A Comparison of Four Pretrained Sentiment Analyzer Toolkits and A Simple Untrained Model

Computer Science Dept. Wilson McGill U00723224, Stilyan Dimitrov U00762724

Intro (why we should care):

Our project is a comparison of various pretrained sentiment analysis toolkits that are free and available to the public, along with the creation of a sentiment analysis tool compared to those as well. The goal was to give those looking for sentiment analysis tools guidelines for how different toolkits handle certain datasets and what they could expect for outputs and results using said tools. We also wanted to compare this to creating your own sentiment analyzer, to see if this was an efficient process for those just needing results. Our focus was largely on those working in the business world who would benefit from a sentiment analyzer (on their own products or elsewhere) and who would like to avoid the prepackaged and paid analyzers focused on this demographic.

Data & preprocessing:

Data Sets Used For Comparison:

Negative Reviews of Hydro Flasks – Filtered by 1 star only. These are for the most part all extremely negative reviews, they are all in English and scraped from amazon. Keep in mind, many 1 star reviews may contain positive or neutral things about a product, however the downsides of the product were a dealbreaker and resulted in the one star review. We should expect these reviews to be slightly higher in positive emotions in proportion to the positive reviews.

Positive Reviews of Hydro Flasks – Filtered by 5 star only. These are extremely positive reviews that have nearly nothing negative to say, they are all in English and scraped from Amazon.

Neutral Texts – A combination of Physics and Mathematics research papers. These were taken from various public domain scientific paper databases and were read over to ensure no extremely high emotions. Keep in mind however, these may contain positive sentiments due to experiment results or scientific discoveries so we may expect slightly positive but overall neutral sentiments.

Trick Review – A tricky Hydro Flask review that is positive but contains negative words. A review that describes why a product is good by stating how they dislike products dissimilar from this one. For instance: I hate carrying heavy large water bottles, so this small and compact one works great for me!

Models:

All sentiment analysis was done with python:

1. TextBlob and NLTK model

Popular and easy to set up pretrained model. Very simple but robust and works very easily with no fluff. However, it lacked other features that may be useful. Took any given text and

outputted a -1 to 1 value based on the sentiment of said text.

1. BERT model

Used the Transformers Library. Less popular but easy to set up and use pretrained model with many quality of life features. However, data needed to be truncated so may need formatting. Took any given text, (certain datasets required truncation) and output a 0-4 stars rating based on the texts sentiment.

2. VADER

Vader is an easy-to-set-up model used for testing sentiment analysis that is sensitive to both the polarity of positive/negative as well as the intensity of emotions. Vader is trained to understand that words like ‘love’, ‘enjoy’ etc. convey a positive sentiment and is also praised for understanding the basic context of words such as ‘did not love’ as a negative sentiment. Its also said to be intelligent enough to take the emphasis of capitalization and punctuation into account such as ‘LOVE IT!!!’.

3. ROBERTA

RoBERTa, short for “Robustly Optimized BERT Approach”, is a variant of BERT. Much like its predecessor, roBERTa is a transformer-based language model that is easy to set up. One key difference between roBERTa and BERT is that roBERTa was trained on a much larger dataset. Unlike BERT, RoBERTa uses a dynamic masking approach meaning, it varies in the words it masks. This means when a sentence is revisited during training, different words may be masked each time. This approach contributes to RoBERTa’s ability to generate more accurate responses across diverse texts.

4. Non Pretrained Model

This model was a very simple example of a sentiment analysis tool built from scratch. It started by formatting the datasets and parsing it into arrays that could be used to analyze the words. It then omitted “STOP” words, or words that did not contain any emotional significance to the sentiment. From there it used an “Emotions” file to find words in the text and what emotion that word signified. After this it would count the number of times a word with a particular emotion appeared and create a chart that plotted this. Allowing the user to see the dominant sentiments within a given text.

Experiment settings:

- The experiment settings for our tests were fairly simple. We set up each of our models in either pycharm or google collab and ran each model on all 4 datasets, recording the result for each one and noting how it performed. In the case where the results may not satisfy expectations we double checked and reran the data and noted any changes (but this never resulted in new results). All of our code and datasets is stored on a github repository which is linked below.
- https://github.com/wrmcgill/NLP-Sentiment-Analyzer-Comparison-Proj

Results:

Results for TextBlob Model:

- First Is the raw output of the Textblob model for each dataset. This is given in a number ranging from -1 to 1.
    - Positive Reviews: 0.22813
    - Negative Reviews: -0.01851
    - Neutral Texts: 0.07334
    - Tricky Review: 0.14761
- We can see from these results that overall the TextBlob model got the sentiment correct for our testing but with some other surprising developments. Overall Textblob was very conservative with its sentiment ratings across the board, none of which

exceeded .3 in positive or negative scores. We also see a very neutral negative review score which largely indicates it put heavy emphasis on the positive remarks made in some of the negative reviews. We can also see that the tricky review did result in a slightly lower sentiment when compared to the positive reviews, indicating that the negative words did play a part in the sentiment calculations.

Results from the BERT Model:

- First is the raw output of the Model for each dataset. This is a number ranging from 0 to 4 with 0 being the most negative sentiment and 4 being positive.
    - Positive Reviews: 4
    - Negative Reviews: 1
    - Neutral Texts: 3
    - Tricky Review: 4
- With these results it is clear the BERT models did very well on these datasets. The positive reviews got a full 4 on the positive sentiment score which is to be expected given the dataset. The negative reviews received a 1 which also falls within expectations, due to the nature of negative reviews possibly containing positive information about the product it could be expected that this would receive a 1 rather than a 0 score. Neutral received a 3 which is near the middle but slightly positive again to be expected given the dataset. This Model also passed the tricky review perfectly determining the review was in fact fully positive even though it contained negative words.

Results from VADER Model:

- First is the raw output of the Model for each dataset. The Vader output score consists of four scores that represent the sentiment for “neg” - negative, “nue” neutral, “pos” - positive, and finally compound which is the overall sentiment of the text. All normalized to range between -1 and +1 representing the extremes of negative and positive
    - The negative reviews set gives: {'neg': 0.102, 'neu': 0.814, 'pos': 0.084, 'compound': -0.9953}
    - The neutral reviews set gives: {'neg': 0.031, 'neu': 0.917, 'pos': 0.053, 'compound': 0.9991}
    - The positive reviews set gives: {'neg': 0.026, 'neu': 0.713, 'pos': 0.261, 'compound': 1.0}
    - The tricky reviews set gives: {'neg': 0.183, 'neu': 0.575, 'pos': 0.243, 'compound': 0.3191}
- From the result for the negative dataset, we can infer that while the model gives a low probability of the text being negative (.102), the compound and final score show a probability of (-0.99) suggesting a high probability of negative sentiment. The Neutral dataset on the other hand gives a stand-alone probability of .91 neutral, but a compound of 0.99 which is unexpected for neutral reviews. The positive gies a stand-alone probability of 0.26 for positive but a perfect positive compound score, indicating a very positive sentiment. Finally, the tricky review gives a compound of 0.31 indicating that its a slightly positive sentiment.

Results from ROBERTA Model:

- First is the raw output of the Model for each dataset. This is a number ranging from 0 to 1 with this range representing the percentage likelihood of the date being negative, neutral, or positive (in this order).
    - Positive Sentiments: [[0.002760493429377675,

0.026426563039422035, 0.9708129167556763]]

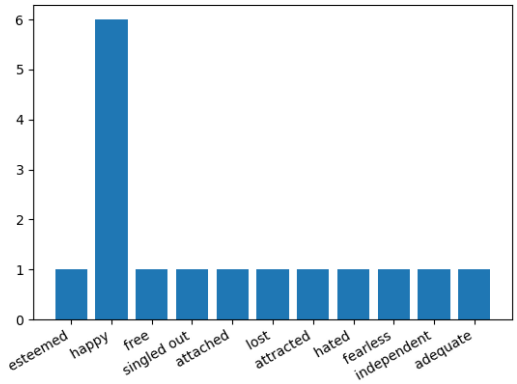
- Negative Sentiments: [[0.8989202976226807, 0.08838112652301788, 0.012698599137365818]]
- Neutral Sentiments: [[0.08287899941205978, 0.7881227731704712, 0.12899823486804962]]
- Tricky Sentiments: [[0.022000707685947418, 0.07985534518957138, 0.8981439471244812]]

- From the results, we can see the probability distribution for positive sentiment is Neg:0.276% Neut:2.64%, and Pos:97.08% which shows high confidence of an overall positive sentiment. The distribution for the negative dataset however has only a negative probability of 89.89% with 8.84% in neutral showing that while it is still confident the output is negative it has misread some of the inputs. The neutral sentiment distribution suggests that the model believes the input text is most likely neutral with a confidence of 78.812%. Finally, for the tricky text, the model predicts a high positive sentiment of 89.8%, suggesting that it's a positively inclined sentence.

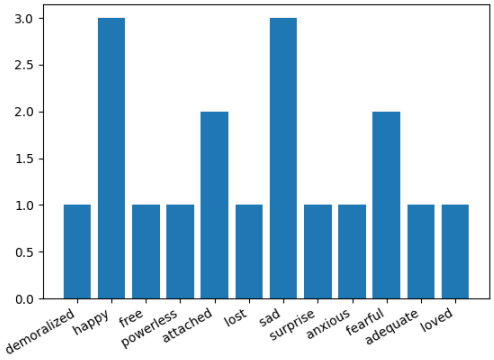
Results from Untrained Model:

- First is the raw output of the model for each dataset. This is a table displaying all the emotions detected and the frequency at which they were detected.

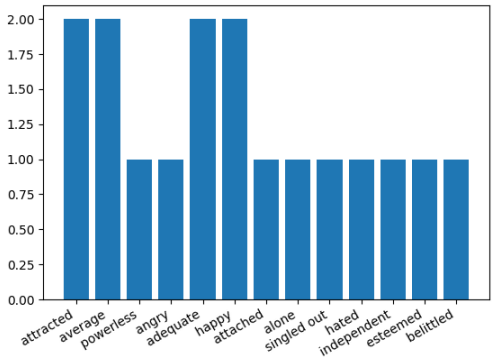
- Positive Reviews



- Negative Reviews:



- Tricky Review:



- Tricky Review:

- With these results we can see mixed accuracy based on each dataset. The positive review did very well, with a variety of emotions present but the highest frequency being happy, which is to be expected. The negative reviews did not have as great of results, with a high frequency of happy and sad it would be unclear what the sentiment was without already knowing. The neutral texts did well however, a variety of emotions with similar frequencies would indicate a neutral sentiment overall. And the tricky review results were interesting, it did get the sentiment of happy correct, but due to the small dataset it only found the one emotion once within the text. This makes this review largely luck based.

Takeaways:

- TextBlob: Leans very conservative with our datasets but is accurate. We found it was reluctant to give strong sentiment in any category, though it did almost pass the tricky review.

- BERT: Very accurate and consistent given the simple output format. Easy to understand output and passed tricky test with ease.

- VADER: Detailed output give good data to analyze. Seemed to default to positive when large neutral sentiment is present though. Did not do well on tricky test.

- ROBERTA: Gives mixed output with no distinct compound score. Still very accurate and passed tricky test with only slightly lower positive value.

- Non Pretrained Model: Consistent good scoring but becomes inconsistent for negative reviews, and largely favors positive sentiments. Smaller datasets may also lead to luck based results.

Discussion and Takeaways:

The testing performed on each of these pretrained models and the untrained one resulted in varying results highlighting the pros and cons for each method of sentiment analysis. After reviewing the results and understanding why some models performed the way they did we can make some conclusions.

- First, for a business setting we found BERT to be the most effective tool for gathering sentiment. It did well in all categories, output a data type that is easy to read for those unfamiliar with sentiment analysis and did not flinch at the tricky review test.
- Next, we concluded that Textblob and Roberta could be effective as well with some caveats, textblob is very conservative in its estimate so keep this in mind. And Roberta does not give a compound score, making it difficult to read the sentiment for those unfamiliar with it.
- Finally we found Vader and the untrained model not as suitable for the target demographic. Vader due to its

mixed results and poor output readability, and the untrained model due to its difficulty of setting up and lack of features available to the pretrained models.

Overall this project provided valuable information to those looking to use sentiment analysis within their own projects or businesses and we hope this proves useful to them.

Contributions:Stilyan Dimitrov

- Wilson McGill U00723224:
    - Set up the BERT model, TextBlob model and ran them against each of the datasets.
    - Set up the untrained model and ran on each of the datasets.
    - Created Git repository and set up for code and datasets to be inputting into it.
    - Scraped amazon for review datasets, found neutral texts for neutral dataset and wrote tricky review for testing.
- Stilyan Dimitrov U00762724
    - Set up Vader and RoBERTa models
    - Added them to the GitHub repository
    - Ran them on all four of the datasets.

References:

Kian Kenyon-Dean, Eisha Ahmed, Scott Fujimoto, Jeremy Georges-Filteau, Christopher Glasz, Barleen Kaur, Auguste Lalande, Shruti Bhanderi, Robert Belfer, Nirmal Kanagasabai, Roman Sarrazingendron, Rohit Verma, and Derek Ruths. 2018. Sentiment Analysis: It's Complicated!. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1886–1895, New Orleans, Louisiana. Association for Computational Linguistics.

Gabriel Roccabruna, Steve Azzolin, and Giuseppe Riccardi. 2022. Multi-source Multi-domain Sentiment Analysis with BERT-based Models. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 581–589, Marseille, France. European Language Resources Association.

Shuai Fan, Chen Lin, Haonan Li, Zhenghao Lin, Jinsong Su, Hang Zhang, Yeyun Gong, JIan Guo, and Nan Duan. 2022. Sentiment-Aware Word and Sentence Level Pre-training for Sentiment Analysis. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4984–4994, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.