Comparison and fine-tuning of methods for Financial Sentiment Analysis

Leonardo Colacicchi

Department of Data Science and Artificial Intelligence

Maastricht University

January 2022

Abstract

Financial sentiment analysis presents a significant challenge in the field of text classification partly due to a lack of reliable large enough labelled datasets as well as the nature of its domain specific language. This research reviews some of the most successful approaches to solve this task and test their performance on two datasets, consisting of financial news headlines and user tweets related to the stock market. Extensive experiments on Google's Bidirectional Encoder Representations from Transformers (BERT) architecture and its variations were performed to achieve state of the art accuracy of 87.1% on the financial PhraseBank dataset by combining a fine-tuned model with a regularized logistic regression classifier.

1 Introduction

In the field of finance, there exist two schools of thought about the functioning of markets, be it stocks, commodities or any publicly tradable good. The efficient market hypothesis states that markets are inherently rational and therefore any new information about a company is almost instantly reflected into its stock price. In contrast, the field of behavioural finance supports the idea that investors in the market are in fact irrational, and emotion and psychology play an important role in what price a certain asset is trading at. With the second perspective in mind, a growing interest in being able to capture the general sentiment about a company in a fast, accurate and automatized way has developed among investors in the market trying to navigate the vast amount of information available to them, from news articles to balance sheets, earnings reports as well as social media. While extensive research exists investigating all the aforementioned information sources, in this study the primary focus is put on the techniques for sentiment analysis of short text in the form financial news headlines and tweets related to the stock market. Existing sentiment analysis techniques are generally trained and measured on common language such as the one found in reviews of Amazon products or movies on iMDB. Performance in such settings is not easy to replicate in any domain making use of specific language which is foreign to a model, although this is especially the case in the financial domain due to the large number of homonynms used, which are words that have the same spelling but differ in meaning. Some of the many examples of these words are 'bear' or 'bull', which are usually neutral in polarity but indicate a falling and rising stock market respectively; 'share' also indicates a positive sentiment in most contexts while having a neutral connotation in a financial setting. Another difficulty in the field of financial sentiment analysis is the relative lack of large

datasets publicly available to train the models, which is partly due to the relatively high expertise required to be able to accurately label financial texts. A technique that has seen growing success in overcoming these issues in the field of Natural Language Processing is the use of pre-trained models and transfer learning. By combining the knowledge acquired by a model through pre-training on large amounts of data with the training required for any given down-stream task, impressive results can be obtained without the need for domain specific lexicons or very large domain specific datasets [1]. Given the success of this approach in a variety of NLP down-stream tasks this paper focuses on experimenting different variations of one of the most recent and successful of these pre-trained models in the form of the Bidirectional Encoder Representations from Transformers (BERT) [2] model. More specifically the following research questions will be addressed:

1. What is the best performing approach to classify text in the financial domain?
2. Can we improve the performance of a pre-trained language model for financial text classification?
3. Which language model is better suited for fine-tuning on financial text data?
4. What is the effect of training set size on the classifier's performance?
5. What is the effect of fine-tuning on a data-set when classifying a different one?

In addition to a variety of experiments presented in Section 4, the main contribution of this research to the existing methodologies was to train different classifiers on top of a fine-tuned BERT model, finding that regularization has a noticeable effect on evaluation accuracy and F1 scores, increasing the state of the art of both by 1%. The paper is structured as follows. In Section 2 an overview of existing related research in the field of financial sentiment analysis is given. Section 3 presents the datasets on which this work was based, followed by an outline of the methods that were used (Section 4). Then, the experiments conducted to answer the research questions are illustrated in Section 5, concluding with a summary of the results and possibilities for further research on the topic (Section 6).

2 Related work

Given the concrete implications of achieving an effective method for classifying texts associated with the financial markets, substantial research has been made in order to solve this task. Existing methods can be mainly divided into three categories: lexicon-based approaches, machine learning or statistical methods, and deep learning methods. Lexicon-based methods involve the calculation of the orientation of a text based on the semantic orientation of words or phrases present in the text [3]. In [4] Loughran and McDonald show that lexicons developed for other ares of study perform poorly when applied to the financial domain, and thus created a finance specific lexicon by examining word usage in a large sample of financial documents during the period 1994–2008 to more precisely represent word sentiment. The LM lexicon has since become predominant in the field of financial text classification [5]. For instance in [6] the authors use the LM lexicon to find that the journalists of the WallStreetJournal's "Abreast of the Market" column who use a more pessimistic tone have a significant effect on a declining stock market the following day. As illustrated in [7] however, even with domain specific lexicons, a common problem in the field of sentiment analysis is that the context in which a word is used may cause the polarity of the word to differ from the one previously assigned to it in the lexicon. The use of labelled data to train machine learning models can help improve on this and other aspects of the task. In [8] authors present a hybrid model, using both a lexicon and a classifier for the classification of short financial text that tries to overcome this issue by incorporating phrase-structure and interactions between financial concepts. A machine learning approach is used in [9], where a Naive Bayes classifier is used to classify companies' quarterly and annual filings. More recently, the use of deep learning to train large language models has seen growing interest and success in an array of different NLP related tasks. In [10] the authors demonstrate the use of BERT for the classification of a collection of long form texts such as newspaper articles and blog posts, achieving an F1 score of 72.5% using only 572 training samples.

Another example of a pre-trained model used for financial sentiment analysis is presented in [11], where authors experiment with a variety of methods to show RoBERTa outperforming traditional methods such as Naive Bayes or Support Vector Machines. A transfer learning approach has also been investigated on the same dataset on which this work is based on; in [12] the author, uses Google's BERT language model as a basis for further training on financial news headlines to achieve state of the art accuracy, a result on which we expand in this research, as will be shown in Section 4.

3 Datasets

The following section presents the two datasets used to perform our investigation.

3.1 Financial news headlines

The Financial PhraseBank is a dataset consisting of English news headlines of companies listed on the OMX Helsinki exchange, first presented in [8]. It was annotated by a team of 16 people with background in finance, economics or accounting, who classified each sentence based on whether it would have a positive, neutral or negative effect on the company's stock price. The data is partitioned based on the agreement level of the annotators regarding each class label. There are four partitions of the dataset representing sentences on which 50%, 66%, 75% and 100% of annotators agree on the label provided, although for the sake of our experiments the dataset was used in its entirety. The label distribution can be seen in Table 1.

3.2 Financial tweets

The second dataset is a collection of tweets related to the stock market collected by a user on kaggle.com [13]. Each sentence is classified into either positive or negative sentiment, with the count of each class illustrated in Table 1. Given the unofficial source of the data and no other literature against which to test our model's performance, the majority of the experiments were conducted on the Financial PhraseBank, with the Financial tweets data being used mostly for comparison.

Label	Fin. News	Fin. Tweets
Negative	604	2106
Neutral	2879	-
Positive	1362	3685

Table 1: Datasets' label distribution.

4 Methods

Section 4 presents the methods used to conduct the analysis. These can be divided into two approaches: feature extraction from pre-trained language models, and fine-tuning of those same models on our datasets. The language models used were all variations of Google's Bidirectional Encoder Representations from Transformers (BERT) architecture, for which a brief overview is illustrated here.

4.1 Overview of the models

4.1.1 BERT

BERT is a language representation model that makes use of the now ubiquitous Transformer architecture. In its base version, the one used in this research, it consists of a stack of 12 encoder layers, each of which is made up of a self attention mechanism and a fully connected feed-forward neural network [14], with a hidden size of 768. The total number of parameters in this configuration is 110M. BERT is pre-trained on two tasks: next sentence prediction, where given a pair of sentences, the model needs to assess whether one follows the other, and masked language modelling, in which 15% of its input tokens are masked and need to be inferred by the model. This allows BERT to retain a unified architecture for a wide range of NLP downstream tasks, with very little difference between the pre-trained and fine-tuned architecture. For tokenization, BERT uses WordPiece embeddings [15], with every input sequence having

a special [CLS] token added as its first token, which is used as the aggregate sequence representation for classification tasks.

4.1.2 DistilBERT

DistilBERT [16] is a pre-trained language model based on the original BERT configuration that aims at retaining as much of the original model's performance as possible while significantly reducing its size, and therefore processing time, during inference. It achieves this by training a smaller model using knowledge distillation, a compression technique in which the smaller model learns to reproduce the behaviour of a larger model by training it on the latter's soft output. In the standard configuration used in this research, DistilBERT uses 6 layers for a total of 66M parameters.

4.1.3 RoBERTa

RoBERTa [17] is another BERT based language model developed by Facebook researchers to study the effect of hyperparameter tuning and training set size. The main differences between RoBERTa and the original model are the following: (1) the batch size during training is increased from 256 to 8k; (2) masking of a sentence is done dynamically each time it is fed as input to the model rather than statically during the pre-processing of the data and (3) the Next Sentence Prediction (NSP) task is removed during pre-training. Despite removing the NSP task the authors show that classification performance is not affected when tested on the Stanford Sentiment Treebank (SST-2) benchmark dataset.

4.1.4 FinBERT

FinBERT [12] is a BERT based language model fine-tuned for sentiment classification in the financial domain. The author uses the original BERT model and further pre-trains it on a subset of the TRC2 corpus, a collection of 1.8M news articles published by Reuters between 2008 and 2010, by filtering for keywords related to finance. The model is then fine-tuned on the Financial PhraseBank by adding a dense layer to the last hidden state of the CLS token.

4.2 Feature extraction

In order to perform classification using a pre-trained language model most approaches that were found in the literature add a dense layer to the CLS token of the last layer and train the weights of the whole model on the desired down-stream task. Instead, one of the methods explored in this analysis was to experiment with feature extraction, where training is performed by using the output vector of the pre-trained language model as input to another classifier, keeping the weights of the original model unchanged. In addition to using the CLS token as our 768 length output vector for sentence representation, mean and max pooling of all the tokens in the last layer was tested. Experiments were conducted to verify which of the following classifiers would yield the best performance: (1) a logistic regression (LR) model using the Limited-memory Broyden–Fletcher–Goldfarb–Shanno (lbfgs) solver with number of max iterations set to 3000 and L2 regularization; (2) a random forest classifier (RF) using 100 estimators and Giny impurity as its splitting criterion and (3) a support vector machine (SVM) model using the RBF kernel. All of the above mentioned experiments were performed on both datasets.

4.2.1 Fine-tuning

Fine-tuning of all the models was also implemented to test for their performance. This was done primarily on the Financial PhraseBank dataset where the results could be compared to previous literature. Figure 1 illustrates the different approaches compared with each other. All models were trained using an AdamW optimizer with a training rate of 1e-3, over 3 epochs and using a batch size of 8. Training on a higher number of epochs was also tested but resulted in worse validation performance due to overfitting. Training was done on a Tesla P100 with 16GB of memory. Additionally, the financial tweets dataset was fine-tuned with the same configuration, henceforth referred to as tweetBERT, and compared to the classification performance of BERT fine-tuned to the Financial PhraseBank to test how well the model's performance transfers between datasets of the same

Model	Accuracy	F1	Loss
BERT	74.2	74.0	67.7
DistilBERT	76.1	75.7	56.9
RoBERTa	68.6	65.4	70.1
FinBERT	87.1	87.1	43.8

Table 2: Language models' performance on the financial PhraseBank dataset. The vector of the CLS token was used as input to a Logistic Regression classifier. All measures are calculated using 10-fold cross validation.

domain. Finally, ablation studies were performed to test the effect of training size on all models' performance. The data of training sizes of [300, 600, 1200, 2400, 4845] were sampled by maintaining the same proportion of each class as in the original dataset.

5 Experiments

In this section the results of the experiments aimed at answering the research questions from Section 1 are presented. The data was split in 80% for training and 20% for testing. Since the Financial PhraseBank data is a multi-class classification problem, the F1 scores are calculated for each label and then weighted according to the number of true instances for each label, in order to account for class imbalance in the data.

5.1 Classifiers & pooling strategy (RQ1 & RQ2)

The first method tested was to use the different language models, without fine-tuning, to explore which of them performed best on financial text data in their pre-trained configuration. To this end, the models average pooled output sentence representation was used as input for the three different classifiers in order to test their performance. Table 2 illustrates the results on the Financial PhraseBank dataset. Unsurprisingly FinBERT significantly outperforms other models, which is to be expected since it was fine-tuned on the very dataset that was used here. More worthy of notice is the fact that DistilBERT, despite its reduced size, actually slightly outperforms the original BERT on all three measures when used in its original configuration. In Table 3 the performance of the three different classifiers can be seen when used on both datasets and with the original architecture. Accuracy is generally better on the Financial PhraseBank dataset, but the model seems to perform better in terms of F1 scores for the Financial tweets data. This is surprising considering that the unorthodox and often grammatically inconsistent language used in a social media context was not part of the original model's pre-training. The three different pooling strategies mentioned in Section 4 were also tested on all four models with their pre-trained weights. The results for BERT are presented in Table 4, which shows that using the mean of the tokens in the last hidden layer of the model actually yields the best results on all three measures. This was consistent across all models except for FinBERT, where its CLS token representation slightly outperformed the other pooling strategies. This is expected since FinBERT was the only model that was fine-tuned on financial text classification data already. Comparing the results to the previous literature on the Financial PhraseBank we find that combining a regularized logistic regression classifier to the last hidden layer's CLS token of the FinBERT model yields a 1% improvement on the state of the art classification performance achieved by the standard FinBERT implementation (Table 5). The main driver behind the increased performance is to be attributed to the regularization term added during the training of the classifier, since it was found to significantly outperform the same model without regularization. This together with a very low training loss suggests that the main limitation behind the original FinBERT's performance was overfitting.

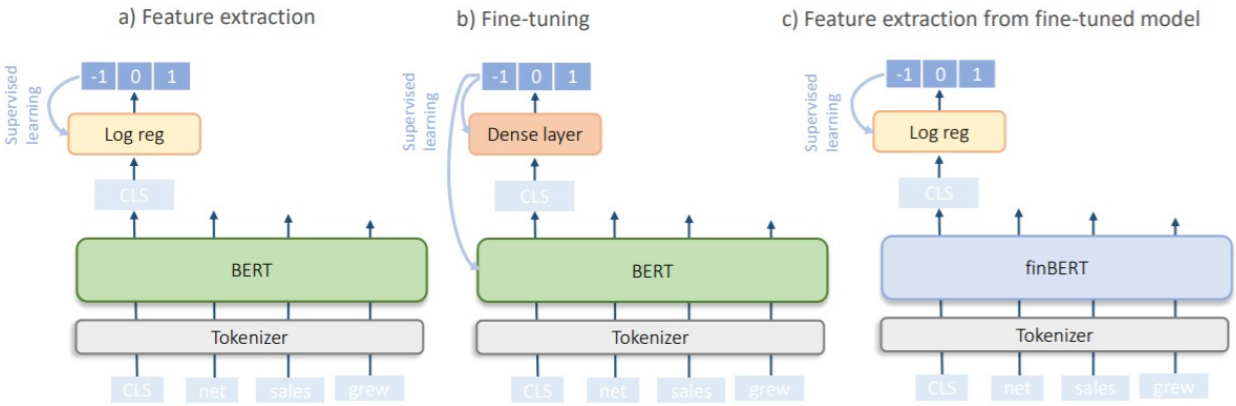

Figure 1: An illustration of the three approaches used in the experiments, with c) outperforming previous state of the art accuracy and F1 scores.

	Fin. News			Fin. Tweets		
Classifier	Accuracy	F1	Loss	Accuracy	F1	Loss
LR	77.1	76.7	56.0	70.7	77.7	59.1
RF	69.9	65.9	69.0	67.1	78.1	60.4
SVM	76.8	75.5	53.3	71.6	79.8	56.9

Table 3: BERT's performance using different classifiers' on its output. All measures were calculated using 10-fold cross validation

Pooling	Accuracy	F1	Loss
CLS Token	77.1	76.7	56.0
Mean pooling	77.3	76.9	54.9
Max pooling	72.1	70.5	66.4

Table 4: Different pooling strategies' performance on the Financial PhraseBank dataset. All measures are calculated using 10-fold cross validation.

Model	Accuracy	F1	Loss
LPS	71.0	71.0	-
stand. FinBERT	86.0	84.0	37.0
FinBERT + LR	87.1	87.1	43.8

Table 5: Comparison of previous literature with the best performing implementation in this research. LPS stands for Linear Phrase Structure and is the model created by the authors of the Financial PhraseBank.

5.2 Fine-tuned performance (RQ3)

Another approach that was tested was to fine-tune the three variations of the original BERT model to the Financial PhraseBank to verify two things: (1) how much of an improvement over the pre-trained configuration this would yield and (2) to test which model is better suited for learning in this downstream task regardless of its pre-training configuration. Results are shown in Table 6. FinBERT is omitted since it was already fine-tuned on the same dataset. In contrast to their pre-training accuracy, in which DistilBERT outperformed the original BERT architecture, when fine-tuned, the latter's increased size and modelling power results in a 2% better accuracy and F1 score, as well as a 7% lower cross-entropy loss compared to DistilBERT.

Model	Accuracy	F1	Loss	Runtime
BERT	86.5	86.6	59.8	609
DistilBERT	84.8	84.8	66.6	393
RoBERTa	73.5	72.1	63.1	755

Table 6: Fine-tuned performance of the three language models on the Financial PhraseBank data with runtime of training in seconds

Model	Accuracy	F1	Loss
BERT + LR	70.7	77.7	59.1
FinBERT	71.57	78.47	63.67
tweetBERT	82.53	82.5	84.9

Table 7: Model performance when classifying Financial tweets.

5.3 Transfer learning between datasets (RQ4)

To answer the fourth research question, the model fine-tuned on the Financial PhraseBank data was used to classify the Financial tweets dataset to test how transferable its classification performance is. This was compared to the model obtained by directly fine-tuning on the twitter data (Table 7). Transfer learning between the two datasets did not seem to have a substantial effect, with the model fine-tuned on the Financial PhraseBank performing comparably with the pre-trained BERT when trying to classify the Financial tweets data.

5.4 Effect of training set size (RQ5)

Finally, the effect of different training set sizes was tested on each model's cross-entropy loss performance. In Figure 2 it is possible to see how, while initially dropping considerably, the loss actually increases in all three models as the training set increases past 2000 samples. This is most probably caused by the inclusion of the more difficult to classify sentences as the test set increases in size, since the evaluation loss steadily decreases as expected when only the sentences with 100% annotator agreement are considered.

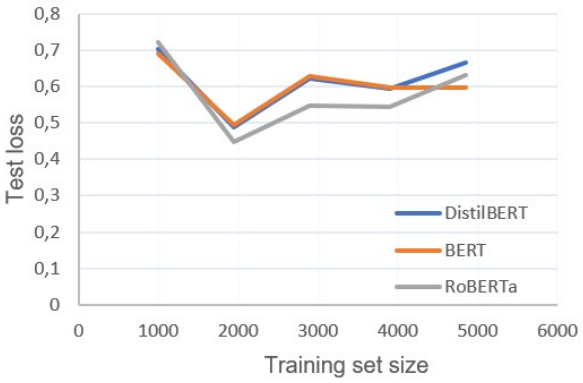

Figure 2: Test loss of the three models with respect to training set size.

6 Conclusion and future work

In this research we explored different approaches to classification by implementing four variations of the BERT language model, combining their pre-trained configuration with an additional classifier as well as fine-tuning them to our down-stream task. The main finding is that using regularization with logistic regression classifier on top of the previous state of the art FinBERT model yields a 1% improvement in both accuracy and F1 scores. A number of additional conclusions can be drawn from our experiments. Firstly, using a pre-trained language model is an effective way to perform text classification in the financial domain even without further fine-tuning, with BERT's performance surpassing the accuracy and F1 scores of the hybrid lexicon and supervised learning approach used in the Financial PhraseBanks' original paper. Secondly, different pooling strategies can yield slightly better results on all three measures when used with a language model's pre-trained weights, with mean pooling of the last layers' tokens achieving a 2% lower loss and slightly better accuracy and F1 scores compared to using the CLS token representation. Future experiments could explore the effect of several other pooling approaches, such as using the output of the second to last layer instead of

the last one or by averaging over the last four layers. Additional research can also still be performed on fine-tuning strategies, by studying the effect of training only a subset of the layers or by using gradual unfreezing. In terms of practical applications an interesting future research could test whether the sentiment that was predicted by the model presented here actually had the expected effect on the financial markets of the following day.

References

[1] Han, Xu, Zhengyan Zhang, Ning Ding, Yuxian Gu, Xiao Liu, Yuqi Huo, Jiezhong Qiu et al. "Pre-trained models: Past, present and future." AI Open, 2021.

[2] Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805, 2018.

[3] Turney, Peter D. "Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews." arXiv preprint cs/0212032, 2002.

[4] Loughran, Tim, and Bill McDonald. "When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks." The Journal of finance 66, no. 1: 35-65, 2011.

[5] Kearney, Colm, and Sha Liu. "Textual sentiment in finance: A survey of methods and models." International Review of Financial Analysis 33: 171-185,2014.

[6] Dougal, Casey, Joseph Engelberg, Diego Garcia, and Christopher A. Parsons. "Journalists and the stock market." The Review of Financial Studies 25, no. 3: 639-679, 2012.

[7] Wilson, Theresa, Janyce Wiebe, and Paul Hoffmann. "Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis." Computational linguistics 35, no. 3: 399-433, 2009.

[8] Malo, Pekka, Ankur Sinha, Pekka Korhonen, Jyrki Wallenius, and Pyry Takala. "Good debt or bad debt: Detecting semantic orientations in economic texts." Journal of the Association for Information Science and Technology 65, no. 4: 782-796, 2014.

[9] Li, Feng. "The information content of forward-looking statements in corporate filings—A naïve Bayesian machine learning approach." Journal of Accounting Research 48, no. 5: 1049-1102, 2010.

[10] Sousa, Matheus Gomes, Kenzo Sakiyama, Lucas de Souza Rodrigues, Pedro Henrique Moraes, Eraldo Rezende Fernandes, and Edson Takashi Matsubara. "BERT for stock market sentiment analysis." In 2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI), pp. 1597-1601. IEEE, 2019.

[11] Zhao, Lingyun, Lin Li, Xinhao Zheng, and Jianwei Zhang. "A BERT based sentiment analysis and key entity detection approach for online financial texts." In 2021 IEEE 24th International Conference on Computer Supported Cooperative Work in Design (CSCWD), pp. 1233-1238. IEEE, 2021.

[12] Araci, Dogu. "Finbert: Financial sentiment analysis with pre-trained language models." arXiv preprint arXiv:1908.10063, 2019.

[13] "Stock market sentiment dataset" https://www.kaggle.com/yash612/stockmarket-sentiment-dataset, 2019

[14] Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. "Attention is all you need." In Advances in neural information processing systems, pp. 5998-6008, 2017.

[15] Wu, Yonghui, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun et al. "Google's neural machine translation system: Bridging the gap between human and machine translation." arXiv preprint arXiv:1609.08144, 2016.

[16] Sanh, Victor, Lysandre Debut, Julien Chaumond, and Thomas Wolf. "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter." arXiv preprint arXiv:1910.01108, 2019.

[17] Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. "Roberta: A robustly optimized bert pretraining approach." arXiv preprint arXiv:1907.11692, 2019.

[18] Yosinski, Jason, Jeff Clune, Yoshua Bengio, and Hod Lipson. "How transferable are features in deep neural networks?." arXiv preprint arXiv:1411.1792, 2014.

[19] Clark, Kevin, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. "What does bert look at? an analysis of bert's attention." arXiv preprint arXiv:1906.04341, 2019.

[20] Mishev, Kostadin, Ana Gjorgjevikj, Irena Vodenska, Lubomir T. Chitkushev, and Dimitar Trajanov. "Evaluation of sentiment analysis in finance: from lexicons to transformers." IEEE Access 8: 131662-131682, 2020.