Data Fusion for Better Fake Reviews Detection

Alimuddin Melleng **Anna-Jurek Loughrey** **Deepak P**
Queen's University Belfast, UK
alimuddinmllg@gmail.com, a.jurek@qub.ac.uk, deepaksp@acm.org

Abstract

Online reviews have become critical in informing purchasing decisions, making the detection of fake reviews a crucial challenge to tackle. Many different Machine Learning based solutions have been proposed, using various data representations such as n-grams or document embeddings. In this paper, we first explore the effectiveness of different data representations, including emotion, document embedding, n-grams, and noun phrases in embedding format, for fake reviews detection. We evaluate these representations with various state-of-the-art deep learning models, such as a BILSTM, LSTM, GRU, CNN, and MLP. Following this, we propose to incorporate different data representations and classification models using early and late data fusion techniques in order to improve the prediction performance. The experiments are conducted on four datasets: Hotel, Restaurant, Amazon, and Yelp. The results demonstrate that a combination of different data representations significantly outperforms any single data representation.

1 Introduction

The internet has become an essential tool for people in their daily lives, serving not only for work-related purposes but also personal entertainment, particularly in searching for products or services. Traditional methods of promoting businesses have become outdated, with social media and online marketing emerging as more efficient ways to engage with customers globally. As a result, organizations and businesses compete to persuade people to purchase or use their products or services, sometimes resorting to negative practices such as promoting fake reviews.

These biased, manipulated and misleading activities impact both customers and businesses, as prospective buyers rely on online user-generated reviews to make informed purchasing decisions and gain insights from others' experiences with products or services of their interest. Meanwhile, businesses depend on reviews for valuable feedback and maintaining a positive reputation. The presence of inauthentic and low-quality reviews raises concerns about their trustworthiness and poses challenges for consumers and businesses in the digital marketplace.

Malicious users frequently post fake reviews (FRs) to deceive customers by promoting or demoting products or specific retailers intentionally. FR authors may manipulate customer choices in favor of companies they are affiliated with or against competitors, making FRs a lucrative business. According to a Harvard Business School report (Luca and Zervas, 2016), the percentage of fake reviews on Yelp increased from 5% in 2006 to 20% in 2013, making detecting FRs a crucial challenge to tackle.

Unlike traditional text analytics, which focuses on domains such as labeling news stories or grouping disease reports based on severity, FR mitigation methods directly confront FR authors' intentions, resulting in a unique gamification dynamic. This requires data-driven FR solutions to rely on more general or higher-level data representations instead of simple lexical ones based on words, phrases, and sentences. FR filters using higher-level, generic features are expected to be more robust and resistant to straightforward workarounds by FR authors, such as word and phrase replacements. Moreover, higher-level features may display limited volatility across domains, making FR detection methods based on them more adaptable across different domains.

In this study, we present a comprehensive assessment of different data representations constructed using embeddings for the critical task of detecting FRs. Our analysis delves into the exploration of a range of deep learning models, as well as the application of various data fusion techniques, in order

Proceedings of Recent Advances in Natural Language Processing, pages 730–738
Varna, Sep 4–6, 2023
https://doi.org/10.26615/978-954-452-092-2_079

to develop an effective approach to combating FR problem. The central emphasis of our study is on the utilization of different data representations to enhance performance of our detection methods. To ensure the validity and reliability of our findings, we implement and analyze four distinct datasets, each specifically designed for the purpose of detecting FR in the digital landscape.

2 Related Work

FR detection was first introduced by Jindal and Liu (2008), who explain that people are influenced by reading reviews, which affects their purchasing decisions. They categorize FR into three types: untruthful reviews, brand reviews, and non-reviews. The problem of automated FR detection gained a lot of attention in recent years. Various solutions using different data representations with different machine learning learning algorithms have been explored.

Wang et al. (2018) studied n-gram combinations and test Naïve Bayes (NB) and Support Vector Machine (SVM) classifiers on the Yelp dataset. Bathla et al. (2022) suggested extracting noun phrases for fake review detection, arguing that spammers often modify aspect sentiments due to their limited product knowledge. In recent years, word and document embeddings have gained popularity as data representations for FR detection (Hajek et al., 2020; Javed et al., 2021; Taneja and Kaur, 2021). Hajek et al. (2020) proposed combining bag-of-words, emotion, and word embeddings representations for document and sentence-level representations.

Some work explored ensemble learning methods for detecting FR in recent years. Javed et al. (2021) proposed an ensemble learning framework that relied on three different models trained (CNN textual, CNN non-textual, CNN behavioral). Taneja and Kaur (2021) focused on fake feedback detection with ensemble classification, training three different classifiers using the labeled CloudArmor dataset and combining their results using the soft voting ensemble method. Gutierrez-Espinoza et al. (2020) employed three ensemble learning techniques (Boosting, Bagging, and Stacking) with four different classifiers on their ”Restaurant Dataset”.

While many studies explored different data representations, to our knowledge, no study uses various embedding data representations such as document-level, n-grams, emotion, and noun phrases embedding for FR detection and also in combination with different machine learning algorithms. We hypothesise that different data representations provide complementary information and hence combining them can improve the FR detection process. We explore different data fusion approaches including early fusion performed via data concatenation and late fusion with application of ensemble learning techniques. Ensemble learning allows to combine the predictions of different models to reduce the impact of individual model biases and errors, resulting in more robust and reliable predictions. By employing data fusion in FR detection task, we aim to leverage the strengths of individual data representation and deep learning algorithms to improve overall performance.

3 Methodology

In this section, we first discuss the different data representations explored in this study. These include the review document level, emotions, noun phrases, unigram, bigram, trigram, a combination of unigram and bigram (bigrams), and a combination of unigram, bigram, and trigrams. All of these are represented as embedding vectors. Furthermore, we discuss the fusion techniques that are integrated with five deep learning algorithms, namely Bi-LSTM, LSTM, GRU, CNN, and MLP.

To evaluate the performance of our proposed method, we utilize k-fold cross-validation with a k value of 15. We report the final average F1 score for each model.

3.1 Data representation

Several studies demonstrate that embeddings outperform other data representations, such as TF-IDF, bag of words, and n-gram, in capturing the context and semantics of words (Pennington et al., 2014; Qaiser and Ali, 2018; Wu and Yuan, 2018; Marcińczuk et al., 2021). Unlike traditional methods (TF-IDF), which represent each word as a sparse vector, embeddings capture the semantic relationships between words and represent them in a dense vector space (Abubakar et al., 2022; Pennington et al., 2014). This ability of embeddings to capture the meaning and context of a word in a sentence is crucial in several natural language processing tasks. There are many studies that have shown the effectiveness of embeddings in various NLP tasks. For instance, Mikolov et al. (2013) demonstrated that word embeddings outperform traditional methods like TF-IDF in sentiment analysis and named entity recognition tasks. Pennington et al. (2014) also reported that embeddings outperformed other methods in tasks such as sentiment analysis, text classification, and language modeling.

Motivated by the above, in our work we convert each data representation into its embedding space. We use pre-trained ROBERTA (Liu et al., 2019) embedding with an embedding dimension of 1024 to obtain the embedding of all data representations. The pre-trained model we used is roberta-large-nli-stsb-mean-tokens[1]. The embedding is converted using SentenceTransformers Library[2]. This study employed eight different data representations: Document level embedding, Noun Phrase embedding, Emotion Embedding, Unigram Embedding, Bigram Embedding, Trigram Embedding, a combination of Unigram and Bigram (uni_big) Embedding, and a combination of Unigram, Bigram, and Trigram (uni_big_tri) Embeddings.

3.1.1 Document embedding

Embeddings, in the form of word, sentence, paragraph, character, and document embeddings, are increasingly popular methods for representing data in the field of fake review detection. In this study, we employ document-level embedding as our chosen data representation. This form of representation has been utilized effectively in previous works, as demonstrated by Li et al. (2015), Ren and Ji (2017), and Hajek et al. (2020). These studies underscore the potential and versatility of document-level embeddings in addressing the challenges associated with fake review detection. For each review, each sentence is first pre-processed and then RoBERTa pre-trained model is used to generate the sentence embedding. The reviews is converted into document-level embeddings by averaging all sentences embeddings.

3.1.2 Noun Phrase Embedding

Previous studies (Ong et al., 2014; Samha et al., 2015; Xue et al., 2019; Bathla et al., 2022), have explored the use of noun phrases in FR task. Noun phrases are defined as opinion features that represent the subject or object of a sentence in a review. To extract noun phrases, we employ the Spacy library's noun chunking algorithm, which uses a rule-based approach to identify contiguous sequences of words that represent a noun phrase. All extracted noun phrases are converted into embeddings using SentenceTransformer and then averaged. Consequently, a single noun phrase embedding vector is constructed for each document.

3.1.3 Emotion Embedding

Emotion plays a vital role in fake review detection, as demonstrated by several previous studies (Melleng et al., 2019; Zeng et al., 2019; Peng and Zhong, 2014). For instance, Zeng et al. (2019) argue that FR tend to exhibit more intense emotions than genuine ones, as fake reviewers fabricate emotions not based on actual experiences (Kim et al., 2015). In our study, we also consider emotion as a feature for detecting FR.

To represent emotion in our study, we utilize DepecheMood's emotion lexicon (Staiano and Guerini, 2014). For each review, we first extract all words that match any word from the lexicon. All identified words are then converted into embeddings using SentenceTransformer. The resulting embeddings are averaged to obtain the final emotion embedding representation of the review. This approach enables us to capture the sentiment and emotional tone of the review, which can be informative in distinguishing fake from genuine reviews.

3.1.4 N-grams

We incorporates n-gram features, including unigram, bigram, and trigram, and combinations of these features, inspired by previous works (Wang et al., 2018; Javed et al., 2021). To extract n-grams, we use a process similar to the one used for noun phrase extraction, with pre-processing steps such as removing stop words, punctuation, special characters, and converting all text to lowercase. In addition, we create a combination of unigram and bigram (bigrams) and a combination of unigram, bigram, and trigram (trigrams) by concatenating the final extraction of unigrams with bigrams. After the extraction process, we convert the features into embeddings, taking their average for the final output. The use of n-gram features enables us to capture the local context of a word and the relationships between words within a given sequence. This approach is effective in many NLP tasks, including fake review detection.

[1] https://huggingface.co/sentence-transformers/roberta-large-nli-stsb-mean-tokens (visited on 14/08/2023)

[2] https://www.SBERT.net (visited on 14/08/2023)

3.2 Data Fusion

We implement two different data fusion strategies, namely early and late fusion. With the early fusion we perform data concatenation. Concatenation involves merging all eight representations discussed above into a single representation, which is then used to train a single model. As the late fusion approach we implement ensemble learning for combining models trained with different data representations and different deep learning algorithms. Several studies have shown that ensemble models can provide better overall prediction accuracy in comparison to single classification models, and avoid overfitting (Wei et al., 2019; Gutierrez-Espinoza et al., 2020; Hajek et al., 2020). With ensemble learning a collection of different classification models (i.e. base classifiers) is first trained. Following this, the prediction made by all base classifiers are combined accordingly based on the chosen ensemble strategy. Two ensemble strategies are explored in this study: majority voting and stacking (Hajek et al., 2020). The majority voting strategy outputs the label with the highest number of votes from the collection of base classifiers predictions (Yao et al., 2021). This strategy is popular due to its simplicity (Wei et al., 2019; Yao et al., 2021). SVM and Random Forest are chosen as the meta-classifiers in the stacking model.

4 Experimental Results and Discussion

We assume that different data representations may contain complementary information that are useful for FR detection. We are going to investigate whether this is the case and whether combining them may provide better performance. There are two research questions that will be addressed in this study.

1. Does any of the data representations provide optimal performance across different machine learning models (MLMs) and different datasets in FR detection task?

2. Can data fusion improve FR detection performance and which data fusion technique is the most effective in FR detection task?

4.1 Experimental Setup

We implement five deep learning models, which include Bi-LSTM, LSTM, GRU, CNN and MLP for FR detection. Bi-LSTM accesses long-range context in both input directions, widely used in NLP tasks. Our Bi-LSTM model comprises an Input layer, a Reshape layer, a Bidirectional LSTM layer, and two Dense layers. The LSTM model includes an Input layer, a Reshape layer, an LSTM layer, and two Dense layers. The CNN model consists of an Input layer, a Reshape layer, a Conv1D layer, a MaxPooling1D layer, a Flatten layer, and two Dense layers. GRU, similar to LSTM, has fewer parameters, making it faster to train. Our GRU model uses the same settings as the LSTM model. MLP consists of an Input layer, two hidden Dense layers, and an output layer, commonly used for supervised learning tasks. All models are trained using binary cross-entropy as the loss function, 'adam' as the optimizer, and f1 score as the metric. All our experiment use K-Fold cross validation with K=15.

4.2 Dataset

In the experiment, four distinct datasets are utilized: Amazon, Restaurant, Yelp, and Hotel datasets

The Amazon[3] dataset comprises 21,000 reviews, balanced between fake and genuine reviews. The Hotel dataset includes 1,600 reviews, with 800 fake and 800 genuine reviews[4]. The Restaurant dataset, developed by Gutierrez-Espinoza et al. (2020) consists of 110 reviews. The Yelp dataset, sourced from Rayana and Akoglu (2015), features reviews from restaurants in NYC. This dataset initially contains 358,922 reviews, with 322,062 genuine and 36,860 FR. To address the imbalance, some restrictions are applied to the Yelp dataset: only reviews containing more than 3 sentences and fewer than 30 sentences are considered. This results in a final Yelp dataset of 50,000 reviews.

4.3 Results

In this study, our objective is to investigate the validity of the hypothesis that various representations contain distinct information that can contribute to improved the task of FR detection. We conducted a series of experiments to ascertain whether combining these representations indeed leads to better results. We divided our work into several experiments that allow us to answer our research questions.

Experiment 1: In this experiment each data representation is evaluated with each deep learning

[3]https://www.kaggle.com/lievgarcia/amazon-reviews (visited on 14/08/2023)

[4]https://myleott.com/op-spam.html (visited on 14/08/2023)

model. The objective is to understand the performance of each data representation and model independently. The results obtained for all four datasets are presented in Figure 1 where rows refer to different data representations and columns represent different learning algorithms. The blue cell represent an average performance obtained by each data representation (with different learning algorithms) and each learnign algorithm (with different data representations).

The four images in Figure 1 show the performance of different models on different data representations for four datasets: Hotel, Restaurant, Amazon, and Yelp. The first image (a) shows that all models perform well on the Hotel dataset, with an overall model mean accuracy of 0.806. The highest-performing data representation is full_review or document embedding, followed by trigram and uni_bi_tri. LSTM achieves the highest average F1 score, while GRU, MLP, and BILSTM perform slightly worse.

In the Restaurant dataset, the overall model mean F1 score is 0.671. The highest-performing data representation is trigram, followed by uni_big and unigram. MLP achieves the highest average F1 score and LSTM is the second best, while GRU performs slightly worse. It is worth noting that some data representations, such as unigram, perform poorly on this dataset.

The third table shows that the overall model mean F1 score for the Amazon dataset is 0.616. The highest-performing data representation is unigram, followed by uni_big and emotion. MLP achieves the highest F1 score, while BILSTM performs slightly worse based on their average F1 score. Again, some data representations, such as noun phrase, perform poorly on this dataset.

Finally, the fourth table shows that all models perform well on the Yelp dataset, with an overall model mean F1 score of 0.672. The highest-performing data representation are uni_big and trigram, followed by full_review. BILSTM, and MLP achieve the highest F1 value, while CNN and GRU perform slightly worse.

Looking at the overall performance across all datasets, the highest-performing data representation is uni_big, followed by trigram and uni_bi_tri. However, we are not able to identify a single data representation which is optimal for all datasets. This presents additional motivation for implementing data fusion strategy, which addresses the problem of selecting the best representation for each dataset. MLP and LSTM consistently perform well across datasets, while BILSTM and GRU have more mixed results. The best-performing model overall is LSTM, followed by MLP.

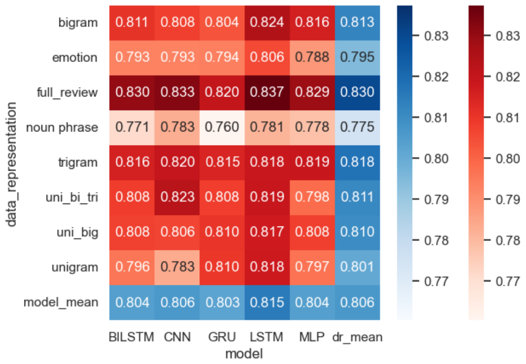
Experiment 2: In this experiment, our objective is to understand whether combining different data representation via concatenation (early fusion) or ensemble learning (late data fusion) can improve FR detection task.

The Figure 2 show the results obtained by three different ensemble techniques (Majority Voting, Stacking + Random Forest, Stacking + SVM) applied with five different learning algorithms for training base classifiers (Bi-LSTM, LSTM, GRU, MLP, CNN) and the concatenated data representations applied with the same five learning algorithms.

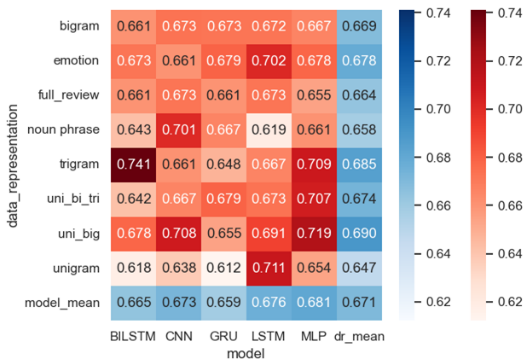
Comparing data fusion against individual data representations. In order to answer our second research questions, we compare the results from Figure 2 with the results obtained by individual data representations from Figure 1. Hotel Dataset: We can see that combining all data representations via Stacking with FR obtained better performance that any of the individual data representation across all five learning algorithms. Concatenating all data obtained better results than any of the individual data representation for 4 out of 5 learning algorithms (all apart from CNN).

In the Restaurant dataset, an enhancement in performance across all models is observed when employing ensemble learning techniques compared to utilizing individual data representations, particularly with the use of the Random Forest stacking strategy. The only exception is the Majority Vote method, which yields results below those of individual data representations. The concatenation method of combining all data representations seems to provide better results than individual data representation for two out of the five learning algorithms (GRU and CNN). However, the performance of the remaining models (BILSTM, LSTM, MLP) seems to decrease with concatenation compared to some individual data representations.

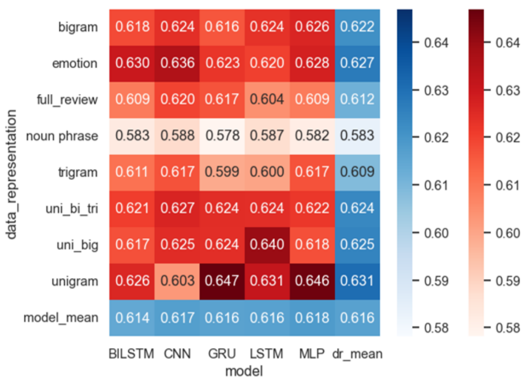
Moving to Amazon dataset, based on the given tables, it is evident that ensemble learning techniques, especially with the implementation of Random Forest stacking, provide a significant improvement in performance compared to individual data representations across all models. The stacking



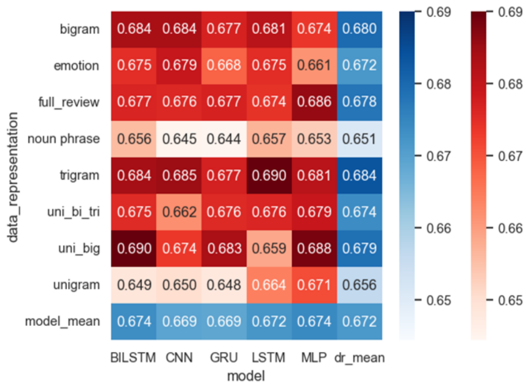
(a) Hotel dataset



(b) Restaurant dataset

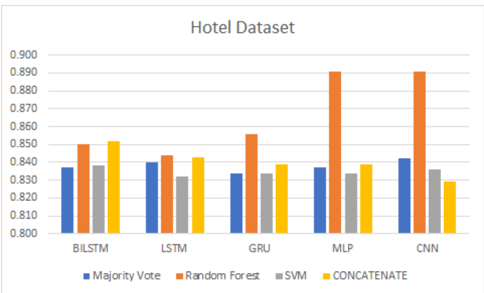


(c) Amazon dataset

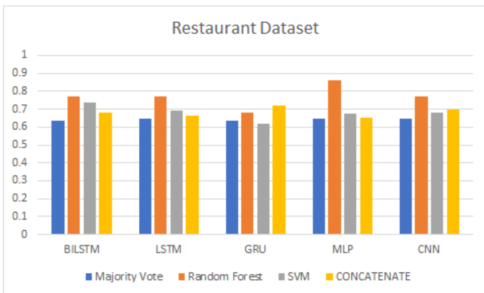


(d) Yelp dataset

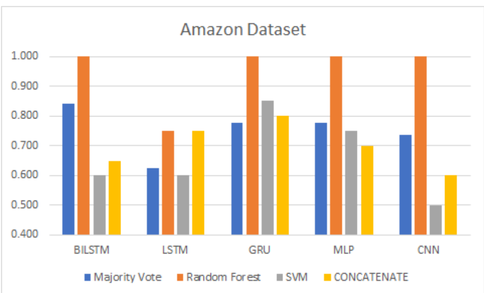
Figure 1: Data representation and deep learning model results



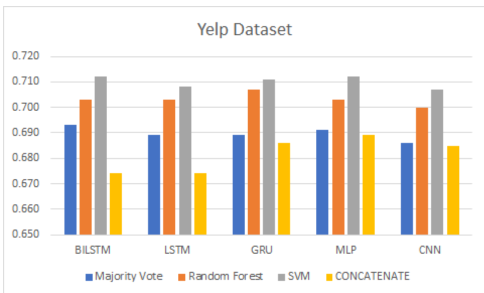
(a) Hotel dataset



(b) Restaurant dataset



(c) Amazon dataset



(d) Yelp dataset

Figure 2: Early and late data fusion result obtained for all four datasets

with Random Forest technique consistently results in higher scores than any single data representation in each model. When inspecting the performance of concatenation, it also generally outperforms individual data representations. Specifically, it yields better results than any individual data representation in 4 out of 5 learning algorithms, with CNN being the only exception.

When examining Yelp dataset, it's clear that ensemble learning techniques generally outperform single data representation models. In particular, Random Forest and SVM stacking methods consistently yield better results than any individual data representation for each of the five models used (BILSTM, CNN, GRU, LSTM, MLP).

However, an interesting trend to note is that the concatenation method, while generally providing improved performance, does not outperform all individual data representations across the five models. For instance, in the BILSTM model, 'uni_big' data representation performs better than the concatenation method.

Comparing different data fusion strategies Looking at the Hotel dataset (Figure 2a), we can observe that Stacking applied with Random Forest achieves the optimal performance for the majority of the learning algorithms. It also performs significantly better than any other fusion methods when applied with MLP and CNN obtaining F1 score of 0.891. The remaining methods have similar F1 scores, with values ranging from 0.837 to 0.842. However, the ensemble approach (Random Forest) for MLP and CNN models perform better, with F1 scores of 0.891. Comparing the performance of the ensemble learning methods with that of the concatenation method, we can see that CNN and MLP models achieve higher F1 scores with the ensemble approach, while the BILSTM models perform better with the concatenation method.

Moving on to the Restaurant dataset (Figure 2b), we see that the Majority Vote model has the lowest F1 score across all models and techniques with values ranging from 0.636 to 0.649. Similarly like with the Hotel dataset, Stacking with Random Forest applied with MLP learning algorithm achieves the highest F1 score of 0.864. BILSTM, LSTM, and CNN have the same F1 score for ensemble stacking with Random Forest with F1 value 0.733. The highest score for Concatenating method achieves 0.718 with GRU model. Comparing the ensemble learning methods with the concatenation method, we can see that ensemble learning perform better than concatenating approach.

On the Amazon dataset (Figure 2c), the Majority Vote, BILSTM, GRU, and MLP models have relatively high F1 scores, with values ranging from 0.737 to 0.842. The LSTM performs relatively poorly, with F1 scores below 0.7. Interestingly, Stacking with Random Forest achieves the highest F1 score across all learning algorithms, with a full score of 1, except for LSTM. In contrast, Stacking with SVM performs poorly, with the lowest F1 score for CNN with 0.500. For the concatenation method, the highest F1 score is achieved by the GRU model with a value of 0.800. The performance of ensemble learning methods on this dataset is better than the concatenating approach.

For the Yelp dataset (Figure 2d) we can see that the early fusion approach consistently achieves the lowest performance across all learning algorithms. At the same time, the two Stacking methods obtain the highest f1 score in all the cases with SMV applied as the meta-lerning algorithm being slightly better than with the Random Forest.

The results of the ensemble learning methods for the Hotel, Restaurant, Amazon, and Yelp datasets show variations in the F1 scores for the different models and ensemble methods. In general, the early fusion method was not as effective as the late fusion approaches for improving the F1 score. Overall, the Random Forest ensemble methods and MLP model performed well in most of the datasets.

5 Conclusion

In conclusion, this study provides a comprehensive evaluation of different embedding data representations for detecting FR. By employing various deep learning algorithms, we investigate the effectiveness of different embedding data representations, including document, n-grams, emotion, noun phrase. Additionally, we apply ensemble learning techniques to improve the detection performance further. Our experiments on four distinct datasets demonstrate that the combination of different data representations significantly enhances the performance of FR detection, outperforming single data representations.

Looking forward, future work can explore the integration of additional data representations and feature engineering techniques to improve the detection accuracy further. For instance, using attention mechanisms and transformers in neural networks could help to identify important parts of the review text and capture the contextual information more effectively. Additionally, incorporating user and product information may provide additional insights and improve the detection of FR.

Acknowledgment

The authors thank LPDP (Indonesia Endowment Fund for Education) for funding this research.

References

Haisal Dauda Abubakar, Mahmood Umar, and Muhammad Abdullahi Bakale. 2022. Sentiment classification: Review of text vectorization methods: Bag of words, tf-idf, word2vec and doc2vec. *SLU Journal of Science and Technology*, 4(1&2):27–33.

Gourav Bathla, Pardeep Singh, Rahul Kumar Singh, Erik Cambria, and Rajeev Tiwari. 2022. Intelligent fake reviews detection based on aspect extraction and analysis using deep learning. *Neural Computing and Applications*, 34(22):20213–20229.

Luis Gutierrez-Espinoza, Faranak Abri, Akbar Siami Namin, Keith S Jones, and David RW Sears. 2020. Ensemble learning for detecting fake reviews. In *2020 IEEE 44th Annual Computers, Software, and Applications Conference (COMPSAC)*, pages 1320–1325. IEEE.

Petr Hajek, Aliaksandr Barushka, and Michal Munk. 2020. Fake consumer review detection using deep neural networks integrating word embeddings and emotion mining. *Neural Computing and Applications*, 32(23):17259–17274.

Muhammad Saad Javed, Hammad Majeed, Hasan Mujtaba, and Mirza Omer Beg. 2021. Fake reviews classification using deep learning ensemble of shallow convolutions. *Journal of Computational Social Science*, pages 1–20.

Nitin Jindal and Bing Liu. 2008. Opinion spam and analysis. In *Proceedings of the 2008 international conference on web search and data mining*, pages 219–230.

Seongsoon Kim, Hyeokyoon Chang, Seongwoon Lee, Minhwan Yu, and Jaewoo Kang. 2015. Deep semantic frame-based deceptive opinion spam analysis. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, pages 1131–1140.

Luyang Li, Wenjing Ren, Bing Qin, and Ting Liu. 2015. Learning document representation for deceptive opinion spam detection. In *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data: 14th China National Conference, CCL 2015 and Third International Symposium, NLP-NABD 2015, Guangzhou, China, November 13-14, 2015, Proceedings 14*, pages 393–404. Springer.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Michael Luca and Georgios Zervas. 2016. Fake it till you make it: Reputation, competition, and yelp review fraud. *Management Science*, 62(12):3412–3427.

Michał Marcińczuk, Mateusz Gniewkowski, Tomasz Walkowiak, and Marcin Bedkowski. 2021. Text document clustering: Wordnet vs. tf-idf vs. word embeddings. In *Proceedings of the 11th Global Wordnet Conference*, pages 207–214.

Alimuddin Melleng, Anna Jurek-Loughrey, and Padmanabhan Deepak. 2019. Sentiment and emotion based representations for fake reviews detection. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 750–757.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Toan Ong, Michael Mannino, and Dawn Gregg. 2014. Linguistic characteristics of shill reviews. *Electronic Commerce Research and Applications*, 13(2):69–78.

Qingxi Peng and Ming Zhong. 2014. Detecting spam review through sentiment analysis. *J. Softw.*, 9(8):2065–2072.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Shahzad Qaiser and Ramsha Ali. 2018. Text mining: use of tf-idf to examine the relevance of words to documents. *International Journal of Computer Applications*, 181(1):25–29.

Shebuti Rayana and Leman Akoglu. 2015. Collective opinion spam detection: Bridging review networks and metadata. In *Proceedings of the 21th acm sigkdd international conference on knowledge discovery and data mining*, pages 985–994.

Yafeng Ren and Donghong Ji. 2017. Neural networks for deceptive opinion spam detection: An empirical study. *Information Sciences*, 385:213–224.

Amani Khalaf Samha, Yuefeng Li, and Jinglan Zhang. 2015. Aspect-based opinion mining from product reviews using conditional random fields. In *Data Mining and Analytics: Proceedings of the 13th Australasian Data Mining Conference [Conferences in Research and Practice in Information Technology, Volume 168]*, pages 119–128. Australian Computer Society.

Jacopo Staiano and Marco Guerini. 2014. Depechemood: a lexicon for emotion analysis from crowd-annotated news. *arXiv preprint arXiv:1405.1605*.

Harsh Taneja and Supreet Kaur. 2021. An ensemble classification model for fake feedback detection using proposed labeled cloudarmor dataset. *Computers & Electrical Engineering*, 93:107217.

Xinyue Wang, Xianguo Zhang, Chengzhi Jiang, and Haihang Liu. 2018. Identification of fake reviews using semantic and behavioral features. In *2018 4th International Conference on Information Management (ICIM)*, pages 92–97. IEEE.

Shuang Wei, Dongqi Yang, Wenyu Zhang, and Shuai Zhang. 2019. A novel noise-adapted two-layer ensemble model for credit scoring based on backflow learning. *IEEE Access*, 7:99217–99230.

Haoying Wu and Na Yuan. 2018. An improved tf-idf algorithm based on word frequency distribution information and category distribution information. In *Proceedings of the 3rd International Conference on Intelligent Information Processing*, pages 211–215.

Hao Xue, Qiaozhi Wang, Bo Luo, Hyunjin Seo, and Fengjun Li. 2019. Content-aware trust propagation toward online review spam detection. *Journal of Data and Information Quality (JDIQ)*, 11(3):1–31.

Jianrong Yao, Yuan Zheng, and Hui Jiang. 2021. An ensemble model for fake online review detection based on data resampling, feature pruning, and parameter optimization. *IEEE Access*, 9:16914–16927.

Zhi-Yuan Zeng, Jyun-Jie Lin, Mu-Sheng Chen, Meng-Hui Chen, Yan-Qi Lan, and Jun-Lin Liu. 2019. A review structure based ensemble model for deceptive review spam. *Information*, 10(7):243.