Bulls, Bears, and Sentiment:
Comparing Sentiment Analysis Models in Financial Text

Alexander Gow
Professors Elkins and Chun

Abstract

The object of this research is to examine the differences between three Sentiment models. Each is designed differently and is intended to be used for different functions. The results should uncover discrepancies in the way each model understands the sentiment rooted within different types of financial text. Ultimately, this shines a light on how we can understand the way humans write and read about finance given different contexts. The results will show that a fine-tuned financial sentiment algorithm, FinBERT, may be better suited for financial text analysis that has been written in a more professional manner than the casual language used in the news and media. Other researchers may want to take my findings into account when doing their own study, as the type of model they use should be determined, at least in part by the source of their text data.

Introduction

A relatively new technology, Sentiment Analysis, has proven to be a powerful tool for analyzing and interpreting the information embedded in written text or spoken words that have been transcribed. Driven by Artificial Intelligence (AI), these machine learning algorithms harness a human's ability to read and when working in conjunction with advanced data retrieval techniques, they can be used to process massive amounts of textual data faster than hundreds of humans could combined. In the early 2010s, these analytical methods were making their way into the financial sector. DCM Capital, a London-based hedge fund attempted to base investment decisions on the results of Sentiment Analysis on Twitter. However, their efforts went virtually unnoticed as the public and potential investors failed to recognize the potential of the technology and the firm ultimately sold its platform at 65% of its breakeven price. As we progressed through the end of the decade and dove further into the age of big data, the technology advanced and wall street took notice. Always looking for an edge on the market, the information hidden in the text of news articles, social media, and financial statements can provide hedge funds and individual investors with the all-important source of alpha.

With the advent and rapid growth in the popularity of code-related knowledge-sharing websites, such as Kaggle and Github, average people are being given the opportunity to wield the analytical powers of complex Natural Language Processing (NLP) algorithms. As a result, the number of scholarly articles written about the application of sentiment analysis to financial markets and investment decisions has skyrocketed. Much of the academic literature focuses on the application of the new technologies for predicting movements within the stock market. On the other hand, the literature surrounding the evaluation of different sentiment-based NLPs appeared to be very sparse. That is what I seek to do with this project. My research focuses on analyzing the ways in which three different sentiment analysis techniques measure the sentiment of financial texts. With such a huge quantity of attention being given to this amazingly powerful, it becomes challenging to make sense of what distinguishes the different models and why they will produce different results. Moreover, new sentiment models are being posted to the internet pretty much every day, and the extent to which each model will produce accurate estimates for the sentiment of the text you feed it is often unknown. This is because the ways in which the model is trained, in other words, the sources of the text used for training the model are critical to determining the way the model measures sentiment. Moreover, I hope to uncover specific sentences and words within financial texts that my three models of interest interpreted differently., ultimately, uncovering the reasons for why we see certain models perform more effectively. Furthermore, using three different types of financial text may help shed light on the ways in which humans talk and write about finance based on the target audience of their work.

References

P. Uhr, J. Zenkert and M. Fathi, "Sentiment analysis in financial markets A framework to utilize the human ability of word association for analyzing stock market news reports," *2014 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 2014, pp. 912-917, doi: 10.1109/SMC.2014.6974028.

Data and Models

The three models that I will be comparing are Vader, RoBERTa-large, and FinBERT. Each of these models will produce different sentiment scores given the same text due to the training and fundamental structure of the model.

Vader

VADER (Valence Aware Dictionary for sEntiment Reasoning) is a model that measures both the polarity of the sentiment, whether the text is positive or negative, and the strength or intensity of the emotion being expressed in the text. Although the field of sentiment analysis is broad, the variety of approaches can be reduced to either lexical or machine learning-based techniques. Likely to be the simplest of the three models, VADER exemplifies the lexical approach, using a dictionary of words with sentiments assigned to them. The VADER model categorizes the words of each sentence and then applies a set a of heuristics to assign an intensity of the sentiment based on the context in which the word is used, ultimately creating an overall sentiment score for each sentence. The heuristics used to estimate the intensity of sentiment include: punctuation, capitalization, degree modifiers, shift in polarity due to "but", and examining the tri-gram before a lexical feature to capture polarity negation.

RoBERTa

The RoBERTa is a transformers model that has been trained in a self-supervised way on a large selection Wikipedia pages, news articles, books, and more. The fact that RoBERTa is trained on data unlabeled by humans makes it very effective for analyzing a wide variety of texts. Moreover, the RoBERTa transformer model differs from classic recurrent neural networks and autoregressive models because the model will randomly hide 15% of the input words and then attempt to predict them. This enables RoBERTa to have a bidirectional understanding of the sentence.

FinBERT

FinBERT a language model base on BERT, was created specifically for the task of sentiment analysis on financial text. Researchers found that the type of language used and ways in which financial texts are written can often cause language models trained on general text data to fail to retain the same accuracy when tested with financial text. Specialized financial sentiment lexicons have existed for a while, such as the Loughran and McDonald library. However, because they are lexical-based, the analysis falls short and cannot unpack the semantic meaning rooted deeply in financial text.

Data

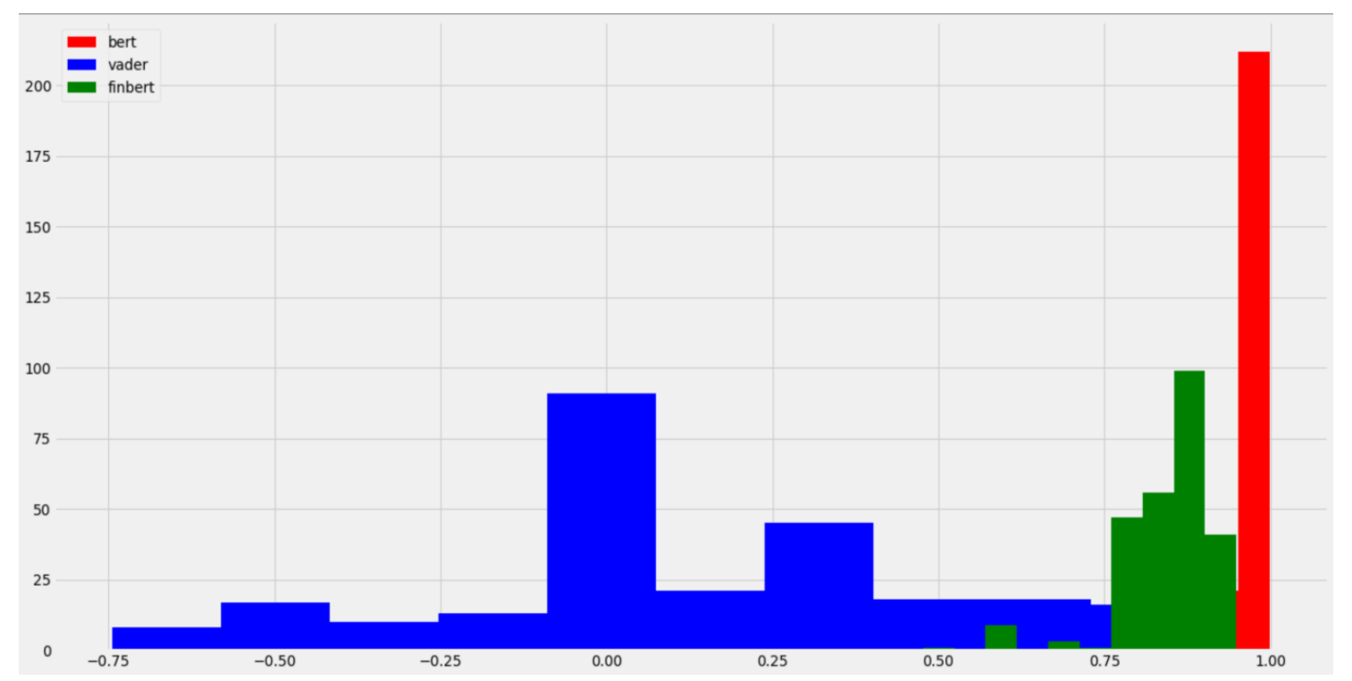
In order to produce some robust results with this analysis and to understand the different ways the models interpret the sentiment of financial text I chose to use three different types of financial text. The first type is news articles that either directly discuss the state of the stock market or are related to the wall street environment. The articles were taken from a variety of sources, such as CNBC, Fortune, and Reuters. The second type of text data is taken from a quasi-analyst report that covers a financial analysis and valuation of Tesla. The final type of data is text taken from the risk factors and management's discussion and analysis (MD&A) sections of Tesla's 10-K SEC filing reported at the beginning of this year. I wanted to use distinctly different types of financial text in order to possibly expose differences in the style of writing used for different purposes. Additionally, I hope to find discrepancies between the ways the models understand the different styles of financial text.

Methods

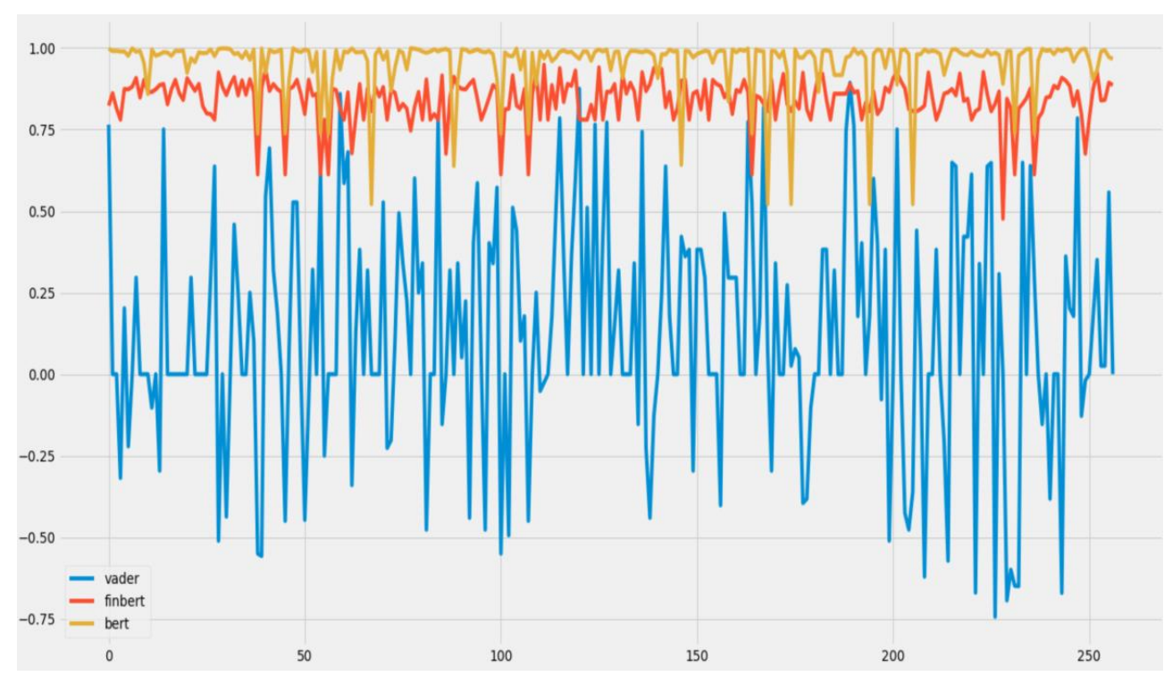
With the guidance and technical assistance from Professor Chun, who has done extensive work with regards to sentiment analysis, I will first take the raw text from each data source and clean each it using TextHero's clean function. Then the cleaned text is separated by sentences. This is done so the models can produce a sentiment score for each line which allows us to understand how the sentiment conveyed by the writers changes from line to line. After the sentiment scores are produced by each model for all lines in all three type of financial text, I will then identify the points at which the sentiment models disagree the most from one another. Ultimately, I will manually take a closer look at the actual text associated with those points of disagreement and elaborate on reasons for why the models disagree with each other about the sentiment.

Results

News

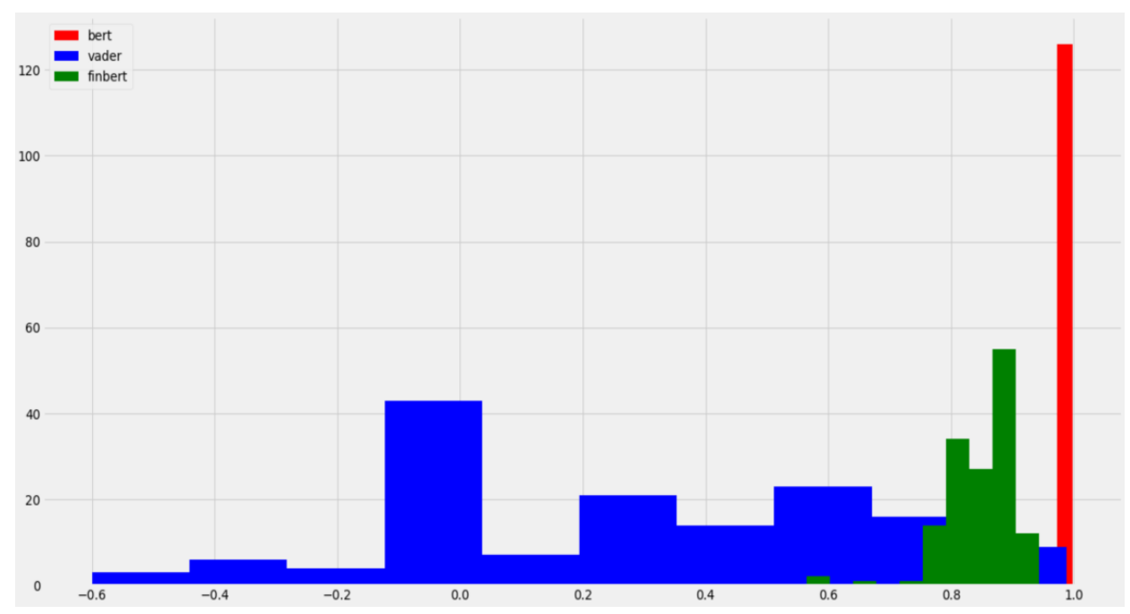


The histogram displayed on the right shows the distribution of the sentiment scores produced by all three models after being run on the financial news text. The general distributions of the scores for each model will show some consistency between the three sources of text. However, the fact that the scores produced by the RoBERTa-large sentiment model are clearly more centered just around 1. It is interesting to note that the VADER model is the only model that found negative sentiment. The line plot shown here displays the line plot of the sentiment scores for the news text with each point representing the sentiment of one line of text. This more clearly shows the similarity in estimates from the RoBERTa (labeled 'bert') model and the FinBERT model. The similarities between the two should not be surprising given that both have very similar cores.

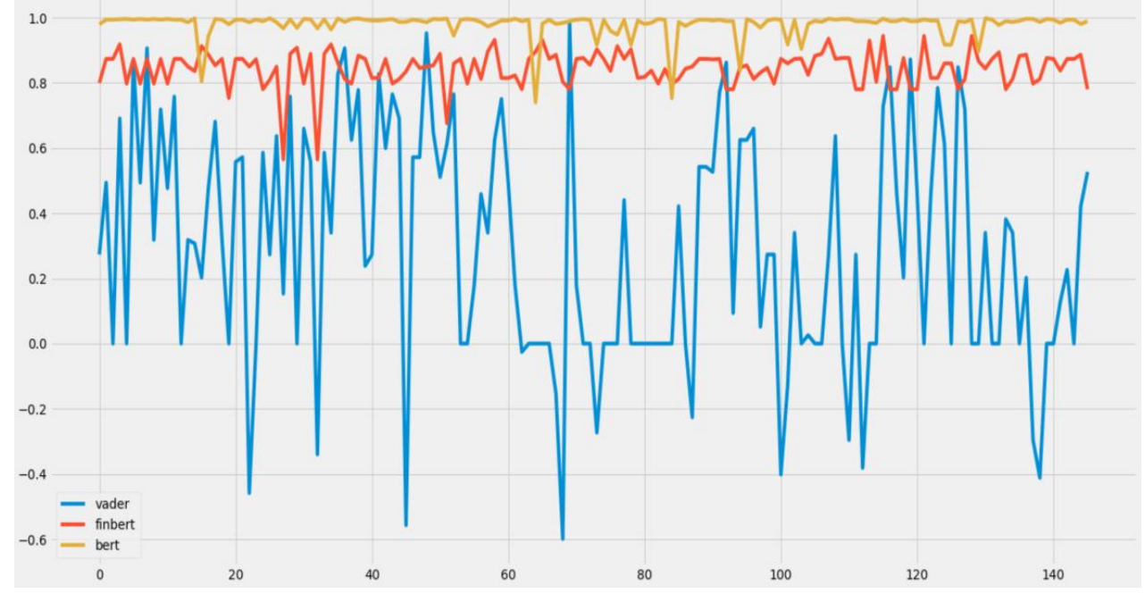


I imagine that the drastic difference between the VADER and the other models is likely due to the fact that the VADER model utilizes a lexical approach while the other two use a machine learning-based algorithm.

Tesla 10-K

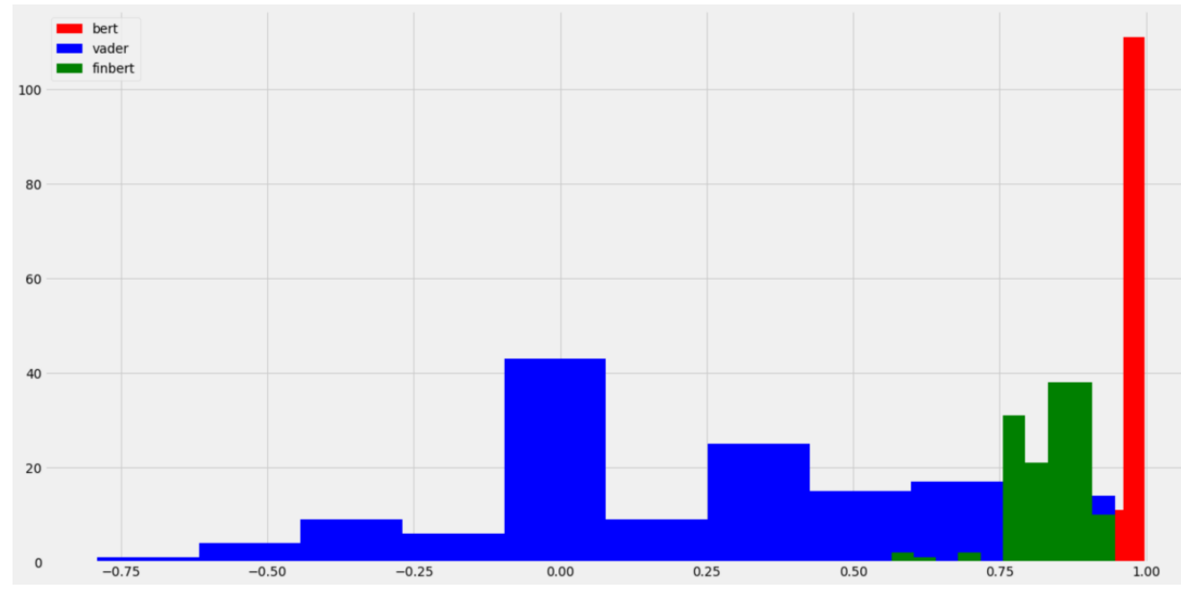


The distribution of sentiment scores for the text gather from Tesla's 10-K sec filing shows a fairly similar pattern to the one produced for financial news. However, we do see some slight differences. Moreover, the distribution of the VADER and FinBERT sentiments is more negative than what we saw in the financial news text. This could be explained by the possibility that financial news is commonly reported using more upbeat and positive language. On the other hand, the risk factors and MD&A sections of the 10-K specifically discuss potential threats or obstacles facing the company. Due to the nature of what is being discussed, it should be of no surprise that we find a slightly less positive sentiment.



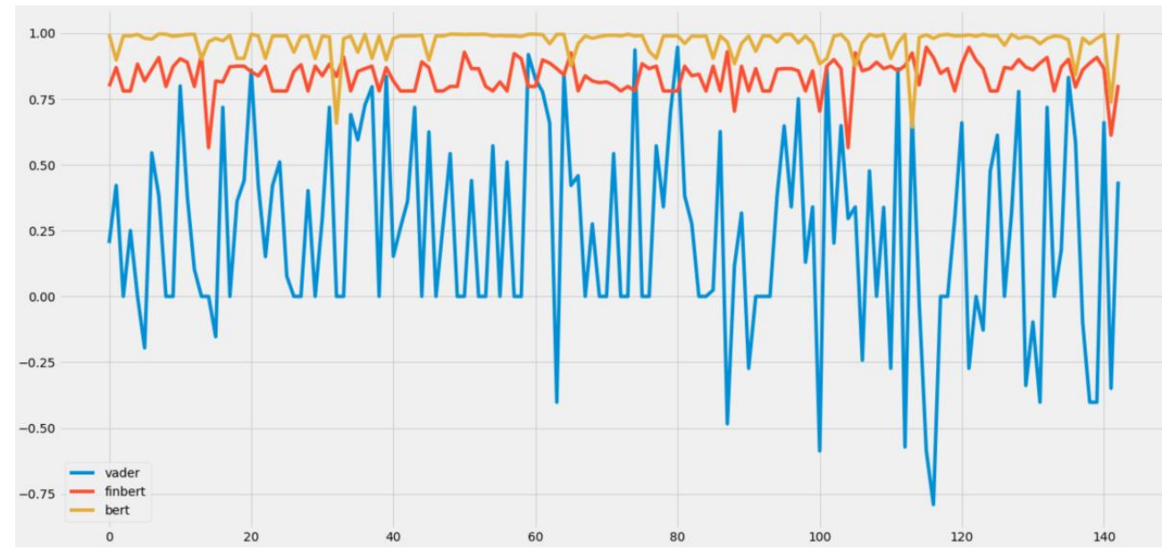
The line plot of the sentiment scores produced by the three models based on the 10-K shows slightly more overlap between the VADER and FinBERT models in comparison to the news articles. More interestingly, the plot sheds important light on points where the models actually produce opposite changes in sentiment based on the same line of text. I will dive deeper into an explanation for why this might occur by examining the text itself in the conclusion section

Financial Analysis and Valuation of Tesla Inc.



Again, we see another very familiar distribution of sentiment scores from the quasi-analyst report text. Moreover, the fact that the distribution of sentiment estimates for all three models remains very consistent over three very different types of financial text proves that the models understand the sentiment rooted in text very differently. Additionally, the results of this analysis show that using the traditional BERT model for understanding the meaning hidden within the sentiment of financial text will not be very informative because it is heavily positively biased.

Results (continued)



It is really interesting to note that the VADER model produces nearly zero negative sentiment scores in comparison to the other two texts. Because the test is a financial analysis and valuation, the writing is filled with language surrounding growth and opportunity which may be causing the basic lexical model to overestimate the positivity within the text.

Conclusion

Taking a closer look at the points of disagreement between the models and the text that produces the sentiment polarity may help uncover some of the practical differences. Lines 111-116 of the analyst report text appeared to produce the most inconsistent sentiment scores across all three models. Moreover, line 113 caused the greatest change in sentiment, measured by summing the changes in sentiment scores of all three models from the previous line. Lines 112-114 read:

- 112) These are rather negative indicators for the future potential of the company.
- 113) SGA (selling, general and administrative) expenses, on the other hand, have reduced in 2019, after four years of growth.
- 114) Due to this reduction, Tesla Inc. shows a positive EBIT in 2019, whilst for the four years before that, it was negative.

Line 113 produced really interesting results, the VADER model estimated the sentiment at a positive 0.7 while rating lines 112 and 114 negatively. Furthermore, the FinBERT and RoBERTa models actually produced opposite changes from 112→113→114. The RoBERTa model, which primarily estimated the sentiment of the text in the high .90's, estimated the 113 to be .643, substantially less positive than the scores it produced for lines 112 and 114. The FinBERT model estimated that the sentiment actually improved from 112→113 and then decreased in positivity from 113–114. I am surprised to find that the lexical model and ML model that has been fine tuned for financial text agree on the change in direction while the traditional RoBERTa model dissents. Furthermore, the finishing lines of the text where the writer discusses how risky assumptions are made in order to estimate future potential growth. Moreover, the company of Tesla struggled greatly to make money in the years prior to this text being written. This might explain why we see more negative results.

Moving on to the News text, the results showed that lines 226 through 229 produced the most disagreement amongst the models. Lines 226-229 read:

- 226) But if they get really, really angry, then ... you'll see managers go short Starbucks and go long, say, Dunkin' Brands or McDonald's ."
- 227) Conclusions Cramer joked on "Squawk on the Street" about going long Dine Brands ' Applebee's and shorting Apple in the same fashion, but he admitted he was being facetious.
- 228) "Still, the point I was making stands," he said.
- 229) "Right now we have a president who doesn't seem to care about what American companies he hurts, including Apple, while trying to get China to change its behavior.

FinBERT rated line 228 as one of the least positive sentences out of the entire text, with a sentiment score of 0.477 while the VADER model rated it as completely neutral with a value of 0. Because we are dealing with news articles, the text is presented in a more casual context, and therefore FinBERT may not be well suited to accurately estimate the sentiment of financial news. Moreover, because the articles are supposed to serve as an entertainment piece as well as a source of information, there is more room for the authors/writers to convey human emotion. This may be the cause of why FinBERT estimates this rapid difference in sentiment.

The sentiment estimates of Tesla's 10-K show the least amount of consistent disagreement among the models, meaning that the lines that the VADER and FinBERT disagree upon are not the same as the lines that FinBERT and the RoBERTa-large model disagree on. Lines 32 and 33 produced the most polarized sentiment scores, line 32 produced the greatest difference in sentiment between both BERT models, and line 33 causing the greatest total change in sentiment across all models. Lines 32 and 33 read:

- 32) However, we operate in a cyclical industry that is sensitive to trade, environmental and political uncertainty, all of which may also be compounded by any future global impact from the COVID-19 pandemic.
- 33) Moreover, as additional competitors enter the marketplace and help bring the world closer to sustainable transportation, we will have to continue to execute well to maintain our momentum.

The more negative results produced by these lines should come as no surprise given that the text is discussing uncertainty and instability.