Literature Review of VADER and Test Data Results

Emma Nash and Kinzlee Changfoot

Department of Linguistics, Simon Fraser University

LING 250 / SDA 250: Computational Text Analysis

Professor Yifang Yuan

November 14, 2024

Introduction

Sentiment Analysis is an NLP technique in which it assigns texts a sentiment score by determining if they are positive, negative, or neutral in nature. Another name for this technique is Opinion Mining and has a wide range of uses such as in Market Intelligence to see how consumers are reacting with products, to research on political figures and what their stances are on current issues. A tool that is commonly used and very helpful for sentiment analysis is VADER (Valence Aware Dictionary for sEntiment Reasoning), which is a lexicon and rule-based sentiment analyzer which determines the sentiment score of the text by assigning each word a positive, negative, or neutral value. VADER takes data to perform the sentiment analysis, so the data that was tested comes from the Canadian House of Commons debate on Housing. Our literature review will go over pre-existing data to better understand how VADER works and how it processes our test data to produce sentiment scores.

How Vader Works

VADER has several key steps that went into developing its capabilities of providing accurate sentiment analysis. The first step takes pre-existing lexicons, adding sentiment assigned words, and finally, combining them to create one lexicon. Another step within this process to create the lexicon was by taking commonly used words and phrases in social media and assigning them a sentiment score. The range in scores vary between -4 and 4, being the most negative to the most positive respectively. The second step adds a human element into understanding the sentiment scores. In other words, humans would evaluate the added words and phrases into the lexicon, reaffirming and adjusting the sentiment scores that were assigned. The third step implements heuristic rules, which allows VADER to understand and properly analyze common occurrences in social media used to convey emotion. Elements such as emoticons, capitalizations, punctuation, negations, conjunctions, and degree modifiers are all common in informal messages seen within social media platforms. The final step is developing VADER's ability to calculate and return a compound score between -1 and +1 to represent the overall positive, negative, or neutrality of a given text (Youvan, p. 7)

VADER was specifically designed to combat the struggles of sentiment analysis on social media, though its capabilities and use also extend to other areas such as analyzing reviews and researching political figures' stances on hotly debated topics. One issue that may still arise with social media is the use of irony and sarcasm as those are tonal issues that may need more context to be able to give a proper sentiment score.

Test Data Evaluation

The text used for this assignment were transcripts of statements said by Justin Trudeau speaking in the House of Commons. The statements were his contributions to the discussion of housing with other members of parliament. We first cleaned up the data by having our text data undergo tokenization, lemmatization, and removing stop words, then we were able to put our text data through sentiment analysis with the use of VADER. Our sentiment score for tokenized words was a 0 which would mean that the overall sentiment of the text is very neutral (Miller). Though this score does not accurately reflect the sentiment score because it has not been fully cleaned up to use. The second step was to test the sentiment score on stop words which gave an overall sentiment score of 0.128 which leans to a more positive sentiment but not by much. The third step was to test the sentiment score on the lemmatized text which yielded a score of 0.1275. This is slightly more negative than our results using stop words. The last step was performing the sentiment analysis on the cleaned-up data which gave a sentiment score of -0.0521, this is a huge leap from the previous score. However, this score is reasonable as with the previous steps, the sentiment score tends to trend downwards.

An aspect to our assignment that may prevent accurate sentiment scores is testing VADER on the cleaned-up version of the data. As previously mentioned, VADER takes all linguistic characteristics into account when calculating the sentiment scores. Therefore, all the steps to clean up the data may be impacting it. However, since the data is speech to text as it comes from discussions within the House of Commons, cleaning the data may not make a difference anyways. VADER is trained to perform specifically for social media data, but its performance also seems to be adequate when used on data on political debates.

Our Test Data

For our test data we used several statements on the topic of housing in Canada that was discussed by Prime Minister Justin Trudeau in the House of Commons on November 6th, 2024. The House of Commons is where all the Members of Parliament come to explain the decisions that they have made to all of the other members. With that being said, since Justin Trudeau is the Prime Minister as well as the data being used is him speaking in the House of Commons, his explanations and questions to other members of the parliament his tone seems to be very neutral without using a sentiment analyzer. Because we chose to use VADER there was no use for creating a training set of data for our program to learn, but our test data consist of fourteen different text entries made by Justin Trudeau. As stated in the previous section, the sentiment score for the test data is at -0.0521 which means that while it does lean more negative, it is very neutral which does seem accurate. This is because political discussions do tend to lean neutral as to avoid being too negative or too positive which in turn does affect the sentiment analysis that VADER does by showing that our data is very neutral.

Relevant Assignment Details

Work Distribution

Name (Last, First)	Student ID	Section contributed	Section edited	Other contributions
Nash, Emma	301452134	Section 1 + 2	Section 2	-
Changfoot, Kinzlee	301425282	Section 1 + 3	Section 3	-

Github Link: https://github.com/kinz52/Assignment-2

References

Miller, I. (n.d.). *Vader sentiment analysis (with examples)*. Hex. https://hex.tech/templates/sentiment-analysis/vader-sentiment-analysis/#:~:text=Positive%3A%20It%20depicts%20the%20positive,being%20the%20strong%20positive%20sentiment.

Search: Document: "/debates/2024/11/6/" "housing" "Trudeau" party: "liberal." openparliament.ca. (n.d.). https://openparliament.ca/search/?q=Document%3A+%22%2Fdebates%2F2024%2F11%2F6%2F%22+%22housing%22+%22Trudeau%22+Party%3A+%22Liberal%22

Youvan, D. C. (2024, June 23). *Understanding Sentiment Analysis with VADER: A Comprehensive Overview and Application*. Research Gate. https://www.researchgate.net/profile/Douglas-Youvan/publication/381650914_Understanding_Sentiment_Analysis_with_VADER_A_Comprehensive_Overview_and_Application/links/6678a048d21e220d89cbc00a/Understanding-Sentiment-Analysis-with-VADER-A-Comprehensive-Overview-and-Application.pdf