Enhanced Multi-Class Detection of Fake News

Chih-Yuan Li[1], Soon Ae Chun[2], James Geller[3]

[1,3]New Jersey Institute of Technology, [2]City University of New York – College of Staten Island
{cl524, james.geller}@njit.edu, soon.chun@csi.cuny.edu

Abstract

The spread of fake news has emerged as a critical challenge in the digital era. Confusion and conflict can arise if people mistake fake news for real news. Thus, advanced detection methodologies are desired. This paper aims to identify fake news, while addressing the issue of class imbalances. We employ multi-class fake news detection, an advanced methodology beyond traditional binary classification. We highlight CNN's better performance over the baseline BERT model in the literature, with improvements in accuracy, precision, recall, and F1-Score. We uniquely experimented with four model variants: CNN and BERT with both trainable embeddings and BERT embeddings. Our experiment demonstrates CNN's effectiveness in identifying text patterns. To address class imbalances, we experimented with three different balancing methods. Our study includes fine-tuning ChatGPT for multi-class classification. The result indicates notable limitations in ChatGPT's automated classification, which highlights the complexities of AI-based categorization. Our findings demonstrate the CNN model's efficiency and effectiveness, and show the intricacies of fake news detection. These insights confirm the need for advanced AI methodologies in combating misleading information.

Introduction

The proliferation of fake news has become a serious issue in the digital age, posing significant challenges in public health, political discourse, economic activities, etc. (Shushkevich et al. 2023; Vosoughi et al. 2018). Readers could mistake fake news for real news, or have less access to authentic information; likely causing confusion of citizens and conflicts in society (Lazer et al. 2018). There are currently robust detection models in the literature (Li et al, 2022; Bojjireddy et al., 2021). However, previous work falls short in addressing the nuanced and complex nature of fake news as using binary detection as the paradigm. In this paper, we follow work on multi-class fake news detection. That work adds to the true/false (or real/fake) dichotomy the additional categories 'partially false' and 'other.'

Our work delves into the application and comparative analysis of CNN model (Yamashita et al. 2018) and BERT model (Devlin et al. 2018) for multi-class fake news detection. We chose CNN based on its robustness in pattern recognition (Bao et al. 2021) and proven success in NLP tasks (Zhu et al. 2021). CNN's ability to identify patterns is important for detecting subtle cues of fake news (Hu et al. 2020). On the other hand, BERT is famous for capturing contextual information (Vaswani et al. 2017). By comparing these models, we aim to unveil their respective strengths and limitations in fake news detection.

The issue of class imbalance often leads to skewed model training and biased outcomes (Johnson & Khoshgoftaar, 2019). We address the imbalance in dataset with three different methods: class weight adjustments (Pedregosa et al. 2012), Synthetic Minority Over-sampling Technique (SMOTE) (Chawla et al., 2002), and SMOTE combined with under-sampling (He & Garcia, 2009). The methodologies aim to create a balanced training environment, ensuring the model is learning effectively from all types of news instead of only the majority.

We also compare the CNN & BERT model with ChatGPT with and without fine-tuning to assess ChatGPT's effectiveness in categorizing news content into multiple classes. This provides critical insights into the capabilities and limitations of AI-driven classification models of multiclass imbalanced fake news.

Related Work

Fake news detection often uses machine learning approaches focused on binary classification. These studies employed traditional feature extraction methods such as TF-IDF, combined with machine learning algorithms such as SVM and Naïve Bayes (e.g., Conroy et al. 2015; Shu et al. 2017), or developed a range of machine learning detection models for performance comparison (e.g., Bojjireddy et al. 2021).

The use of CNNs and RNNs, including LSTM networks, resulted in improving the models' capabilities of pattern recognition (Wang 2017). This advanced the SOTA from

surface-level analysis to a deeper, contextual understanding. The advent of transformer-based models such BERT, further advanced text classification models. BERT's ability to understand contextual information substantially advanced the field of fake news detection (Li et al. 2022). By employing CNN-based, LSTM-based models, and BERT-based models across multiple datasets, Kozik et al. (2023) shows the robustness and versatility of NLP-driven hybrid models in fake news detection.

To address the challenge of class imbalances in training datasets, (Alghamdi et al. 2022) proposed an innovative architecture, BERT-CNN-BiLSTM. Moreover, the application of cross-validation (Refaeilzadeh et al. 2009) has become essential in evaluating model performance.

The exploration of AI tools such as ChatGPT in news categorization represents a novel and promising approach (Iqbal et al. 2023). While the tools offer automated content analysis, they face challenges in processing complex narratives; highlighting the limitations of current AI technologies in fake news recognition such as multi-class classification (Bhowmik et al. 2023).

Methods

Dataset

We used the data of (Shahi et al. 2021). They crawled multiple fact-checking sites (Shahi & Majchrzak, 2022). It contained total 1263 news articles containing titles and contents. Each news article has a label from 'False,' 'True,' 'Partially False,' or 'Other', denoting 'fake news", "non-fake news", "partially fake", or "other than the three" classes. (Shahi et al. 2021; Shahi & Nandini, 2020). **Table 1** shows the dataset distribution of each class label. The majority class (fake/false) exceeds 45%, while the non-fake (True) records are 16.7%, the Partially False (P.F.) 28% and the other class is 9%. The imbalanced multi-class dataset is more accurately reflecting the reality of different degrees of fake news (Shushkevich et al. 2023).

Table 1: Data Label, Label Definition and Data Counts

Class Label (Definition)	Total
False (The main claim is untrue.)	578
True (The main claim is true.)	211
Partially False (A mixture of true and false claim.)	357
Other (None of above as lacking claim evidence.)	117
Total data records	1263

Deep Learning Models

We selected CNN model for the multi-label classification. CNN shows its simplicity and efficiency, with limited computational resources (Taneja & Vashishtha, 2022). In addition, CNN excels in local feature extraction, and identifying indicative n-gram patterns, critical for fake news detection (Bao et al. 2021). CNN's architecture includes a trainable word embedding layer, transforming tokenized words into dense vector representations. The embeddings are optimized during training to capture semantic relationships (Chollet & Allaire, 2018).

Figure 1 shows our CNN training process, beginning with separate tokenization of title and text, followed by embedding each token into 128-dimensional vectors. The embeddings are processed through Conv1D layers to extract features from word sequences. GlobalMaxPooling1D layers condense the features, which are subsequently merged using a concatenation layer. A dropout layer is applied before and after a dense layer with ReLU activation to mitigate overfitting. A softmax-activated dense layer yields a probability distribution over predicted classes.

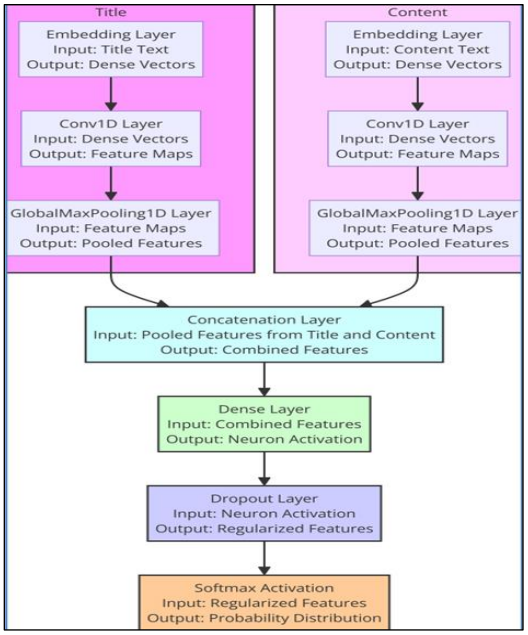

Figure 1: CNN Model

We compared our CNN model with a BERT model (Köhler et al. 2022). BERT's transformer encoder (Vaswani et al. 2017) processes the entire input sequence at a time, allowing the model to process the contextual information of each word in relation to its neighboring words. We compare the two models with different embeddings, with trainable embedding and BERT embedding.

Addressing Class Imbalance

Our dataset imbalance (**Table 1**) can undermine the model's reliability and robustness. We experimented with CNN and BERT models with three methods to address the issue.

1) *Class Weight Adjustments* (Pedregosa et al. 2012) assigns higher/lower weights to minority/majority classes during training. This ensures that models focus more on the

under-represented class(es) without modifying the data.

2) *Synthetic Minority Over-Sampling Technique (SMOTE)* (Chawla et al., 2002) creates new instances by interpolating between existing minority class instances to balance the class distribution.

3) *A combination of SMOTE with under-sampling* (He & Garcia, 2009), first applies SMOTE to oversample the minority, and then randomly removes instances from the majority class.

Fine-Tuning ChatGPT

Given the imbalanced dataset, traditional labeling methods were deemed insufficient. ChatGPT should provide a context-aware approach to class labeling. We conducted two experiments with ChatGPT labelling. First, for each input news item, we asked the ChatGPT 3.5 model to classify the news with the prompt: "Please classify the above text as 'partially false', 'False', 'other', or 'TRUE'." Second, we fine-tuned the ChatGPT 3.5 model with the training and validation sets. Among the 4-classification labels in our dataset, "Other" has the least amount of data items 117 (**Table 1**). To maintain the class balance, we combined the training and validation sets, and used 100 items for each class for training/validation with an 80/20 split. We then performed prediction on test data, and compared with the actual labels to evaluate the model's performance.

Results

We first compared our CNN model (using trainable embedding) with the BERT model of (Köhler et al. 2022). We used the given training and validation set for model training and validation, while the literature used 90% of the training set for training and 10% of the training set for validation. Our training parameters were: batch size 32, 50 epochs with early stopping, and a learning rate 1e-3. The training parameters in BERT model were batch size 8, 10 epochs and a learning rate of 3e-5. Our CNN model had better precision and recall, while it didn't improve the accuracy and F1-Score **(Table 2)**. These results were obtained prior to balancing and without cross-validation, and provide a baseline scenario.

We also compared the base models CNN and BERT with embeddings: each with trainable embedding or BERT embedding. Trainable embeddings are learned as part of the model training process and evaluated for their adaptability and fine-tuning capability specific to our dataset, while BERT embeddings are pretrained with large language dataset and leverages contextual richness in data representations. BERT models, especially with trainable embeddings, perform better in terms of recall, F-1, and overall accuracy **(Table 2)**. This experiment does not address class imbalances, nor does it apply cross validation.

Table 2: Comparative Performance Analysis of CNN and BERT Models with Different Embeddings.

Model (embedding)	Precision	Recall	F-1	ACC.
CNN (Trainable)	0.65	0.45	0.37	0.45
CNN (BERT)	0.38	0.47	0.41	0.47
BERT (Köhler et al. 2022)	0.44	0.44	0.42	0.56
BERT (Trainable)	0.55	**0.49**	**0.46**	**0.49**
BERT (BERT)	**0.67**	0.48	0.44	0.48

Addressing Class Imbalance and Overfitting

Class imbalance in our dataset is a significant concern. We experimented with (1) Class Weight Adjustments, (2) SMOTE, and (3) combining SMOTE with under-sampling.

Table 3: Performance Analysis of CNN and BERT Models with Different Embeddings, addressing class imbalances using Class Weight Adjustments.

	Precision	Recall	F-1	ACC.
CNN (Trainable Emb.)	**0.57**	**0.55**	**0.54**	**0.56**
CNN (BERT Emb.)	0.4	0.35	0.32	0.35
BERT (no Emb.)	0.36	0.34	0.34	0.34
BERT (Trainable Emb.)	0.43	0.37	0.38	0.37

After addressing class imbalances using Class Weight Adjustments, CNN with Trainable Embeddings shows an improvement across all metrics except for precision (**Table 3**). This indicates the model benefits significantly from class weight adjustments. On the other hand, both BERT-based models have degraded performance after addressing class imbalances. This could have several reasons. First, the BERT models might be more sensitive to the changes in class distribution. Second, the initial imbalance might favor BERT models in some way, and correcting this imbalances causes overfitting. The model might become too specialized in handling the minority class, thus sacrificing performance for the majority class. Overall, CNN with trainable embedding has the best performance across all metrics (bold) after weight adjustment.

In **Table 4**, we show the results of the three different imbalance techniques with 10-fold cross validation. With class weight adjustments and 10-fold cross-validation, the performance of the CNN model with trainable embeddings improved across all metrics from **Table 3**. This improvement underscores the sensitivity of the CNN model with trainable embeddings to the specific subsets of data used for training and validation, highlighting the importance of cross-validation (Yazıcı & Gures 2023).

The experimentation with SMOTE involved generating new examples for each minority class instance by creating a randomly weighted average with one of its k-nearest neighbors (k = 5) (Chawla et al. 2002). Although SMOTE shows performance enhancements, the results are slightly lower compared to class weight adjustment alone (**Table 4**). This suggests that SMOTE might introduce synthetic noise and outliers, potentially distorting the true data distribution.

Further combining SMOTE with under-sampling, which involved randomly removing instances from the majority class, resulted in even lower performance (**Table 4**). This combination appears to lead to information loss while possibly adding noise from synthetic minority class samples. Therefore, this two-step data manipulation seems less effective than the direct approach of adjusting class weights.

Table 4: Comparative Analysis of CNN+Trainable Embedding with(out) addressing class imbalance and 10-fold cross validation.

	Cross Validation	No Cross Validation
No Imbalance Handling	Precision: 0.55 Recall: 0.58 F1-Score: 0.51 Accuracy: 0.58	Precision: 0.54 Recall: 0.43 F1-Score: 0.34 Accuracy: 0.43
Class Weight Adjustments	**Precision: 0.63** **Recall: 0.60** **F1-Score: 0.58** **Accuracy: 0.60**	Precision: 0.57 Recall: 0.55 F1-Score: 0.54 Accuracy: 0.56
SMOTE	Precision: 0.59 Recall: 0.57 F1-Score: 0.56 Accuracy: 0.57	Precision: 0.56 Recall: 0.56 F1-Score: 0.55 Accuracy: 0.56
SMOTE with Under-Sampling	Precision: 0.41 Recall: 0.44 F1-Score: 0.39 Accuracy: 0.44	Precision: 0.59 Recall: 0.38 F1-Score: 0.35 Accuracy: 0.38

Evaluation of ChatGPT with Fine-Tuned Model

ChatGPT's efficacy in multi-class classification reveals both strengths and areas for improvement. **Table 5** shows the performance of the fine-tuned ChatGPT (FT) model compared with ChatGPT without any fine-tuning (Base). The fine-tuned model demonstrates a precision of 61% in classifying 'False' news, with a high recall rate of 87%. This indicates a reliable capability to identify 'False' news and capture a significant majority of instances. In contrast, 'Partially False' showcases a precision of 16%, a recall of 40%, and an F1 score of 0.23, suggesting challenges in the model's ability to accurately discern news with mixed true and false statements. For 'TRUE' news, the fine-tuned model achieves a 100% precision rate; however, a recall of merely 3% signals a shortfall in detecting true news, culminating in a disappointing F1 score of 0.06. The 'Other' category registers no precision or recall, revealing a gap in classifying those outside the true-false spectrum. The overall accuracy of the model is 49%, indicating that nearly half of the articles are correctly categorized. Comparatively, the base GPT-3.5 model without fine-tuning presents a more balanced performance across categories. It maintains higher precision for 'False' (88%) and 'TRUE' (79%) categories, with corresponding recall rates of 62% and 66%. However, it struggles with the 'Other' and 'Partially False' categories. The overall accuracy of the base model is 58%, outperforming the fine-tuned model's accuracy.

These results underscore the challenges that ChatGPT faces in multi-class fake news detection. While fine-tuning has enhanced the model's sensitivity to certain categories, it has also introduced trade-offs in precision and overall accuracy. The fine-tuned model's overall effectiveness is limited, underscoring the need for further advancements in this domain.

Table 5: Performance of ChatGPT models. ('FT' for Fine-Tuned model; 'Base' without 'FT')

	Precision (FT / Base)	**Recall** (FT / Base)	**F-1** (FT / Base)	**ACC.** (FT / Base)
False	0.61 / 0.88	0.87 / 0.62	0.72 / 0.72	0.49 / 0.58
Other	0.00 / 0.06	0.00 / 0.23	0.00 / 0.10	
P.F.	0.16 / 0.16	0.40 / 0.29	0.23 / 0.21	
True	1.00 / 0.79	0.03 / 0.66	0.06 / 0.72	

Discussion and Conclusions

Our findings confirm that news content presents a combination of fact and fiction (Lazer et al. 2018), posing challenges for automated fake news detection. Our evaluation of ChatGPT multi-class labeling revealed this challenge in accurately categorizing news items across four classes. This finding highlights the current weaknesses of AI in text analysis and draws attention to the need for more sophisticated AI solutions.

As the boundary between fake and real information becomes more complex, it is important for readers to have software tools to help them identify trustworthy new items. We demonstrated the superiority of CNN models over BERT for multi-class classification on Shahi's dataset. Furthermore, of the three methods for balancing training data we experimented with, Class Weight Adjustment worked best with CNN, achieving an accuracy of 0.60. Finally, we showed that ChatGPT is not the perfect solution for news classification.

Future Work

Future work will focus on developing hybrid AI models that integrate the strengths of CNN and BERT, aiming to leverage CNN's robustness at pattern recognition with BERT's contextual analysis capabilities.

Another future project involves addressing the class imbalances by exploring advanced techniques for data augmentation and sampling strategies. Additionally, we will seek collaborations to integrate AI tools into media literacy programs to help the public with discerning misinformation. Overall, we aim to advance the technological aspects of fake news detection, and contribute to a more informed. Our goal is to create tools and methodologies that empower individuals to critically evaluate the information they encounter.

References

Alghamdi, J., Lin, Y., & Luo, S. (2022). Modeling Fake News Detection Using BERT-CNN-BiLSTM Architecture. 2022 IEEE 5th International Conference on Multimedia Information Processing and Retrieval (MIPR), 354–357.

Bao, T., Ren, N., Luo, R., Wang, B., Shen, G., & Guo, T. (2021). A BERT-Based Hybrid Short Text Classification Model Incorporating CNN and Attention-Based BiGRU.

Bhowmik, S., Sultana, S., Sajid, A. A., Reno, S., Manjrekar, A. (2023). Robust multi-domain descriptive text classification leveraging conventional and hybrid deep learning models. Int. j. inf. tecnol. (2023).

Bojjireddy, S., Chun, S. A., & Geller, J. (2021). Machine Learning Approach to Detect Fake News, Misinformation in COVID-19 Pandemic. DG.O2021: The 22nd Annual International Conference on Digital Government Research, 575–578.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research, 16, 321-357.

Chen, X., Cong, P., & Lv, S. (2022). A Long-text Classification Method of Chinese News based on BERT and CNN.

Chollet, F., & Allaire, J. J. (2018). Deep Learning with R. Manning Publications. ISBN 9781617295546. 360 pages

Conroy, N., Rubin, V. L., & Chen, Y. (2015). Automatic deception detection: Methods for finding fake news. Proceedings of the Association for Information Science and Technology, 52.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. North American Chapter of the Association for Computational Linguistics.

He, H., & Garcia, E. A. (2009). Learning from Imbalanced Data. IEEE Transactions on Knowledge and Data Engineering, 21(9), 1263–1284.

Hu, S., Gao, Y., Niu, Z., Jiang, Y., Li, L., Xiao, X., Wang, M., Fang, E. F., Menpes-Smith, W., Xia, J., Ye, H., & Yang, G. (2020). Weakly Supervised Deep Learning for COVID-19 Infection Detection and Classification From CT Images. IEEE Access, 8, 118869–118883.

Iqbal, A., Shahzad, K., Khan, S.A. and Chaudhry, M.S. (2023), "The relationship of artificial intelligence (AI) with fake news detection (FND): a systematic literature review", Global Knowledge, Memory and Communication, Vol. ahead-of-print No. ahead-of-print.

Johnson, J.M., Khoshgoftaar, T.M. Survey on deep learning with class imbalance. J Big Data 6, 27 (2019).

Köhler, J., Shahi, G. K., Struß, J. M., Wiegand, M., Siegel, M., Mandl, T., & Schütz, M. (2022). Overview of the CLEF-2022 CheckThat! Lab: Task 3 on Fake News Detection. Conference and Labs of the Evaluation Forum.

Kozik, R., Pawlicka, A., Pawlicki, M., Choraś, M., Mazurczyk, W., & Cabaj, K. (2023). A Meta-Analysis of State-of-the-Art Automated Fake News Detection Methods. IEEE Transactions on Computational Social Systems, 1–11.

Lazer, D. M. J., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., Metzger, M. J., Nyhan, B., Pennycook, G., Rothschild, D., Schudson, M., Sloman, S. A., Sunstein, C. R., Thorson, E. A., Watts, D. J., Zittrain, J. L. (2018). "The Science of Fake News," Science, vol. 359, no. 6380, pp. 1094–1096.

Li, C., Chun, S., Geller, J. (2022). Stemming the Tide of Fake News about the COVID-19 Pandemic. Proceedings of the 35th International Florida Artificial Intelligence Research Society Conference, FLAIRS.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Édouard Duchesnay. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12(85), 2825–2830.

Refaeilzadeh, P., Tang, L., Liu, H. (2009). Cross-Validation. In: LIU, L., ÖZSU, M.T. (eds) Encyclopedia of Database Systems. Springer, Boston, MA.

Shahi, G. K., & Nandini, D. (2020). FakeCovid - A Multilingual Cross-domain Fact Check News Dataset for COVID-19. ArXiv, abs/2006.11343.

Shahi, G. K., Dirkson, A., Majchrzak, T. A. (2021). An exploratory study of COVID-19 misinformation on Twitter, Online Social Networks and Media, Volume 22, (2021), 100104, ISSN 2468-6964,

Shahi, G. K., Majchrzak, T.A. (2022). AMUSED: An Annotation Framework of Multimodal Social Media Data. In: Sanfilippo, F., Granmo, OC., Yayilgan, S.Y., Bajwa, I. S. (eds) Intelligent Technologies and Applications. INTAP 2021. Communications in Computer and Information Science, vol 1616. Springer, Cham.

Shahi, G. K., Struß, J. M., & Mandl, T. (2021). CT-FAN-21 corpus: A dataset for Fake News Detection [Data set]. Zenodo.

Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake News Detection on Social Media: A Data Mining Perspective. SIGKDD Explor. Newsl., 19(1), 22–36.

Shushkevich, E., Alexandrov, M., Cardiff, J. (2023) Improving Multiclass Classification of Fake News Using BERT-Based Models and ChatGPT-Augmented Data. Inventions 2023, 8, 112.

Taneja, K., & Vashishtha, J. (2022). Comparison of Transfer Learning and Traditional Machine Learning Approach for Text Classification.

Vaswani, A., Shazeer, N. M., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is All you Need. Neural Information Processing Systems.

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. Science, 359(6380), 1146-1151.

Wang, W. Y. (2017). "Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection. In R. Barzilay & M.-Y. Kan (Eds.), Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) (pp. 422–426). Association for Computational Linguistics.

Yamashita, R., Nishio, M., Do, R.K.G., Togashi, K. (2018) Convolutional neural networks: an overview and application in radiology. Insights Imaging 9, 611–629 (2018).

Yazıcı, İ., & Gures, E. (2023). A Novel Approach for Machine Learning-based Load Balancing in High-speed Train System using Nested Cross Validation. 2023 10th International Conference on Wireless Networks and Mobile Communications (WINCOM), 1–6.

Zhu, X., Wang, J., & Zhang, X. (2021). YNU-HPCC at SemEval-2021 Task 6: Combining ALBERT and Text-CNN for Persuasion Detection in Texts and Images.