An N-gram-Based BERT model for Sentiment Classification Using Movie Reviews

Tina Esther Trueman
Department of Computer Science
University of the People, United States
tina.trueman@uopeople.edu

Ashok Kumar Jayaraman
Information Science and Technology
Anna University, Chennai, India
jashokkumar83@auist.net

Gayathri Ananthakrishnan
Department of Information Technology
VIT University, Vellore, India
gayathri.a@vit.ac.in

Erik Cambria
Computer Science and Engineering
Nanyang Technological University, Singapore
cambria@ntu.edu.sg

Satanik Mitra
Department of ISE
Indian Institute of Technology, Kharagpur
satanikmitra@iitkgp.ac.in

***Abstract*—An abundance of product reviews and opinions is being produced every day across the internet and other media. Sentiment analysis analyzes those data and classifies them as positive or negative. In this paper, a classification model is proposed for n-gram sentiment analysis using BERT. Specifically, the large IMDB movie review dataset is used that contains 50K instances. This dataset is tokenized and encoded into unigrams, bigrams, and trigrams and their combinations such as unigram and bigram, bigram and trigram, and unigram, bigram, and trigram. The proposed BERT model employs on these extracted features. Then, this model is evaluated using the F1 score and its micro, macro, and weighted-average scores. The model shows comparable results to state-of-the-art methods for all n-gram features. In particular, the model achieves 94.64% highest accuracy for the combination of bigram and trigram features, and 94.68% unigram, bigram, and trigram features than other n-gram features.**

***Index Terms*—Sentiment classification, Deep learning, Transformers, BERT, N-gram features.**

I. Introduction

Thanks to the explosion of social media, companies often have to deal with mountains of customer feedback. Therefore, sentiment analysis is useful for quickly gaining insights from the large volumes of text data [1]. It also helps organizations to measure the ROI of their marketing campaigns and improve their customer service [2]. Since sentiment analysis offers insights to organizations for understanding their customer's emotions, they can be conscious of any crisis to come well in time and manage it appropriately [3]. Many statistical models can be used to achieve this task [4], [5]. With the advancement in deep learning, neural network architectures have shown a decent improvement in performance in solving several natural language processing (NLP) tasks like language modeling, text classification, machine translation, etc. [6]. In 2018, Google introduced the transformer model [7], which is used as transfer learning in various NLP tasks with state-of-the-art performance ever since. Transfer learning is a mechanism in which a deep learning model is trained on a large dataset. Then, it is used to perform similar tasks on another dataset. Such a model is known as a pre-trained model.

In this paper, the Bidirectional Encoder Representations from Transformers (BERT) model [8] is used. It has big neural network architecture with a huge number of parameters. Practically, training a BERT model on a small dataset from scratch would leads in overfitting. Hence, a pre-trained BERT model is used that has already been trained on a huge dataset. This model is then fine-tuned on a relatively smaller dataset for the sentiment classification task. After the recognition and popularity of the BERT model, researchers used this model on various NLP tasks such as document classification, recommendation systems, and question and answering. However, most of them have targeted binary sentiment classification. The pre-trained BERT model can be fine-tuned with just one additional output layer to create progressive, state-of-the-art models for a broad range of NLP tasks. In particular, the BERT pre-trained models become fast, easy, and powerful to use for various downstream tasks, it is likely to give promising results in different Sentiment Datasets that are chosen as well. Most of the existing works have focused on unigrams. In this work, the BERT transformer is focused with N-gram feature representation. The contribution of this paper is listed out as follows.

- Addresses sentiment classification task for movie reviews with context-independent features.
- Employs BERT-Based transformer model with N-gram features such as unigrams, bigrams, trigrams, and their combinations of features.
- The proposed N-gram-based BERT model achieves the best result than existing models in terms of precision, recall, and F1 scores.

The rest of this paper is structured as follows: Section II discusses related works in sentiment analysis; Section III introduces the BERT-based model for IMDB movie reviews; Section IV discusses experiment results; finally, Section V offers concluding remarks.

II. Related work

Sentiment analysis is used in various applications such as tourism [9], finance [10], healthcare [11], social network analysis [12], and social media monitoring [13]. Wang et al. [14] studied sentiment and topic classification using bigram features. Their study indicated that: the bigram word features consistently improves the performance for sentiment analysis tasks; the NB performs well for the task of short snippet sentiment and SVM performs well for the task of longer snippet sentiment; identified a simple NB and SVM variants performs well on various datasets. Tripathy et al. [15] performed the sentiment classification for the movie reviews dataset using n-gram features. They employed the NB (Naïve Bayes), SVM (Support Vector Machine), Stochastic Gradient descent (SGD), and maximum entropy (ME) classifiers on the n-gram features and the combination of n-gram features. The authors indicated that the system accuracy is decreased for the increased level of n-gram features such as trigram, four-gram, and five-gram. Their results show that the combination of unigram and bigram achieves a better result. Fang et al. [16] presented a multi-task learning model to improve the performance of stance prediction. In particular, the authors performed both supervised and unsupervised models for multiple NLP tasks. They achieved 91.2% accuracy for the sentiment analysis task using unigram features. Vashishtha et al. [17] proposed an unsupervised method using n-gram features for the task of sentiment analysis. This method formulates phrases, computes opinion scores, and opinion polarity using the fuzzy linguistic method. In particular, the authors used k-means clustering with fuzzy entropy filter to extract keyphrases that are significant for sentiment analysis. Cambria et al. used neurosymbolic AI for sentiment analysis [18].

Moreover, Das et al. [19] studied the unstructured text with n-gram features and TF-IDF features. They performed the MNB, NB, SVM, DT, RF, and KNN on these two features. Their results indicated that the LR achieves an accuracy of 90.47% using bigram features and the SVM machine achieves an accuracy of 91.99% using TF-IDF features. Ali et al. [20] developed a hybrid model for the sentiment classification task. This model combines convolutional neural network (CNN) and long short-term memory (LSTM) networks. In this model, the authors achieved 89.2% accuracy using the IMDB movie review dataset. Wang et al. [28] proposed a convolutional recurrent neural network for the text modeling task. Their study indicated that the proposed hybrid model strengths the semantic understandings of the text. Especially, the authors achieved 90.39% accuracy for the IMDB movie reviews dataset. Tian et al. [29] implemented an attention-aware bidirectional gated recurrent unit (BiGRU) framework for the sentiment analysis task. The authors incorporated interaction between words using pre-attention BiGRU and extracted the predicted features using post-attention. Their results show that the attention-aware BiGRU model achieves 90.3% accuracy. Rauf et al. [21] determined human emotions from the IMDB movie reviews using the BERT model. The authors achieved 89.90% accuracy for the sentiment analysis task. Alaparthi et al. [22] investigated the sentiment analysis task using LR classifier, lexicon-based, LSTM, and BERT. Their study achieved 92.31% accuracy using the BERT model. Furthermore, Ekbal and Bhattacharyya [23] solved the problem of resource scarcity in sentiment analysis using a high-resource language. The authors used multi-task multilingual framework, which transfers knowledge and maps their semantic meaning between different languages. Especially, the authors extracted character n-grams to generate vectors.

Ashok Kumar et al. [24] studied the n-gram features for Abilify drug user reviews using supervised learning methods. The authors indicated that the TF-IDF-based n-gram features achieve a better result. Bhuvaneshwari et al. [25] introduced Bi-LSTM with a self-attention-based CNN model for subjectivity identification. The authors' used pre-trained word embedding with n-gram features to capture context information between words and sentences. Arevalillo-Herrez et al. [26] adopted the Dual Intent and Entity Transformer for the task of sentiment analysis using the Rasa NLU open-source tool kit. The authors achieved a performance of 90.7% for the IMDb dataset. Especially, their study indicated that the n-gram features with traditional machine learning and deep learning models are not performing well.

Srikanth et al. [27] investigated deep belief neural networks to analyze sentiment in COVID-19 tweets. The authors used a different combination of preprocessing techniques to investigate sentiment in tweets using n-gram features. In summary, the existing researchers studied the n-gram sentiment analysis task using a bag of words, TF-IDF, and context-dependent features. Therefore, this research paper considers context-independent features of texts with N-grams. In this context, the n-gram-based BERT model is proposed with contraction word mapping for sentiment analysis.

III. The proposed methods

An n-gram-based BERT pre-trained model is proposed for sentiment classification using the large IMDB movie reviews as shown in Fig. 1. The proposed model is split into four main subgroups, namely, input data, pre-processing, n-gram characterization, and BERT pre-trained fine-tuning model.

A. Input data

The IMDB movie review dataset [30] is used for the n-gram sentiment analysis task. This dataset contains 50K movie reviews, which are categorized into 25K positive reviews and 25K negative reviews. For instance, the review "It is a funny film, and it doesn't make you smile. What a pity!! It's a simply painful film. The story is presented without a goal" is labeled as negative sentiment. Similarly, the review "I like the whole film and everything in it. I almost felt like watching my friends and me on screen. This movie is a pure masterpiece, very creative and original" is associated with positive sentiment.

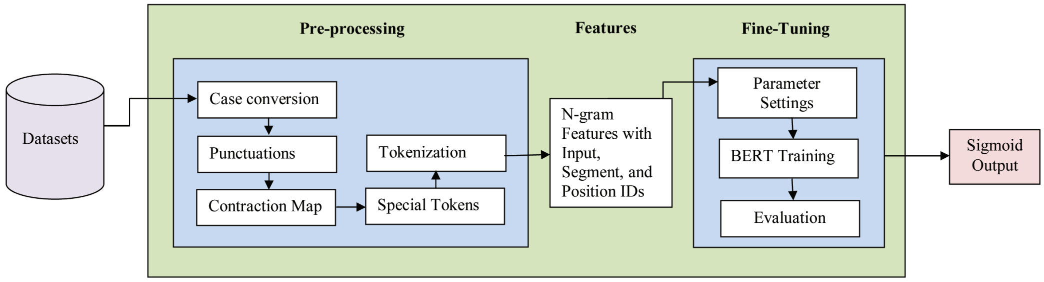

Fig. 1. Architecture Diagram of N-gram-Based BERT model for Sentiment Analysis

TABLE I
CONFUSION MATRIX FOR N-GRAM-BASED BERT MODEL

Dataset	N-grams	1G		2G		3G		1G+2G		2G+3G		1G+2G+3G	
	Class	N	P	N	P	N	P	N	P	N	P	N	P
Training	N	20154	96	20157	93	20132	118	20166	84	20151	99	20146	104
	P	76	20174	84	20166	81	20169	78	20172	76	20174	74	20176
Validation	N	2119	131	2108	142	2096	154	2131	119	2111	139	2102	148
	P	133	2117	105	2145	71	2179	134	2116	116	2134	111	2139
Testing	N	2357	143	2349	151	2343	157	2359	141	2378	122	2346	154
	P	153	2347	141	2359	131	2369	154	2346	146	2354	112	2388

* N-Negative sentiment, P-Positive sentiment

TABLE II
THE PERFORMANCE OF THE UNIGRAM (1G) AND BIGRAM (2G)

Class	Unigram (1G)									Bigram (2G)								
	Training (%)			Validation (%)			Testing (%)			Training (%)			Validation (%)			Testing (%)		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Negative	99.62	99.53	99.58	94.09	94.18	94.14	93.90	94.28	94.09	99.59	99.54	99.56	95.26	93.69	94.47	94.34	93.96	94.15
Positive	99.53	99.62	99.58	94.17	94.09	94.13	94.26	93.88	94.07	99.54	99.59	99.56	93.79	95.33	94.56	93.98	94.36	94.17
Macro	99.58	99.58	99.58	94.13	94.13	94.13	94.08	94.08	94.08	99.56	99.56	99.56	94.52	94.51	94.51	94.16	94.16	94.16
Micro	99.58	99.58	99.58	94.13	94.13	94.13	94.08	94.08	94.08	99.56	99.56	99.56	94.51	94.51	94.51	94.16	94.16	94.16
Weighted	99.58	99.58	99.58	94.13	94.13	94.13	94.08	94.08	94.08	99.56	99.56	99.56	94.52	94.51	94.51	94.16	94.16	94.16

* P-Precision, R-Recall, F1-F1 Score

TABLE III
THE PERFORMANCE OF THE TRIGRAM (3G), AND UNIGRAM AND BIGRAM (1G+2G)

Class	Trigram (3G)									Unigram and Bigram (1G+2G)								
	Training (%)			Validation (%)			Testing (%)			Training (%)			Validation (%)			Testing (%)		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Negative	99.60	99.42	99.51	96.72	93.16	94.91	94.70	93.72	94.21	99.61	99.59	99.60	94.08	94.71	94.40	93.87	94.36	94.12
Positive	99.42	99.60	99.51	93.40	96.84	95.09	93.78	94.76	94.27	99.59	99.61	99.60	94.68	94.04	94.36	94.33	93.84	94.08
Macro	99.51	99.51	99.51	95.06	95.00	95.00	94.24	94.24	94.24	99.60	99.60	99.60	94.38	94.38	94.38	94.10	94.10	94.10
Micro	99.51	99.51	99.51	95.00	95.00	95.00	94.24	94.24	94.24	99.60	99.60	99.60	94.38	94.38	94.38	94.10	94.10	94.10
Weighted	99.51	99.51	99.51	95.06	95.00	95.00	94.24	94.24	94.24	99.60	99.60	99.60	94.38	94.38	94.38	94.10	94.10	94.10

* P-Precision, R-Recall, F1-F1 Score

TABLE IV
THE PERFORMANCE OF THE BIGRAM AND TRIGRAM (2G+3G), AND UNIGRAM, BIGRAM AND TRIGRAM (1G+2G+3G)

Class	Bigram and Trigram (2G+3G)									Unigram, Bigram and Trigram (1G+2G+3G)								
	Training (%)			Validation (%)			Testing (%)			Training (%)			Validation (%)			Testing (%)		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Negative	99.62	99.51	99.57	94.79	93.82	94.30	94.22	95.12	94.67	99.63	99.49	99.56	94.98	93.42	94.20	95.44	93.84	94.63
Positive	99.51	99.62	99.57	93.88	94.84	94.36	95.07	94.16	94.61	99.49	99.63	99.56	93.53	95.07	94.29	93.94	95.52	94.72
Macro	99.57	99.57	99.57	94.34	94.33	94.33	94.64	94.64	94.64	99.56	99.56	99.56	94.26	94.24	94.24	94.69	94.68	94.68
Micro	99.57	99.57	99.57	94.33	94.33	94.33	94.64	94.64	94.64	99.56	99.56	99.56	94.24	94.24	94.24	94.68	94.68	94.68
Weighted	99.57	99.57	99.57	94.34	94.33	94.33	94.64	94.64	94.64	99.56	99.56	99.56	94.26	94.24	94.24	94.69	94.68	94.68

* P-Precision, R-Recall, F1-F1 Score

TABLE V
COMPARISON OF THE PROPOSED MODEL

Authors	Methods	1G	2G	3G	1G+2G	2G+3G	1G+2G+3G
Wang and Manning [14]	MNB	83.55	86.59	-	-	-	-
	SVM	86.95	89.16	-	-	-	-
	NBSVM	88.29	91.22	-	-	-	-
Tripathy et al. [15]	NB	83.65	84.06	70.53	86.00	83.82	86.23
	MaxEnt	88.48	83.22	71.38	88.42	82.94	83.36
	SVM	86.97	83.87	70.16	88.88	83.63	88.94
	SGD	85.11	62.36	58.40	83.36	58.74	83.36
Das et al. [19]	LogisticRegression	-	90.47	-	-	-	-
Vashishtha et al. [17]	SentiScore+Fuzzy entropy+k-means	-	-	-	68.60	53.60	69.10
Wang et al. [28]	Conv-RNN	90.39	-	-	-	-	-
Tian et al. [29]	Attention-Aware BiGRU	90.30	-	-	-	-	-
Ali et al. [20]	CNN-LSTM	89.20	-	-	-	-	-
Fang et al. [16]	MTransSAN	91.20	-	-	-	-	-
Rauf et al. [21]	BERT	89.90	-	-	-	-	-
Alaparthi et al. [22]	BERT	92.31	-	-	-	-	-
Proposed	N-grams+BERT+CM	**94.08**	**94.16**	**94.24**	**94.10**	**94.64**	**94.68**

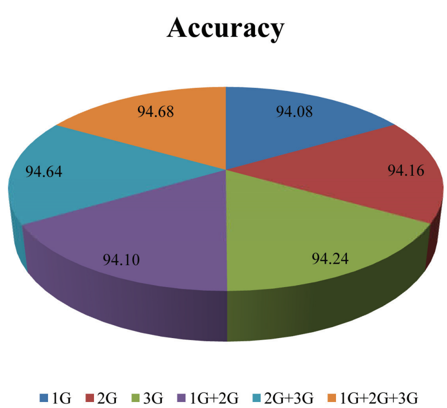

Fig. 2. The obtained result of the BERT model with N-grams

B. Pre-processing

The following preprocessing steps are performed on the movie reviews before feeding them into the BERT model [19]. First, the punctuations are removed except for the single and double quotes, and periods. Second, all reviews are converted from the upper case to the lower case. Third, the Special tokens [CLS] and [SEP] are added at the appropriate positions [7], [8]. Finally, the contraction map is applied for expanding short words like aren't into are not.

C. N-Grams Features

The preprocessed reviews are tokenized using the Word-Piecetokenizer. It breaks the words into their prefix, root, and suffix to handle unseen words better. In particular, the Word-Piecetokenizer is used to create n-gram features such as unigram (1G), bigram (2G), trigram (3G), unigram and bigram (1G+2G), bigram and trigram (2G+3G), and unigram, bigram, and trigram (1G+2G+3G) features [14], [15]. The n-gram defines a continuous sequence of n tokens from a given review. Moreover, the model training using n-gram features gives a pretty good idea of the 'probability' of the occurrence of a word after a certain word.

D. BERT pre-trained fine-tuning model

BERT is a new language representation model developed by Google [8] and excels in natural language processing tasks since it is trained from a large corpus. It overcomes the problem present in other language models that are learning either from the left or right only. BERT learns from both directions and hence has been very successful at natural language prediction. BERT is pre-trained on a large corpus of unlabeled text including the entire Wikipedia and BookCorpus of 3,300 million words. BERT uses random masking to predict the next word during the training phase. BERT learns the context of a word left and right at the same instant. The two variants of BERT are BERT Base and BERT Large. Both models are Encoder-only blocks derived from the original transformer model. The BERT Base consists of 12 layers (transformer layers) and 12 attention heads with 110 million parameters, and the BERT Large consists of 24 layers (transformer layers) and 16 attention heads with 340 million parameters. Each encoder layer has self-attention and feed-forward layers. Self-attention relates positions with each other through queries, keys, and values. The feed-forward is used to normalizes the output units and learn backpropagation. In this work, the BERT base is used for n-gram sentiment analysis using IMDB movie reviews.

IV. RESULTS AND DISCUSSION

The n-gram-based BERT model is implemented for the sentiment analysis task. Specifically, the large IMDB movie review dataset is used that contains 25K positive reviews and 25K negative reviews. This data was pre-processed using case conversion, punctuation, and contraction map. Then, the IMDB movie review dataset is divided into training (40500), validation (4500), and testing (5000) using stratified sampling. Later, the n-gram features are created for 1G, 2G, 3G, and 1G and 2G, 2G and 3G, and 1G, 2G, and 3G. The BERT base model was employed on these n-gram features. It uses 512 sequence length, 20000 maximum word features, 3 epochs, and 2e-5 one-cycle learning rate.

Table I shows the confusion matrix of the n-gram features for training, validation, and testing respectively. Tables I, II, and III show the performance of the unigram, bigram, and trigram individually as well as unigram and bigram, bigram and trigram, and unigram, bigram, and trigram together based on the precision, recall, F1 score, and its micro, macro, and weighted averages [31], [32]. In these tables, the training dataset achieves 100% accuracy for all n-gram features, and the validation dataset achieves 94% for 1G, 1G+2G, 2G+3G, and 1G+2G+3G features and 95% accuracy for 2G and 3G features, and the testing dataset achieves 94% accuracy for 1G, 2G, 3G, and 1G+2G features and 95% accuracy for 2G+3G and 1G+2G+3G features. Overall, the combination of bigram and trigram, and unigram, bigram, and trigram features achieve the highest accuracy of 95%. Table V compares our proposed model with other models. Our model performs comparatively better than other state-of-the-art models (Figure 2). In particular, our model improves 2% accuracy for 1G features, 3% accuracy for 2G features, 23% accuracy for 3G features, 5% accuracy for 1G+2G features, 11% accuracy for 2G+3G features, 6% accuracy for 1G+2G+3G features. Moreover, the performance of the n-gram-based BERT model is not compared with the researchers who performed only with 25K reviews. Therefore, the proposed model seems to outperform well for all n-gram features than other existing models. The limitation of the proposed method takes longer training time and weight updates based on the big corpus size. It also needs more computation cost.

V. Conclusion

Online movie reviews are involved in the promotion and box-office revenue collection of a movie among people. There, it is one of the most influential processes in the film industry. In this work, an n-gram-based BERT model is performed for the task of sentiment classification using the IMDB movie reviews dataset. The dataset was pre-processed into the format of BERT where it accepts input, segment, and position. In particular, a list of n-gram features is created such as unigrams, bigrams, trigrams, and their combination of features. Then, the BERT-based model was employed on these n-gram features for the task of sentiment analysis. This paper mainly focused on the context-independent features of n-grams. The obtained results indicate a better result for all n-gram features than other existing models. In particular, the highest accuracy is achieved from the combination of bigram and trigram features (96.64%), and unigram, bigram, and trigram features (94.68%) than other n-gram features. These indicate that higher-order n-gram features significantly improve the accuracy. In future works, the n-gram features can be studied with gender information using graph neural networks-based transformers and quantum machine learning approaches.

Acknowledgment

This work was supported by the UGC (University Grants Commission), Government of India under the National Doctoral Fellowship.

References

[1] Shayaa, S., Jaafar, N. I., Bahri, S., Sulaiman, A., Wai, P. S., Chung, Y. W., ... & Al-Garadi, M. A. (2018). Sentiment analysis of big data: Methods, applications, and open challenges. IEEE Access, 6, 37807-37827.

[2] Yom-Tov, G. B., Ashtar, S., Altman, D., Natapov, M., Barkay, N., Westphal, M., & Rafaeli, A. (2018, April). Customer sentiment in web-based service interactions: Automated analyses and new insights. In Companion Proceedings of the The Web Conference 2018 (pp. 1689-1697).

[3] Oneto, L., Bisio, F., Cambria, E., Anguita, D. (2016). Statistical Learning Theory and ELM for Big Social Data Analysis. IEEE Computational Intelligence Magazine 11(3), 45-55.

[4] Cambria, E., Schuller, B., Liu, B., Wang, H., Havasi, C. (2013). Statistical Approaches to Concept-Level Sentiment Analysis. IEEE Intelligent Systems 28 (3), 6-9.

[5] Ragusa, E., Gastaldo, P., Zunino, Cambria, E. (2020). Balancing Computational Complexity and Generalization Ability: A Novel Design for ELM. Neurocomputing 401, 405-417.

[6] Chaturvedi, I., Ong, Y., Tsang, I., Welsch, R., Cambria, E. (2016). Learning Word Dependencies in Text by Means of a Deep Recurrent Belief Network. Knowledge-Based Systems 108, 144-154.

[7] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. In Advances in neural information processing systems (pp. 5998-6008).

[8] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

[9] Guerreiro, C., Cambria, E., Nguyen, H. (2019). Understanding the Role of Social Media in Backpacker Tourism. Proceedings of ICDM Workshops, 530-537.

[10] Merello, S., Ratto, A., Oneto, L., Cambria, E. (2019). Ensemble Application of Transfer Learning and Sample Weighting for Stock Market Prediction. Proceedings of IJCNN.

[11] Mondal, A., Cambria, E., Das, D., Bandyopadhyay, S. (2017). MediConceptNet: An Affinity Score Based Medical Concept Network. Proceedings of FLAIRS, 335-340.

[12] Chandra, P., Cambria, E., Hussain, A. (2012). Clustering Social Networks Using Interaction Semantics and Sentics. Advances in Neural Networks, 379-385.

[13] Rosso, P., Bosco, C., Damiano, R., Patti, V., Cambria, E. (2016). Emotion and sentiment in social and expressive media: Introduction to the special issue. Information Processing and Management 52(1), 1-4.

[14] Wang, S. I., & Manning, C. D. (2012, July). Baselines and bigrams: Simple, good sentiment and topic classification. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) (pp. 90-94).

[15] Tripathy, A., Agrawal, A., & Rath, S. K. (2016). Classification of sentiment reviews using n-gram machine learning approach. Expert Systems with Applications, 57, 117-126.

[16] Fang, W., Nadeem, M., Mohtarami, M., & Glass, J. (2019, November). Neural multi-task learning for stance prediction. In Proceedings of the Second Workshop on Fact Extraction and VERification (FEVER) (pp. 13-19).

[17] Vashishtha, S., & Susan, S. (2021). Highlighting keyphrases using senti-scoring and fuzzy entropy for unsupervised sentiment analysis. Expert Systems with Applications, 169, 114323.

[18] Cambria, E., Liu, Q., Decherchi, S., Xing, F., Kwok, K. (2022). SenticNet 7: A Commonsense-based Neurosymbolic AI Framework for Explainable Sentiment Analysis. Proceedings of LREC, 3829-3839.

[19] Das, M., Kamalanathan, S., & Alphonse, P. (2020). A Comparative Study on TF-IDF Feature Weighting Method and its Analysis using Unstructured Dataset.

[20] Ali, N. M., Abd El Hamid, M. M., & Youssif, A. (2019). Sentiment analysis for movies reviews dataset using deep learning models. International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol, 9.

[21] Rauf, S. A., Qiang, Y., Ali, S. B., & Ahmad, W. (2019). Using BERT for Checking the Polarity of Movie Reviews. International Journal of Computer Applications, 975, 8887.

[22] Alaparthi, S., & Mishra, M. (2020). Bidirectional Encoder Representations from Transformers (BERT): A sentiment analysis odyssey. arXiv preprint arXiv:2007.01127.

[23] Ekbal, A., & Bhattacharyya, P. (2022). Exploring Multi-lingual, Multi-task, and Adversarial Learning for Low-resource Sentiment Analysis. Transactions on Asian and Low-Resource Language Information Processing, 21(5), 1-19.

[24] Ashok Kumar, J., Abirami, S., & Trueman, T. E. (2022). An N-Gram Feature-Based Sentiment Classification Model for Drug User Reviews. In Artificial Intelligence and Evolutionary Computations in Engineering Systems (pp. 277-297). Springer, Singapore.

[25] Bhuvaneshwari, P., Rao, A. N., Robinson, Y. H., & Thippeswamy, M. N. (2022). Sentiment analysis for user reviews using Bi-LSTM self-attention based CNN model. Multimedia Tools and Applications, 81(9), 12405-12419.

[26] Arevalillo-Herrez, M., Arnau-Gonzlez, P., & Ramzan, N. (2022). On adapting the DIET architecture and the Rasa conversational toolkit for the sentiment analysis task. IEEE Access.

[27] Srikanth, J., Damodaram, A., Teekaraman, Y., Kuppusamy, R., & Thelkar, A. R. (2022). Sentiment Analysis on COVID-19 Twitter Data Streams Using Deep Belief Neural Networks. Computational Intelligence and Neuroscience, 2022.

[28] Wang, C., Jiang, F., & Yang, H. (2017, August). A hybrid framework for text modeling with convolutional RNN. In Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining (pp. 2061-2069).

[29] Tian, Z., Rong, W., Shi, L., Liu, J., & Xiong, Z. (2018, August). Attention aware bidirectional gated recurrent unit based framework for sentiment analysis. In International Conference on Knowledge Science, Engineering and Management (pp. 67-78). Springer, Cham.

[30] Maas, A., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., & Potts, C. (2011, June). Learning word vectors for sentiment analysis. In Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies (pp. 142-150).

[31] Grandini, M., Bagli, E., & Visani, G. (2020). Metrics for multi-class classification: an overview. arXiv preprint arXiv:2008.05756.

[32] Alejo, R., Antonio, J. A., Valdovinos, R. M., & Pacheco-Snchez, J. H. (2013, June). Assessments metrics for multi-class imbalance learning: A preliminary study. In Mexican Conference on Pattern Recognition (pp. 335-343). Springer, Berlin, Heidelberg.