A study on over-reliance on word-level information for fine-grained emotion analysis and a novel method for improvement

Dinghong Wang
New York University
dw2618@nyu.edu

Jing Tao
New York University
jt3149@nyu.edu

Abstract

Sentiment analysis has long been a key problem in natural language processing. In this paper, we studied GoEmotions dataset which consists of multi-labeled comments in order to understand how BERT interpret the emotions encoded in sentences. In addition, we proposed a solution to modify the training sample text data and further fine-tune the BERT model on these modified text. The result model achieves the same performance while 1 earning more context-level knowledge other than word-level information.

1 Introduction

The study of sentiment analysis begins in 1990's. However, computer-aided sentiment analysis begins mostly after 2004 (Mäntylä et al., 2018). Pretrained models such as BERT have become very trendy since it comes out in 2019 due to its extraordinary power in extracting features automatically from text and benefiting the downstream tasks. Since the sentence embeddings are generated automatically instead of manually and given that for a BERT-base model, the sentence embeddings are vectors of size 768. They are non-interpretable to humans. Previous works (Shi et al., 2016; Adi et al., 2017; Conneau et al., 2018) shows that sentence embeddings can capture syntactic properties like word order and semantic properties like the tense of the main verb or the number of the subject. These findings give rise to further questions on whether these sentence information are taken advantages of when sentence embeddings are used for downstream tasks. In this paper, we will focus on the use of context information of the sentence embeddings in the task of sentiment analysis.

2 Dataset

The dataset used is GoEmotions(Demszky et al., 2020), which consists of over 58k manually annotated sample texts from English Reddit comments, labeled for 27 emotion categories or Neutral. Each sample can have one or more labels therefore making it very hard to predict precisely. In the original paper, the author provides a strong baseline with a BERT-based model which achieves an average F-1 score of .46 across all the proposed taxonomy, leaving much room for improvement.

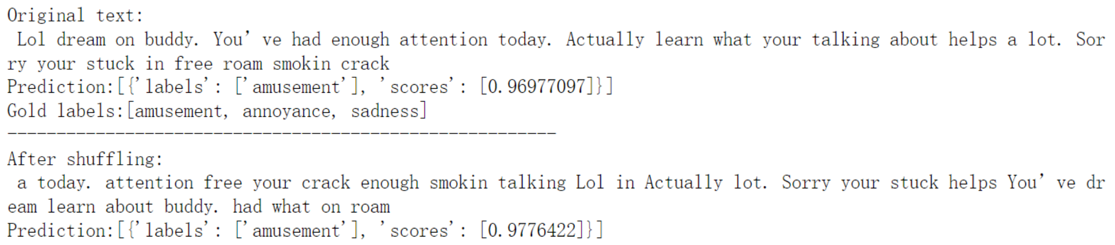
Texts are preprocessed through length filtering, sample balancing and named entity masking. After the preprocessing, texts are given to human raters. The author makes sure that each final label of a sample has at least two human raters agree on. And the labels that do not meet this requirement are left out.

Notably, the 27 emotions are shown to be highly dissociable from each other by applying Principal Preserved Component Analysis. The paper provides a great variety of emotions, which can enormously benefit the downstream tasks that require a more detailed analysis of the emotions.

The author splits the 58k samples into training, development and test set with ratio 80:10:10. And we will be using the same dataset as well.

3 Method and Experiment

In this section, we will introduce experiments designed to show evidence and feasibility that a BERT model uses little context knowledge for sentiment prediction. The purpose of the first experiment is to show that when the context information are gone, the model remains the same level of performance and the most reasonable explanation is that the model uses mostly word-level information. However, it is in doubt whether a model can reach similar level of performance given a complicated dataset like GoEmotions. So in experiment two, we try to extract only word-level features and train a model to show that it is totally feasible.



```
Original text:
 Lol dream on buddy. You’ ve had enough attention today. Actually learn what your talking about helps a lot. Sor
ry your stuck in free roam smokin crack
Prediction:[{’labels’: [’amusement’], ’scores’: [0.96977097]}]
Gold labels:[amusement, annoyance, sadness]
------------------------------------------------------
After shuffling:
 a today. attention free your crack enough smokin talking Lol in Actually lot. Sorry your stuck helps You’ ve dr
eam learn about buddy. had what on roam
Prediction:[{’labels’: [’amusement’], ’scores’: [0.9776422]}]
```

Figure 1: Comparison between predictions on sample before and after shuffling

3.1 Predicting on Shuffled Sentences

We replicate the BERT based emotion prediction model baseline provided in the GoEmotions paper by fine-tuning with the same hyperparameters. One thing we notice that is different from GoEmotions original paper(Demszky et al., 2020) is the training epoch. In the original paper, the author claims that they found that training for 4 epochs is necessary for learning the data and training for more epochs will result in overfitting. Our observation is that training for more epochs makes the model learn more about the context even though it does not improves the performance in terms of loss, accuracy or F-1 score. We have chosen a BERT model trained with 5 epochs and a BERT model trained with 10 epochs for illustration.

We first set the random seed for reproducing and randomly shuffle the order of each word in each sample text. As shown in the figure 11, the shuffled phrase becomes totally non-interpretable for human readers. We can conclude that the context information is gone. We make the BERT models predict on the whole development set and test set with all the texts shuffled. Ideally, the loss of context information should be detrimental to the predicting model. So we will be comparing the prediction accuracy of the same model on the same text before and after shuffling and checking if the prediction results remain unchanged even though the text is shuffled which we name this evaluation metrics as Prediction Similarity. We acknowledge that the result can be very different given different random seeds. So we set 3 random seeds and calculate the mean of Prediction Accuracy and Prediction Similarity.

The result shows that both the BERT(5 epochs) and BERT(10 epochs) do not see significant decrease in prediction accuracy given that the context is gone. But for prediction similarity over two-thirds of the predictions by the BERT(5 epochs) remain unchanged after shuffling the text while half of the predictions by BERT(10 epochs) are unchanged. It indicates that the 10 epochs version relies slightly more on context than the 5 epochs version. The almost non decreasing performance of both versions just implies that models are barely affected by the loss of context information. The result is sufficient to conclude that those models do not rely on the context to predict sentiments.

Dev set	Accuracy before shuffle	Accuracy after shuffle	Prediction Similarity
BERT(5 epochs)	**45.11**	**44.51**	**67.66**
BERT(10 epochs)	41.91	40.71	50.68

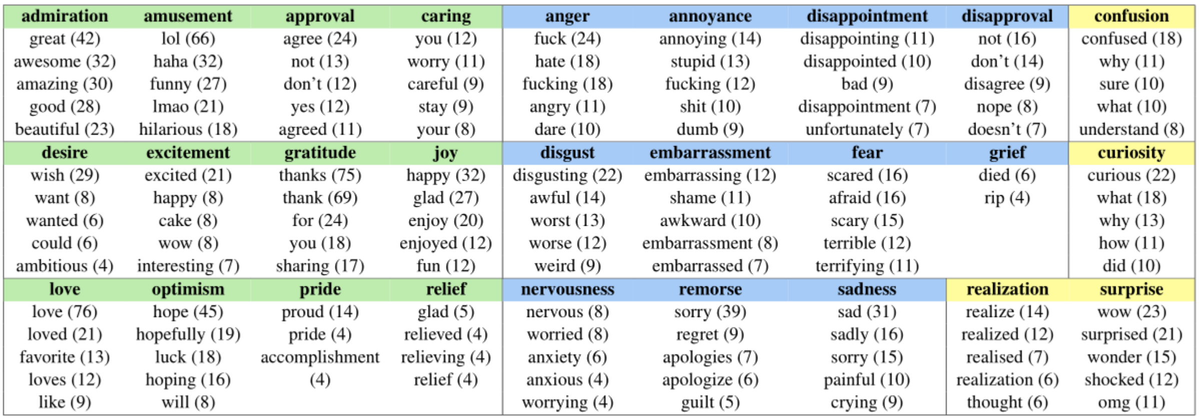
Table 1: Comparison for accuracy before and after shuffling on Dev set (5.4k examples)

Test set	Accuracy before shuffle	Accuracy after shuffle	Prediction Similarity
BERT(5 epochs)	**47.44**	**45.58**	**69.10**
BERT(10 epochs)	41.96	41.24	49.51

Table 2: Comparison for accuracy before and after shuffling on Test set ((5.4k examples))

3.2 Predicting with word vectors and non-neural method

We used term frequency–inverse document frequency(TFIDF) to vectorize the text. The TFIDF value of a word is the product of term frequency and inverse document frequency. The inverse document frequency is a weighting factor of the term frequency designed to raise importance of terms that occur rarely across the documents and decrease weight for terms like ”the” that has high term frequency but occurs everywhere across the documents. For terms, we choose to extract terms by using unigram, bigram and trigram models in order to capture the negations and short-phrases as it is in the text.



admiration	**amusement**	**approval**	**caring**	**anger**	**annoyance**	**disappointment**	**disapproval**	**confusion**
great (42)	lol (66)	agree (24)	you (12)	fuck (24)	annoying (14)	disappointing (11)	not (16)	confused (18)
awesome (32)	haha (32)	not (13)	worry (11)	hate (18)	stupid (13)	disappointed (10)	don't (14)	why (11)
amazing (30)	funny (27)	don't (12)	careful (9)	fucking (18)	fucking (12)	bad (9)	disagree (9)	sure (10)
good (28)	lmao (21)	yes (12)	stay (9)	angry (11)	shit (10)	disappointment (7)	nope (8)	what (10)
beautiful (23)	hilarious (18)	agreed (11)	your (8)	dare (10)	dumb (9)	unfortunately (7)	doesn't (7)	understand (8)
desire	**excitement**	**gratitude**	**joy**	**disgust**	**embarrassment**	**fear**	**grief**	**curiosity**
wish (29)	excited (21)	thanks (75)	happy (32)	disgusting (22)	embarrassing (12)	scared (16)	died (6)	curious (22)
want (8)	happy (8)	thank (69)	glad (27)	awful (14)	shame (11)	afraid (16)	rip (4)	what (18)
wanted (6)	cake (8)	for (24)	enjoy (20)	worst (13)	awkward (10)	scary (15)		why (13)
could (6)	wow (8)	you (18)	enjoyed (12)	worse (12)	embarrassment (8)	terrible (12)		how (11)
ambitious (4)	interesting (7)	sharing (17)	fun (12)	weird (9)	embarrassed (7)	terrifying (11)		did (10)
love	**optimism**	**pride**	**relief**	**nervousness**	**remorse**	**sadness**	**realization**	**surprise**
love (76)	hope (45)	proud (14)	glad (5)	nervous (8)	sorry (39)	sad (31)	realize (14)	wow (23)
loved (21)	hopefully (19)	pride (4)	relieved (4)	worried (8)	regret (9)	sadly (16)	realized (12)	surprised (21)
favorite (13)	luck (18)	accomplishment (4)	relieving (4)	anxiety (6)	apologies (7)	sorry (15)	realised (7)	wonder (15)
loves (12)	hoping (16)		relief (4)	anxious (4)	apologize (6)	painful (10)	realization (6)	shocked (12)
like (9)	will (8)			worrying (4)	guilt (5)	crying (9)	thought (6)	omg (11)

Figure 2: high-frequency words associated with different sentiments

For the model, we choose the Label Powerset model which is one of the few machine learning models that supports outputting multi-labels for each input. The LP method transforms a multi-label problem into a multi-class problem by creating a binary classifier for every combination of labels presented in the training set.

The prediction result of training a label powerset model with TFIDF word vectors inputs on the test set is an accuracy of 44.49% and a F-1 score of 0.48. These numbers exceed the number achieved by the BERT based model. The result shows that only word-level information is needed for a model to achieve the same performance as the BERT model on the same dataset. And it is possible that the BERT model can be using mostly word-level information.

3.3 Solution

To make the model learn from the context, our intuition is to force the model to learn from the non-obvious samples. By obvious sample, we refer to the samples that just starts with ”thank you” and are labeled as gratitude. There are a great amount of these examples and the model can predict the label correctly with great confidence (in terms of probability). However, when it comes to the samples that the gratitude is expressed more implicit, the model fails most of the times. Hence, our hypothesis is that those keywords have dominant weights and we decide to remove them for each label to enforce the model to give some weight to other information. We arbitrarily decide to remove words that occur over 20 times (or the sum of occurrences of derivations of a single word e.g. apologize, apology) under a label. It is not reasonable to remove the word on those very short sentences like ”Thank you sir”. So we only remove the words when the text length is greater than the medium length (15 words). In this way, we believe there will still be sufficient information for sentiment analysis without the keywords as a human reader can read the sentiment correctly most of the times.

4 Conclusion

We removed the keywords as described above on the training set and fine-tuned a BERT model on that. The batch size is 16, learning rate 5e-5 which are just the same hyperparameters used in the GoEmotions paper.

Test set	Accuracy before shuffle	F-1 score	Accuracy after shuffle	Prediction Similarity
BERT	41.96	0.46	**41.24**	**49.51**
Context Learner	**44.48**	**0.48**	28.39	37.29

Table 3: Comparison between BERT and ContextLearner on four metrics

For comparison, we choose the 10 epochs version of BERT model as baseline and named our model ContextLearner in the hope of it will learn to predict from the context information. As the result in the table shows, the performance in terms of accuracy and F-1 score are slightly better than the baseline. And there is a significant drop in accuracy when the texts are shuffled which implies that the loss of context indeed negatively affected the ContextLearner's ability to predict correctly. And the percentage of predictions being the same is further decreased from one half to one third. The numbers are convincing that removing keywords for each label is a feasible way of forcing the model to learn from the context.

Text	Label(s)	Prediction(s)
I'm really sorry about your situation :(Although I love the names Sapphira, Cirilla, and Scarlett!	sadness	love, remorse
Kings fan here, good luck to you guys! Will be an interesting game to watch!	excitement	excitement, optimism
Boomers ruined the world	neutral	annoyance, disappointment
Now I'm wondering if [NAME] drinks, and if he's ever been inebriated during one of his deals.	surprise	curiosity, surprise
I totally thought the same thing! I was like, oh honey nooooo!	neutral	realization

Figure 3: Examples of Mismatch from other related work (Singh et al., 2021)

5 Related work

Gargi and his collaborators obtain the state-of-the-art results on dataset GoEmotionswith 0.54 accuracy and 0.52 F-1 score on the test set(Singh et al., 2021) . They use BERT as the base model and employ a multi-tasking framework attempting to model the semantic meaning of emotion classes through their definitions while training the model for emotion classification. In their paper, they presented examples of mismatch by their model which is shown in figure 3. We observe that the common trait of these mismatched examples is that you can find links between the high frequency keywords we mentioned in figure 2 in text and their corresponding labels as we have highlighted them with the same color. This implies that other BERT-based models even the ones that reach state-of-the-art results fail by some extent to learn from the context in this dataset.

We have shown that the the BERT model does not learn from context situation occurs in this dataset and some future work can be to explore whether the problem remains in other corpora as well.

6 Acknowledgment

We thank the original authors of Goemotions and teaching assistants and professor for reviewing our project.

References

Yossi Adi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg. 2017. Fine-grained analysis of sentence embeddings using auxiliary prediction tasks.

Nurudin Alvarez-Gonzalez, Andreas Kaltenbrunner, and Vicenç Gómez. 2021. Uncovering the limits of text-based emotion detection. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2560–2583, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136, Melbourne, Australia. Association for Computational Linguistics.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. GoEmotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Mika V. Mäntylä, Daniel Graziotin, and Miikka Kuutila. 2018. The evolution of sentiment analysis—a review of research topics, venues, and top cited papers. *Computer Science Review*, 27:16–32.

Xing Shi, Inkit Padhi, and Kevin Knight. 2016. Does string-based neural MT learn source syntax? In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1526–1534, Austin, Texas. Association for Computational Linguistics.

Gargi Singh, Dhanajit Brahma, Piyush Rai, and Ashutosh Modi. 2021. Fine-grained emotion prediction by modeling emotion definitions.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.