Computer Science and Information Technologies
Vol. 4, No. 1, March 2023, pp. 43~49
ISSN: 2722-3221, DOI: 10.11591/csit.v4i1.p43-49

Fine grained irony classification through transfer learning approach

Abhinandan Shirahattii[1], Vijay Rajpurohit[2], Sanjeev Sannakki[2]
[1]Department of Computer Science, KLE Society's Dr. M. S. Sheshgiri College of Engineering and Technology, Visvesvaraya Technological University, Belagavi, Karnataka, India
[2]Department of Computer Science, KLS Gogte Institue of Technology, Visvesvaraya Technological University, Belagavi, India

Article Info

Article history:

Received Aug 7, 2022
Revised Dec 23, 2022
Accepted Jan 5, 2023

Keywords:

Bidirectional long-short term memory
Encoder
Irony
Natural language processing
Sentiment analysis
Transformer

ABSTRACT

Nowadays irony appears to be pervasive in all social media discussion forums and chats, offering further obstacles to sentiment analysis efforts. The aim of the present research work is to detect irony and its types in English tweets We employed a new system for irony detection in English tweets, and we propose a distilled bidirectional encoder representations from transformers (DistilBERT) light transformer model based on the bidirectional encoder representations from transformers (BERT) architecture, this is further strengthened by the use and design of bidirectional long-short term memory (Bi-LSTM) network this configuration minimizes data preprocessing tasks proposed model tests on a SemEval-2018 task 3, 3,834 samples were provided. Experiment results show the proposed system has achieved a precision of 81% for not irony class and 66% for irony class, recall of 77% for not irony and 72% for irony, and F1 score of 79% for not irony and 69% for irony class since researchers have come up with a binary classification model, in this study we have extended our work for multiclass classification of irony. It is significant and will serve as a foundation for future research on different types of irony in tweets.

Corresponding Author:

Abhinandan Shirahatti
Department of Computer Science, KLE Society's Dr. M. S. Sheshgiri College of Engineering and Technology, Visvesvaraya Technological University
R.C Nagar 2nd Stage Belagavi Karnataka, India
Email: abhinandans2010@gmail.com

1. INTRODUCTION

Irony has indeed been demonstrated to be ubiquitous in social media, offering major challenge to sentiment analysis field [1]. It is a cognitive phenomenon in which affect-related features play a significant role. In data driven world data is increasing exponentially day by day [2]. Machine learning and deep learning algorithms are playing a vital role in massive data analysis and knowledge extraction. The broad use of creative and metaphorical expressions like irony and sarcasm is common in user-generated content on social media sites like Twitter and Facebook [3]. Irony is the use of language that traditionally means the contrary to express one's meaning, usually for amusing or emphatic effect. Despite their significant distinctions in connotation, the phrases sarcasm and irony are frequently interchanged. The precision of irony identification is crucial in marketing research. Because irony usually causes polarity inversion, failing to acknowledge it may result in poor sentiment classification findings [4]. Intelligence services must be able to identify irony in order to separate perceived risks from ironic statements. Irony identification is indeed a complex problem especially relative to most natural language processing (NLP) tasks. Irony manifests itself in the form of polarized feeling, which is common on Twitter. For example “I really love this year’s summer;

weeks and weeks of awful weather". In this example, irony results from a polarity inversion between two evaluations, the literal evaluation "I really love this year's summer" is positive, while the intended one, which is implied by the context "weeks and weeks of awful weather", is negative.

2. RELATED WORK

Transformers have the ability to learn longer-term dependence, but in the context of language modelling, they are constrained by a fixed-length context. Transformer-XL (extra long), which extends the length of learning dependency without interfering with sequential coherence [5]. Bidirectional encoder representations from transformers (BERT) is intended to train deep bidirectional representations from unlabeled text by reinforcing on both left and right context simultaneously in all levels [6]. Dai and Le presents two ways for improving sequence learning using recurrent networks that employ unlabeled text input. The first method is to anticipate what will happen next in a series. The second method is to utilize a sequence auto encoder, to scans the supplied sequence and to predicts it again [7]. Howard and Ruder proposes an efficient transfer learning approach that may be used to in any NLP tasks [8]. Provided a deep bidirectional language model that has been pertained on huge text corpora and will be put directly on top of the current model, considerably improving performance in subsequent NLP tasks [9]. Suggestion data mining is an emerging and demanding topic of NLP that aims to follow user recommendations on web forums [10]. Proposed 8×8 encoder and decoder layer with an attention mechanism that aids in parallel processing and reduces training time. The attention mechanism aids in paying special attention to each word and its position in the sentence [11]. Xie *et al.* offer an n-dimensional linkage approach for incorporating aspect relationships into deep neural networks for aspect value estimation [12].

3. PROPOSED METHODOLOGY

The Proposed bidirectional long-short term memory-DistilBERT (BiLSTM-DistilBERT) framework consist a stack of layers like sentence embedding, transformer, BiLSTM [13], concatenate, pooling and finally softmax classification layers [14], [15], as shown in Figure 1. The sentence is represent as $S = \{ s_1, s_2, s_3 \ldots s_n \}$ is embedded into the pre-trained DistilBERT transformer layer followed by BiLSTM recurrent neural network [16]. Pooling mechanism is used to the representation of concatenated tensor value of DistilBERT and BiLSTM outputs and finally routed through a fully connected softmax-layer.

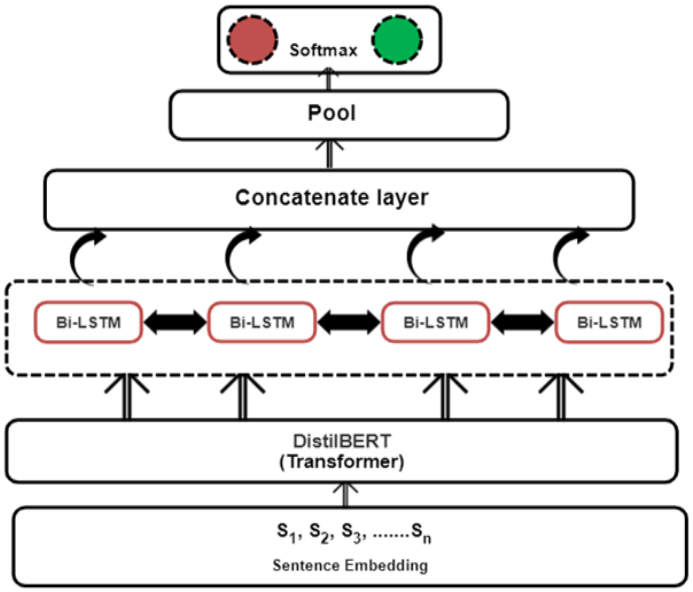

Figure 1. The Proposed BiLSTM-DistilBERT framework

The preceding insight supports our proposed model intuition, as per our observations the pertained deep neural networks play a vital role in NLP downstream tasks. Finally we proposed an end to end model that employs transfer learning by selecting a pre-trained DistilBERT uncased model for our base and adding

additional BiLSTM recurrent neural network to extend the model [17], [18]. This is effective because the pre-trained model's weights contain information representing a high-level understanding of the English language, so we can build on that general knowledge by adding additional layers whose weights will come to represent task-specific understanding of what makes a tweet irony or non-irony [19], [20]. The Hugging Face Transformers library makes transfer learning very approachable, as our general workflow can be divided into four main stages, namely input embedding, defining a model architecture, training classification layer Weights, fine-tuning DistilBERT and training all weights. Hugging Face application programming interface makes it extremely easy to convert words and sentences into sequence of tokens and these tokens are get converted into tensors by the text vectorization class, finally these tensors are fed into our model. Once we instantiate our tokenizer object, we can then go about encoding our training, validation, and test sets in batches using the tokenizer's. batch_encode_plus() method. Important arguments set as part of training are max_length to controls the maximum number of words to tokenize in a given text. Padding or truncation to adjust input according to max_length. Attention mask help the model to decide on which tokens to pay more attention and what all need to ignore thus, including the attention mask as an input to our model helps us to increase the model performance. As pertained model is extended by BiLSTM recurrent neural network units are capable to capture the long range dependencies among the tokens through this proposed model can learn the semantics of each inputs with respect to the specific task. The output of LSTM units get concatenated and passed through a feedforward network with maximum kernel size followed by pooling layer and as output softmax layer uses the softmax function to squash the vector of arbitrary real-valued scores.

4. RESULT AND DISCUSSION

The Proposed model is used Keras [21], is an open-source software library that provides a Python interface for artificial neural networks. Keras acts as an interface for the Tensor Flow library. In binary classification challenge, tweets are classified as irony or not irony. For binary classification, we trained model with 30 epochs, Adam optimizer and sparse categorical cross entropy loss function [22].

4.1. DataSet

SemEval-2018 task 3: irony detection in English tweets, shared task on irony detection [23]: given a tweet, automatic NLP systems should determine whether the tweet is ironic (task A) and which type of irony (if any) is expressed (task B). The ironic tweets were collected using irony-related hashtags (i.e. #irony, #sarcasm, #not) and were subsequently manually annotated to minimize the amount of noise in the corpus. For both tasks, a training corpus of 3,834 tweets was provided, as well as a test set containing 784 tweets. Table 1 represents the irony and not irony samples for binary classification task. Table 2 shows the dataset splitting ration for training, validation and testing for multi-class classification task.

Table 1. Binary classification dataset splitting ration for training, validation and testing

	Training	Validation	Testing
Not irony	1,545	369	455
Irony	1,506	395	329

Table 2. Multi-class classification dataset splitting ration for training, validation and testing

	Training	Validation	Testing
Not irony	1,534	382	473
Clash irony	1,088	295	164
Situtional	263	53	85
Others	168	54	62

4.2. Experimental results

SemEval 2018 irony dataset has only training and testing samples. So, we have divided training samples into training and validation set in ratio of 80:20 [24]. Table 2 shows dataset splitting ration for training, validation and testing phases. We have achieved maximum training accuracy of 98% and validation accuracy of 69%. On testing samples, we have achieved precision of 81% for not irony class and 66% for irony class, recall of 77% for not irony and 72% for irony and 79% F1 score for not irony and 69% irony class. Table 3 shows precision, recall and F1 score for testing samples for binary classification. Our model is performing better in classifying not irony tweets as compare to irony class.

Table 3. Precision, recall and F1 score for testing samples for binary classification

	Precision	recall	F1 score
Not irony	81	77	79
Irony	66	72	69

Figure 2 shows accuracy and loss for training and validation dataset during training, as shown in Figure 2 training loss is always high compare to validation loss. Figure 3 shows confusion matrix for binary classification. Total of 168 samples of not irony class are classified as irony class and total of 108 samples visa-versa. Figure 4 shows area under the curve (AUC) and receiver operating characteristics (ROC) curve for irony binary classification. AUC-ROC curve shows performance of classification model under various threshold settings. AUC of our model is 0.72.

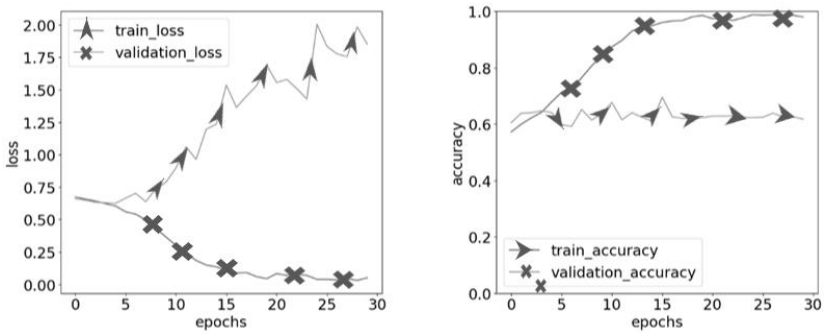

Figure 2. Training and Validation loss and accuracy for 30 epochs in Binary classification

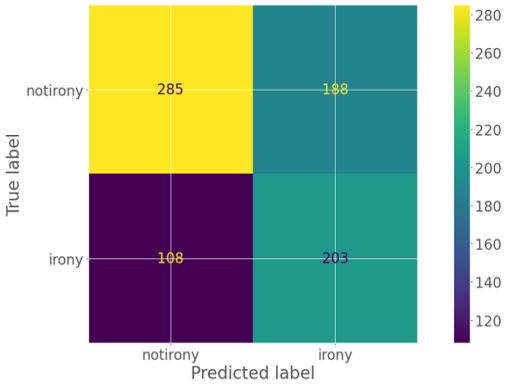

Figure 3. Confusion matrix for binary classification

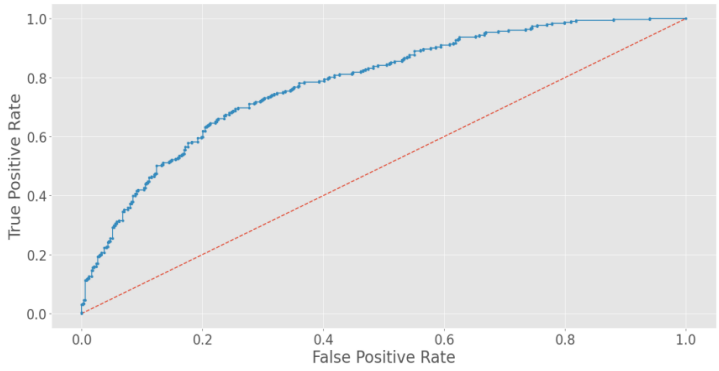

Figure 4. AUC-ROC curve for irony binary classification

In multi-class classification challenge, tweets are classified as three irony categories namely, clash irony, situational irony, and others and irony. Since multi-class dataset is derived from binary classification challenge by categorizing irony tweets, multi-class dataset is not balanced. For multi-class classification, we trained model with 30 epochs, Adam optimizer and sparse categorical cross entropy loss function.

Table 4, shows testing phase results multi-class classification, as shows in the Table 4, proposed model achieves the F1 score of 84% for not irony, 18% for clash irony and 12% for situational. Figure 5, shows accuracy and loss for training and validation dataset during training. We have achieved maximum training accuracy of 87% and validation accuracy of 66%. On testing samples, we have achieved precision of 73% for not irony class, 57% for clash irony class, 80% for situational irony class, recall of 99% for not irony and 10% for clash irony, 6% for situational irony and 84% F1 score for not irony and 18% clash irony and 12% for situational irony. Our model is performing better in classifying not irony tweets as compare to different irony classes. Other type of irony class tweets not properly classified. Figure 6 shows confusion matrix for multi class irony classification. Proposed model is classified total of 553 samples as not irony class, 111 samples as clash irony, 56 samples as situational irony and 36 samples are as otherirony.

Table 4. Testing phase results for different phases for all four classes

	Precision	recall	F1 score
Not irony	73	99	84
Clash irony	57	10	18
Situtional	80	06	12
Others	00	00	00

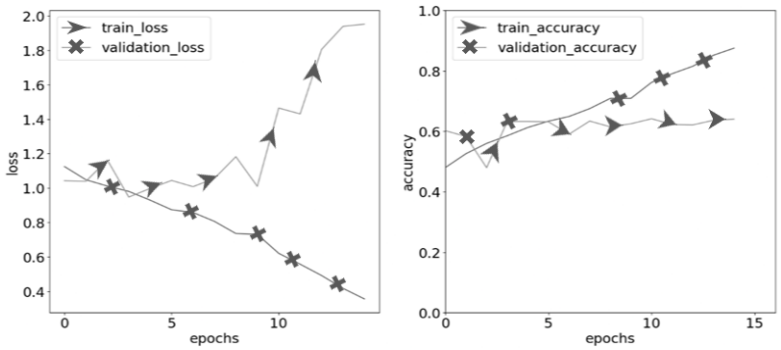

Figure 5. Loss and accuracy for training and validation samples during training phase of multi class irony classification

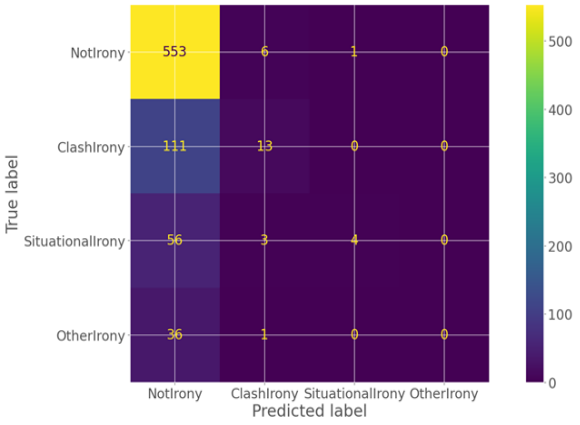

Figure 6. Confusion matrix for multi class irony classification

4.3. Evaluation metrics

In this research work we used confusion matrix to evaluate the performance of the proposed model for fine-grained irony classification task on SemEval-2018 task 3. In (1) to (3) are used to compute hyper

parameters of proposed hybrid neural network model, which might impact classification performance [25]. F1-Score is a measure combining both precision and recall. It is generally described as the harmonic mean of the two.

$$\text{Precision}(P) = \frac{TruePositive}{TruePositive+FalsePositive} \quad (1)$$

$$\text{Recall}(R) = \frac{TruePositive}{TruePositive+FalseNegative} \quad (2)$$

$$\text{F1} - \text{Score} = 2 * \frac{P*R}{P+R} \quad (3)$$

5. CONCLUSION AND FUTURE SCOPE

In this research work we proposed a BiLSTM-DistilBERT hybrid neaural network model to address fine-grained irony classification task on SemEval-2018 task 3 dataset. Transformers are used to minimize the data preprocessing and feature extraction tasks. Through transfer learning approach our proposed BiLSTM-DistilBERT model achieves state-of-the-art results over the SemEval-2018 task 3 dataset. Also in future, instead of DistilBERT transformers other type of transformers could be used to extract features from tweets. Also classification models such as basic supervised machine learning algorithm support vector machine could be used.

ACKNOWLEDGEMENTS

We would like to express our gratitude to all of our coworkers who contributed to this research, both in terms of their knowledge and other forms of support, at the Faculty of Computer Science, KLS Gogte Institute of Technology, affiliated to Visvesvaraya Technological University.

REFERENCES

[1] M. Deckert, M. Schmoeger, M. Geist, S. Wertgen, and U. Willinger, "Electrophysiological correlates of conventional metaphor, irony, and literal language processing – an event-related potentials and eLORETA study," *Brain and Language*, vol. 215, p. 104930, Apr. 2021, doi: 10.1016/j.bandl.2021.104930.

[2] Z. L. Chia, M. Ptaszynski, F. Masui, G. Leliwa, and M. Wroczynski, "Machine learning and feature engineering-based study into sarcasm and irony classification with application to cyberbullying detection," *Information Processing and Management*, vol. 58, no. 4, p. 102600, Jul. 2021, doi: 10.1016/j.ipm.2021.102600.

[3] J. Á. González, L. F. Hurtado, and F. Pla, "Transformer based contextualization of pre-trained word embeddings for irony detection in Twitter," *Information Processing and Management*, vol. 57, no. 4, p. 102262, Jul. 2020, doi: 10.1016/j.ipm.2020.102262.

[4] S. Zhang, X. Zhang, J. Chan, and P. Rosso, "Irony detection via sentiment-based transfer learning," *Information Processing and Management*, vol. 56, no. 5, pp. 1633–1644, Sep. 2019, doi: 10.1016/j.ipm.2019.04.006.

[5] Z. Dai, Z. Yang, Y. Yang, J. Carbonell, Q. V. Le, and R. Salakhutdinov, "Transformer-XL: attentive language models beyond a fixed-length context," in *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, 2020, pp. 2978–2988, doi: 10.18653/v1/p19-1285.

[6] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," in *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, 2019, vol. 1, pp. 4171–4186.

[7] A. M. Dai and Q. V Le, "Semi-supervised sequence learning," *Advances in Neural Information Processing Systems*, vol. 28, 2015, [Online]. Available: https://proceedings.neurips.cc/paper/2015/file/7137debd45ae4d0ab9aa953017286b20-Paper.pdf.

[8] J. Howard and S. Ruder, "Universal language model fine-tuning for text classification," in *ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*, 2018, vol. 1, pp. 328–339, doi: 10.18653/v1/p18-1031.

[9] M. E. Peters *et al.*, "Deep contextualized word representations," in *NAACL HLT 2018 - 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, 2018, vol. 1, pp. 2227–2237, doi: 10.18653/v1/n18-1202.

[10] R. A. Potamias, A. Neofytou, and G. Siolas, "NTUA-ISLab at SemEval-2019 task 9: mining suggestions in the wild," in *NAACL HLT 2019 - International Workshop on Semantic Evaluation, SemEval 2019, Proceedings of the 13th Workshop*, 2019, pp. 1224–1230, doi: 10.18653/v1/s19-2215.

[11] Y. Wu *et al.*, "Google's neural machine translation system: bridging the gap between human and machine translation," *arXiv preprint arXiv:1609.08144*, 2016, doi: 10.48550/arXiv.1609.08144.

[12] H. Xie, W. Lin, S. Lin, J. Wang, and L. C. Yu, "A multi-dimensional relation model for dimensional sentiment analysis," *Information Sciences*, vol. 579, pp. 832–844, Nov. 2021, doi: 10.1016/j.ins.2021.08.052.

[13] S. Brahma, "Improved sentence modeling using suffix bidirectional LSTM," *arXiv preprint arXiv:1805.07340*, 2018, doi: 10.48550/arXiv.1805.07340.

[14] S. Ilić, E. Marrese-Taylor, J. Balazs, and Y. Matsuo, "Deep contextualized word representations for detecting sarcasm and irony," in *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, 2018, pp. 2–7, doi: 10.18653/v1/w18-6202.

[15] C. Wu, F. Wu, S. Wu, J. Liu, Z. Yuan, and Y. Huang, "THU NGN at SemEval-2018 task 3: Tweet irony detection with densely connected LSTM and multi-task learning," in *NAACL HLT 2018 - International Workshop on Semantic Evaluation, SemEval 2018 - Proceedings of the 12th Workshop*, 2018, pp. 51–56, doi: 10.18653/v1/s18-1006.
[16] R. A. Potamias, G. Siolas, and A. Stafylopatis, "A robust deep ensemble classifier for figurative language detection," in *Communications in Computer and Information Science*, vol. 1000, Springer International Publishing, 2019, pp. 164–175.
[17] C. Baziotis *et al.*, "NTUA-SLP at SemEval-2018 task 3: tracking ironic Tweets using ensembles of word and character level attentive RNNs," in *NAACL HLT 2018 - International Workshop on Semantic Evaluation, SemEval 2018 - Proceedings of the 12th Workshop*, 2018, pp. 613–621, doi: 10.18653/v1/s18-1100.
[18] G. Alwakid, T. Osman, M. El Haj, S. Alanazi, M. Humayun, and N. U. Sama, "MULDASA: multifactor lexical sentiment analysis of social-media content in nonstandard arabic social media," *Applied Sciences (Switzerland)*, vol. 12, no. 8, p. 3806, Apr. 2022, doi: 10.3390/app12083806.
[19] K. Cortis and B. Davis, "Over a decade of social opinion mining: a systematic review," *Artificial Intelligence Review*, vol. 54, no. 7, pp. 4873–4965, Jun. 2021, doi: 10.1007/s10462-021-10030-2.
[20] E. Savini and C. Caragea, "Intermediate-task transfer learning with BERT for sarcasm detection," *Mathematics*, vol. 10, no. 5, p. 844, Mar. 2022, doi: 10.3390/math10050844.
[21] S. Nitish, H. Geoffrey, K. Alex, S. Ilya, and S. Ruslan, "Dropout: a simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, pp. 1929–1958, 2014.
[22] H. Luo, T. Li, B. Liu, and J. Zhang, "Doer: dual cross-shared RNN for aspect term-polarity co-extraction," in *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, 2020, pp. 591–601, doi: 10.18653/v1/p19-1056.
[23] S. Chen, Y. Wang, J. Liu, and Y. Wang, "Bidirectional machine reading comprehension for aspect sentiment triplet extraction," *35th AAAI Conference on Artificial Intelligence, AAAI 2021*, vol. 14A, no. 14, pp. 12666–12674, May 2021, doi: 10.1609/aaai.v35i14.17500.
[24] C. Zhang, Q. Li, D. Song, and B. Wang, "A multi-task learning framework for opinion triplet extraction," in *Findings of the Association for Computational Linguistics Findings of ACL: EMNLP 2020*, 2020, pp. 819–828, doi: 10.18653/v1/2020.findings-emnlp.72.
[25] Y. Yin, F. Wei, L. Dong, K. Xu, M. Zhang, and M. Zhou, "Unsupervised word and dependency path embeddings for aspect term extraction," *IJCAI International Joint Conference on Artificial Intelligence*, vol. 2016-Janua, pp. 2979–2985, 2016.

BIOGRAPHIES OF AUTHORS

Abhinandan Shirahatti is Assistant Professor in Computer Science and Engineering at KLE DR MSSCET, Belagavi, India and he is currently pursuing PhD studies at the Visvesvaraya Technological University, Belagavi, India in the Department of Computer Science and Engineering. He received his Master and Bachelor of Engineering degrees from the Visvesvaraya Technological University, Belagavi, India in 2012 and 2010, respectively. He can be contacted at email: abhinandans2010@gmail.com.

Dr. Vijay Rajpurohit Professor and Head, Department of CSE at GIT, Belgaum, Karnataka, India. Received B.E. in Computer Science from KUD Dharwad, M. Tech from N.I.T.K Surathkal, and PhD from Manipal University His research interest include image processing, cloud computing, and data analytics. He has good number of publications. He can be contacted at email: vijaysr2k@yahoo.com.

Dr. Sanjeev Sannakki Professor in the department of CSE at GIT, Belgaum, and Karnataka, India. He received his B.E. in ECE from KUD Dharwad in 2009, M. Tech and PhD from VTU, Belagavi. His research interest include image processing, cloud computing, computer networks, and data analytics. He has good number of publications. He is the reviewerfor a few international journals. He can be contacted at email: sannakkisanjeev@gmail.com.