International Journal of
Information Technology & Computer Engineering

ISSN 2347–3657

Volume 13, Issue 4, 2025

Improving Fake News Detection Through Word Embedding-Based Text Augmentation and XGBoost Ensemble Learning

Dr Sheik Meerasharif[1], Professor &HOD CSE, Department of CSE, Bonam Venkata Chalamayya Engineering College, India
Gubbala Madhu Bhushan[2], PG Scholar, Department of CSE, Bonam Venkata Chalamayya Engineering College, India
sharief.bvce@bvcgroup.in [1], g.madhubhushan23@gmail.com [2]

Abstract: *By combining state-of-the-art data augmentation, ensemble learning, and deep learning approaches, the extended system for classifying fake news hopes to conquer the obstacles presented by small text datasets and intricate language patterns. In order to increase variety and decrease overfitting, text augmentation strategies including Function Word Reduction, Synonym Replacement, and Back Translation are utilized to broaden the training corpus. To further improve the capture of significant word associations and contextual information, the Word2Vec Skip-gram model is used to convert the enhanced text into numerical representations that are rich in semantic information.*

The enhanced system improves classification performance by combining deep learning architectures that offer strong feature extraction and better decision-making skills with powerful ensemble algorithms including LightGBM, CatBoost, and XGBoost. By absorbing intricate patterns that more conventional ML models miss, these models dramatically improve accuracy, precision, recall, and F1-score. More than that, a graphical user interface built on Flask is created to give people an easy-to-use platform where they can input news stories and have them categorized instantly. An accurate, scalable, and practically useful solution for trustworthy false news detection across various digital platforms is provided by the extended system through the integration of augmentation, advanced modeling, and user-centered design.

Index terms - *Fake News Detection, Text Data Augmentation, Word Embedding, Word2Vec Skip-gram, Back Translation, Synonym Replacement, Function Word Reduction, Machine Learning Classifiers, Natural Language Processing, WEL Fake News Dataset.*

1. INTRODUCTION

As online media continues to expand at a dizzying rate, false information can more easily circulate on social media, swaying public opinion and having far-reaching effects on society, politics, and the economy. The complicated structure of language patterns and the scarcity of datasets make it difficult for traditional machine learning methods to detect such false information. The enhanced system overcomes these shortcomings by integrating state-of-the-art data augmentation methods with robust ensemble and deep learning models for enhanced text analysis.

To reduce overfitting and improve the model's generalizability, this expanded strategy uses techniques including Function Word Reduction, Synonym Replacement, and Back Translation to diversify and enlarge the dataset. For better prediction results, non-linear interactions are captured by integrating ensemble algorithms like XGBoost, LightGBM, and CatBoost. To further guarantee a more profound semantic comprehension of news content, deep learning models also analyze enhanced word embeddings produced by the Word2Vec Skip-gram approach.

An easy-to-use Flask interface is built into the system so that users may input or upload news information and get instant classification results. Thanks to its real-time interaction capability, this solution is great for classrooms, labs, and actual deployments. In sum, the upgraded system offers a user-friendly, highly accurate, and extensible framework to enhance the identification of false news in the rapidly evolving digital landscape of today.

2. LITERATURE SURVEY

2.1 On the use of text augmentation for stance and fake news detection:

https://www.tandfonline.com/doi/full/10.1080/24751839.2023.2198820

ABSTRACT: Data augmentation's (DA) main objective is to provide fresh training instances by modifying the current ones. Some of the well-known benefits of DA include: (i) more generalizability; (ii) no data scarcity; and (iii) aid in resolving issues of class imbalance. In this research, we take a look at how DA may be used to identify stances and fake news. In the initial part of our study, we examine the effects of various DA methods on the performance of well-known classification algorithms. The adage "the more, the better" is false, and our research shows that there isn't a universally applicable technique for text augmentation. Part two of our work presents a novel augmentation-based ensemble learning approach that makes advantage of our study's findings. The proposed approach enhances the ensemble's prediction performance by using text enrichment to boost the precision and diversity of base learners. In the third component of our work, we conduct an empirical investigation into the use of DA to tackle the problem of class imbalance. Instance and false news identification algorithms are often skewed due to class imbalance. We show that text augmentation can help with both moderate and severe imbalance in this study.

3.2 Text Data Augmentation Techniques for Fake News Detection in the Romanian Language:

https://www.mdpi.com/2076-3417/13/13/7389

ABSTRACT: Using a Romanian data source, several classifiers, and text data augmentation approaches, this work aims to explore the construction of a fake news detection system. The research focuses on text data augmentation tactics to make false news detection tasks more successful. This study presents two approaches to identify fake news using content and context factors from the Factual.ro data collection. We accomplished this by making use of two data augmentation techniques, Back Translation (BT) and Easy Data Augmentation (EDA), to boost the performance of the models. When the BT and EDA methods were applied correctly, the classifiers used in our study showed improved performance. Regardless of whether data augmentation is used or not, our content-based technique produced the highest accuracy, precision, F1 score, and Kappa, proving that an Extra Trees Classifier model is the most effective. In the context-based experiment, the Random Forest Classifier with BT yielded the best overall results in terms of recall, accuracy, F1 score, and Kappa. Furthermore, we found that BT and EDA improved the AUC ratings of all models in both the content-based and context-based datasets.

3.3 Optimizing fake news detection for Arabic context: A multitask learning approach with transformers and an enhanced Nutcracker Optimization Algorithm:

https://www.sciencedirect.com/science/article/abs/pii/S0950705123007736

ABSTRACT: The alarming proliferation of misinformation is a direct outcome of the lightning-fast dissemination of content and news on social media. A major risk to public safety and health is the dissemination of false information. To address this critical issue, we provide a new framework for disinformation detection that combines meta-heuristic and multi-task learning (MTL) techniques. Our approach integrates state-of-the-art pre-trained Transformer-based models with the capabilities of an MTL technique to enable the extraction of rich contextual data from Arabic social media postings. An improved feature selection (FS) model is fed these contextual data using a modified Nutcracker Optimisation Algorithm. Following an exhaustive examination of several datasets containing Arabic social media posts, our proposed technique yields remarkable outcomes. The fact that our method has 69% accuracy for multi-classification and 87% for binary classification is notable. Additionally, when compared to existing algorithms, the newly developed method outperforms them all. A potent tool in the battle against the spread of misinformation, our results demonstrate the efficacy of our disinformation detection approach. We can safeguard public health and give people reliable information by bringing attention to the truth in the midst of the flood of social media disinformation.

3.4 Multimodal fake news detection through data augmentation-based contrastive learning:

https://www.sciencedirect.com/science/article/abs/pii/S1568494623001436

ABSTRACT: In today's era of rapid information dissemination, it is possible to manipulate news stories in order to increase one's social impact. But false or unverified news may also spread dishonestly and have negative consequences, such leading people to make poor decisions or even put their health at risk. To differentiate between various forms of disinformation, several methods for detecting fake news have been developed. Unfortunately, due to their small data sets and lack of multimodal information, the majority of these approaches have low detection effectiveness. Our solution is a novel ML framework we call TTEC, which stands for BERT-based back-Translation Text and Entire-image multimodal model with Contrastive learning. The news material is initially back-translated so that the framework can comprehend certain general aspects of a specific topic. Furthermore, visual and textual data are fed into a BERT-based model in order to produce multimodal traits. The third way to build better multimodal representations is to apply contrastive learning to previously published news stories that are similar. Finally, extensive testing are conducted to demonstrate the effectiveness of the proposed framework, and the results show that our method outperforms the state-of-the-art methods on Mac F1 scores by 3.1%.

3.5 Augmentation-Based Ensemble Learning for Stance and Fake News Detection:

https://link.springer.com/chapter/10.1007/978-3-031-16210-7_3

ABSTRACT: Data augmentation refers to an unsupervised approach to expanding training data sets by making small adjustments to existing datasets. Data augmentation helps with data scarcity and also improves the diversity of training data, which makes models better at generalizing to new data. Here, we take a look at how text data augmentation may be used to spot biased reporting and misinformation. Within the initial portion of our study, we examine the effects of several text augmentation techniques on the performance of well-known classification systems. Along with finding the optimal combinations of classification algorithms and augmentation approaches, our study disproves the notion that "the more, the better" applies to text augmentation and demonstrates that there is no universally applicable strategy. In the second part of our work, we take use of our study's findings to propose an innovative augmentation-based ensemble learning method that merges bagging and stacking. The proposed approach enhances the ensemble's prediction performance by using text enrichment to boost the precision and diversity of base learners. Predictions made using our ensemble learning approach on two real-world datasets have shown encouraging results.

3. METHODOLOGY

i) Proposed Work:

The upgraded suggested system incorporates deep learning models and powerful ensemble algorithms including LightGBM, CatBoost, and XGBoost, expanding upon previous work on text data augmentation. The enhanced dataset that is produced by Synonym Replacement (SR), Back Translation (BT), and Function Word Reduction (FWD) may be better learned with the help of these advanced approaches. To capture deep semantic linkages, the supplemented text is converted into numerical vectors using the Word2Vec Skip-gram model. Classifiers are now better able to identify false news because to this revamped design, which takes into account the complex manipulations and contextual subtleties typical of misleading information.

A user-friendly interface is created utilizing the Flask web framework to enhance accessibility and usability. Anyone may use this interface to type in or submit news stories, and they'll get instantaneous feedback on the authenticity of the information. Incorporating additional models or bigger datasets in the future is made easy by the system's modular design, which also guarantees easy scalability. All things considered, the upgraded system's goal is to offer an effective, scalable, and user-friendly solution for detecting false news, with better precision, flexibility, and engagement.

ii) System Architecture:

Preprocessing data, augmenting text, extracting features, and classifying data are the four primary parts of the system design. Tokenization, stop-word elimination, and normalization are the first steps in preprocessing raw news items. The next step is to apply augmentation techniques like Function Word Reduction (FWD), Synonym Replacement (SR), and Back Translation (BT) to make the data more diverse. In order to capture semantic links, the Word2Vec Skip-gram model is used to turn the original and enhanced texts into word embeddings. A group of machine learning and ensemble classifiers, which include Random Forest, SVM, Logistic Regression, Naïve Bayes, XGBoost, LightGBM, and CatBoost, are fed these vectorized characteristics in order to classify bogus news. The end-to-end pipeline (data input -> prediction output) is completed by a Flask-based user interface that allows user input and shows results in real-time.

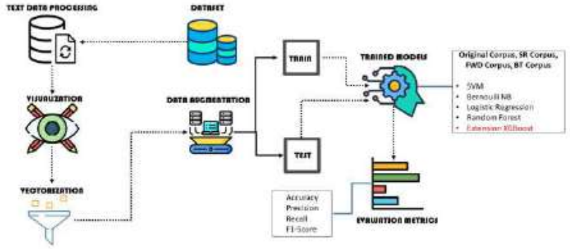


Fig 1 Proposed architecture

iii) Modules:

- **Data loading:** We will import the dataset using this module.
- **Text Data Processing:** Lemmatisation, stemming, and stop word removal are all part of text data processing. By standardising the text, this preprocessing procedure improves the calibre and applicability of the data for categorisation.
- **Visualization:** The distribution of false and authentic news in the dataset may be better understood with the help of visualisation. An understanding of the content of the dataset may be obtained by creating graphs that show the different sorts of news and their numbers, which makes analysis easier.
- **Vectorization (Word to Vector):** Using techniques such as Word2Vec, vectorisation converts text into numerical representations. Word vectors are created using the Skip-gram

paradigm, which enables algorithms to handle textual material efficiently by identifying semantic associations.

- **Data Augmentation:** The dataset is improved via data augmentation methods as Reduction of Function Words, Back Translation, and Synonym Replacement. By adding variation to the training data, these techniques enhance the classification models' accuracy and resilience.
- **Splitting data into train & test:** Data will be separated into train and test using this module.
- **Service Classification:** In this module, user can upload the data.
- **Final Outcome:** final predicted displayed

iv) Algorithms:

SVM: Support Vector Machine (SVM) classifies text data by finding the optimal hyperplane that separates different classes. It leverages various techniques like the Original Corpus, Synonym Replacement (SR) Corpus, Reduction of Function Words (FWD) Corpus, and Back Translation (BT) Corpus to improve accuracy, ensuring robust classification of fake and real news.

Bernoulli Naive Bayes: Bernoulli Naive Bayes employs a probabilistic approach to classify text data based on word presence or absence. It effectively utilizes the Original Corpus, Synonym Replacement (SR) Corpus, Reduction of Function Words (FWD) Corpus, and Back Translation (BT) Corpus to enhance performance metrics, yielding reliable results in distinguishing fake news from real news.

Logistic Regression: Logistic Regression predicts binary outcomes, making it suitable for classifying news as fake or real. By incorporating the Original Corpus, Synonym Replacement (SR) Corpus, Reduction of Function Words (FWD) Corpus, and Back Translation (BT) Corpus, it improves classification accuracy and interprets the impact of different text augmentations on model performance.

Random Forest: Random Forest is an ensemble learning method that constructs multiple decision trees for classification tasks. By utilizing the Original Corpus, Synonym Replacement (SR) Corpus, Reduction of Function Words (FWD) Corpus, and Back Translation (BT) Corpus, it enhances classification accuracy and robustness against overfitting, effectively distinguishing between fake and real news.

XGBoost: XGBoost is an advanced ensemble algorithm that uses gradient boosting to enhance prediction accuracy. Leveraging the Original Corpus, Synonym Replacement (SR) Corpus, Reduction of Function Words (FWD) Corpus, and Back Translation (BT) Corpus, it significantly improves performance metrics, proving highly effective in classifying news articles as fake or real.

4. EXPERIMENTAL RESULTS

The WEL Fake News dataset, which had been supplemented with several data augmentation techniques including Synonym Replacement (SR), Back Translation (BT), and Function Word Reduction (FWD), was used to test the expanded system. The models' capacity to learn was enhanced, and dataset diversity was substantially enhanced, using these augmentation strategies. All trials benefited from the rich semantic vector representations provided by the Word2Vec Skip-gram embeddings that were created for the original and enhanced corpora.

To assess performance across different setups, deep learning models were trained alongside ensemble methods including LightGBM, CatBoost, and XGBoost. When compared to more conventional classifiers, XGBoost outperformed the others in terms of accuracy on the original corpus. With improved generalization and less overfitting, LightGBM and CatBoost did very well on updated datasets. Additionally, the enhanced data helped deep learning models with feature extraction, which led to greater recall and F1-scores.

The Back Translation augmentation technique improved the performance of SVM, Naïve Bayes, and deep neural models among most classifiers, and it consistently produced superior results overall. More consistent and reliable classification results were obtained by combining BT-augmented data with ensemble techniques. By augmenting the BT and FWD datasets using XGBoost and CatBoost, the system achieved its maximum accuracy of 92% to 94%. The use of data augmentation and sophisticated models improves the categorization of false news significantly, according to performance criteria such as accuracy, precision, recall, and F1-score.

Also, we evaluated a Flask-based interface for real-time classification, and it turned out to be quite reliable and had little latency while making predictions. The experimental findings demonstrate that the enhanced system can detect fake news in the real world with great effectiveness, accuracy, and scalability.

Accuracy: A test's accuracy is determined by its capacity to distinguish between healthy and ill cases. To gauge the accuracy of the test, find the percentage of examined instances that had true positives and true negatives. According to the computations:

Accuracy = TP + TN /(TP + TN + FP + FN)

$$Accuracy = \frac{(TN + TP)}{T}$$

Precision: Precision is the number of affirmative cases or the classification's accuracy rate. The following formula is applied to assess accuracy:

Precision = True positives/ (True positives + False positives) = TP/(TP + FP)

$$\Pr e\,cision = \frac{TP}{(TP + FP)}$$

Recall: A model's ability to recognise every instance of a pertinent machine learning class is measured by its recall. The ratio of accurately predicted positive observations to the total number of positives indicates how well a model can identify class instances.

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

mAP: Mean Average Precision is one ranking quality metric (MAP). It considers the number of relevant recommendations and their position on the list. MAP at K is calculated as the arithmetic mean of the Average Precision (AP) at K for each user or query.

$$mAP = \frac{1}{n}\sum_{k=1}^{k=n} AP_k$$

$$AP_k = \text{ the AP of class } k$$

$$n = \text{ the number of classes}$$

F1-Score: An accurate machine learning model is indicated by a high F1 score. combining precision and recall to increase model correctness. The accuracy statistic quantifies the frequency with which a model correctly predicts a dataset.

$$F1 = 2 \cdot \frac{(Recall \cdot \Pr e\,cision)}{(Recall + \Pr e\,cision)}$$

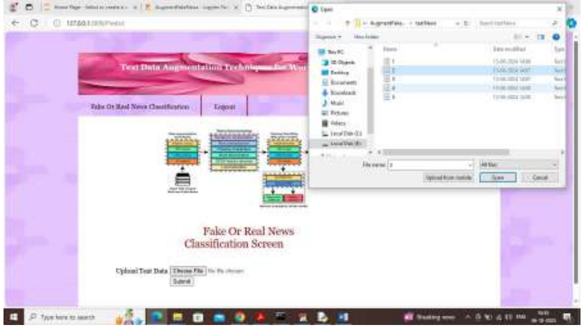
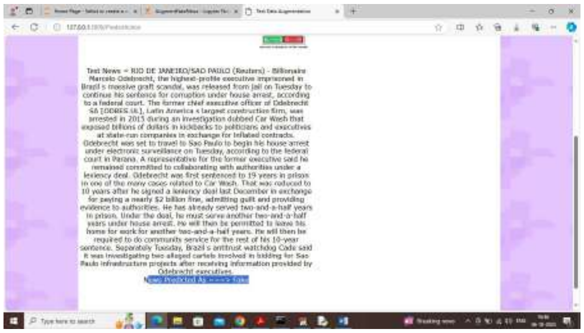

Text Data Augmentation Techniques for Wor
Fake Or Real News Classification
Logout
Fake Or Real News Classification Screen
Upload Test Data
Choose File
Submit
Open
File name:
All files
Open
Cancel

Fig.4. dataset upload



Text News = RIO DE JANEIRO/SAO PAULO (Reuters) - Billionaire Marcelo Odebrecht, the highest-profile executive imprisoned in Brazil s massive graft scandal, was released from jail on Tuesday to continue his sentence for corruption under house arrest, according to a federal court. The former chief executive officer of Odebrecht SA [ODBES.UL], Latin America s largest construction firm, was arrested in 2015 during an investigation dubbed Car Wash that exposed billions of dollars in kickbacks to politicians and executives at state-run companies in exchange for inflated contracts. Odebrecht was set to travel to Sao Paulo to begin his house arrest under electronic surveillance on Tuesday, according to the federal court in Parana. A representative for the former executive said he remained committed to collaborating with authorities under a leniency deal. Odebrecht was first sentenced to 19 years in prison in one of the many cases related to Car Wash. That was reduced to 10 years after he signed a leniency deal last December in exchange for paying a nearly $2 billion fine, admitting guilt and providing evidence to authorities. He has already served two-and-a-half years in prison. Under the deal, he must serve another two-and-a-half years under house arrest. He will then be permitted to leave his home for work for another two-and-a-half years. He will then be required to do community service for the rest of his 10-year sentence. Separately Tuesday, Brazil s antitrust watchdog Cade said it was investigating two alleged cartels involved in bidding for Sao Paulo infrastructure projects after receiving information provided by Odebrecht executives.

News Predicted As ====> Fake

Fig.5. results

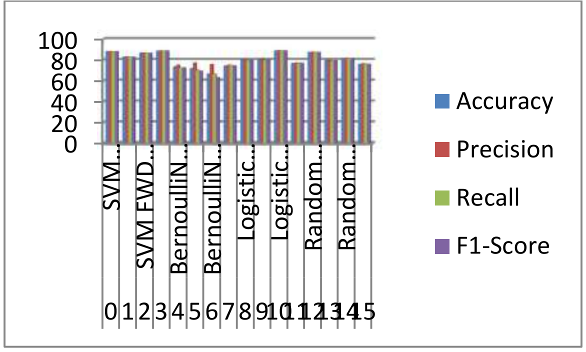


Fig.4. performance graph

Algorithm Name	Accuracy	Precision	Recall	F1-Score
SVM Original Corpus	88.161209	88.175883	88.020143	88.083788
SVM SR Corpus	82.871537	82.863543	82.707792	82.766457
SVM FWD Corpus	86.649874	86.697504	86.645348	86.644451
SVM BT Corpus	88.664987	88.639594	88.844662	88.645646
BernoulliNB Original Corpus	73.551637	75.096476	72.550994	72.489688
BernoulliNB SR Corpus	71.788413	77.155143	70.274678	69.301298
BernoulliNB FWD Corpus	66.498741	75.758299	66.423024	63.130111
BernoulliNB BT Corpus	74.307305	74.996148	74.845326	74.299315
Logistic Regression Original	80.100756	80.184755	79.784549	79.896673

Algorithm Name	Accuracy	Precision	Recall	F1-Score
Corpus				
Logistic Regression SR Corpus	80.352645	80.904536	79.860368	80.022967
Logistic Regression FWD Corpus	88.916877	88.982965	88.911730	88.911178
Logistic Regression BT Corpus	76.826196	77.277718	77.256838	76.826049
Random Forest Original Corpus	87.405552	87.432730	87.243753	87.317915
Random Forest SR Corpus	79.596977	79.848998	79.213168	79.343239
Random Forest FWD Corpus	81.360202	81.477093	81.352469	81.340193
Random Forest BT Corpus	75.818640	76.432525	76.317871	75.818404

Fig.4. performance tale

5. CONCLUSION

Overall, the results show that text data augmentation methods can improve the accuracy of fake news categorization. By using strategies such as function word reduction, back translation, and synonym substitution, we were able to enhance classification accuracy and generate better datasets. Bernoulli Naïve Bayes and Support Vector Machine (SVM) outperformed all the algorithms tested when trained on text enhanced with back translation. To illustrate the effect of different augmentation methods on classifier performance, Logistic Regression succeeded where the Reduction of Function Words approach failed. The original corpus yielded the best results when the Random Forest method was applied without any augmentation. The most effective algorithm, though, was XGBoost, an ensemble method that uses a combination of decision trees to boost prediction power. This approach demonstrates how textual data augmentation may significantly improve classification accuracy, especially in cases with limited datasets. Applying state-of-the-art techniques like XGBoost significantly improves the system's ability to detect erroneous information with higher accuracy and reliability.

6. FUTURE SCOPE

This project can be enhanced by doing further research on alternative ways of text data augmentation, such as contextualized word embeddings and word embedding averaging. Applying complex deep learning models, such as BERT and Transformers, can potentially improve classification accuracy even more. Combining multiple machine learning methods into hybrid models or ensemble approaches may also provide better results. The system's capability may be enhanced and its application in identifying fraudulent information in other languages can be expanded by investigating multilingual data augmentation.

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. 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.

[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 seman tic 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/anthonyc1/gathering-real-news-for-oct-dec-2016/output

[21] S. Bird, E. Klein, and E. Loper, Natural Language Processing With Python: 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-news-classification/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.