Hybrid Deep Learning Based Model on Sentiment Analysis of Peer Reviews on Scientific Papers

Ritika Sarkar[1], Prakriti Singh[1], Mustafa Musa Jaber[2], Shreya Nandan[1], Shruti Mishra[1(✉)], Sandeep Kumar Satapathy[1], and Chinmaya Ranjan Pattnaik[3]

[1] School of Computer Science and Engineering, Vellore Institute of Technology, Chennai, Vandalur-Kelambakkam Road, Chennai, Tamil Nadu, India
{ritika.sarkar2019,prakriti.singh2019, shreya.nandan2019}@vitstudent.ac.in, shrutim2129@gmail.com
[2] Department of Medical Instruments Engineering Techniques, Al-Farahidi University, Baghdad 10021, Iraq
mustafa.musa@alfarahidiuc.edu.iq
[3] Department of Computer Science and Engineering, Ajay Binay Institute of Technology (ABIT), Cuttack, Odisha, India

Abstract. The peer review process involved in evaluating academic papers submitted to journals and conferences is very perplexing as at times the scores given by the reviewer may be poor in contrast with the textual comments which are in a positive light. In such a case, it becomes difficult for the judging chair to come to a concrete decision regarding the accept or reject decision of the papers. In our paper, we aim to extract the sentiment from the reviewers' opinions and use it along with the numerical scores to correlate that in order to predict the orientation of the review, i.e., the degree of acceptance. Our proposed methods include Machine learning models like Naive Bayes, Deep learning models involving LSTM and a Hybrid model with BiLSTM, LSTM, CNN, and finally Graph based model GCN. The dataset is taken from the UCI repository consisting of peer reviews in Spanish along with other parameters used for judging a paper. Bernoulli's Naive Bayes was the model that fared the highest amongst all the approaches, with an accuracy of 75.61% after varying the parameters to enhance the accuracy.

Keywords: Peer reviews · Sentiment analysis · Natural Language Processing · AI algorithms

1 Introduction

Sentiment analysis is the way in which we detect whether a statement indicates a positive, negative or neutral emotion. It is the method of imparting some emotional intelligence to the machines with the help of Artificial Intelligence algorithms and Natural Language Processing. A positive emotion indicates that the person who spoke or wrote the text can

S. Nandan Mohanty et al. (Eds.): ICISML 2022, LNICST 471, pp. 109–116, 2023.
https://doi.org/10.1007/978-3-031-35081-8_9

be happy, a negative emotion shows he/she may be angry or sarcastic, and a neutral one indicates that he/she is indifferent. The application of sentiment analysis in businesses and industries is done as an indicator of how well the products and commodities are received in the market by using customer feedback. Sometimes, the number of criteria for judging the sentiment of the text is increased and mapped to a scale of five in order to gain more insights. The knowledge of emotions from written text is especially helpful while drafting notices or emails in the corporate and academic world.

Over the years, sentiment analysis has been applied to use cases like social media monitoring like Twitter sentiment analysis, obtaining the voice of the customer in major customer-centric businesses, monitoring brands, and market research. In this paper, we are going to deviate from these market and customer-centric analyses and move towards the analysis of paper reviews. Sentiment analysis of paper reviews is a complex and important domain which has not been explored greatly. The comments given by peer reviews are seldom considered in the final decision for publishing the article or paper in a journal or conference. Hence, we attempt to derive the sentiment from these reviews in order for them to be considered as well as the final accept or reject decision of a paper.

We aim to employ machine learning methods like the different types of Naive Bayes, namely Gaussian [11], Multinomial [12], Bernoulli's [13] and Complement [14], deep learning algorithm LSTM [15], Graph-based algorithm GCN [16] and draw comparisons between them, finally identifying the most suited algorithm for the paper reviews domain in sentiment analysis.

2 Literature Review

Chakraborty et al. [1] provide a comprehensive evaluation of implicit aspect sentiments in the peer-reviewed content of works submitted to/published at one of the leading machine learning conferences – ICLR. The paper holds the upper hand over other publications in the same category by creating a non-pre-existing database through the annotation of around 25000 reviews of data. The downside of the generated model was that it could only attain a maximum accuracy of 65%. The accuracy only suffered more as a result of removing the features. In Keith et al. [2] the classification of 382 reviews of research articles presented at an international conference was accomplished utilizing supervised and unsupervised methodologies along with a hybrid technique. The hybrid approach, the HS-SVM, is more robust than most, relative to the number of classes, which is among the paper's merits. The paper's shortcoming is that the dataset examined was quite limited, rendering the approach's viability uncertain. Furthermore, when additional classes were introduced, the performance began to deteriorate. Kang et al. [3] outline the data gathering strategy and reflect on observed occurrences in peer reviews, as well as NLP tasks centered around the dataset. The paper's key benefit is that its most optimum classifier consistently beats the majority model, exhibiting up to a 22% reduction in error. Their models are inadequately nuanced to judge the quality of the work reported in a given publication; this could imply that some of the features they specify are associated with favorable papers, or that reviewers' opinions are persuaded by them. Anta et al. [4] utilizes a corpus of Spanish tweets and presents a near examination of various methodologies and grouping procedures for these problems. The information is preprocessed utilizing strategies and apparatuses proposed in the literature, together with others

explicitly proposed here that consider the qualities of Twitter. Then, popular classifiers have been used. (In particular, all classifiers of WEKA have been assessed. Aue et al. [5] surveys four different approaches to customizing a sentiment analysis system to a new target domain in the absence of large amounts of labeled data. The paper bases the experiments on data from four different domains. After establishing that Naive Bayes classification results in poor classification accuracy, they compare results obtained by using each of the four approaches and discussing their advantages, disadvantages, and performance. Baccianella et al. [6] present SENTIWORDNET 3.0 which is a lexical resource used in sentiment classification and opinion mining applications. It is a result of when wordnet synsets are automatically annotated to positive, negative, or neutral sentiments. The sentiwordnet 3.0 gives 20% more accuracy than the 1.0 version. Through comparative analysis of the above literature, the proposed work introduces exhaustive implementations using three learning paradigms of AI, which is the first of its kind in the analysis of paper reviews. The method achieves good accuracy despite the small dataset, in contrast to the above works [1–3, 5].

3 Proposed Method

The implementation of the proposed method is outlined in Fig. 1 on a dataset of paper reviews, taken from the UCI repository which consists of anonymous reviews on papers submitted to an international conference on computer science. The dataset contains most of the reviews in Spanish and a few in English in a JavaScript Object Notation format. The reviews are then translated from Spanish to English using Microsoft Azure's Translator Text API by creating a resource in Cognitive Services. The translated JSON dataset is converted into CSV and the non-null textual values of the reviews are utilized for the sentiment analysis task. After pre-processing the reviews, they are fed to the four Naive Bayes algorithms and the LSTM models. For the graph model, first a graph is created out of the cleaned corpus which is then fed to the Graph Convolution network. The accuracies on the test set are compared for this task.

4 Methodology

4.1 Dataset Description

The original dataset from UCI has fields like timespan, paper id, preliminary decision, and the following fields for each review for a paper: review id, text and remarks in Spanish or English, language, orientation, evaluation, confidence in JSON format. Out of these fields, we use the text as our feature input to the classifier, and the preliminary decision as our class label.

4.2 Pre-processing

Data preprocessing is the procedure for prepping raw data to be used in a machine learning model. In this paper, all basics of data preprocessing [21] have been covered in order to make the data suitable for all the proposed models. Exploratory Data Analysis is done

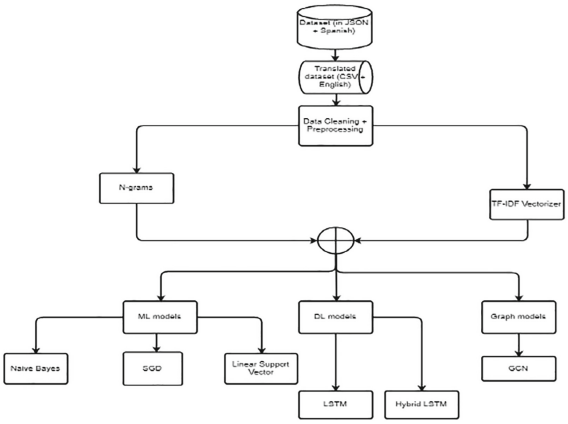

Fig. 1. A flowchart of the working methodology.

for visualization in order to get a better understanding of data such as removing null and duplicate values if any and performing feature reduction. Papers having review 0 are only permitted to be fed into the model so that data is more balanced. After noticing the drastic class imbalance in our dataset (173 total samples), we came to a consensus to provide sentiment value '0' to all those papers having the preliminary decision as "accept", 115 in count, and '1' to those having the preliminary decision "reject", "probably reject" and "no decision", 56 in total, which balanced it a little. Furthermore, we perform data cleaning [21] like stopwords removal, tokenization, lemmatization, etc., in order to increase the model's accuracy. Consequently, we reach at 109 for positive sentiment '0' and 55 for the negative sentiment '1'.

4.3 Implementation

Implementation of Machine Learning Models. Sentiment analysis is a machine learning technology [18, 19] that looks for polarities in texts, ranging from positive to negative. For sentiment analysis categorization, Naive Bayes is a relatively simple collection of probabilistic techniques that assigns a probability that a specific word or phrase should be regarded as positive or negative [11]. In mathematical terms, in order to find the probability of y given input features X we use Eq. 1.

$$p(y/X) = \left[p(X/y) * p(y)\right]/p(X) \tag{1}$$

In this paper, we have tried to implement various variations of the Naive Bayes algorithm, namely Multinomial, Gaussian, Bernoulli, and Complement, in order to check which one gives us the highest accuracy. Multinomial Naive Bayes [12], after plane counter vectorization, only gives us 56.10% accuracy which is not desirable by far. Hence, we further try to improve the accuracy by tweaking the n-grams range but the maximum we achieve is 73.17%. On implementation of the other variations, Bernoulli Naive Bayes [13] technique gave us the highest accuracy of 75.61%. Even alternative approaches like TF-IDF vectorization and other algorithms like SGD and Linear SVM

[17] (on both TF-IDF and CV fitted data) weren't able to reach an accuracy higher than 75%. Also, in addition to the accuracy scores, precision, recall and f1 scores were also calculate to see whether the two kinds of vectorizations would provide any difference in these parameters but they turned up to be the exact same.

Implementation of Deep Learning Model. Deep learning models are used in extracting abstract features, increasing performance measures and performing analytical tasks with the help of neural networks. We have used the LSTM (Long Short-Term Memory), which is a special kind of RNN (recurrent neural network) capable of perceiving long-term dependencies, for sentiment analysis consisting of 3 layers: Embeddings, LSTM and Dense with Softmax. After training the model on the dataset for a total of 15 epochs, the model is used to predict on the test set to measure the accuracy scores. The results show that the f1-score for the negative sentiments is 0.85 whereas, for the positive ones, it is 0. We perform re-sampling with substitution next to make a layer without any weights; it duplicates the data so that it may be utilized in the generation of the model followed by the convolutional layer. Test-set needs to be separated before up-sampling as it creates multiple copies of the same data. Up-sampling reduces the class imbalance and improves the f1-score i.e., 0.07 for negative sentiments and 0.49 for positive sentiments. On running a few more epochs, the f1-score improves greatly, as shown in Table 1.

Implementation of Graph-Based Model. Text_gcn [10], where a large heterogeneous text graph is constructed considering the number of nodes as the corpus size in addition to the number of unique words in the corpus, is used on our dataset. The words are one hot encoded and are inputted to the GCN. An edge connects two nodes based on the occurrence of the word in the whole corpus, and it is added between two nodes if their pointwise mutual information (PMI) is positive, as that indicates a high correlation between the words. The weight of the edges is taken as TF-IDF and PMI of the words. For a graph $G = (V, E)$, where V is the set of nodes and E is the set of edges, let A be an adjacency matrix, with the degree matrix $D = \sum_j (A_{ij})$. The feature matrix H for the $(l + 1)$th GCN layer is given by Eq. 2.

$$H^{(l+1)} = \sigma(D^{-1/2} A D^{-1/2} H^{(l)} W^{(l)}) \tag{2}$$

where W is the weight matrix and σ is the ReLu activation function. The GCN model consists of two GCN layers [16] and a final Dense layer with softmax activation. The model allows the passage of messages to a maximum of two nodes away and helps in capturing the semantic relationships between the words in the reviews.

Implementation of Hybrid Model. The proposed ensemble model is a hybrid of bidirectional LSTM and CNN for generating the final feature representation. CNN, for its ability to extract features from the text and LSTM/BiLSTM, for maintaining the sequential control between words and having the ability to overlook unnecessary words utilizing the forget gate, are combined as it uses the strengths of the two to give the accuracy. The CNN extracted features are then fed to the LSTM as input. The model consists of the Embedding layer, Spatial Dropout, a BiLSTM, a LSTM, 2 blocks of 1D CNN and Average Pooling, Dropout, Flatten and Dense layers. In the convolution layer, filters act as n-gram finders; each filter looks for a particular class of n-grams and appoints them high scores. The identified grams with most elevated score pass the max pooling

function. The convolution layer applies the Rectified Linear Unit to replace the negative output with a 0 in order to remove the non-linearity of the model. Optimization is used to change the attributes such as weights, and learning rate in order to reduce the losses. The model uses Adaptive Moment Estimation optimizer that uses learning rate to optimize the network that converges quickly. The BiLSTM layer maintains the sequential order between the information. It permits connecting the links between the past inputs and outputs. The input of this layer is the connection of the max pooling outputs.

5 Results Analysis

The Machine Learning [18] models gave us various accuracies through the implementation of several Naive Bayes variants, the highest being 75.61% achieved by Bernoulli Naive Bayes approach. The LSTM model, after up-sampling the class imbalance gave an F1 score of 0.69 for negative sentiments and 0.37 for positive sentiments. The GCN model trained for 200 epochs gives an accuracy of 73.171% on the test set which consisted of 25% of the dataset considered for the study. Figure 2 visualizes the training accuracy and loss for the Hybrid model. After training for 100 epochs, it is seen that the training accuracy oscillates between 0.97 and 1.0 and gives a f1-score of 0.73 for negative sentiments and 0.44 for positive sentiments.

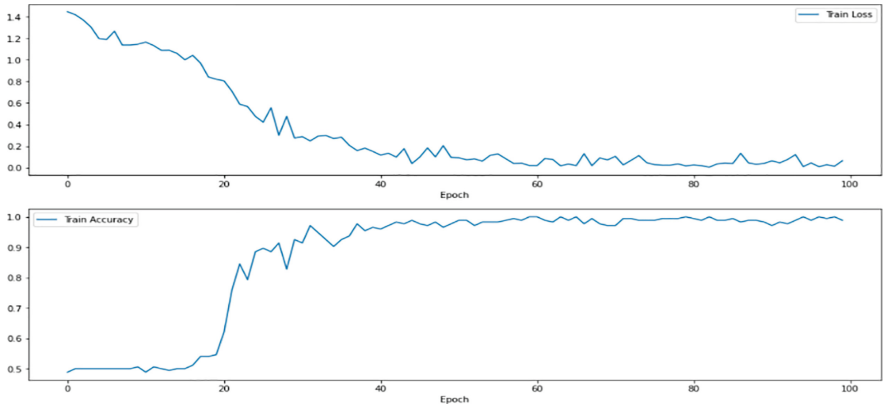

Fig. 2. Plot of the training of the Hybrid model

Table 1 summarizes the accuracy scores of the respective models created using the mentioned algorithms. It is observed that Bernoulli's Naive Bayes gives the highest accuracy out of all the models. Since the dataset is small with only 163 non-null observations, the ML models are observed to perform better. The Hybrid model is proposed to give better results for a larger test set, as it is observed that on increasing the size of the test set the accuracy increases. The GCN model is also very promising on this kind of problem. Hence, we put forward the Bernoulli's Naive Bayes as the best model in our study on this peer reviews dataset.

Table 1. Result comparison of all the Models.

Sl. No.	Model name	Accuracy (in %)
1a	Multinomial NB	73.17
1b	Bernoulli's NB	75.61
1c	Complement NB	73.17
1d	Linear Support Vector CV	73.17
1e	Stochastic Gradient Descent CV	73.17
2	Hybrid Model	63.00
3	GCN	73.17

6 Conclusion and Future Work

Sentiment analysis of peer reviews on conference papers is very useful for automating the task of the judging chair in considering both the numerical scores and the textual comments given by the reviewers and capturing any conflicts between the scores and comments. To make a competent system, we performed graded multilingual sentiment analysis, which is a complex task. Hence Microsoft's Translator Text which has free academic access has been used for the translation of the reviews. A number of pre-processing techniques have been experimented and the one giving the best accuracy in the classifiers has been recorded. It is observed that the Machine learning Bernoulli's Naive Bayes classifier using N-grams vectorizer has given the best accuracy. In a future study, we aim to compare the predicted sentiments with the numerical scores given by the reviewers and correlate them to the final decision taken by the judging chair.

References

1. Chakraborty, S., Goyal, P., Mukherjee, A.: Aspect-based sentiment analysis of scientific reviews. In: Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020, pp. 207–216. Association for Computing Machinery, New York, NY, USA (2020)
2. Keith, B., Meneses, C.: A hybrid approach for sentiment analysis applied to paper reviews (2017)
3. Kang, D., et al.: A dataset of peer reviews (PeerRead): collection, insights and NLP applications. In: NAACL 2018 (2018)
4. Fernández Anta, A., Morere, P., Chiroque, L.F., Santos, A.: Techniques for sentiment analysis and topic detection of Spanish tweets: preliminary report. In: Spanish Society for Natural Language Processing Conference (SEPLN 2012), September 2012
5. Aue, A., Gamon, M.: Customizing sentiment classifiers to new domains: a case study. In: Proceedings of Recent Advances in Natural Language Processing (RANLP), vol. 1, no. 3.1, p. 2-1, September 2005
6. Baccianella, S., Esuli, A., Sebastiani, F.: SentiWordNet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In: Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 2010) (2010)

7. Shi, H., Zhan, W., Li, X.: A supervised fine-grained sentiment analysis system for online reviews. Intell. Autom. Soft Comput. **21**(4), 589–605 (2015)
8. Tang, D., Qin, B., Liu, T.: Document modeling with gated recurrent neural network for sentiment classification. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pp. 1422–1432, September 2015
9. Zhang, L., Ghosh, R., Dekhil, M., Hsu, M., Liu, B.: Combining lexicon-based and learning-based methods for Twitter sentiment analysis. HP Laboratories, Technical report HPL-2011 89, pp. 1–8 (2011)
10. Yao, L., Mao, C., Luo, Y.: Graph convolutional networks for text classification. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, no. 01, pp. 7370–7377, July 2019
11. Kök, H., İzgi, M.S., Acılar, A.M.: Evaluation of the artificial neural network and Naive Bayes Models trained with vertebra ratios for growth and development determination. Turk. J. Orthod. **34**(1), 2 (2021)
12. Kibriya, A.M., Frank, E., Pfahringer, B., Holmes, G.: Multinomial Naive Bayes for text categorization revisited. In: Webb, G.I., Yu, X. (eds.) AI 2004. LNCS (LNAI), vol. 3339, pp. 488–499. Springer, Heidelberg (2004). https://doi.org/10.1007/978-3-540-30549-1_43
13. Singh, M., Bhatt, M.W., Bedi, H.S., Mishra, U.: Performance of Bernoulli's Naive bayes classifier in the detection of fake news. Mater. Today Proc. (2020)
14. Seref, B., Bostanci, E.: Sentiment analysis using Naive Bayes and complement Naive Bayes classifier algorithms on Hadoop framework. In: 2018 2nd International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT), pp. 1–7. IEEE, October 2018
15. Zhou, C., Sun, C., Liu, Z., Lau, F.: A C-LSTM neural network for text classification. arXiv preprint arXiv:1511.08630 (2015)
16. Zhang, S., Yin, H., Chen, T., Hung, Q.V.N., Huang, Z., Cui, L.: GCN-based user representation learning for unifying robust recommendation and fraudster detection. In: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 689–698, July 2020
17. Satapathy, S.K., Jagadev, A.K., Dehuri, S.: An empirical analysis of training algorithms of neural networks: a case study of EEG signal classification using Java framework. In: Jain, L.C., Patnaik, S., Ichalkaranje, N. (eds.) Intelligent Computing, Communication and Devices. AISC, vol. 309, pp. 151–160. Springer, New Delhi (2015). https://doi.org/10.1007/978-81-322-2009-1_18
18. Mishra, S., Mishra, D., Satapathy, S.K.: Fuzzy frequent pattern mining from gene expression data using dynamic multi-swarm particle swarm optimization. In: 2nd International Conference on Computer, Communication, Control and Information Technology (C3IT 2012), Published in Journal Procedia Technology, vol. 4, pp. 797–801, February 2012
19. Chandra, S., Gourisaria, M.K., Harshvardhan, G.M., Rautaray, S.S., Pandey, M., Mohanty, S.N.: Semantic analysis of sentiments through web-mined Twitter corpus. In: Proceedings of the International Semantic Intelligence Conference 2021 (ISIC 2021), New Delhi, India, 25–27 February 2021. CEUR Workshop Proceedings, vol. 2786, pp. 202, 122–135 (2021). CEUR-WS.org