CS447 Literature Review: Sentiment Analysis

Shyam Shah
shyam3@illinois.edu

December 6, 2021

Abstract

This paper will discuss the current issues faced by modern day sentiment analysis techniques.

1 Introduction

Sentiment analysis is a type of natural language processing in which the goal is to identify the attitude or general feelings that the text portrays. An example of a practical use for sentiment analysis is analyzing comments and reviews of a product to determine whether users generally like it or not. However, natural language processing in general is a difficult task to get right all the time. So the purpose of this paper will be to talk about some of the sentiment analysis techniques used today, some of the things that they struggle with, and some things being done to improve them. This will be done by summarizing and discussing four different articles.

2 Background

This section is to give some more information on some of the concepts and terminology used in this paper.

The first is the long short-term memory model (**LSTM**). A neural network is a statistical model that is commonly used for regression or classification purposes. It takes some input and generates some output which can be used for many machine learning tasks. However sometimes we may want to use the context of previous outputs of a neural network to help make generate the next outputs. An LSTM is a type of neural network that does this, and is able to maintain context over large gaps which may be expensive/impossible to do with a traditional recurrent neural network. A bidirectional LSTM helps provide additional context to the model by feeding the data through the LSTM twice, once in each direction, so that context from before and after a given word can be used.

Next there is the bidirectional encoder representation from transformers (**BERT**) [1] model. This model leverages the same benefits of a bidirectional LSTM except it uses transformers. Transformers are different from a traditional neural network because they have an attention mechanism which allows them to not have to process an input sentence from start to finish, allowing for parallelization and therefore reduced training time.

[1] https://towardsdatascience.com/bert-explained-state-of-the-art-language-model-for-nlp-f8b21a9b6270

ELMo [2] is a bi-attentive classification network that uses both word and character level embeddings. Word representations are first passed through a feed forward layer and then through a sequence to sequence network with biattention. Then the output and the original representations are passed through another sequence to sequence network.

A **bag-of-words** classifier uses a representation of sentences based on the occurrence (or lack of occurrence) of words in the entire dataset's vocabulary. These are then used to train any kind of statistical classifier, such as a logistic regression model.

Now some of the metrics used to measure performance of models will be discussed. First is the **F-score.** The F-score is the harmonic mean of the precision and the recall, where precision is the proportion of positive identifications that were actually correct, and recall is the proportion of actual positives that were identified correctly.

Next is the Matthews Correlation Coefficient (**MCC**) [3]. This is a metric useful for when you have imbalanced data as it requires obtaining good results proportionally to the number of positive and negative elements in all of the confusion matrix categories to be high. The confusion matrix categories are true positives, false positives, true negatives, and false negatives.

3 Paper 1: A Case Study on Social Media

3.1 Motivation

The first article I will be talking about is called 'Multilingual Connotation Frames: A Case Study on Social Media for Targeted Sentiment Analysis and Forecast' Rashkin et al. (2017). This article was chosen to first get an idea of some modern techniques for sentiment analysis through a case study and how it can be used to forecast sentiment while also seeing some of it's error analysis. This specific article is a case study on social media.

3.2 Summary

The goal of this paper was to demonstrate a large scale analysis on public sentiment using the social media platform Twitter and a framework called multilingual connotation frames. The paper first defines connotation frames as a framework to help encode relations implied by a predicate towards it's arguments. Arguments would include the reader and the writer as well as the agent and theme of the phrase. So for example, if you had the sentence 'teenager survives Boston Marathon bombing', then the agent would be the sympathetic victim, and the theme would be the bombing. The implied relations from the writer to the agent would be positive (sympathetic) whereas the implied relations from the writer to the theme would be negative (hardships). Similarly there are implied relations between the reader and the agent and theme. Multilingual connotation frames are an extension of this framework to also include other languages, which in the case of this paper is 10 European languages including Polish, Finnish, and Russian.

The paper then briefly states two benefits of multilingual connotation frames which are enabling targeted sentiment analysis, and allowing the ability to study more sentiments including nuanced

[2] https://www.analyticsvidhya.com/blog/2019/03/learn-to-use-elmo-to-extract-features-from-text/
[3] https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-019-6413-7

ones. The definition of targeted sentiment is given as being able to label how a specific source entity feels about a specific target entity. The case study was used to demonstrate these two new utilities.

The data used was gathered over a two week period surrounding the Brussels attack of March 2016 across various languages. The process of getting the multilingual connotation frame consists of first getting the most likely translation for a word from it's source language to English, and then using the connotation frame for the English word. Then these connotation frames were used to get the sentiment expressed towards named entities (such as 'Obama' or 'Google') by countries by aggregating data based on where the user is located. The writers then used this data to forecast changes in sentiment via long short-term memory models (LSTMs 2) and used this to observe trends over time.

Error analysis was done by comparing the predictions made by these LSTMs versus LSTMs trained with just English data. The findings were that the predicted probabilities followed the same trends as the true probabilities, but was not able to predict any of the true sudden spikes.

3.3 Discussion

While this paper should not be seen as the only modern way to do sentiment analysis, it still demonstrated a problem that it can face, which is the inability to predict sudden changes in sentiment based on real world events. The data used for the analysis was the week before and after the Brussels attacks (as well as the day of the attacks) for a total of 1.2 million pieces of data. It is likely that even if they increased this number exponentially, their models would not have had done any better in predicting the sudden sentiment changes, because they would not have the ability to predict events happening in the real world.

Similarly, if a company is trying to predict how users would react to their product, it would be very difficult to predict changes in sentiment if, for example, a scandal involving the company's CEO was revealed. This is not an issue specific to sentiment analysis though, as humans would also be unable to predict a real world event happening, so it would be unrealistic to expect sentiment analysis to get to a point where it can handle these situations.

4 Paper 2: Assessing and Probing Sentiment Classification

4.1 Motivation

This paper is called 'Assessing and Probing Sentiment Classification' Barnes et al. (2019). The first paper talked about how to forecast sentiment, but this paper is focused more on the actual sentiment classification part and will provide concrete examples of the issues faced.

4.2 Summary

This paper looks at sentences that were mis-classified by many modern sentiment classifiers and then uses this dataset of sentences and the linguistic phenomena that makes them hard to classify to analyze the performance of of the classification methods.

The experiment done in this paper used six different English datasets with existing sentiment annotations. These datasets were taken from various domains such as news, blogs, reviews, and

social media. The models used include BERT, ELMo, a bidirectional LSTM and a bag-of-words classifier. 2

A total of 836 sentences were classified incorrectly by all four models. The phenomena responsible for the most mis-classified sentences will be discussed briefly below.

The first is mixed polarity. This is when both positive and negative sentiment are expressed in the sentence. Nearly one third of these sentences contained a 'but' clause making them hard to annotate for humans and for classifiers. Next there is non-standard spelling and hashtags. The paper gives the example '#imtiredof this SNOW and COLD weather!!!' which is easy to classify for humans, but the use of the # symbol as well as 'I'm tired of' being combined into one word makes it hard for a classifier to deal with. Lastly there are idioms, which rarely appear multiple times in the training corpus, which is what makes them hard to classify. Some of the other less common phenomena include negation, world knowledge, irony and emojis.

The writers then experimented by re-training the BERT model, which was the best performing one overall, but using phrases instead of sentences. The motivation behind this was to give that phrases would give the model access to more compositional information which was identified as one way to address some of the phenomena discussed above. The overall accuracy improved slightly from 53% to 55.1%. There was also a noticeable improvement in the negation and world knowledge sentences, but accuracy got worse in the irony sentences. However irony is not one of the phenomena that can be improved on with compositionality so this did support the theory that phrase level annotations help improve the ability to classify compositional sentiment.

4.3 Discussion

This article had a really comprehensive list of exactly what modern techniques used for sentiment analysis struggle with. The most prominent phenomenon was the mixed polarity one which is another example of something that even humans will struggle with. For example, if you're given the sentence 'I am upset, but I am also happy', it would be difficult do determine this as positive or negative sentiment, even if we had additional context about it. The paper touched on the difficulty to get an accurate annotation for some of these sentences, but it would be interesting to determine whether there actually is a 'correct' annotation for these types of sentences if they are this ambiguous.

It was also interesting to see the amount of variation in model accuracy depending on the dataset used. For example, the BERT model had accuracy as low as 53% and as high as 84.2%. This brings up the idea of whether we could use prior knowledge on the topic of the training corpus to help further increase accuracy. One downside of this approach would be that finding the right representation of the prior knowledge without introducing any bias could be an entirely new problem that would be difficult to solve.

5 Paper 3: Financial Sentiment Analysis: An Investigation into Common Mistakes and Silver Bullets

5.1 Motivation

This paper is called "Financial Sentiment Analysis: An Investigation into Common Mistakes and Silver Bullets" Xing et al. (2020). It is similar to the previous one, except that it focuses on financial

sentiment analysis. The reason behind picking this article is to now see whether reducing the scope of the problem (to just finance in this case) has any affect on the issues faced and to be able to compare the two.

5.2 Summary

This paper focuses on financial sentiment analysis, which is when you classify a piece of text to figure out whether there will be a bull or bear market. One thing that makes financial sentiment analysis harder than regular sentiment analysis is the lack of available training data. With financial sentiment analysis, an issue with a model could cost users money, and so understanding why accuracy is low is just as important as improving it.

This paper used lexicon, machine learning, and deep learning based approaches. Lexicon based approaches use words that are already annotated to determine the sentiment of a sentence, whereas the machine and deep learning based approaches are similar to the ones discussed in the previous paper 4. Two datasets were used, one to cover the business domain, and one to cover the finance domain so that they can be compared to see if the finance domain led to worse results.

The MCC, F-score 2, and accuracy were on average lower for the finance domain. It was also noticed that each class of model (lexicon, machine learning, and deep learning) seemed to have the same type of errors, indicating that errors were model specific. All three metrics used did not rank the models within the classes the same either. One of the lexicon based approaches was using a lexicon specifically geared towards finance, yet it ranked the worst among them so even a domain specific lexicon did not have any noticeable benefits.

The errors were then analyzed and were broken into six groups of linguistic phenomena like the previous paper. The first is irrealis mood which is when a certain situation or action is not known to have happened at the moment the speaker is talking. Second is rhetoric which includes things like personification and sarcasm. Third is dependent opinions the sentiment represents that of a third person, and not the writer. Fourth is unspecified aspects where there may be negative sentiment in a sentence, but it is directed towards an aspect that may not be the topic of the sentence and vice versa. Fifth is unrecognized words, and the last one is external references.

5.3 Discussion

Many of the linguistic phenomena discussed here came up in the previous paper as well. For example unrecognized words and external references have came up in both papers. These two are worth mentioning because they are both relatively easy for humans to classify. They both also are not things that one would traditionally think of when they think of sentiment analysis so there may be a way to split them apart and treat them as their own problems. So for example, you could first 'translate' text with unrecognized words like Twitter hashtags into equivalent text and then do sentiment analysis on this text.

6 Paper 4: A Human-in-the-loop Error Detection Framework for Sentiment Analysis

6.1 Motivation

This paper is called "A Human-in-the-loop Error Detection Framework for Sentiment Analysis" Liu et al. (2021). Now that we have looked at two papers and their common pitfalls with sentiment analysis, we will wrap up by looking at a paper that discusses an approach for error detection to see if it can address these pitfalls.

6.2 Summary

Many users of sentiment analysis models use them as a black box, so when they do run into errors, it can reduce their confidence in the system. Currently human error analysis is used to correctly label these errors, but this is time and effort consuming. So researchers are looking for more proactive error detection that does not require as much human intervention. The framework proposed by this paper is to first analyze local feature contributions for a black box model, aggregate them for global-level feature contributions, have humans evaluate the top ranked features for errors and then integrate this to send users alerts for errors. This will be discussed in more detail below.

The local feature contributions and their corresponding explanations are used to why the model made a specific decision. Unigrams are used as they are small units of text that humans can easily interpret, and each unigram in a phrase is given a contribution score to rank which words are most responsible for a given decision.

The global feature contributions are an aggregation of the local contributions. A global aggregation is taken by getting the sentiment of a phrase without each word using the black box model and ranking the words change the prediction the most.

The human intervention step is where humans take the top N globally contributing features and their sentiment labels and flag any that seem erroneous. So in this case the human is given a list of unigrams. This is where choosing unigrams has some benefits, as it is easier to evaluate unigrams versus bigrams/trigrams and it is more likely to find an erroneous unigram than it is to find an erroneous bigram/trigram.

The last step would be to use the results to flag bad predictions. The paper discusses the use of a metric called the erroneous score to avoid wrongly penalizing unigrams. This score is a normalized version of the accumulated error contributions induced by the problematic feature. A threshold value is picked to determine when to accept a unigram as erroneous.

This framework was able to identify errors with precision ranging from 67.7% to 90.8%, varying based on the value chosen as the threshold. But as the threshold is increase, the number of errors found decreased so there is a trade off.

6.3 Discussion

The framework proposed here appears to be a good way to start addressing some of the phenomena discussed in previous papers. It would still struggle with unrecognized words as they are unlikely

to show up in the top N features analyzed by humans, but it seems like it could help with external references by helping associate an appropriate sentiment with an object. One issue here is that it becomes subjective based on the human doing the analysis. For example, if an artist came up for a positive sentiment, some people may flag that as erroneous while others may not.

Also the fact that unigrams were used for this approach could make it hard to identify phrases that may show signs of both positive or negative sentiment, or ironic/sarcastic phrases. Using bi-grams/trigrams may help make this more effective but there would be a trade off in the effort required.

7 Conclusion

The purpose of this paper was to take a look at modern sentiment analysis, what it is not capable of doing currently, and what is being done to improve. This was done by first looking at a case study to introduce sentiment analysis, then looking at various linguistic phenomena that are pitfalls for sentiment analysis, even in different domains, and finally a technique that could help address these phenomena. Some of the phenomena discussed that could be improved on are non-standard spelling/unrecognized words, real world references, idioms, and rhetoric. It seems that human interference will be required to help address these, either by annotating more data or by looking for errors with existing models. It is unrealistic to expect models to correctly classify sentences and phrases that humans can not classify either but researchers have noticed and have started to look into solutions for the rest of the phenomena as evident by the papers discussed.

References

Jeremy Barnes, Lilja Øvrelid, and Erik Velldal. 2019. Sentiment analysis is not solved! assessing and probing sentiment classification. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 12–23, Florence, Italy. Association for Computational Linguistics.

Zhe Liu, Yufan Guo, and Jalal Mahmud. 2021. When and why a model fails? a human-in-the-loop error detection framework for sentiment analysis. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Papers*, pages 170–177, Online. Association for Computational Linguistics.

Hannah Rashkin, Eric Bell, Yejin Choi, and Svitlana Volkova. 2017. Multilingual connotation frames: A case study on social media for targeted sentiment analysis and forecast. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 459–464, Vancouver, Canada. Association for Computational Linguistics.

Frank Xing, Lorenzo Malandri, Yue Zhang, and Erik Cambria. 2020. Financial sentiment analysis: An investigation into common mistakes and silver bullets. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 978–987, Barcelona, Spain (Online). International Committee on Computational Linguistics.