Analyzing DistilBERT for Sentiment Classification of Banking Financial News

Varun Dogra, Aman Singh, Sahil Verma, Kavita, N. Z. Jhanjhi, and M. N. Talib

Abstract In this paper, the sentiment classification approaches are introduced in Indian banking, governmental and global news. The study assesses state-of-art deep contextual language representation, DistilBERT, and traditional context-independent system, TF-IDF, on multiclass (positive, negative, and neutral) sentiment classification news-events. The DistilBERT model is fine-tuned and fed into four supervised machine learning classifiers Random Forest, Decision Tree, Logistic Regression, and Linear SVC, and similarly with baseline TF-IDF. The findings indicate that DistilBERT can transfer basic semantic understanding to further domains and lead to greater accuracy than the baseline TF-IDF. The results also suggest that Random Forest with DistilBERT leads to higher accuracy than other ML classifiers. The Random Forest with DistilBERT achieves 78% accuracy, which is 7% more than with TF-IDF.

Keywords Sentiment classification · DistilBERT · TF-IDF · Machine Learning classifiers · and Transformers

V. Dogra · A. Singh
School of Computer Science and Engineering, Lovely Professional University, Phagwara, India

S. Verma (✉) · Kavita
Department of Computer Science and Engineering, Chandigarh University, Mohali, India
e-mail: sahilverma@ieee.org

Kavita
e-mail: kavita@ieee.org

N. Z. Jhanjhi
School of Computer Science and Engineering, Taylor's University, Subang Jaya, Malaysia
e-mail: NoorZaman.Jhanjhi@taylors.edu.my

M. N. Talib
Papua New Guinea University of Technology, Lae, Papua New Guinea
e-mail: muhammad.talib@pnguot.ac.pg

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2021

S.-L. Peng et al. (eds.), *Intelligent Computing and Innovation on Data Science*,
Lecture Notes in Networks and Systems 248,
https://doi.org/10.1007/978-981-16-3153-5_53

5 Conclusion and Future direction

This paper goals to classify the sentiments of banking news-events amongst three classes positive, negative and neutral. DistilBERT with fine-tuning on this sentiment classification task is compared with traditional TF-IDF. The key objective was to observe how much this task can be benefitted from the deep contextual pre-trained language representation. It is found that DistilBERT has performed better than TF-IDF with all four machine learning classifiers. The DistilBERT with Random Forest has achieved 78% accuracy, 7% better than Random Forest with TF-IDF. This is also found that Random Forest with DistilBERT has performed better than other classifiers like Logistic Regression, Decision Tree, and Linear SVC. The precision and recall for all classes were also higher with DistilBERT as compared to TFIDF.

In conclusion, despite much longer training times and memory requirements, when a model's transfer capacity is a priority, it is worth choosing contextual neural models over traditional approaches. More training data and test data may be obtained for future work to generalize sentiment classification findings in the banking news domain. To compare the effects more conveniently, different tests may be performed. Moreover, dictionary-based rules may be created for each class, positive, negative, and neutral, to support the classification with DistilBERT or other pre-trained transformer-based models.

References

1. Omotosho BS, Tumala MM (2019) A text mining analysis of Central Bank Monetary Policy Communication in Nigeria
2. Verma I, Dey L, Meisheri H (2017) Detecting, quantifying and accessing impact of news events on Indian stock indices. In: Proceedings of the international conference on web intelligence, pp 550–557
3. Kaya M, Fidan G, Toroslu IH (2012) Sentiment analysis of turkish political news. In: 2012 IEEE/WIC/ACM international conferences on web intelligence and intelligent agent technology, vol 1. IEEE, pp 174–180
4. Yu L, Wu J, Chang P, Chu H (2013) Knowledge-based systems using a contextual entropy model to expand emotion words and their intensity for the sentiment classification of stock market news. Knowl Based Syst 41:89–97
5. Mikolov T, Chen K, Corrado G, Dean J (2013) Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781
6. Pennington J, Socher R, Manning CD (2014) Glove: global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pp 1532–1543
7. Azar PD (2009) Sentiment analysis in financial news. Doctoral dissertation, Harvard University
8. Sanh V, Debut L, Chaumond J, Wolf T (2019) DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108
9. Nguyen TH, Shirai K, Velcin J (2015) Sentiment analysis on social media for stock movement prediction. Expert Syst Appl 42(24):9603–9611

10. Schumaker RP, Chen H (2009) A quantitative stock prediction system based on financial news. Inf Process Manag 45(5):571–583
11. Xia R, Zong C, Hu X, Cambria E (2013) Feature ensemble plus sample selection: domain adaptation for sentiment classification. IEEE Intell Syst 28(3):10–18
12. Jing LP, Huang HK, Shi HB (2002) Improved feature selection approach TFIDF in text mining. In: Proceedings of the international conference on machine learning and cybernetics, vol 2. IEEE, pp 944–946
13. Mikolov T, Karafiát M, Burget L, Černocký J, Khudanpur S (2010) Recurrent neural network based language model. In: Eleventh annual conference of the international speech communication association
14. Zhou P, Qi Z, Zheng S, Xu J, Bao H, Xu B (2016) Text classification improved by integrating bidirectional LSTM with two-dimensional max pooling. arXiv preprint arXiv:1611.06639.
15. Sujatha R, Chatterjee JM, Jhanjhi NZ, Brohi SN (2021) Performance of deep learning vs machine learning in plant leaf disease detection. Microprocess Microsyst 80:103615
16. Devlin J, Chang MW, Lee K, Toutanova K (2018) Bert: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.
17. Elagamy MN, Stanier C, Sharp B (2018) Stock market random forest-text mining system mining critical indicators of stock market movements. In: 2018 2nd international conference on natural language and speech processing (ICNLSP). IEEE, pp 1–8
18. Batra I, Verma S, Malik A, Ghosh U, Rodrigues JJ, Nguyen GN, Mariappan V (2020) Hybrid logical security framework for privacy preservation in the green internet of things. Sustainability 12(14):5542
19. Batra I, Verma S, Alazab M (2020) A lightweight IoT-based security framework for inventory automation using wireless sensor network. Int J Commun Syst 33(4):e4228
20. Hochreiter S (1997) JA1 4 rgen Schmidhuber, Long short-term memory. Neural Comput 9(8)