# Hybrid sentiment analysis system to extract language bias in news media

Sijie Sally Song

## Abstract

The modern news media is often biased. By pushing a certain perspective through its news narrative, biased news outlets use their tremendous influence to manipulate public perception of information. Amongst other methods of swaying their audience, strategic word choices are frequently incorporated in news reports. This method is usually successful, the core reason being the audience's unawareness of most news articles' biased nature. It is, therefore, the aim of this study to detect and quantify the affective states in the language used in modern news articles using a novel sentiment analysis model.

VADER (Valence Aware Dictionary and sEntiment Reasoner) is a lexicon and rule-based sentiment analysis tool specially adapted to analyzing short social media texts. In this paper, an error analysis of VADER is performed in a transductive transfer learning scenario; the pre-trained tool is then optimized using a Multinomial Naive Bayes classifier machine learning-based model trained on tweets for detecting and quantifying sentiments in news articles. This will be done manually after analysis of the outputs and combining the strengths of each model to create an improved hybrid system. This novel model, which would give insight into the severity of bias in major global news outlets for later analysis, would be presented and evaluated.

## Keywords

Sentiment analysis, media bias, VADER, transductive transfer learning scenario, multinomial Naive Bayes

## 1 Introduction

The present-day news media exerts great influence over public opinion on issues being reported on. This power is often abused when news reports stray from objective facts and instead attempt to promote certain viewpoints over others.

Media bias can be accomplished through multiple means, such as the omission of information or the emphasis on certain views over others. It is difficult, however, to quantify information bias as there is no clear measure for the level of information transparency in journalism. Therefore tone bias, as it is relatively more detectable, usually functions as an indicator of bias for readers.

Biased language in news media is defined as strategically selecting words that carry specific nonobjective attitudes with the intent to induce an affective response in the reader. Sentiment carried by the word choices is often subtle, increasing the difficulty of recognizing the biased tone of the article, thus making readers more prone to being persuaded. Computer algorithms, on the other hand, are capable of detecting bias in language and evaluating its severity on a fixed criterion.

Sentiment analysis (also referred to as opinion mining) is a natural language processing (NLP) task that uses data mining and text analysis to extract subjective attitudes from a body of text. A common use of sentiment analysis is extracting sentiments from social media content to recognize public opinion on certain topics or products. Specially designed for this purpose, the sentiment analysis lexicon VADER (Valence Aware Dictionary and sEntiment Reasoner) is designed to evaluate short social media texts. This model, with its unique approach of employing pure manual human analysis in its lexicon generation, has the highest performance scores out of 11 state-of-the-practice sentiment analysis tools on four different domains and remains one of the most reliable sentiment analysis models (Hutto and Gilbert, 2014). However, this model's performance lowers in other domains. Out of the four domains tested on, its performance score is the lowest on New York Times editorials, which also happens to be the domain with the longest text. Its overall precision, overall recall, and overall F1 score on New York Times editorials is 0.69, 0.49, and 0.55 respectively, compared to its performance on a dataset of social media text (tweets), which happens to be 0.99, 0.94 and 0.94 respectively.

Naive Bayes (NB) is a conditional probability classifier established on the basis of Bayes' theorem (its exponential form shown below):

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

The theorem calculates the probability of event A occurring under the condition of event B having occurred, under the naive assumption that all features in P(A) and P(B) are mutually independent. It then finds the maximum probability (y) through this equation where P(x) is a constant:

$$y = argmax_y[P(y) * \prod_{i=1}^{n} P(x_i|y)]$$

In this paper, the VADER tool that is especially attuned to sentiments in social media will be combined with a Multinomial Naive Bayes algorithm trained on tweets to create a novel hybrid system to analyze news articles for language bias. The model development will be based on an analysis of the model in a transductive transfer learning scenario, where the performance of the VADER tool is first analyzed on a familiar domain, a dataset of New York Times editorials for error analysis, and is then tested on an unfamiliar domain, a dataset of recent online news articles, its performance compared with the Naive Bayes model's performance on the same dataset. After testing and comparing the models, the most successful characteristics of each individual model are combined and integrated.

## 2 Related Work

### 2.1 Historical Background

Sentiment analysis has been one of the most rapidly developing fields of computer science. To understand the historical origin of sentiment analysis, Mäntylä, Mika V. et al. (2017) used text mining and qualitative coding to analyze 6996 papers from Scopus to generate a literature review on the evolution of sentiment analysis. Findings were that sentiment analysis was first utilized in text subjectivity analysis by the computational linguistics community in the 1990s, as well as for studying public opinion at the beginning of the 20th century. However, sentiment analysis research only gained notable popularity in 2004, as 99% of the papers from Scopus were from after 2004. Pang, Bo and Lillian Lee (2007) attributed the following factors as causes for this sudden increase in sentiment analysis research:

- the rise of machine learning methods in natural language processing and information retrieval;
- the increased supply of subjective texts on the World Wide Web, due to its popularity gain and, specifically, the exponential growth of online review sites;
- recognition of researchers for the wide range of applications and intellectual exploration that the area offers.

It was also found by Mäntylä, Mika V. et al. (2017) that the usage of sentiment analysis has expanded significantly in recent years - the most studied sentiment analysis tasks have evolved from interpreting online product reviews for commercial purposes to sentiment analysis is now used to analyze texts on social platforms. The study on sentiment analysis has branched out further to connect with other areas of research, such as the stock market, elections, medicine, cyberbullying, software engineering, etc.

## 2.2 Sentiment Analysis

### 2.2.1 General Challenges

To define the characteristics of sentiment analysis, Pang and Lee (2007) identify its regression-like nature which differentiates it from other fact-based text-mining tasks.
They examined factors that make sentiment analysis a difficult task and concluded that it is not only difficult to create a set of keywords assigned to sentiments stochastically, as the complex grammatical rules of the English language could negate the sentiment that a keyword is assigned, but sentiments are sometimes expressed without the use of subjective words, which would deem the lexicon ineffective. In conclusion, sentiment depends heavily on the context. Consequently, it is difficult to quantify the sentiment expressed with a numerical unit of measure.

### 2.2.2 Semantic Orientation (Sentiment Polarity) Classification

In polarity classification, where the task is to classify the text sentiment as either negative or positive, the line between subjective and objective information thins. Pang and Lee (2007) specify examples where inherently objective information can convey subjective meaning based on the context. For example, “battery life is 2 hours” which is objective information but can be interpreted negatively when compared to the general standard for battery life. Some words, such as “democrats” can carry different sentiment polarity depends on the recipient of the information.

Pang and Lee (2007) also suggest other approaches to sentiment polarity classification that may be taken to create or assist a model for sentiment polarity classification. These approaches are summarized below:

- Related categories: categorizing text on other categories besides polarity that may be helpful in ranking polarity, such as categorizing reasons behind a certain sentiment (“I don’t like this computer because of its short battery life”) or ranking with comparative polarity (“I like this product less than the one last year”) or using outcome polarity such as recovery/death in medical journals to predict possible polarity.
- Rating inference (ordinal regression): analyzing with data that provide the text to analyze along with the author’s self-rating on the sentiment expressed by their text (i.e. product reviews where

users rate their satisfaction from one to five stars). This sort of analysis can arguably constitute a slightly different category - ordinal regression.

- Agreement: agreement detection (determining whether two pieces of text agree with each other)

Yu and Hatzivassiloglou (2003) approached sentiment valence in a 3-class classification task - classifying a text as either subjective, neutral, or objective. They used 3 approaches: 1. Measuring the similarity of sentences to other sentences labeled as either objective or subjective based on words, phrases, and WordNet synsets. The hypothesis is that subjective sentences within a certain topic will be more similar to other subjective sentences than to objective sentences; 2. Multiple Naive Bayes classifiers, each using a different subset of the features; 3. Log-likelihood ratio to classify words based on log-likelihood average scores:

$$L(W_i, POS_j) = log\left(\frac{\frac{Freq(W_i, POS_j, ADJ_p)+\varepsilon}{Freq(W_{all}, POS_j, ADJ_p)}}{\frac{Freq(W_i, POS_j, ADJ_n)+\varepsilon}{Freq(W_{all}, POS_j, ADJ_n)}}\right).$$

where $W_i$ is a word in the sentence, $ADJ_p$ is positive seed word set, $ADJ_n$ is negative seed word set, $POS_j$ is part of speech collocation frequency ratio with $ADJ_p$ and $ADJ_n$ and is $\varepsilon$ is a smoothing constant (0.5).

A comparison of machine learning methods used for sentiment analysis was created by Pang, Bo et al. (2020), analyzing the performance of three methods. The first algorithm they examined was a Naive Bayes (NB) classifier, consisting of relative-frequency estimation of P(c) and P($f_i$ | c), and add-one smoothing. Its exponential form is as shown below:

$$P_{\mathrm{NB}}(c \mid d) := \frac{P(c)\left(\prod_{i=1}^{m} P(f_i \mid c)^{n_i(d)}\right)}{P(d)}.$$

The second algorithm that was reviewed was the maximum entropy classification, with its exponential form shown below:

$$P_{\mathrm{ME}}(c \mid d) := \frac{1}{Z(d)} \exp\left(\sum_i \lambda_{i,c} F_{i,c}(d, c)\right),$$

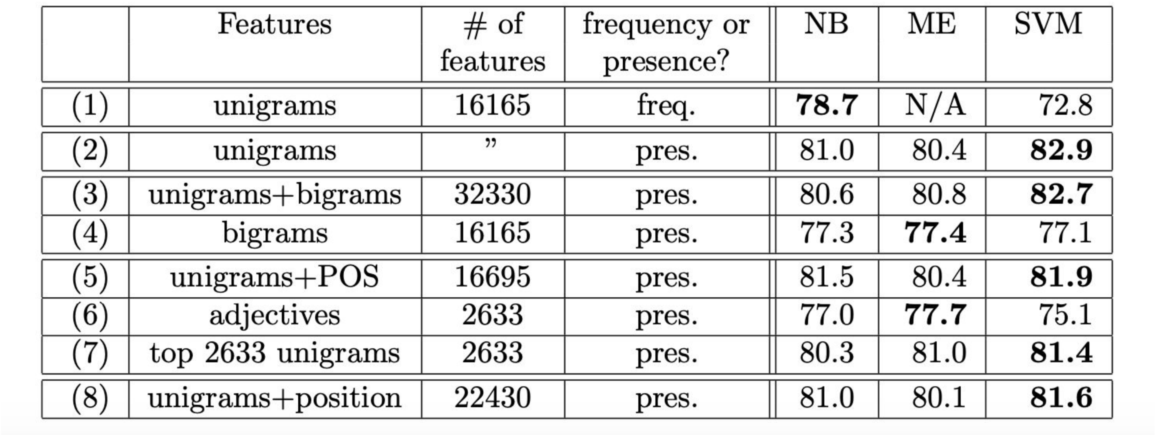
where Z(d) is a normalization function. $f_{i,c}$ is a feature/class function for feature $f_i$ and class c, defined as follows:

$$F_{i,c}(d, c') := \begin{cases} 1, & n_i(d) > 0 \text{ and } c' = c \\ 0 & \text{otherwise} \end{cases}.$$

The third algorithm reviewed was support vector machines (SVMs), which are large-margin, instead of probabilistic classifiers, unlike the previous two classifiers. Its exponential form is shown below:

$$\vec{w} := \sum_j \alpha_j c_j \vec{d_j}, \;\; \alpha_j \geq 0,$$

The accuracies of the algorithms are compared in Figure 1:



| | Features | # of features | frequency or presence? | NB | ME | SVM |
|---|---|---|---|---|---|---|
| (1) | unigrams | 16165 | freq. | **78.7** | N/A | 72.8 |
| (2) | unigrams | ” | pres. | 81.0 | 80.4 | **82.9** |
| (3) | unigrams+bigrams | 32330 | pres. | 80.6 | 80.8 | **82.7** |
| (4) | bigrams | 16165 | pres. | 77.3 | **77.4** | 77.1 |
| (5) | unigrams+POS | 16695 | pres. | 81.5 | 80.4 | **81.9** |
| (6) | adjectives | 2633 | pres. | 77.0 | **77.7** | 75.1 |
| (7) | top 2633 unigrams | 2633 | pres. | 80.3 | 81.0 | **81.4** |
| (8) | unigrams+position | 22430 | pres. | 81.0 | 80.1 | **81.6** |

*Figure 1: Average three-fold cross-validation accuracies, in percent (%). Numbers in* ***bold*** *represent the best performance for a given setting (row).*

### 2.2.3 Sentiment Intensity (Valence-based) Analysis

Hatzivassiloglou and Wiebe (2000) established through studying the effects of dynamic adjectives, semantically oriented adjectives, and gradable adjectives on a subjectivity classifier, that adjective orientation has a prominent effect on the subjectivity of a sentence. They created a novel method that statistically combines two indicators of gradability, showing that sets involving dynamic adjectives with positive or negative polarity or gradability are better predictors of subjective sentences than the class of adjectives as a whole.

VADER (Valence Aware Dictionary and sEntiment Reasoner) is a sentiment analyzer that classifies input with regard to both polarity and valence. Created by Hutto and Gilbert (2014) in their research article “VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text”, it is rather unique in its approach: out of all of the sentiment analysis models examined in section 2.2 of this paper, it is the only one that completely strays from a machine learning approach by employing human labor to manually analyze text, which is extremely rare among sentiment analysis fields. It is also the only model reviewed here that is both valence and polarity detecting. Without a machine learning approach, this model is completely rule-based, combining grammatical rules with its human-generated lexicon. Below are some heuristic rules used to assess the sentiment intensity and polarity of a sentence:
1. Punctuation. “I like this.” vs “I like this!”
2. Capitalization. “This place is amazing.” vs “THIS PLACE IS AMAZING.”
3. Degree modifiers (also called intensifiers, booster words, or degree adverbs). “Good” vs “Very good”
4. In the case of the contrastive conjunction “but”, the sentiment of the text following the conjunction is dominant. “I like the food here, but the service isn’t great.”
5. Negation flips the polarity of the sentence 90% of the time. i.e. “The food here isn’t really all that great”.

The output of VADER is a number between -4 to 4, with -4 representing an “extremely negative” sentiment and 4 representing an “extremely positive” sentiment. In this way, VADER measures the valence of its sentiment on a regression scale.

The VADER model is then evaluated against 11 typical state-of-practice benchmarks which include LIWC, ANEW, the General Inquirer, SentiWordNet, and machine learning models relying on Naive Bayes, Maximum Entropy, and Support Vector Machine (SVM) algorithms. The results, as displayed in Figure 4 and 5, show VADER to have the highest classification precision, recall, and F1 accuracy compared to 11 other benchmarks.

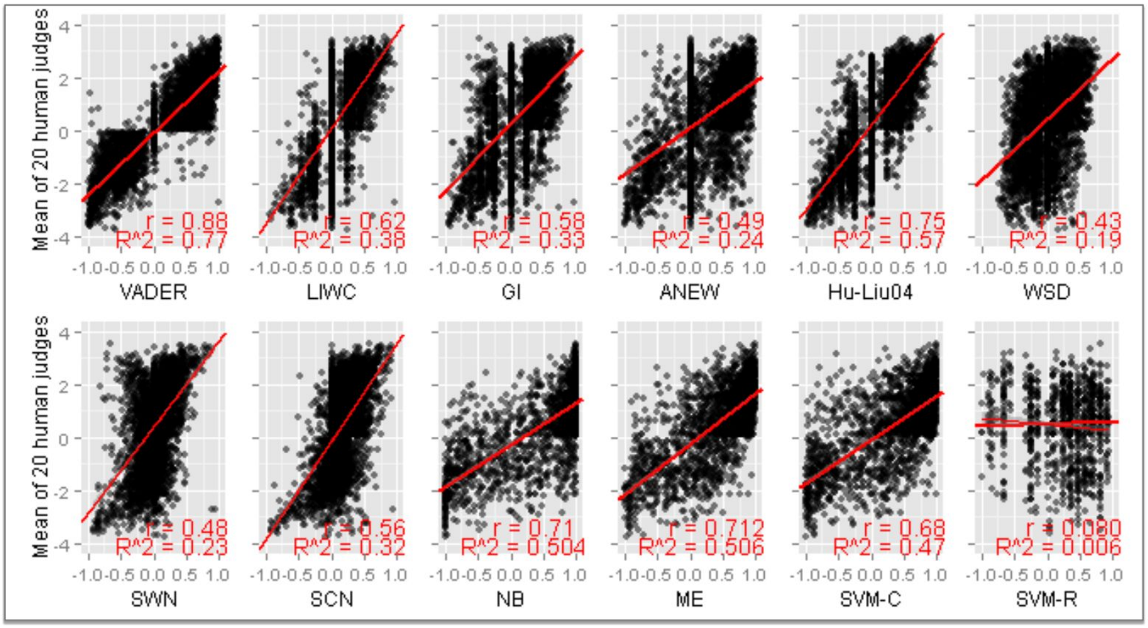
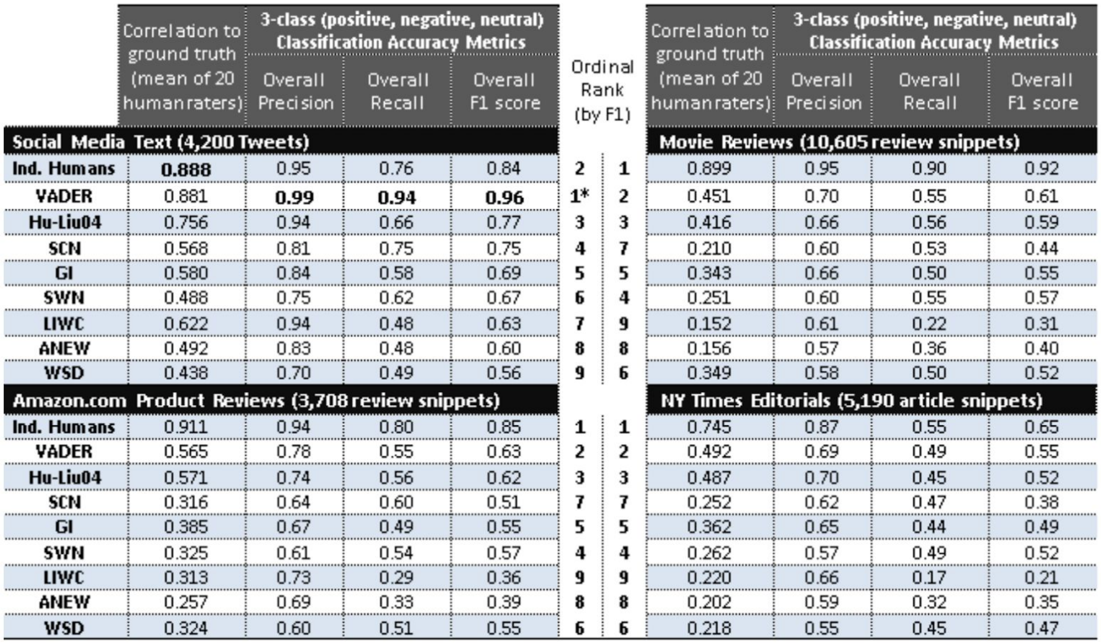


*Figure 2: VADER's classification precision, recall, and F1 accuracy compared to 11 other benchmarks.*



| | Correlation to ground truth (mean of 20 human raters) | 3-class (positive, negative, neutral) Classification Accuracy Metrics | | | Ordinal Rank (by F1) | | Correlation to ground truth (mean of 20 human raters) | 3-class (positive, negative, neutral) Classification Accuracy Metrics | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Overall Precision | Overall Recall | Overall F1 score | | | | Overall Precision | Overall Recall | Overall F1 score |
| **Social Media Text (4,200 Tweets)** | | | | | | | **Movie Reviews (10,605 review snippets)** | | | |
| **Ind. Humans** | **0.888** | 0.95 | 0.76 | 0.84 | 2 | 1 | 0.899 | 0.95 | 0.90 | 0.92 |
| **VADER** | 0.881 | **0.99** | **0.94** | **0.96** | 1* | 2 | 0.451 | 0.70 | 0.55 | 0.61 |
| **Hu-Liu04** | 0.756 | 0.94 | 0.66 | 0.77 | 3 | 3 | 0.416 | 0.66 | 0.56 | 0.59 |
| **SCN** | 0.568 | 0.81 | 0.75 | 0.75 | 4 | 7 | 0.210 | 0.60 | 0.53 | 0.44 |
| **GI** | 0.580 | 0.84 | 0.58 | 0.69 | 5 | 5 | 0.343 | 0.66 | 0.50 | 0.55 |
| **SWN** | 0.488 | 0.75 | 0.62 | 0.67 | 6 | 4 | 0.251 | 0.60 | 0.55 | 0.57 |
| **LIWC** | 0.622 | 0.94 | 0.48 | 0.63 | 7 | 9 | 0.152 | 0.61 | 0.22 | 0.31 |
| **ANEW** | 0.492 | 0.83 | 0.48 | 0.60 | 8 | 8 | 0.156 | 0.57 | 0.36 | 0.40 |
| **WSD** | 0.438 | 0.70 | 0.49 | 0.56 | 9 | 6 | 0.349 | 0.58 | 0.50 | 0.52 |
| **Amazon.com Product Reviews (3,708 review snippets)** | | | | | | | **NY Times Editorials (5,190 article snippets)** | | | |
| **Ind. Humans** | 0.911 | 0.94 | 0.80 | 0.85 | 1 | 1 | 0.745 | 0.87 | 0.55 | 0.65 |
| **VADER** | 0.565 | 0.78 | 0.55 | 0.63 | 2 | 2 | 0.492 | 0.69 | 0.49 | 0.55 |
| **Hu-Liu04** | 0.571 | 0.74 | 0.56 | 0.62 | 3 | 3 | 0.487 | 0.70 | 0.45 | 0.52 |
| **SCN** | 0.316 | 0.64 | 0.60 | 0.51 | 7 | 7 | 0.252 | 0.62 | 0.47 | 0.38 |
| **GI** | 0.385 | 0.67 | 0.49 | 0.55 | 5 | 5 | 0.362 | 0.65 | 0.44 | 0.49 |
| **SWN** | 0.325 | 0.61 | 0.54 | 0.57 | 4 | 4 | 0.262 | 0.57 | 0.49 | 0.52 |
| **LIWC** | 0.313 | 0.73 | 0.29 | 0.36 | 9 | 9 | 0.220 | 0.66 | 0.17 | 0.21 |
| **ANEW** | 0.257 | 0.69 | 0.33 | 0.39 | 8 | 8 | 0.202 | 0.59 | 0.32 | 0.35 |
| **WSD** | 0.324 | 0.60 | 0.51 | 0.55 | 6 | 6 | 0.218 | 0.55 | 0.45 | 0.47 |

*Figure 3: 3-class classification performance as compared to individual human raters and 7 lexicon baselines across four domains.*

## 3 Data

The VADER tool will be obtained from GitHub, titled under vaderSentiment by cjhutto.

The Naive Bayes model is trained on Sentiment140 dataset on Kaggle, the same dataset that is used to develop the VADER tool. The dataset contains 1.6 million tweets which are labeled as one of the three following numbers that represent a polarity: 0 (negative), 2 (neutral), 4 (positive).

The VADER performance and its ground truth on New York Times editorials (opinion articles) will be examined for error analysis, since opinion articles are heavily rhetoric, difficult to analyze and therefore would induce an extensive amount of errors for analysis. This data is also obtained from GitHub under the same directory as that of the VADER tool.

VADER will then be tested on an unfamiliar domain, which is a manually collected small dataset of news articles on recent significant political events that underwent extensive news coverage. Only political events are chosen because other genres of factual news inherently contain polarity (i.e. climate change is bad, stock market surge is good, etc.); political news of topics such as terrorism that inherently contain strong polarity also aren't chosen, since mass reports of these events will express a similar polarity. The dataset consists of 5 snippets that were taken from each of the 45 political news articles, each article is from one news outlet about one selected political event. Each news outlet will have 5 articles in total, each article reporting on a single event.

| Name of news outlet | Country of news outlet |
|---|---|
| Xinhua News | China |
| SOHU | |
| Sina | |
| BBC | United Kingdom |
| Reuters | |
| the Guardian | |
| CNN | The United States of America |
| New York Times | |
| Fox News | |

*Figure 2: 9 influential news outlets selected to evaluate the model's performance*

| Event | Time of the event | Countries involved |
|---|---|---|

| | | |
|---|---|---|
| Black Lives Matter protest | May 2020 - present | U.S., New Zealand, France, the U.K., Ireland, Brazil, Germany, Italy, Poland, Denmark, The Netherlands, Israel, Sudan, |
| Donald Trump’s Impeachment | December 2019 – February 2020 | U.S. |
| 2019-20 Hong Kong protest | June 2019 - present | Hong Kong, China, U.S. (alleged), Taiwan |
| US-China Trade War | January 2018 - present | U.S., China |
| Brexit | June 2016 - January 2020 | The U.K., EU |

*Figure 4: 5 recent significant political events in the dataset*

## 4 Manual analysis of VADER

### 4.1 Error analysis on VADER

VADER’s performance on New York Times editorials is significantly lower than its performance on social media text. To discover the cause for this, an error analysis is performed on the ground truth of the model. VADER’s ground truth is the average human rating of twenty human raters collected in an anonymous survey. The ground truth on New York Times editorials is provided by Hutto and Gilbert (2004). Below are the characteristics observed from these human ratings of texts:

1. Some neutral, factual sentences are rated as having a sentiment intensity. For example, “On Dec. 19, around noon, New York City's Union Square Business Improvement District plans the First Santa Claus Ulympics (that's yule-lympics), in which about 50 contestants, in red suits and beards, will compete.” was rated as having a 0.4 positive sentiment. Some questions, holding no sentiment in itself, such as “Have countries threatened with retaliatory tariffs under the new trade law entered into negotiations with the United States?” was rated as moderately negative (-0.55), although the question is neutral, slightly negative at most.
2. On the other hand, some sentences that carry sentiment are rated as neutral. The sentence “It's an economic euphemism with precedent.”, which carries a negative sentiment in reality but was rated as neutral. This may be that the word “euphemism” is a noun that is generally neutral. However, in this context, the word carries feelings of criticism or mockery.
3. Sarcasm isn’t detected. For example, when rating the sentences “Some budget terms are technical necessities, like the distinction between actual outlays and budget authority. Others are gems of political euphemism.”, both sentences were rated as having positive sentiments (0.05 and 0.95, respectively) even though the second sentence actually expresses criticism. This might be because “gems” and “euphemism” are relatively positive words. The sentence “No one has devised a

euphemism of equal elegance for the other side of the ledger - until now." was also rated as 1, extremely positive, even though it was a sarcastic sentence that expresses a negative connotation. There are many cases of subtle sarcasm in the NYT editorials that weren't detected by the algorithm.

4. Consecutive sentences are not analyzed in accordance with each other, although it is often necessary to do so in order to analyze sentiments. For example, the sentences "The new Reagan budget is skimpy on detail and limp on initiatives, but page 2-13 offers a new pearl of circumlocution. The Administration, it says, calls for no cuts in benefit levels for several programs." were given separate sentiment scores, -0.75 and 0.35 respectively. This analysis is incorrect because these two sentences support each other, thus they convey the same sentiment, which is a negative sentiment of relatively high intensity. A correct analysis of one of these sentences cannot be made without the context of the other sentence.

In New York Times editorials containing 500 snippets of news articles each containing about 10 sentences, approximately 5 sentences out of the 10 sentences in the second snippet of the NYT Editorials had their sentiment polarity inappropriately labeled.

| Sequence | Sentence | Polarity score |
|---|---|---|
| 1 | Government budget time means budget language time. | -0.09 |
| 2 | Some budget terms are technical necessities, like the distinction between actual outlays and budget authority. | 0.01 |
| 3 | Others are gems of political euphemism. | 0.23 |
| 4 | The recent prize-winner, which may be the work of an unknown New York state budget writer a few years ago, is revenue enhancement, meaning tax increase. | -0.05 |
| 5 | No one has devised a euphemism of equal elegance for the other side of the ledger - until now. | 0.25 |
| 6 | The new Reagan budget is skimpy on detail and limp on initiatives, but page 2-13 offers a new pearl of circumlocution. | -0.19 |
| 7 | The Administration, it says, calls for no cuts in benefit levels for several programs. | 0.08 |
| 8 | For some others, however, the Administration proposes carefully targeted reforms. | 0.03 |
| 9 | Perfect: a description for spending cuts, even harsh ones, that sounds innocent, progressive, humane. | 0.30 |
| 10 | Is there, somewhere, a government budget in which carefully targeted reforms involve | -0.11 |

| | | |
|---|---|---|
| | spending more money instead of less? | |

*Figure 5: a 10-sentenced example of error spotting in sentences from the ground truth on New York times editorials with sentiment scores rated by 20 human raters. Ratings have been processed to range from -1 (extremely negative) to 1 (extremely positive). Correct ratings are highlighted in green, incorrect ones in red*

Since the ground truth for New York times editorials contains a noteworthy amount of inaccuracies, the analysis of VADER in figure 3 may be incorrect. An error analysis on VADER's performance on New York Times editorials is therefore performed again on the same 10 sentences.

| **Sentence sequence** | **Sentence** | **Polarity** |
|---|---|---|
| 1 | Government budget time means budget language time. | 0.00 |
| 2 | Some budget terms are technical necessities, like the distinction between actual outlays and budget authority. | 0.10 |
| 3 | Others are gems of political euphemism. | 0.00 |
| 4 | The recent prize-winner, which may be the work of an unknown New York state budget writer a few years ago, is revenue enhancement, meaning tax increase. | 0.08 |
| 5 | No one has devised a euphemism of equal elegance for the other side of the ledger - until now. | 0.06 |
| 6 | The new Reagan budget is skimpy on detail and limp on initiatives, but page 2-13 offers a new pearl of circumlocution. | 0.00 |
| 7 | The Administration, it says, calls for no cuts in benefit levels for several programs. | 0.15 |
| 8 | For some others, however, the Administration proposes carefully targeted reforms. | 0.03 |
| 9 | Perfect: a description for spending cuts, even harsh ones, that sounds innocent, progressive, humane. | 0.06 |
| 10 | Is there, somewhere, a government budget in which carefully targeted reforms involve spending more money instead of less? | 0.03 |

*Figure 6: a 10-sentenced example of error spotting in sentences from a dataset of New York times editorials with sentiment (polarity and subjectivity) scores rated by VADER. Ratings have been processed to range from -1 (extremely negative) to 1 (extremely positive). Correct ratings are highlighted in green, incorrect ones in red*

As seen in figure 6, VADER labeled the polarity of 2 out of the example 10 sentences correctly, its accuracy much lower than that of the ground truth. This remains true throughout all of the snippets in the NYT Editorials dataset.

Through the error analysis performed on a dataset of snippets in NYT Editorials, it can be observed that the VADER lexicon itself has four main shortcomings:

1. Lack of vocabulary

The VADER lexicon has a relatively extensive vocabulary, yet there are words that aren't labelled in the lexicon; usually ones that are uncommonly used or particularly formal, since the lexicon is designed to analyze short, casual social media texts containing mostly every-day language. Therefore, quite frequently, VADER fails to label the sentiment of a certain word, "abomination", for example, that doesn't exist in the lexicon.

2. General lack of context

Phrases that are related to historical or cultural contexts, such as "The Statue of Liberty", "concentration camp", or "Garden of Eden" aren't detected, even though some of these phrases are often used in an attempt to express a sentiment. Idioms such as "big headed" are also not detected by the system.

3. Lack of inter-textual context

VADER considers each sentence individually without the context of other sentences. Therefore, sentences that refer to each other, such as "The democrats are optimistic. Other parties, however, feel the opposite way.", are always given an incorrect score.

4. Lack of understanding of rhetorical devices

VADER's lack of inter-textual context leads to its inability to detect rhetorical devices. The detection of rhetorical devices is heavily dependent on the style of a text (e.g. Humorous irony is expected to be seen in a text with a consistently humorous tone), which can't be extracted without a comprehensive analysis of the whole text. Without inter-textual context, one can't tell whether a sentence like "Well that's just great." is sarcasm or not - it refers back to the last sentence. Not only sarcasm - other common rhetorical devices such as antanagoge (using a sentence to negate the sentiment of previous sentence(s), used in groups of sentences such as "It was a blip on the radar of protest movements. It would fade away like Occupy Wall Street. With no clear structure and no strong leader, some said, it was bound to fail, especially when the infighting began. But still it rises -- and polarizes."

## 5 Learning approach

The drawbacks of the VADER system listed above are unavoidable in a strictly rule-based system as it is incredibly difficult to manually design heuristics that encompass all of the subtleties involved in human languages. Combining the tool with a probabilistic machine learning model would amend for potential flaws in the VADER system.

Along with VADER, the built-in multinomial Naive Bayes Classifier in NLTK is used to analyze 5 snippets that were taken from each of the 45 political news articles (see Section 3: Data). The comparison of their outputs is shown in figure 7.

The polarity output from the Naive Bayes (NB) model is separated into two scores for positive and negative sentiments. These two scores are combined, allowing the positive and negative scores to negate each other (etc. if positive score is 0.6 and negative score is 0.4, the sentence would be positive with an intensity of 0.2), for a compound score of the overall sentiment polarity in the sentence.

| Sentence | NB | | | VADER | | |
|---|---|---|---|---|---|---|
| | Rating | Correct? | | Rating | Correct? | |
| | | Polarity | Intensity | | Polarity | Intensity |
| One year ago on Tuesday, hundreds of thousands of demonstrators in Hong Kong gathered for a march that became the start of the semi autonomous Chinese city's biggest political crisis and the broadest expression of public anger with Beijing in decades. | 0.99 | No | Yes (highly intense) | -0.83 | Yes | Yes (highly intense) |
| In the months that followed, protesters filled the city's streets, broke into the local legislature and vandalized it, staged sit-ins at the airport, and turned a university campus into a fiery battleground. | 0.65 | No | Yes (highly intense) | -0.83 | Yes | Yes (highly intense) |
| Earlier this year, the demonstrations quieted amid the coronavirus pandemic. | -0.33 | No | No | 0.00 | Yes | Yes (neutral) |
| But Beijing's push to impose national security laws over the territory has prompted some protesters to return to the streets. | 0.63 | No | No | -0.26 | Yes | Yes (slightly intense) |
| Their presence is a reminder that many thorny issues — including the demonstrators' demands for greater official accountability — remain unresolved. | 0.99 | No | Yes (highly intense) | 0.10 | No | No |

*Figure 7: an example of an article snippet. Five sentences chosen from the article being analyzed by the model. "Hong Kong Protests, One Year Later" from New York Times. Ratings range from -1 (extremely negative) to 1 (extremely positive).*

Through observing their outputs on the news dataset, it is concluded that the VADER model classifies polarity at a relatively high accuracy rate - much higher than the Naive Bayes classifier. The accuracy has improved significantly compared to its analysis on NYT Editorials since news articles traditionally do not use as many rhetoric devices that may be difficult for VADER to detect.

A notable result is that in almost every article snippet, the Naive Bayes classifier is generally more proficient at scoring the subjectivity intensity of the text. An example of this can be seen in the snippet analyzed in figure 7.

Therefore, to optimize the performance of both models, the polarity classification function of the VADER lexicon - the superior function - is combined with the subjectivity intensity detection of the Naive Bayes classifier.

## 6 Discussions

Through manual analysis on the novel model's performance on the news articles dataset (see section 3: Data), it can be concluded that the novel hybrid system model is relatively more successful in quantifying bias and sentiment in news articles than each of the individual models tested.

VADER's performance is accurate in its polarity mostly because it is a lexicon with labeled words; the potential reason for the Naive Bayes model's success in predicting sentiment intensity is that there is usually a pattern in sentiment intensity: while sentiment polarity can easily be negated or flipped, the intensity for the sentiment often happens to be consistent throughout the text, making it easier to predict through probabilistic approaches yet difficult to detect through a lexicon of

The Naive Bayes model's performance could have been improved significantly if the training data belonged to the same domain as its testing data; having been trained on a dataset of tweets, due to the lack of availability of labeled news article datasets, the model's function would have potentially been disrupted by its transductive transfer learning scenario.

Further work on sentiment analysis would include the assistance of a part-of-speech tagging model to improve inter-textual context - an important aspect towards better sentiment analysis.

## References

"Cjhutto/Vadersentiment". Github, 2020, https://github.com/cjhutto/vaderSentiment/blob/master/vaderSentiment/vaderSentiment.py. Accessed 13 June 2020.

"Hong Kong Protests, One Year Later". Nytimes.Com, 2020, https://www.nytimes.com/2020/06/09/world/asia/hong-kong-protests-one-year-later.html. Accessed 26 June 2020.

Hong Yu and Vasileios Hatzivassiloglou. "Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences.” In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2003.

Hutto, Clayton, and Eric Gilbert. "VADER: A Parsimonious Rule-Based Model For Sentiment Analysis Of Social Media Text". Undefined, 2014, p. ., https://www.semanticscholar.org/paper/VADER%3A-A-Parsimonious-Rule-Based-Model-for-Analysis-Hutto-Gilbert/bcdc102c04fb0e7d4652e8bcc7edd2983bb9576d. Accessed 13 June 2020.

"Introduction To Machine Learning With Python". O’Reilly Online Learning, 2020, https://www.oreilly.com/library/view/introduction-to-machine/9781449369880/. Accessed 26 June 2020.

Mäntylä, Mika V. et al. "The Evolution Of Sentiment Analysis—A Review Of Research Topics, Venues, And Top Cited Papers". Computer Science Review, vol 27, 2018, pp. 16-32. Elsevier BV, doi:10.1016/j.cosrev.2017.10.002. Accessed 12 June 2020.

Pang, Bo and Lillian Lee. “Opinion Mining and Sentiment Analysis.” Found. Trends Inf. Retr. 2 (2007): 1-135.

Pang, Bo et al. "Thumbs Up? Sentiment Classification Using Machine Learning Techniques". Cs.Cornell.Edu, 2020, https://www.cs.cornell.edu/home/llee/papers/sentiment.home.html. Accessed 12 June 2020.

"Senate Acquits Trump On Abuse Of Power, Obstruction Of Congress Charges". Fox News, 2020, https://www.foxnews.com/politics/senate-acquits-president-trump-impeachment-vote. Accessed 24 June 2020.

"Sentiment140 Dataset With 1.6 Million Tweets". Kaggle.Com, 2020, https://www.kaggle.com/kazanova/sentiment140. Accessed 13 June 2020.

Vasileios Hatzivassiloglou and Janyce Wiebe. “Effects of adjective orientation and gradability on sentence subjectivity.” In Proceedings of the International Conference on Computational Linguistics (COLING), 2000.