Sentiment Classification: GoEmotions

Lucas Lam

Abstract Emotions are complex; people are complex. Humans have the ability to understand people and distinguish between two seemingly similar, yet clearly nuanced emotions such as “fear” or “worry”. Getting machine’s to understand such difference in human emotion is a challenge, and getting data that can train that model too can be challenging. However, researchers at Stanford, Google, and Amazon have developed GoEmotions, a dataset Reddit data manually labeled 28 unique classes emotions. Developing a model that can classify such fine-grained emotions will go a long way to furthering Natural Language Understanding. In this paper, three models were developed to tackle this classification problem. The first two models, Bag of Words (F1 score: 0.183) and Long-Short Term Memory (F1 score: 0.016), were applied on all 28 emotions. A third model, a Convolutional Neural Network (F1 score: 0.210) was applied on a classification task of 3 classes.

Keywords: Sentiment analysis; Deep Learning; Neural Networks; Natural Language

1. Introduction

Understanding emotions has been a long term goal in Natural Language Processing. There are variety of applications for machines to understand emotion, from detecting harmful online behavior all the way to empathetic chatbots.

In the past, existing datasets that contain labeled data useful for emotion classification have mostly been both small and/or limited in emotion taxonomy. Examples include including news headlines data (Strapparava and Mihalcea, 2007), tweets data (CrowdFlower, 2016; Mohammad et al., 2018), and narrative sequences (Liu et al., 2019), to name a few.

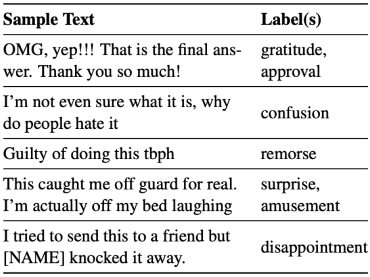
Demszky, Movshovitz-Attias, Ko, Cowen, Nemade, Ravi are researchers at Stanford, Google, and Amazon and have created a manually labeled dataset of 58k Reddit comments, labeled with a comprehensive emotion taxonomy of 28 emotion categories (Demszky, 2020). Though a lot of emotion categories will fall into the larger umbrellas of positive, negative, or neutral emotions, there are nuanced differences between emotions within each broad category that is innate to human understanding. “Annoyance” and “grief” for example, though both are typically associated as negative emotions, are very different feelings and can elicit very different responses from people. Building a model that can accurately classify text into a larger and more fine-grained emotion taxonomy will be huge step forward for Natural language understanding.

1.1 Background

Researchers from the GoEmotions paper (Demszky, 2020) applied two models to this dataset: Fine-Tuned BERT base model and a BiDirectional Long-Short Term Memory with F1 scores of 0.46 and 0.41 respectively. They applied these two models on all 28 classes, as well as a more general 6 class taxonomy presented by Ekman (1992a). More than the model building, however, was the researchers ability to get their hands on a accurately labeled Reddit comments with a large taxonomy of emotions.

1.2 About the Data

One of the main aspects that distinguishes this dataset is the emotion taxonomy. Ekman (1992a) and other researchers in the past have proposed 6 basic emotion categories (joy, anger, fear, sadness, disgust, and surprise), but this dataset expands on these emotion categories into a more fine-grained 27 + Neutral emotional categories. These categories were selected base of the study of distribution of emotion responses to a diverse array of stimuli via computational techniques. They attempt to provide the greatest coverage of emotions expressed in the dataset, coverage in terms of general emotional expression, and limited overlap between emotional categories. Table 1 provides an example of Reddit comments and its corresponding labels.



Sample Text	Label(s)
OMG, yep!!! That is the final answer. Thank you so much!	gratitude, approval
I'm not even sure what it is, why do people hate it	confusion
Guilty of doing this tbph	remorse
This caught me off guard for real. I'm actually off my bed laughing	surprise, amusement
I tried to send this to a friend but [NAME] knocked it away.	disappointment

Table 1: Example annotations from our dataset.

1.3 Selecting Comments

Comments are from 2005 to January 2019, selecting subreddits with at least 10k comments and remove deleted and non-English comments. There is a skew toward offensive language, so the authors have taken measures to curate the data:

- reduce profanity where 10 percent or more of comments include offensive/adult and vulgar tokens

- manually review comments to remove offensive material
- length filtering, comments that are 3-30 tokens long
- sentiment and emotion balancing
- masking, replacing proper names with [NAME]

The resulting dataset is about 54k reddit comments.

2. Methodology

2.1 Data Preparation

The data was randomly split into a test set (80%), test set (10%), and validation set (10%). For comments that were labeled with more than one class, comments were duplicated in the dataset. For example, if the comment "I love NLP" was classified as "love" and "admiration", the comment would be duplicated into two separate rows. Emotions were given an identification number from 0-27, where 27 represented a neutral comment.

Tokenization of the comments, padding, and other pre-processing steps to feed comments into the model was done using NLTK library. Specifically, the "TweetTokenizer" class was used to tokenize comments due to it's ability to recognize more colloquial tokens that appear on social media, tokens such as "lol", "haha", etc.

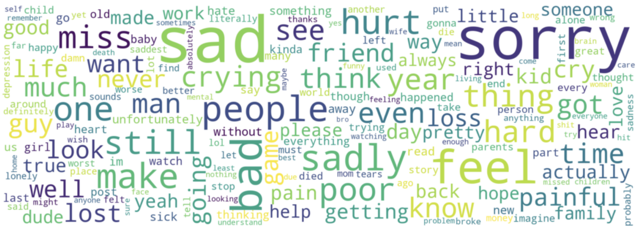
2.2 Word Embeddings

Pre-trained GloVe embeddings (Pennington, 2014) trained on 2 billion Tweets, with 27 billion tokens and 1.2 million words in the vocab were used to train the models. This reduced the number of parameters having to be trained by the model by about 1 million parameters. GloVe embeddings of 100 dimensions were used, and it accounted for about (80%) of the vocabulary seen in the train dataset. For the 20% of data that was unseen in the dataset, embeddings of 0's were given. An analysis of the words missed by GloVe included more colloquial slang, common abbreviations, or emoji's. Dealing with words out of the vocabulary was considered (Kandi, 2018).

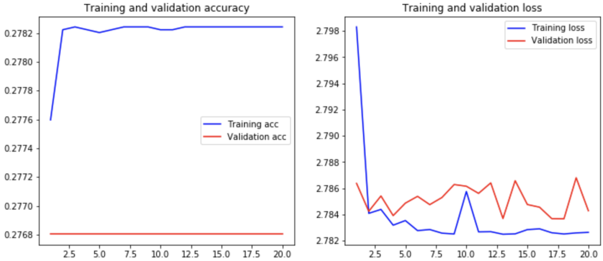
2.3 Models

Three models were applied for classification. The first two models, Bag of Words (F1 score: 0.183) and Long-Short Term Memory (F1 score: 0.016), were applied on all 28 emotions. A third model, a Convolutional Neural Network (F1 score: 0.210) was applied on a classification task of 3 classes, which was the best performing model.

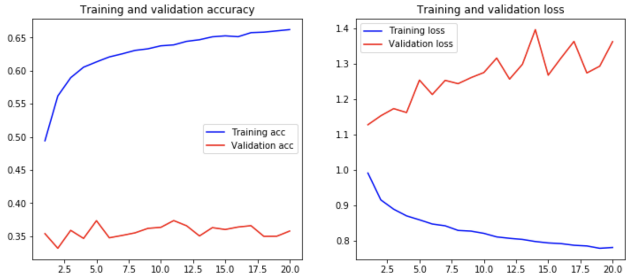
Bag of Words model (BOW) was applied as a baseline model. Simply using occurrence of vocabulary has a low chance of accurate classification, given that all words are orthogonal to each other, hence no context or position awareness. Still, the model was able to achieve an F1 Score of 0.183. From this model, we can use a word cloud to see some of the words that had the highest probability associated with each class. For example, the following is an image of a word cloud associated with the emotion "sadness."



To improve on this model, the Long-Short Term Memory model along with the GloVe embedding attempted to solve the issue of lack of context/-positional awareness that BOW had. The LSTM applied had an state size of 20 and 5 hidden dense layers with a dropout rate of 0.5 after every layer with a softmax output layer. Below are graphs showing training and validation accuracy/loss for the LSTM model



Seeing the results of the LSTM model above, a Convolutional Neural Network (CNN) was applied on 3 more general classes as an attempt to simplify the classification task. Instead of the 28 classes, a more general 3 class sentiment classification was applied with "positive", "negative", and "neutral." Using a CNN to classify emotions in a more broad sense opens the door for further opportunity to feed results into other networks that specifically focus on classifying "positive" emotions and "negative" emotions. The CNN model used had a convolution layer with 40 filters, filter size of 3, and 4 hidden layers with a dropout of 0.3 after each layer with a softmax output layer. GloVe embeddings were also used for his model. Below are graphs showing training and validation accuracy/loss for the CNN model.



3. Results

3.1 Model Performances

BOW had an F1 score of 0.183, and LSTM had an F1 score of 0.016 applied on the classification task of 28 emotions. Given that th original GoEmotions paper (Demszky 2018) fine-tuned a BERT base model and achieved an F1 score of 0.46 for 28 emotions highlights the difficulties of such a classification task.

The F1 score for the LSTM model was pretty low, and from the training/loss graph above its evident that validation accuracy is not at all, despite the moves in validation loss. Training loss and accuracy, however, move as according to expectation.

The CNN model performed a little better, granted that it was only applied on a classification task of 3 classes.

3.2 Limitations

In addition to the limitations of not using state of the art pre-trained deep representations i.e. BERT, there are other factors that could've improved model performance. Espeically with Reddit data, the missing vocabulary that didn't have GloVe embeddings could've served as key words that contribute to the classification task. These missing words can actually contain a lot of context. Whether it be through emojis, colloquial slang that we don't find in other contexts, etc, developing a model that could also robustly represent these missing words would go a long way in improving model performance. Perhaps applying Kadni's LSTM for dealing with out of vocabulary words could help improve model performance.

4. Conclusion

With a fine-grained emotion classification task, applying an LSTM and CNN with GloVe embeddings were able to produce some results. However, a lot of work can still be done to architect a model that can better handle the complexities of such a classification task.

References

Paul Ekman. 1992a. Are there basic emotions? Psychological Review, 99(3):550–553.

Carlo Strapparava and Rada Mihalcea. 2007. SemEval2007 task 14: Affective text. In Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007), pages 70–74, Prague, Czech Republic. Association for Computational Linguistics.

Chen Liu, Muhammad Osama, and Anderson De Andrade. 2019. Dens: A dataset for multi-class emotion analysis. arXiv preprint arXiv:1910.11769.

Demszky, Dorottya et al. “GoEmotions: A Dataset of Fine-Grained Emotions.” ArXiv abs/2005.00547 (2020): n. pag.

Devlin, Jacob et al. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.” NAACL-HLT (2019).