Multi-task Ensemble Learning for Fake Reviews Detection and Helpfulness Prediction: A Novel Approach

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

Abstract

Research on fake reviews detection and review helpfulness prediction is prevalent, yet most studies tend to focus solely on either fake reviews detection or review helpfulness prediction, considering them separate research tasks. In contrast to this prevailing pattern, we address both challenges concurrently by employing a multi-task learning approach. We posit that undertaking these tasks simultaneously can enhance the performance of each task through shared information among features. We utilize pre-trained RoBERTa embeddings with a document-level data representation. This is coupled with an array of deep learning and neural network models, including Bi-LSTM, LSTM, GRU, and CNN. Additionally, we employ ensemble learning techniques to integrate these models, with the objective of enhancing overall prediction accuracy and mitigating the risk of overfitting. The findings of this study offer valuable insights to the fields of NLP and machine learning and present a novel perspective on leveraging multi-task learning for the twin challenges of fake reviews detection and review helpfulness prediction.

1 Introduction

The proliferation of online marketplaces has significantly altered the way consumers purchase goods and services. As part of this transformation, user-generated reviews have become a vital factor in influencing purchasing decisions. However, the increased reliance on these reviews has given rise to an unsettling phenomenon: the spread of deceptive or "fake" reviews. Fake reviews, either overly positive or overly negative, can distort the perceived quality or popularity of products or services, misleading consumers and affecting businesses' reputations.

Simultaneously, the concept of review helpfulness has emerged as another crucial aspect of user-generated reviews. Helpfulness prediction aims to rank and highlight reviews that potential consumers would find most useful. It is based on the premise that not all reviews provide the same value to consumers, and certain reviews are more informative and helpful than others. Accurate helpfulness prediction can thus enhance the shopping experience by guiding consumers towards reviews that offer the most beneficial insights.

An example of helpful and unhelpful review: ***Helpful***: "I purchased this phone two weeks ago and have been using it ever since. The battery life is impressive, and the screen is bright and colourful. The camera produces high-resolution images, especially in night mode, which delivers fantastic results."

Unhelpful: "I bought this phone as a gift for my daughter and she's happy with it. The delivery was quick and the packaging was satisfactory."

Recently, multi-task learning, a paradigm of machine learning, has been recognized as a promising approach to improve the performance of related tasks (Ruder, 2017; Xue et al., 2017; Fan et al., 2018). Multi-task learning operates on the principle that learning multiple tasks simultaneously, leveraging shared representations, can lead to improved generalization by exploiting commonalities and differences across tasks. In the context of fake reviews detection (FRD) and helpfulness prediction (HP), these tasks are closely related as they both involve understanding the content and context of reviews to make predictions.

This study seeks to apply the principles of multi-task learning, combined with ensemble learning strategies, to the tasks of FRD and HP. The objective is to harness the shared information between these tasks to enhance the effectiveness of FRD and the accuracy of HP. The commonalities and inter-task correlations learned in one task can be shared and used to reinforce the feature learning of

Proceedings of Recent Advances in Natural Language Processing, pages 721–729
Varna, Sep 4–6, 2023
https://doi.org/10.26615/978-954-452-092-2_078

the other task, thereby boosting the overall performance of both tasks.

Ensemble learning is incorporated to further optimize the model's performance. It combines predictions from multiple models to generate a final prediction, thereby capitalizing on the strengths of each individual model while mitigating their weaknesses. The utilization of ensemble learning techniques further strengthens the robustness of our approach, enhancing the precision and reliability of our predictions.

To the best of our knowledge, this is the first study that employs a multi-task learning approach integrated with ensemble learning for simultaneous FRD and HP. This paper presents the design, implementation, and evaluation of our proposed multi-task ensemble learning model, providing a novel contribution to the field of online review analysis.

2 Related Work

Fake review (FR), also referred to as fake opinions, deceptive reviews, deceptive opinions, spam reviews, or spam opinion, present a challenge in online platforms. The primary objective of FRD is to determine whether a review is genuine or fraudulent. Over the past decade, myriad studies have endeavored to devise more effective methodologies to uncover these fraudulent reviews. These methodologies leverage a range of techniques, each aiming to optimize the detection performance.

Several studies employ machine learning methodologies such as Support Vector Machines (SVM) (Ott et al., 2011; Mukherjee et al., 2012; Yafeng et al., 2014; Melleng et al., 2019; Wang et al., 2014), Random Forest (Rout et al., 2017; Gutierrez-Espinoza et al., 2020), Naive Bayes (Li et al., 2011), Logistic Regression (Banerjee et al., 2015), and Decision Trees (Gutierrez-Espinoza et al., 2020). On the other hand, some research explores the utility of Deep Learning techniques. These include Long Short-Term Memory (LSTM) networks (Wang et al., 2018), Convolutional Neural Networks (CNN) (Zhao et al., 2018), Bidirectional Long Short-Term Memory (Bi-LSTM) networks (Liu et al., 2020), and Gated Recurrent Units (GRU) (Anass et al., 2020).

Research on online reviews encompasses not just the detection of FRs, but also the evaluation of review helpfulness (Luo and Xu, 2019; Alsmadi et al., 2020), and even the use of reviews for recommendation or ranking based on the helpfulness (Melleng et al., 2021). The examination and understanding of online reviews provide a wealth of insights that can be harnessed to enhance user experiences, refine products and services, and inform business strategies. The advent of machine learning and deep learning techniques has significantly amplified the potential for extracting meaningful information from these reviews. Such information serves as a valuable resource for customers, aiding them in making informed decisions (Bilal et al., 2019). Alsmadi et al. (2020) effectively identified helpful reviews by employing three distinct approaches: a supervised approach (Fasttext, SVM, Bi-LSTM, CNN, RCNN), a semi-supervised approach (RCNN), and a pre-trained model approach (BERT and RoBERTa), using an Amazon dataset across four domains. Their comparative analysis revealed that among all the approaches, the RCNN model demonstrated superior performance.

Although there has been extensive research on online reviews, particularly in the areas of FRD and HP, to the best of our knowledge, no existing work has undertaken the task of combining these two areas of study. Multi-task learning (MTL) have the potential to outperform those focused on single tasks learning (STL). The effectiveness of MTL can be attributed to its capacity to leverage a larger volume of data from various learning tasks, compared to STL models. With access to a more diverse dataset, MTL models are capable of learning more robust and universally applicable patterns for multiple tasks, resulting in the development of more powerful models.

In the realm of MTL for FRD, Hai et al. (2016) have made significant contributions for MTL for FRD for multiple domain datasets. They devised an MTL-Logistic Regression (MTL-LR) model and an advanced variant known as semi-supervised multi-task learning through Laplacian regular logistic regression (SMTL-LLR). This latter model was designed to improve performance with unlabeled data, and it indeed outperformed its MTL-LR counterpart as well as other conventional models such as SVM, LR, and semi-supervised positive-unlabeled (PU) learning.

Meanwhile, Fan et al. (2018) utilized MTL for review helpfulness prediction and star rating regression. They achieved this by employing a CNN model to simultaneously perform two tasks: helpfulness identification and star rating regression.

Their approach incorporated two kinds of input: character-level embeddings and word-level embeddings, extracted from two separate Amazon datasets, namely Amazon Clothes and Electronics.

In a similar work, Liu et al. (2022) proposed a multi-task Dual Attention Recommendation Model (DARMH) for both review helpfulness and rating prediction. This work utilized word embeddings and user ID embeddings from a specific Amazon dataset. The researchers demonstrated that DARMH exhibited a 3.9%-5.4% performance improvement compared to other rating prediction algorithms.

From our investigation, it is apparent that only a limited number of studies have ventured into the application of MTL for the dual challenges of FRD and review helpfulness.

Our research stands out by uniquely integrating MTL with ensemble learning, a strategy that simultaneously addresses these two tasks. We innovatively utilize document-level embeddings—a type of data representation—to exploit shared information and correlations inherent in these tasks, thereby boosting both the detection accuracy of FRs and the prediction precision for review helpfulness.

To the best of our understanding, this research is pioneering in its exploration of an ensemble-based MTL framework, specifically tailored for FRD and HP using document-level embeddings. Consequently, our study marks a significant contribution by comparing results across diverse multi-task ensemble models, thereby highlighting the unique advantages of this novel combination.

3 Methodology

In this section, we propose multi-task learning (MTL) of FRD and HP. In this research, we run MTL on five different algorithms (Bi-LSTM, LSTM, GRU, CNN, and MLP). Two objectives are focused on MTL: implementation of MTL for FRD and HP and MTL-ensemble.

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

3.1 Preprocessing Data

In order to prepare the data for effective analysis and detection, we employ various pre-processing techniques, including stop words removal, lowercasing, stemming, noise removal, normalization, and tokenization (Shan et al., 2021). This crucial step enhances the dataset's quality and reliability, facilitating the extraction of valuable insights from the data (Uysal and Gunal, 2014).

3.2 Feature Representation

In the field of FRD, researchers explore various data representations that can serve as effective features. Multi-dimensional embeddings have been shown to 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 like 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). Ren and Ji (2017) advocate for the use of document-level embedding representation as a feature in detecting FRs, as they found that it yields enhanced results when paired with deep learning techniques. The capacity of embeddings to grasp the meaning and context of words within sentences is crucial for a range of NLP tasks. Multiple studies have validated the effectiveness of embeddings in a variety of NLP tasks. For instance, Mikolov et al. (2013) demonstrated that word embeddings surpass traditional methods such as TF-IDF in sentiment analysis and named entity recognition tasks. Similarly, Pennington et al. (2014) found that embeddings exceeded the performance of other approaches in tasks like sentiment analysis, text classification, and language modeling. In our study, we employ document-level embedding as a feature. To derive the embedding vector, each sentence undergoes conversion via RoBERTa (Liu et al., 2019). We use a pre-trained model for this conversion process: roberta-large-nli-stsb-mean-tokens[1]. The conversion to embeddings is facilitated by the SentenceTransformers Library[2]. By averaging all sentence embeddings, we convert the reviews into document-level embeddings.

3.3 Multi-task Learning (MTL)

Figure 1 illustrates the framework of our proposed model, which integrates two tasks: FRD and HP. The task of FRD aims to discern if a review is

[1]https://huggingface.co/sentence-transformers/roberta-large-nli-stsb-mean-tokens

[2]https://www.SBERT.net

fake or genuine, whereas HP strives to assess the usefulness of a review. Our proposed methodology employs hard parameter sharing, where the hidden layer is shared across all tasks, while maintaining distinct output layers for each task (Vazan et al.). This approach sways the parameters within the shared hidden layer to generalize over all tasks, thereby minimizing the risk of overfitting for each individual task (Ruder, 2017). Unlike STL models, MTL strategies can take advantage of the interrelations between corresponding tasks to discern complex signals indicative of deception. By considering the inter-task relationships, the representations learned in one task can be transferred and utilized to fortify the feature learning in the other task. This results in an enhancement of the overall performance of both tasks through mutual feedback within a single framework (Ma et al., 2018).

3.4 Ensemble

In this study, we apply ensemble learning to amalgamate models trained with various deep learning algorithms for Fake Review Detection (FRD) and Helpfulness Prediction (HP), derived from MTL. Ensemble learning is a machine learning technique intended to enhance the performance of individual models by integrating multiple models, thus facilitating a collaborative learning environment where weaker models learn from the stronger ones (Vazan et al.; Zeng et al., 2019).

Several types of ensemble learning methods exist, including bagging, boosting, stacking, voting, blending, and bootstrap. In this study, we employ two ensemble learning methods: majority voting and stacking. Majority voting, also known as hard voting, is a method in which each model in the ensemble casts a vote for each class for a given test instance, and the class receiving the majority of votes is predicted as the final output. Stacking, on the other hand, combines different models and trains them using another model, known as a meta-classifier. This combination is trained and tested to produce the final prediction (Wolpert, 1992; Yao et al., 2021; Jiang et al., 2021). For stacking, we select Random Forest and SVM as the meta-classifiers.

3.5 Integration of Ensemble Learning in Single-task Learning (STL) and Multi-task Learning (MTL)

In our study, we implement ensemble learning in both STL and MTL models, as depicted in Figure 2.

For the STL model, we construct independent models using our selected classifiers: Bi-LSTM, LSTM, GRU, MLP, and CNN. Each of these models is trained and used to make predictions independently. The predictions are then consolidated using the ensemble methods described in Section 3.4, forming a collective prediction result for the STL model.

Similarly, in the MTL model, we employ the same classifiers to generate predictions for each task (FRD and HP). These task-specific predictions are then combined separately using the ensemble methods, creating an ensemble prediction for each task.

By applying ensemble learning in this way, we aim to enhance the performance of both the STL and MTL models, leveraging the strengths of individual classifiers and mitigating their weaknesses.

4 Experimental Results and Discussion

In this study, we want to investigate whether MTL for FRD and HP may provide better performance. There are three research questions that will be explored.

1. How can MTL learning be effectively applied to simultaneously detect FRs and predict review helpfulness?

2. What impact does the application of MTL have on the F1 score and efficiency of FRD and HP compared to STL methods?

3. How can ensemble learning strategies be integrated into a MTL model to improve the performance of FRD and HP?

4.1 Experimental Setup

Our MTL framework incorporates various deep learning and neural network models, specifically Bi-LSTM, LSTM, GRU, and CNN. The Bi-LSTM model is structured with an Input layer, a Reshape layer, a Bidirectional LSTM layer, and two Dense layers. The LSTM model, on the other hand, includes an Input layer, a Reshape layer, an LSTM layer, and two Dense layers. The CNN model is composed of an Input layer, a Reshape layer, a Conv1D layer, a MaxPooling1D layer, a Flatten layer, and two Dense layers. The GRU model, which is noted for its fewer parameters and consequent faster training time, aligns closely with the LSTM model in terms of its architecture. Lastly,

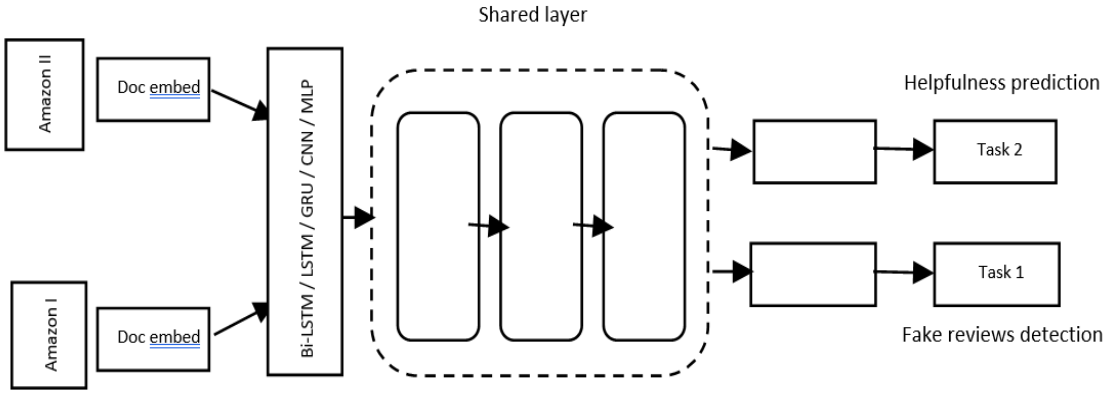

Figure 1: MTL-FR detection and helpfulness prediction

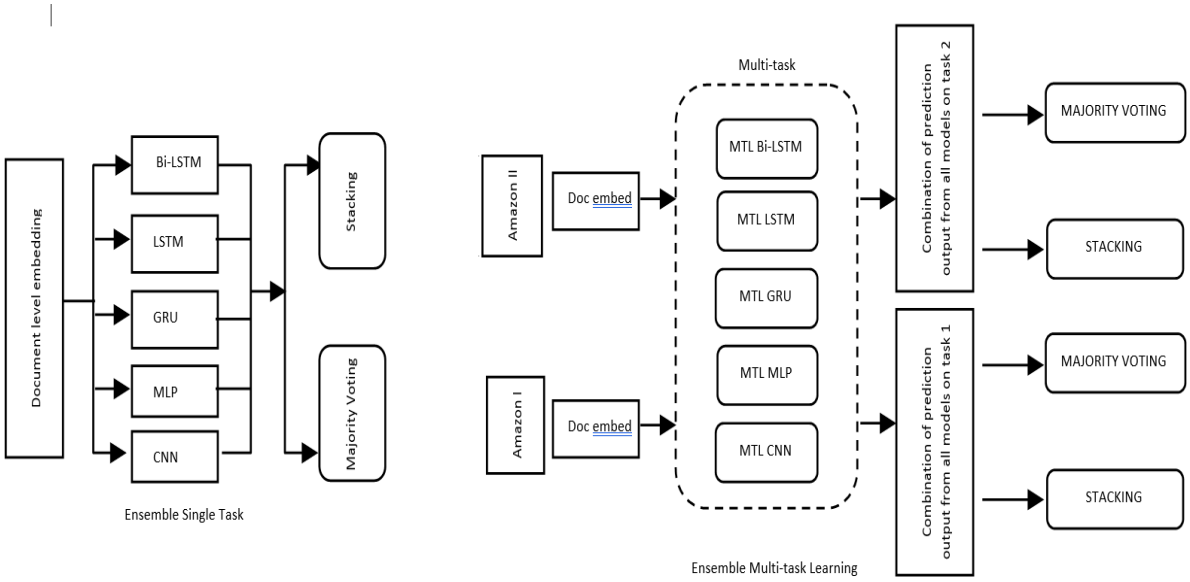

Figure 2: STL and MTL ensemble learning

the MLP model, often utilized for supervised learning tasks, consists of an Input layer, two hidden Dense layers, and an output layer. All these models are compiled using binary cross-entropy as the loss function, 'adam' as the optimizer (Kingma and Ba, 2014; Lu et al., 2019), and the F1 score as the metric for evaluation. To ensure the robustness of our results, we implement K-fold cross-validation with K set at 15 for all models. Additionally, for the Random Forest and SVM models used in the ensemble learning approach, we apply the same 15-fold cross-validation strategy.

4.2 Dataset

The datasets utilized for this experiment are derived from two different Amazon datasets. The first dataset, referred to as Amazon I[3], is used for the task of FRD. The second dataset is another public dataset, denoted as Amazon II[4]. One significant distinction between the two datasets is that the second dataset does not contain helpfulness labels. We generate labels following the methodology outlined in (Alsmadi et al., 2020; Du et al., 2019), where a review is categorized as helpful if it garners at least 70% of the votes, and unhelpful otherwise.

A key limitation encountered during the experiment is that MTL requires inputs and features of identical length. The first dataset, Data 1, comprises approximately 21,000 reviews, with a balanced distribution of fake and non-fake reviews. In

[3]https://www.kaggle.com/lievgarcia/amazon-reviews
[4]http://jmcauley.ucsd.edu/data/amazon/

contrast, the second dataset contains about 300,000 reviews post pre-processing. We set certain pre-processing conditions for the helpfulness review data. Only reviews with a minimum of 5 sentences and no more than 30 sentences are processed. Furthermore, we only consider reviews that have received at least 5 helpfulness votes. The final dataset for helpfulness prediction consists of 20,400 reviews. Since MTL need balance dataset, we balance the first dataset into 20,400 with random sample model.

4.3 Results

This study explores the implementation of MTL for two distinct tasks: FRD and HP. Document-level embedding is utilized as the primary data representation, based on the hypothesis that its use within a MTL context can enhance the model's performance. The study is structured as a series of experiments, each aimed at addressing research questions related to FRD and HP, within the framework of MTL combines with ensemble learning using document-level embeddings.

Initially, we implement both MTL and STL for FRD and HP, conducting an in-depth analysis comparing these approaches. Subsequently, we apply ensemble learning to the results of both MTL and STL models to examine the effectiveness of this method in improving model performance. This investigation provides valuable insights into the potential benefits of using an ensemble approach in combination with MTL for this particular set of tasks.

Experiment 1: The effectiveness of each model is gauged on how well it accomplishes both tasks - FRD and review HP. The performance of MTL and STL is evaluated across multiple metrics to provide a comprehensive assessment. Notably, by comparing the performance of MTL and STL, the potential advantages of performing these tasks simultaneously, as opposed to individually, are elucidated. The results of these experiments offer valuable insights into the effectiveness of MTL in these specific contexts and contribute to the broader understanding of the application of MTL in NLP tasks.

Figure 3 presents a comparison of the performance of five different models—BiLSTM, CNN, GRU, LSTM, and MLP—on two tasks using STL and MTL approaches. The tasks are FRD and HP. The performance metric used in this table is the F1-score.

For the ST approach, BiLSTM achieves the highest F1-score of 0.613 in FRD, while the CNN model outperforms the other models with an F1-score of 0.705 in HP. The lowest F1-scores for ST-FR detection and ST-Helpfulness prediction are obtained by the GRU model (0.604) and BiLSTM model (0.689), respectively.

In the MTL approach, the LSTM model shows the best performance for both FRD and HP, with F1-scores of 0.623 and 0.722, respectively. The lowest F1-scores in MTL-FR detection and MTL-Helpfulness prediction are achieved by the BiLSTM model (0.611) and the MLP model (0.681), respectively.

Comparing the performance of the models between STL and MTL approaches, it can be observed that the MTL approach generally results in improved F1-scores for HP across all models. For FRD, the MTL approach leads to better F1-scores for the CNN, GRU, LSTM, and MLP models, while the BiLSTM model's performance slightly decreases.

Overall, the MTL approach appears to be more effective in enhancing the performance of HP. For FRD, the MTL approach is beneficial for most models, except for the BiLSTM model. The LSTM and CNN models demonstrate stronger performance across both STL and MTL scenarios.

Experiment 2: In this experiment, the objective lies in exploring the potential benefits of an ensemble learning approach in enhancing the performance of both STL and MTL. The premise of the investigation hinges on the assumption that combining results from different models could enhance the predictive capacity of both STL and MTL. By integrating various models in an ensemble method, the goal is to examine if the collective intelligence could outperform the individual models, thereby providing a boost to the performance of both STL and MTL.

Figure 4 presents a comparative analysis of three ensemble methods - Majority Voting, Random Forest, and SVM - applied to STL and MTL for two tasks: FRD and HP.

For the STL-Ensemble FR detection, Majority Voting results in a score of 0.631, Random Forest gives a slightly higher score of 0.634, while SVM substantially lags behind with a score of 0.433. For the STL-Ensemble Helpfulness prediction, the scores are closer together: Majority

Dataset name	Number of review	Fake/Helpful	Non-fake/unhelpful
Amazon I	21,000 reviews	10500 fake	10500 non-fake
Amazon II	20,400 reviews	10200 helpful	10200 unhelpful

Table 1: Review Dataset used in this study

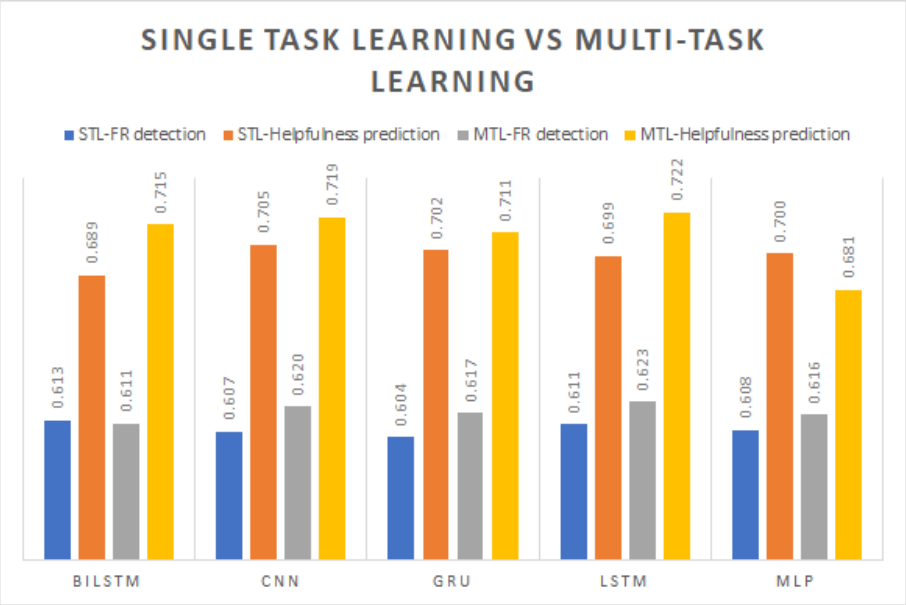

Figure 3: single-task vs multi-task learning results

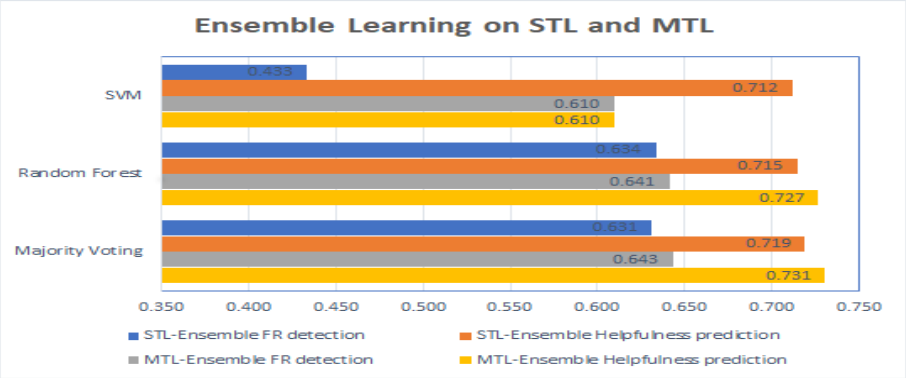

Figure 4: Ensemble Learning on single task vs multi-task learning results

Voting scores 0.719, Random Forest scores 0.715, and SVM scores 0.712.

In the context of MTL-Ensemble, the FR detection scores are generally higher. Majority Voting scores 0.643 and Random Forest scores 0.641, both slightly higher than their STL-Ensemble counterparts. SVM, despite still being the least effective method, improves its score to 0.610. For the MTL-Ensemble Helpfulness prediction, Majority Voting leads with a score of 0.731, followed by Random Forest with 0.727. SVM, however, significantly underperforms with a score of 0.610.

Looking on the results, the Majority Voting and Random Forest methods consistently outperform SVM in both STL and MTL scenarios for FR detection and Helpfulness prediction. Moreover, MTL-Ensemble generally yields superior results compared to STL-Ensemble, suggesting that MTL could be more effective for these tasks.

5 Conclusion

In conclusion, this research offers an in-depth evaluation of the application of MTL for the simultaneous detection of FRs and prediction of review

helpfulness. By focusing on document-level embedding as the sole data representation, a departure from conventional methods, this study presents a streamlined and efficient approach. The findings suggest that MTL consistently outperforms STL in these tasks, illuminating the potential benefits of this method in real-world applications.

In addition to the main task, this study also investigates the use of ensemble learning, based on prediction scores, as a means to enhance the results. The comparative performance of STL and MTL under different ensemble methods underscores the robustness of MTL in this context.

The findings of this study open a promising path for future work, which could explore further optimization of data representations or model architectures. For example, more sophisticated attention mechanisms or transformer models could be employed to better capture and utilize the semantic richness in the reviews. Additional features, such as user and product information, could also be integrated into the model to potentially provide deeper insights and further improve performance in both FRD and HP.

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.

Abdalraheem Alsmadi, Shadi AlZu'bi, Mahmoud Al-Ayyoub, and Yaser Jararweh. 2020. Predicting helpfulness of online reviews. *arXiv preprint arXiv:2008.10129*.

Fahfouh Anass, Riffi Jamal, Mohamed Adnane Mahraz, Yahyaouy Ali, and Hamid Tairi. 2020. Deceptive opinion spam based on deep learning. In *2020 Fourth International Conference On Intelligent Computing in Data Sciences (ICDS)*, pages 1–5. IEEE.

Snehasish Banerjee, Alton YK Chua, and Jung-Jae Kim. 2015. Using supervised learning to classify authentic and fake online reviews. In *Proceedings of the 9th international conference on ubiquitous information management and communication*, pages 1–7.

Muhammad Bilal, Mohsen Marjani, Ibrahim Abaker Targio Hashem, Akibu Mahmoud Abdullahi, Muhammad Tayyab, and Abdullah Gani. 2019. Predicting helpfulness of crowd-sourced reviews: A survey. In *2019 13th International Conference on Mathematics, Actuarial Science, Computer Science and Statistics (MACS)*, pages 1–8. IEEE.

Jiahua Du, Jia Rong, Sandra Michalska, Hua Wang, and Yanchun Zhang. 2019. Feature selection for helpfulness prediction of online product reviews: An empirical study. *PloS one*, 14(12):e0226902.

Miao Fan, Yue Feng, Mingming Sun, Ping Li, Haifeng Wang, and Jianmin Wang. 2018. Multi-task neural learning architecture for end-to-end identification of helpful reviews. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 343–350. IEEE.

Luis Gutierrez-Espinoza, Faranak Abri, Akbar Siami Namin, Keith S Jones, and David RW Sears. 2020. Fake reviews detection through ensemble learning. *arXiv preprint arXiv:2006.07912*.

Zhen Hai, Peilin Zhao, Peng Cheng, Peng Yang, Xiao-Li Li, and Guangxia Li. 2016. Deceptive review spam detection via exploiting task relatedness and unlabeled data. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 1817–1826.

TAO Jiang, Jian Ping Li, Amin Ul Haq, Abdus Saboor, and Amjad Ali. 2021. A novel stacking approach for accurate detection of fake news. *IEEE Access*, 9:22626–22639.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Fangtao Huang Li, Minlie Huang, Yi Yang, and Xiaoyan Zhu. 2011. Learning to identify review spam. In *Twenty-second international joint conference on artificial intelligence*.

Wentao Liu, Weipeng Jing, and Yang Li. 2020. Incorporating feature representation into bilstm for deceptive review detection. *Computing*, 102:701–715.

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*.

Zhen Liu, Baoxin Yuan, and Ying Ma. 2022. A multitask dual attention deep recommendation model using ratings and review helpfulness. *Applied Intelligence*, 52(5):5595–5607.

Peng Lu, Ting Bai, and Philippe Langlais. 2019. Sc-lstm: Learning task-specific representations in multi-task learning for sequence labeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2396–2406.

Yi Luo and Xiaowei Xu. 2019. Predicting the helpfulness of online restaurant reviews using different machine learning algorithms: A case study of yelp. *Sustainability*, 11(19):5254.

Jing Ma, Wei Gao, and Kam-Fai Wong. 2018. Detect rumor and stance jointly by neural multi-task learning. In *Companion proceedings of the the web conference 2018*, pages 585–593.

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 P Deepak. 2021. Ranking online reviews based on their helpfulness: An unsupervised approach. In *RANLP*, pages 959–967.

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*.

Arjun Mukherjee, Bing Liu, and Natalie Glance. 2012. Spotting fake reviewer groups in consumer reviews. In *Proceedings of the 21st international conference on World Wide Web*, pages 191–200.

Myle Ott, Yejin Choi, Claire Cardie, and Jeffrey T. Hancock. 2011. Finding deceptive opinion spam by any stretch of the imagination. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 309–319, Portland, Oregon, USA. Association for Computational Linguistics.

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.

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

Jitendra Kumar Rout, Anmol Dalmia, Kim-Kwang Raymond Choo, Sambit Bakshi, and Sanjay Kumar Jena. 2017. Revisiting semi-supervised learning for online deceptive review detection. *IEEE access*, 5:1319–1327.

Sebastian Ruder. 2017. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*.

Guohou Shan, Lina Zhou, and Dongsong Zhang. 2021. From conflicts and confusion to doubts: Examining review inconsistency for fake review detection. *Decision Support Systems*, 144:113513.

Alper Kursat Uysal and Serkan Gunal. 2014. The impact of preprocessing on text classification. *Information processing & management*, 50(1):104–112.

Milad Vazan, Fatemeh Sadat Masoumi, and Sepideh Saeedi Majd. A deep convolutional neural networks based multi-task ensemble model for aspect and polarity classification in persian.

Chih-Chien Wang, Min-Yuh Day, Chien-Chang Chen, and Jia-Wei Liou. 2018. Detecting spamming reviews using long short-term memory recurrent neural network framework. In *Proceedings of the 2nd International Conference on E-commerce, E-Business and E-Government*, pages 16–20.

Xiaoguang Wang, Xuan Liu, Nathalie Japkowicz, and Stan Matwin. 2014. Ensemble of multiple kernel svm classifiers. In *Advances in Artificial Intelligence: 27th Canadian Conference on Artificial Intelligence, Canadian AI 2014, Montréal, QC, Canada, May 6-9, 2014. Proceedings 27*, pages 239–250. Springer.

David H Wolpert. 1992. Stacked generalization. *Neural networks*, 5(2):241–259.

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.

Wei Xue, Wubai Zhou, Tao Li, and Qing Wang. 2017. Mtna: a neural multi-task model for aspect category classification and aspect term extraction on restaurant reviews. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 151–156.

Ren Yafeng, Yin Lan, and Ji Donghong. 2014. Deceptive reviews detection based on language structure and sentiment polarity. *Journal of Frontiers of Computer Science & Technology*, 8(3):313.

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.

Siyuan Zhao, Zhiwei Xu, Limin Liu, Mengjie Guo, and Jing Yun. 2018. Towards accurate deceptive opinions detection based on word order-preserving cnn. *Mathematical Problems in Engineering*, 2018.