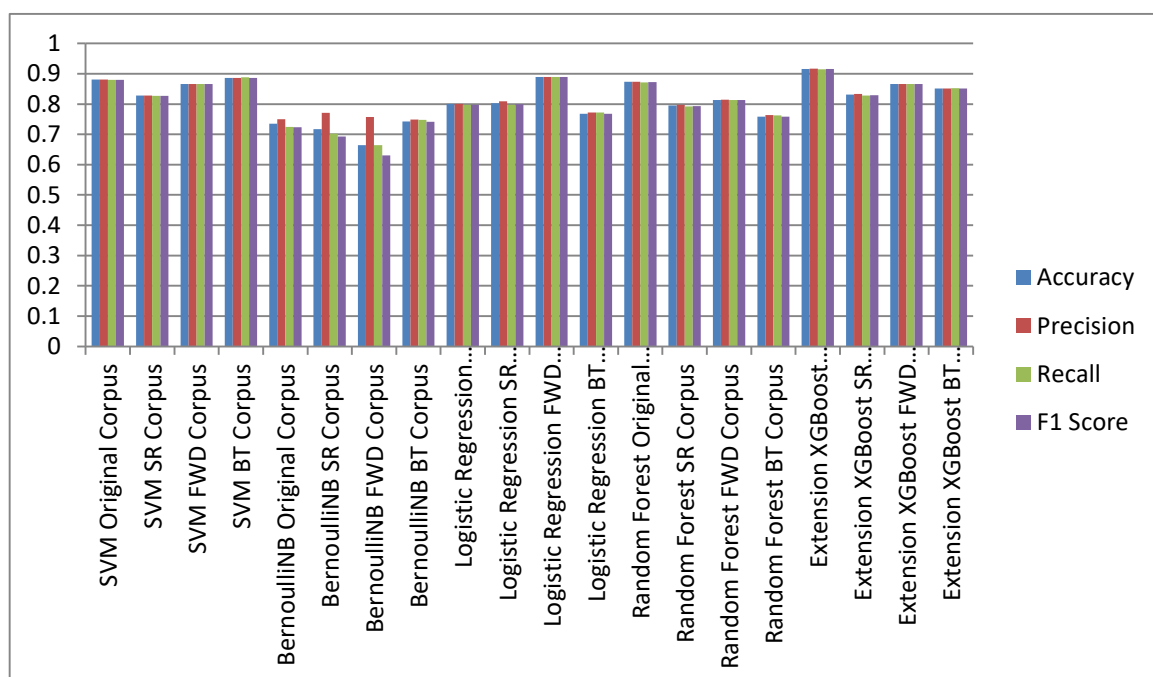
In Table 1, the performance metrics—Accuracy, Precision, Recall and F1 Score—are evaluated for each algorithm. The Extension XGBoost achieves the highest scores. Other algorithms' metrics are also presented for comparison.

Table.1 Performance Evaluation Metrics of Classification

Model	Accuracy	Precision	Recall	F1 Score
SVM Original Corpus	0.881	0.881	0.880	0.880
SVM SR Corpus	0.828	0.828	0.827	0.827
SVM FWD Corpus	0.866	0.866	0.866	0.866
SVM BT Corpus	0.886	0.886	0.888	0.886
BernoulliNB Original Corpus	0.735	0.750	0.725	0.724
BernoulliNB SR Corpus	0.717	0.771	0.702	0.693
BernoulliNB FWD Corpus	0.664	0.757	0.664	0.631
BernoulliNB BT Corpus	0.743	0.749	0.748	0.742

Logistic Regression Original Corpus	0.801	0.801	0.797	0.798
Logistic Regression SR Corpus	0.803	0.809	0.798	0.800
Logistic Regression FWD Corpus	0.889	0.889	0.889	0.889
Logistic Regression BT Corpus	0.768	0.772	0.772	0.768
Random Forest Original Corpus	0.874	0.874	0.872	0.873
Random Forest SR Corpus	0.795	0.798	0.792	0.793
Random Forest FWD Corpus	0.813	0.814	0.813	0.813
Random Forest BT Corpus	0.758	0.764	0.763	0.758
Extension XGBoost Original Corpus	0.916	0.917	0.915	0.916
Extension XGBoost SR Corpus	0.831	0.833	0.828	0.829
Extension XGBoost FWD Corpus	0.866	0.866	0.866	0.866
Extension XGBoost BT Corpus	0.851	0.851	0.852	0.851

Graph.1 Comparison Graphs of Classification



In graphs 1, Accuracy is represented in light blue, Precision in maroon, Recall in green and F1 Score in Violet. In comparison to the other models, the Extension XGBoost shows superior performance across all achieving the highest values. The graphs above visually illustrate these findings.

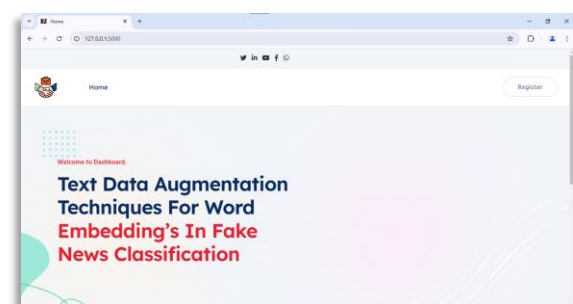
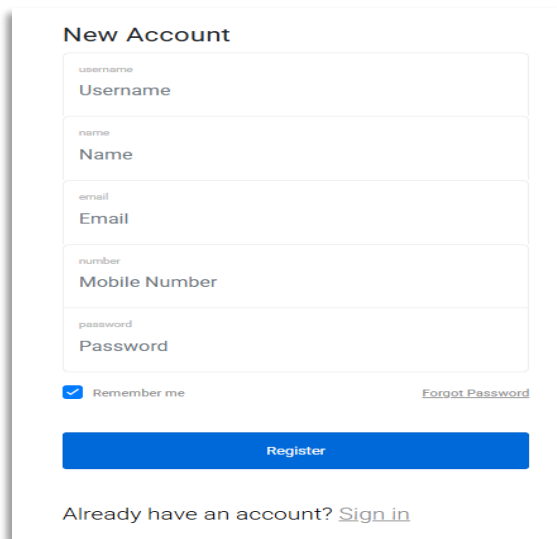


Fig.3 Home Page



In above fig.3 user interface dashboard with text data augmentation techniques for fake news classification and word embedding.



**New Account**

USERNAME  
Username

NAME  
Name

EMAIL  
Email

NUMBER  
Mobile Number

PASSWORD  
Password

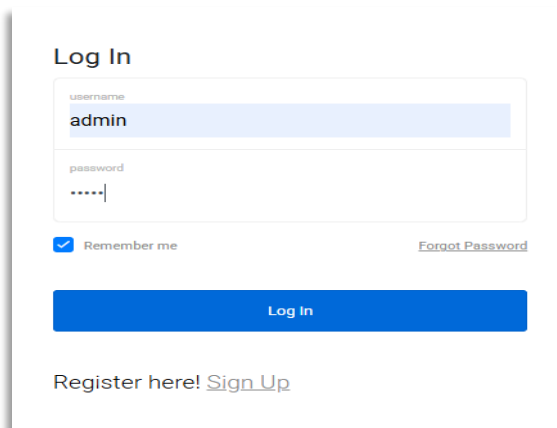
☒ Remember me [Forgot Password](#)

**Register**

Already have an account? [Sign in](#)

Fig.4 Signup Page

In above fig.4 shows a new account registration form with fields for username, name, email, mobile number, and password.



**Log In**

USERNAME  
admin

PASSWORD  
.....

☒ Remember me [Forgot Password](#)

**Log In**

Register here! [Sign Up](#)

Fig.5 Signin Page

In above fig.5 displays a user login form with fields for username and password, including "Remember me" and "Forgot Password" options.

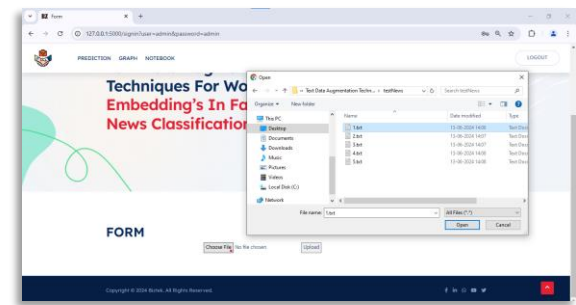


Fig.6 Upload Input File

In above fig.6 shows a BZ form with file upload functionality, navigation tabs, and a logout option.

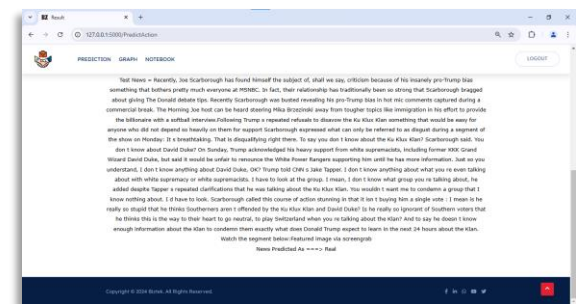


Fig.7 Predicted Result as Real for uploaded input file

In above fig.7 displays the BZ result, showing predicted news as "Real" with navigation tabs and a logout option.

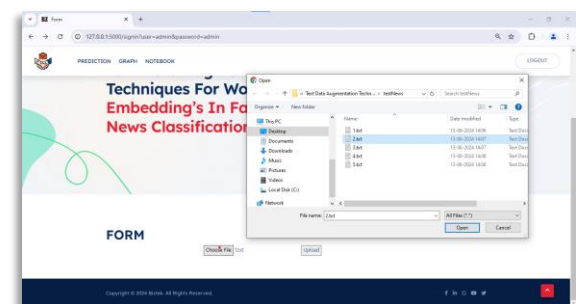


Fig.8 Upload Another Input File

In above fig.8 BZ form with file selection dialog open, showing text files for upload.

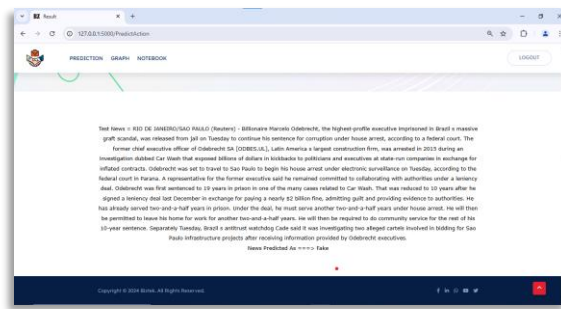


Fig.9 Final Outcome as Fake for uploaded test input file

In above fig.9 BZ form predicts news as "Fake" with navigation tabs and logout option.

## 5. CONCLUSION

In conclusion, this study successfully demonstrates the effectiveness of applying text data augmentation techniques to enhance the performance of fake news classification. By employing methods such as Synonym Replacement (SR), Back Translation (BT), and Reduction of Function Words (FWD), we generated augmented datasets that contributed to improved classification accuracy. Among the algorithms tested, Support Vector Machine (SVM) and Bernoulli Naïve Bayes exhibited the highest performance when trained on Back Translation-augmented text. Logistic Regression performed optimally with the Reduction of Function Words technique, highlighting the impact of different augmentation methods on classifier performance. The Random Forest algorithm, however, delivered the best results using the original corpus without any augmentation. The highest overall accuracy was achieved by XGBoost, an ensemble algorithm that uses gradient boosting to combine decision trees for enhanced predictive power. This system demonstrates that augmenting textual data can significantly improve classification accuracy, particularly in scenarios with limited datasets, and that advanced techniques like XGBoost can further

improve the system's ability to detect fake news with greater reliability and precision.

**Future Scope:** In the future, this project can be expanded by exploring additional text data augmentation techniques, such as Word Embedding Averaging and Contextualized Word Embeddings. Implementing advanced deep learning models, including Transformers and BERT, could further boost classification accuracy and improve the system's ability to capture nuanced semantic relationships. Additionally, incorporating ensemble methods and hybrid models that combine various machine learning algorithms may yield better results. Investigating multilingual data augmentation could also broaden the system's applicability, allowing for fake news detection across different languages and cultural contexts.

## REFERENCES

- [1] 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.
- [2] M. Bucos and G. Țucudean, "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.
- [3] 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.



- [4] 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.
- [5] 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.
- [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: 10.1007/978-3-030-57321-8\_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 synonym extraction using Word2 Vec word embeddings," *Complexity*, vol. 2021, pp. 1–13, Feb. 2021, doi: 10.1155/2021/6627434.
- [11] 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.
- [12] L. F. A. O. Pellicer, T. M. Ferreira, and A. H. R. Costa, "Data augmentation techniques in natural language processing," *Appl. Soft Comput.*, vol. 132, Jan. 2023, Art. no. 109803, doi: 10.1016/j.asoc.2022.109803.
- [13] 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.
- [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] 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.
- [17] 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.

[18] M. I. Marwat, J. A. Khan, M. D. Alshehri, “Sentiment analysis of product reviews to identify deceptive rating information in social media: A SentiDeceptive 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/anthonymc1/gathering-real-news-for-oct-dec-2016/output>