Augmenting NLP Training Data for Improved Fake News Identification

[1] ESARLA CHANDRA SEKHARA SAI, Student in Dept. Of Master of Computer Applications, at Miracle Educational Society Group of Institutions
[2] Bhadagiri Sai Prasad, Miracle Educational Society Group of Institutions
chandrasekharasai1234@gmail.com

ABSTRACT:

With the rapid proliferation of fake news, detection mechanisms are of paramount importance. This project is aimed at exploring text data augmentation and its effect on classification performance for a particular NLP task. For this project, the WELFake dataset was manipulated using three augmentation methods: Synonym Replacement, Back Translation, and Function Word Reduction. The resultant texts were represented using Word2Vec Skip-gram embeddings and were subsequently classified using SVM, Logistic Regression, Naïve Bayes, and Random Forest. The results indicated that Back Translation improved SVM and Naïve Bayes accuracy while Function Word Reduction improved Logistic Regression. The study demonstrates that data augmentation directly increases the diversity of the training dataset, reduces overfitting, and strengthens the semantic representation of the model, thus making the model more effective for fake news classification.

Keywords: NLP, Machine Learning, Fake News

INTRODUCTION

The detection of fake news has become a serious issue due to the rapid spread of fake news and the uptake of digital platforms. The extensive volume and ever-changing nature of online information makes traditional and manual fact verification methods, including rule-based approaches, obsolete. Attempts to resolve that using machine learning techniques have proven insufficient, as these approaches, along with NLP, are limited due to the absence of extensive and balanced datasets. This project attempts to fill this gap through data augmentation by changing the original corpus of training data. We used the methods of Synonym Replacement, Back Translation, and Function Word Reduction to create new datasets, which we processed using Word2Vec embeddings. Afterwards, we input the embeddings into machine learning classifiers to assess changes in detection accuracy. The study aims to prove that augmentation using well-defined strategies improves classification results and strengthens model generalization, making such augmentation crucial in the fight against misinformation.

RELATED WORK

Wei and Zou (2019) developed Easy Data Augmentation (EDA), a collection of techniques, including synonym and random insertion, to aid classification for petite datasets. While these techniques are beneficial, their impact remains quite outshined in larger corpora teeming with rich vocabularies. Marivate et al. (2020) focused on

applying Word2Vec augmentation techniques on fake news datasets. Their study showed that the augmentation of word embeddings proved to capture some subtle but important inter-word relationships, thus improving the accuracy of classification especially in social media contexts. Despite these findings, the computational cost of training large-scale embeddings remains a hurdle. Kobayashi (2018) suggested contextual augmentation, which replaces words in a given sentence with words from a bi-directional language model, thus keeping the sentence coherent. While this technique helped model generalization, it needed highly accurate pre-trained language models to avoid drifts from intended meanings. Al-Matham and Al-Khalifa (2021) created SynoExtractor, a synonym extraction framework designed specifically for NLP workflows. Its performance was better than traditional dictionary methods due to the application of linguistic filters. However, the use of complex synonym databases caused scalability issues. Salah et al. (2023) incorporated augmentation into an ensemble learning architecture to increase classifier robustness toward stance and fake news detection. While improvements with ensemble models were achieved, they came at the cost of high computational resources and required diverse datasets for training.

TABLE1. Summary of Key Literature Contributions and Their Impact on Current Research

Author(s)	Contribution	Impact on Current Research
Wei & Zou (2019)	Introduced EDA techniques (synonym replacement, insertion, deletion)	Inspired the use of simple augmentation strategies in this project to boost small datasets
Marivate et al. (2020)	Applied Word2Vec-based augmentation for fake news classification	Supported this project's choice of Word2Vec Skip-gram embeddings for semantic learning
Kobayashi (2018)	Proposed contextual word replacement using bi-directional language models	Highlighted the importance of maintaining semantic coherence in augmented texts
Al-Matham & Al-Khalifa (2021)	Developed SynoExtractor for synonym filtering with linguistic precision	Reinforced the approach of careful synonym selection to avoid noise in the dataset
Salah et al. (2023)	Combined data augmentation with ensemble learning techniques	Informed the integration of XGBoost as an extended model for accuracy comparison

PROPOSED APPROACH

The approach described here targets enhancement of accuracy of fake news detection with the implementation of sophisticated techniques of text data augmentation alongside embedding models. To solve the issues of sparse training data and lack of consistent meaning, three primary augmentation techniques are employed: Synonym Replacement, Back Translation, and Function Word Reduction. The strategies create diverse yet meaningful forms of the original text samples, thus ensuring dataset diversity while maintaining semantic integrity. The Word2Vec Skip-gram model, which maps words into dense vectors capturing semantic relationships, is used to transform each augmented dataset. Each augmented dataset is then transformed using the Word2Vec Skip-gram model, which converts words into dense vector representations that capture semantic relationships. These embeddings are used to train a set of classification models including Support Vector Machine (SVM), Logistic Regression, Bernoulli Naïve Bayes, and

Random Forest. Engagement is assessed across the models using accuracy, precision, recall, and F1-score. Important outcomes were achieved using Back Translation in conjunction with SVM and Naïve Bayes classifiers, and Function Word Reduction improved results in Logistic Regression. To these results, XGBoost with an ensemble classifier was incorporated, which achieved the best accuracy for the original data set ensemble classifier on the original dataset.

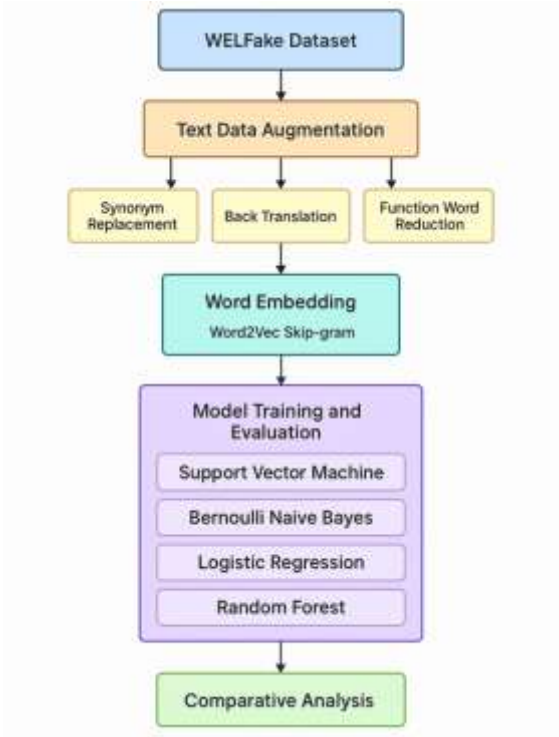


Figure 1: Proposed Fake News classification System

METHODOLOGIES

1. Dataset Selection and Preprocessing: The WELFake News Dataset from Kaggle was used, consisting of labeled news articles classified as "Fake" or "Real." Initial preprocessing involved removing noise, special characters, stop words, and applying lemmatization and stemming. This standardization ensures clean, uniform input for augmentation.

2. Text Data Augmentation Techniques: Three augmentation strategies were implemented:

- **Synonym Replacement (SR):** Words were randomly replaced with their synonyms using the WordNet thesaurus. This increased lexical variety without altering semantic meaning.
- **Back Translation (BT):** Sentences were translated to a foreign language and back to English to introduce structural variations while preserving meaning.
- **Function Word Reduction (FWD):** Non-essential words like articles and prepositions were removed, simplifying text and focusing on content-heavy words.

3. Word Embedding with Word2Vec: To convert text into numerical form, the Word2Vec Skip-gram model was applied. This method captures contextual relationships between words, enabling the models to understand semantic connections and word usage patterns in both original and augmented texts.

4. Model Training and Evaluation: The processed embeddings were used to train four classifiers:

- **Support Vector Machine (SVM)**
- **Bernoulli Naïve Bayes**
- **Logistic Regression**
- **Random Forest**

Each model was evaluated using metrics like accuracy, precision, recall, and F1-score. Additionally, XGBoost was employed as an ensemble technique to further boost performance and benchmark the baseline models.

5. Comparative Analysis: Model performance was compared across different augmentation strategies. Back Translation significantly boosted SVM and Naïve Bayes performance, while Function Word Reduction enhanced Logistic Regression accuracy.

RESULTS

The experimental results demonstrate the effectiveness of applying text data augmentation techniques in improving fake news classification accuracy. Each of the three augmentation strategies Synonym Replacement (SR), Back Translation (BT), and Function Word Reduction (FWD) was evaluated using four classifiers: SVM, Logistic Regression, Bernoulli Naïve Bayes, and Random Forest.

Among all combinations, Back Translation (BT) emerged as the most impactful technique. When paired with SVM, it achieved the highest accuracy across all models, indicating its ability to generate meaningful yet syntactically diverse training samples. Bernoulli Naïve Bayes also performed exceptionally well with BT, benefiting from the preservation of sentence semantics.

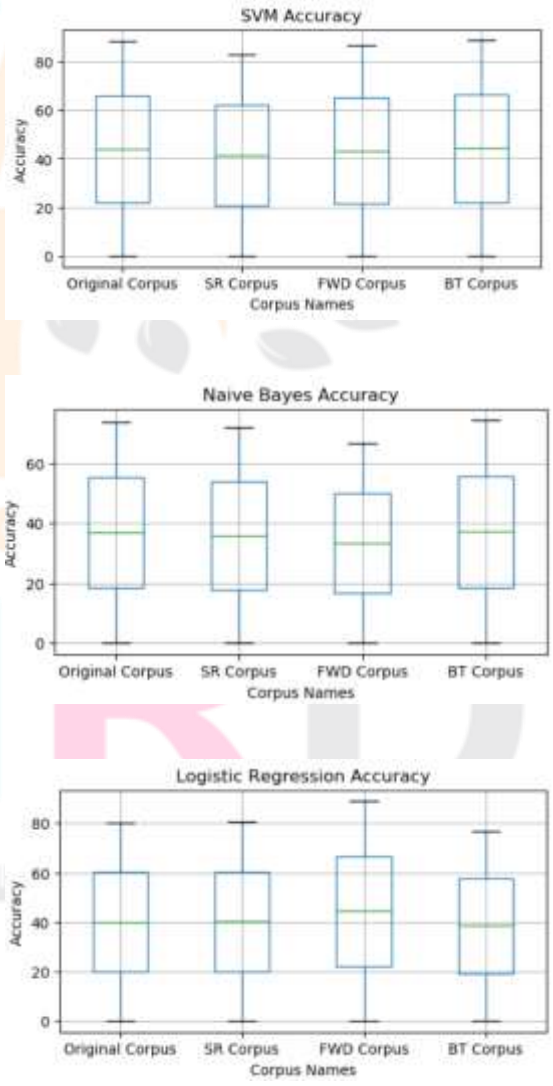
Function Word Reduction (FWD), though more aggressive in reducing text length, helped Logistic Regression classifiers improve their focus on content-rich words, leading to noticeable accuracy gains.

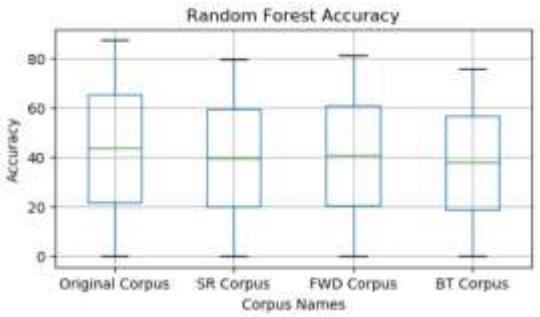
Synonym Replacement (SR) showed modest improvements across models, particularly in smaller data segments, but introduced occasional semantic noise due to inappropriate synonym substitutions.

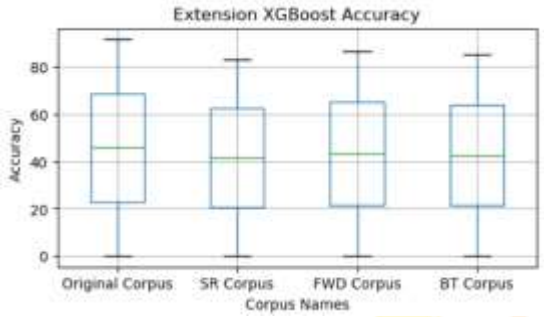
The original dataset, without augmentation, performed best with the Random Forest model, indicating its ability to handle unaltered data variability effectively.

Additionally, XGBoost, used as an ensemble benchmark, outperformed all models on the original data, reaching an impressive accuracy of 91%, validating its robustness in complex classification tasks.

Graphs visualizing accuracy on all 4 augmented techniques







DISCUSSION

The results underscore the significant role of text data augmentation in enhancing fake news classification. Among the applied techniques, Back Translation (BT) demonstrated superior performance by generating syntactically diverse sentences while preserving original semantics. This method improved the generalization capability of models like SVM and Naïve Bayes, which rely on consistent patterns in training data to perform accurate classification.

Function Word Reduction (FWD) proved effective in reducing noise by eliminating non-essential words, thus refining the focus of models like Logistic Regression on core content. However, excessive removal could risk losing contextual meaning, which requires careful calibration of this technique.

Synonym Replacement (SR) provided moderate improvement, though it introduced semantic variability due to context-insensitive substitutions, showing that uncontrolled augmentation might dilute data quality.

The use of Word2Vec Skip-gram embeddings played a pivotal role in capturing semantic relationships among words, ensuring that augmented variations retained meaning during model training. Performance evaluation metrics (accuracy, precision, recall, F1-score) confirmed that augmentation, when applied thoughtfully, boosts classification robustness.

Moreover, the high accuracy achieved by XGBoost on the original dataset highlighted the importance of combining strong base models with ensemble learning strategies.

In conclusion, combining smart augmentation techniques with efficient embeddings and classifiers can significantly improve NLP outcomes in real-world misinformation detection tasks.

CONCLUSION

This project demonstrates that incorporating text data augmentation significantly enhances the performance of fake news classification systems. By applying three strategic augmentation methods Synonym Replacement, Back Translation, and Function Word Reduction the study effectively expanded the training dataset, enabling models to better generalize and reduce overfitting. Among the techniques, Back Translation yielded the highest improvements, especially with Support Vector Machine (SVM) and Bernoulli Naïve Bayes, while Function Word Reduction improved the accuracy of Logistic Regression.

The use of Word2Vec Skip-gram embeddings facilitated deeper semantic understanding, which allowed classifiers to interpret context-rich augmented data more effectively. Furthermore, XGBoost showcased superior performance as an ensemble learner on the original dataset, affirming its strength in complex classification tasks.

Overall, the findings confirm that thoughtful augmentation strategies, combined with appropriate machine learning models and embeddings, can significantly improve fake news detection, offering a reliable framework for future NLP-based misinformation combat systems.

REFERENCES

1. Wei, J. and Zou, K., 2019. Easy data augmentation techniques for boosting performance on text classification tasks. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp.6381–6387.
2. Marivate, V. and Sefara, T., 2020. Improving short text classification through global augmentation methods. *International Cross-Domain Conference on Machine Learning and Knowledge Extraction*, pp.385–399.
3. Kobayashi, S., 2018. Contextual augmentation: Data augmentation by words with paradigmatic relations. *Proceedings of NAACL-HLT*, pp.452–457.
4. Al-Matham, R.N. and Al-Khalifa, H.S., 2021. SynoExtractor: A novel pipeline for Arabic synonym extraction using Word2Vec embeddings. *Complexity*, 2021, pp.1–13.
5. Salah, I., Jouini, K. and Korbaa, O., 2023. On the use of text augmentation for stance and fake news detection. *Journal of Information and Telecommunication*, 7(3), pp.359–375.
6. Bucos, M. and Ţucudean, G., 2023. Text data augmentation techniques for fake news detection in the Romanian language. *Applied Sciences*, 13(13), p.7389.
7. Keya, A.J., Wadud, M.A.H., Mridha, M.F., Alatiyyah, M. and Hamid, M.A., 2022. AugFake-BERT: Handling imbalance through augmentation of fake news using BERT. *Applied Sciences*, 12(17), p.8398.
8. Haralabopoulos, G., Torres, M.T., Anagnostopoulos, I. and McAuley, D., 2021. Text data augmentations: Permutation, antonyms and negation. *Expert Systems with Applications*, 177, p.114769.
9. Dahou, A. et al., 2023. Optimizing fake news detection for Arabic context using multi-task learning and transformer models. *Knowledge-Based Systems*, 280, p.111023.
10. Hua, J. et al., 2023. Multimodal fake news detection through data augmentation-based contrastive learning. *Applied Soft Computing*, 136, p.110125.
11. Feng, S.Y. et al., 2021. A survey of data augmentation approaches for NLP. *arXiv preprint arXiv:2105.03075*.
12. Pellicer, L.F.A.O., Ferreira, T.M. and Costa, A.H.R., 2023. Data augmentation techniques in natural language processing. *Applied Soft Computing*, 132, p.109803.
13. Nazir, S. et al., 2022. Toward the development of large-scale word embedding for low-resourced language. *IEEE Access*, 10, pp.54091–54097.

14. Agirre, E. et al., 2009. A study on similarity and relatedness using distributional and WordNet-based approaches. *NAACL-HLT*, pp.19–27.

15. Hill, F., Reichart, R. and Korhonen, A., 2015. SimLex-999: Evaluating semantic models with (Genuine) similarity estimation. *Computational Linguistics*, 41(4), pp.665–695.

16. Miller, G.A., 1995. WordNet: A lexical database for English. *Communications of the ACM*, 38(11), pp.39–41.

17. Salah, I., Jouini, K. and Korbaa, O., 2022. Augmentation-based ensemble learning for stance and fake news detection. *International Conference on Computational Collective Intelligence*, pp.29–41.

18. Risdal, M., 2016. Getting Real About Fake News. *Kaggle*. [online] Available at: https://www.kaggle.com/code/anthonyc1/gathering-real-news-for-oct-dec-2016/output [Accessed 28 Dec 2023].

19. Khan, J.A. et al., 2022. Valuating requirements arguments in the online user's forum: The CrowdRE-VArg framework. *Software: Practice and Experience*, 52(12), pp.2537–2573.

20. Marwat, M.I., Khan, J.A. and Alshehri, M.D., 2022. Sentiment analysis of product reviews: A SentiDeceptive approach. *KSII Transactions on Internet and Information Systems*, 16(3), pp.830–860.