Deriving Corporate Image from Textual Data using Sentiment Analysis

by

Bilal Majeed, BMath, University of Waterloo, 2018

A Major Research Project
presented to Ryerson University
in partial fulfillment of the requirements for the degree of

Master of Science
in the Program of
Data Science and Analytics

Toronto, Ontario, Canada, 2021

AUTHOR'S DECLARATION FOR ELECTRONIC SUBMISSION OF A MAJOR RESEARCH PROJECT (MRP)

I hereby declare that I am the sole author of this Major Research Paper. This is a true copy of the MRP, including any required final revisions.

I authorize Ryerson University to lend this MRP to other institutions or individuals for the purpose of scholarly research.

I further authorize Ryerson University to reproduce this MRP by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

I understand that my MRP may be made electronically available to the public.

Bilal Majeed

Deriving Corporate Image from Textual Data using Sentiment Analysis

Bilal Majeed

Master of Science 2021

Data Science and Analytics

Ryerson University

Abstract

In this paper, we focus on attempting to understand if a corporation's success can be explained using various online sources: star ratings, reviews, news, social media, and other platforms. From each of these sources, we can extract consumer opinions that may have an impact on a corporation's success. Therefore, we first defined the appropriate sources to extract these opinions, and then determined an effective sampling technique to obtain a representative sample. Once the opinions were extracted, we defined corporate sentiment using different combinations of the extracted sources. Lastly, we used the derived corporate sentiments to measure the degree to which it has an impact on a corporation's popularity and financial performance.

Key words: Sentiment, Blogs, Social Media, News, Reviews, VADER, TextBlob, GoogleAPI

Acknowledgements

I am grateful to Dr. Ozgur Turetken from the Information Technology Management department within the Ted Rogers School of Management for his support and assistance in this project. Dr. Ozgur Turetken was my supervisor for this MRP, and his guidance, support and feedback proved to be invaluable. Furthermore, this project stems off the ideas presented in his research paper in collaboration with Sameh Al-Natour: A comparative assessment of sentiment analysis and star ratings for consumer reviews.

Thank you, Dr. Ozgur Turetken

Table of Contents

1.0 Introduction

This document provides details on the domain, topic, datasets, and research question I chose for my Major Research Project. First, a brief background about the topic, the problem, and datasets is provided. Next, a literature review accompanied by a detailed exploratory analysis was conducted to design effective experiments. Lastly, I outline the methodology and experiments performed, review the results, and provide recommendations for future work.

1.1 Background and Problem Statement

The availability of ratings and reviews, and presence on social media and other platforms may be helpful in determining a company's success. The opinions of consumers from these different bodies of textual data are useful sources of sentiment. Especially with the recent pandemic, these opinions might have an impact on the overall corporate sentiment. Therefore, using the available data, an analysis was conducted to first determine the corporate sentiment from each available source, then the extracted sentiments were combined to generate an overall corporate sentiment. Once the overall corporate sentiment was generated, the correlation with popularity in terms of search volume and two financial performance indicators were used to measure effectiveness of the individual or combination of sentiments by answering the question: Does the calculated corporate sentiment score impact a corporation's popularity and financial performance?

1.2 Dataset

Companies will be selected from different industries to generate a corporate sentiment dataset, i.e., all sources will be summarized to a single 0 – 1 sentiment score and then experiments will be performed.

Available Data:

1- Yelp Star Ratings
2- Yelp Reviews

Extract Data:

1- Tweets (Company #)
2- Blog Reviews
3- News Articles
4- Google Search Trends (Proxy for Popularity)
5- Financial Performance Indicator

1.3 Research Questions

1- Which of the individual or combination of extracted sentiments have the strongest correlation with popularity or the financial performance indicators?
2- How much of the variability in the popularity and the financial performance indicators can be explained using the derived corporate sentiment?

2.0 Literature Review and Data Extraction

Research has effectively shown that online word-of-mouth plays a more important role as compared to traditional WOM in shaping consumer attitudes and behaviors (Bulbul et al., 2014; Chen & Xie, 2008; Cheung & Lee, 2012; Dang et al., 2010; Hu & Chen, 2016; Pavlou & Dimoka, 2006). With more human interactions happening online over the last couple of years, mining and understanding online word of mouth has become more important. Sentiment analysis is the process of utilizing natural language processing to identify and categorize opinions in a particular topic as positive, negative, or neutral. "Sentiment analysis has been used for a range of purposes such as tracking the popularity and desirability of a brand (e.g., Greco & Polli, 2019), identifying product opportunities (e.g., Jeong et al., 2017), studying product launches (e.g., Rathore & Ilavarasan, 2020), and predicting market movement of stocks (e.g., Maqsood et al., 2020)" (Al-Natour et al., 2020). The existing research on sentiment analysis can be extended to determine corporate sentiment, i.e., how people feel about the overall corporation or business. The problem with corporate sentiment is that it is difficult to measure but it may be a strong predictor of a corporation's success (Campbell, 2020). In this paper I attempt to determine if corporate sentiment, derived from multiple online platforms, has an impact on either the popularity of the company and/or the financial performance of the company.

First, the sources must be determined from the available online platforms that provide a holistic image of the corporate sentiment. I aimed to capture areas of the internet where different types of users interact and different styles of writing is utilized, i.e., news, blogs, social media, and store reviews.

Source	Reason
News	Generally objective, formally written by a set of writers, and focus on events
Blogs	Generally subjective, and formally or informally written by a set of writers
Microblogging	Generally subjective, informally written by the general population
Store Reviews	Generally subjective, and formally or informally written by the general population

For each of the selected four sources, I selected a platform that could represent each source accurately by capturing the differing sentiments available.

Source	Platform	Reason
News	Google News	Collection of relevant news from other news websites
Blogs	Huffington Post	Largest blog which covers a wide variety of topics (Adams, 2017)
Microblogging	Twitter	Large business presence with user and business interaction
Store Reviews	Yelp	Has influential power due to the reviews (Medical, 2019)

After the four sources and platforms are chosen, a representative sample of corporations is required to calculate corporate sentiment for further testing for impact. To select this sample, I focused on corporations that have a strong presence in North America and are available in all four sources. Availability of content in all four sources is essential as it will allow for testing of significance for each source. Furthermore, the sample should represent all different types of products and services as defined by economists and market researchers, i.e., search, experience, and credence goods and services. “Each of these categories are established based on ease or difficulty of obtaining information when evaluating a purchase decision” (About: SEC Classification of Goods and Services, n.d.). This sampling technique was also used in paper “A comparative assessment of sentiment analysis and star ratings for consumer review” to gather a representative sample of reviews (Al-Natour et al., 2020).

Product / Service Type	Evaluating a Purchase Decision
Search (S)	Can be evaluated prior to purchase or consumption, e.g., most products
Experience (E)	Can be evaluated only after the product has been purchased or experienced, e.g., personal services
Credence (C)	Difficult or impossible to evaluate even after consumption due to a lack of knowledge to make a realistic evaluation, e.g., professional services

Within each type of good or service (SEC), there exist different industries that need to be accounted for in the experiment, e.g., search goods include technology, clothing, automobiles, banks, restaurants, etc. (About: SEC Classification of Goods and Services, n.d.). The condition placed on the corporations of existing in North America and in the four sources limits the sample from containing an equal number of corporations for each one of the types of goods or services. Furthermore, due to corporations providing a diverse set of products or services, it is difficult to classify a corporation in only one product or service type, e.g., Costco could be put under Search, Experience and Credence because they sell technology, trip packages and vitamins.

Product / Service Type	Corporations	Notes
Search (S)	Apple Store, Microsoft, GameStop, Mercedes-Benz, Tesla Motors, Toyota, Nike, Adidas, Home Depot, Best Buy, Walmart, Target, Costco	Walmart, Target, and Costco offer all 3 types of product / service types, but the majority are search goods
Experience (E)	Hyatt Hotels, Marriott Hotels, Verizon, Rogers, AT&T, Citi Bank, JP Morgan Chase, Wells Fargo Bank, Cineplex, McDonalds, Starbucks, Delta Airlines	Banks are included here due to the personal services offered but they also offer credence services, e.g., insurance

Credence (C)	State Farm Insurance, Allstate Insurance, CVS Healthcare, Berkshire Hathaway	Corporations with offers predominantly in this category are difficult to locate due to not existing in the four sources or an experimental metric is not available

After selecting a representative sample from the four sources, the data had to be gathered through either using an available API or web scraping. There are a few legal and ethical guidelines that must be followed when web scraping to acquire content, i.e., use a public API when available, respect the privacy policy, give proper credit for content, and be responsible with the scraping (Densmore, 2019; Paul, 2020). First, the simplest source to acquire data from was Yelp as data can be requested from their site (https://www.yelp.com/dataset/download). The two files used are the businesses and review files that were merged to acquire reviews for all locations of the selected corporations. Next, Twitter has a public API that can be set up using a free developer account (https://developer.twitter.com/en), which allows 500,000 monthly tweet pulls. The number of tweets that can be extracted is large, e.g., #Apple can have thousands of tweets in a few hours, hence approximately 1000 tweets were extracted for each corporation for each quarter of 2020. Lastly, Huffing Post and Google News were scraped using BeautifulSoup (https://pypi.org/project/beautifulsoup4/) and Newspaper3k (https://newspaper.readthedocs.io/en/latest/) to iterate through pages, find the links to the articles from the HTML code, and extract the headline and descriptions of each article as an API was not available. Once the raw data was collected, simple cleaning was performed as web scraping pulls advertisements on the page as well as promoted articles that are not part of the corporation's blogs or news. For Google News, articles were scraped for 2020 but for Huffington Post, there was a lack of blog posts available so blogs after 2018 were included.

There were a few additional sources that had to be extracted for experimentation and for testing the importance of corporate sentiment: Google Trends, revenue growth, and profit growth. Google Trends will be used as a surrogate for popularity from 2020, which was extracted from the Trends website (https://trends.google.com/trends/?geo=CA). The data from this source is pre-scaled and normalized between 0 and 100 to allow for easier processing of data (https://support.google.com/trends/answer/4365533?hl=en). On the other hand, revenue growth and profit growth were manually extracted from Yahoo Finance's financials tab using the last two annual reports, i.e., late 2019 vs late 2020 or early 2020 vs early 2021 to calculate growth rates (https://ca.finance.yahoo.com/). If financials were not available on Yahoo Finance, income statements were used to gather the two growth rates.

Once the data was acquired and cleaned, sentiment analysis had to be done for each of the four sources: Yelp reviews, tweets, blog summary and news article summary. To get the summaries of the blogs and the news articles, natural language functionality of the Newspaper3k was utilized. The goal is to extract sentiment score for each sentence instead of a sentiment classification because the score provides a more accurate value for comparison and experimentation. Since there are no labels available for this dataset, I had to use unsupervised techniques by using pre-trained models. Since the data is not domain specific, it is safe to use a pre-trained model that is trained on a general corpus (Intellica.AI, 2019). Therefore, I chose to use a few algorithms that will first calculate individual sentiment scores: Google Natural Language API, VADER, and TextBlob (Al-Natour et al., 2020, Intellica.AI, 2019; Terry-Jack, 2019). The algorithms will be further described in the Methodology section of this paper.

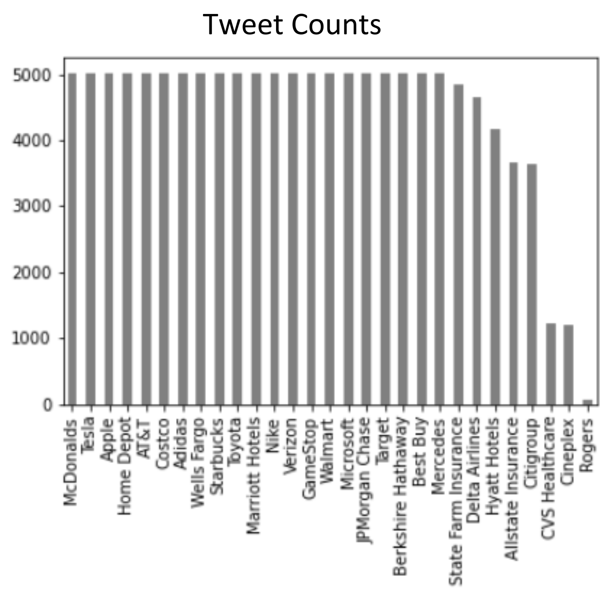
Alogrithm	Strengths	Weaknesses
Google Natural Language API	State-of-the-art transformer model and utilizes latest innovations.	Trained on general corpus, hence does not work well for domain-specific content.
VADER and TextBlob	Rule-based, bag of words approach, words with negated positive words, and captures subtle details.	Out of vocabulary words are not captured, i.e., if word is not included in training, it will receive no sentiment. May cause issues with social media dataset.

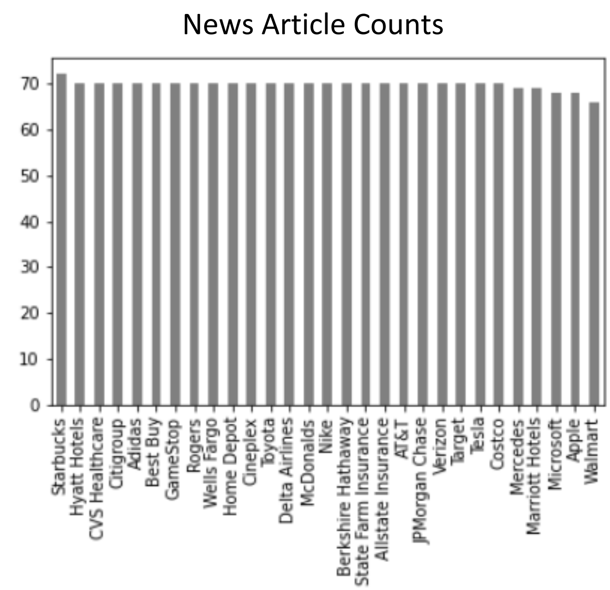
In the article “VADER, IBM Watson or TextBlob: Which is Better for Unsupervised Sentiment Analysis?” by Intellica.AI, a comparison of the three models was done in which they showcased that “some libraries work better at detecting positive sentiments, (while) others work better with negative (sentiments)”. To solve for this problem, the algorithms were then stacked in an ensemble to attempt to make predictions more accurate by accounting for flaws in each algorithm (Seni, 2010). This combination of algorithms was also utilized in the paper “A comparative assessment of sentiment analysis and star ratings for consumer reviews” by Al-Natour and Turetken (2020). Once sentiment scores were acquired, they were combined to represent the sentiment for each corporation by source. Using the sentiments by source, I was then able to experiment and test if any combinations of sources that define corporate sentiment have an impact on popularity or performance.

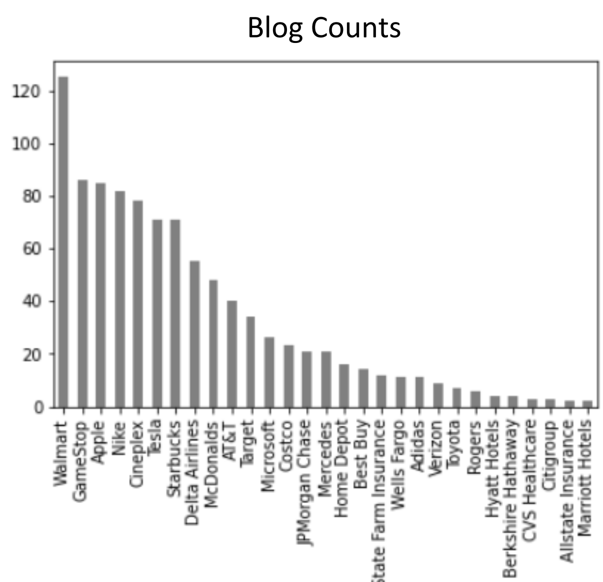
3.0 Exploratory Data Analysis

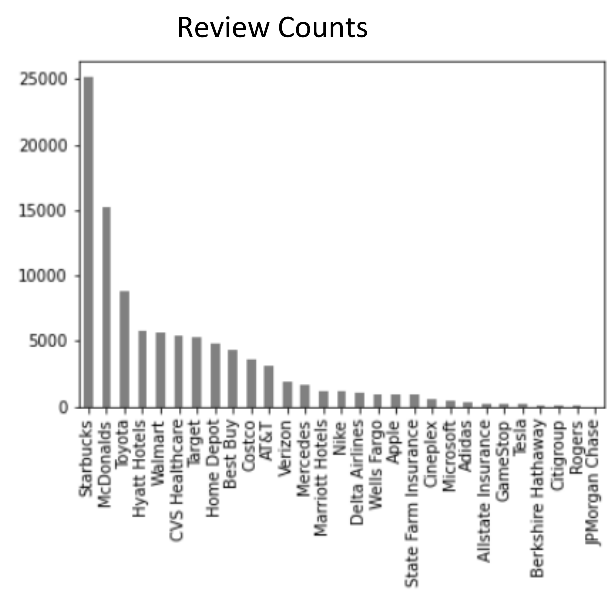
With the corporations selected, the data extracted, and the sentiment analysis techniques selected, it is important to make sure the extracted data is cleaned correctly for techniques to work effectively. Each data source went through similar cleaning procedures: removing any special characters, removing mentions, hashtags without # character were kept as people use them to write a sentence, removing new lines and tab keys, and removing stop words.

First, I checked the counts of content extracted from each source to understand the distribution. For some of the corporations, there seems to be a lack of presence, especially in the Yelp reviews when there is a lack of in person presence, and in the blogs as the selected blog didn't write as much about them. Also, Rogers Communications tweet results were quite low as the tweets are scattered across many hashtags and it is difficult to distinguish which tweets are "Rogers" and the name "Rogers".









To review the data cleaning was done effectively, I generated word clouds to make sure mentions, tags, special characters, and stop words were removed. Looking at the charts below, there are a few more words that have no importance to sentiment that must be removed, e.g., said shows up in two sources as one of the most utilized word but has no importance for sentiment analysis.

Tweet Wordcloud

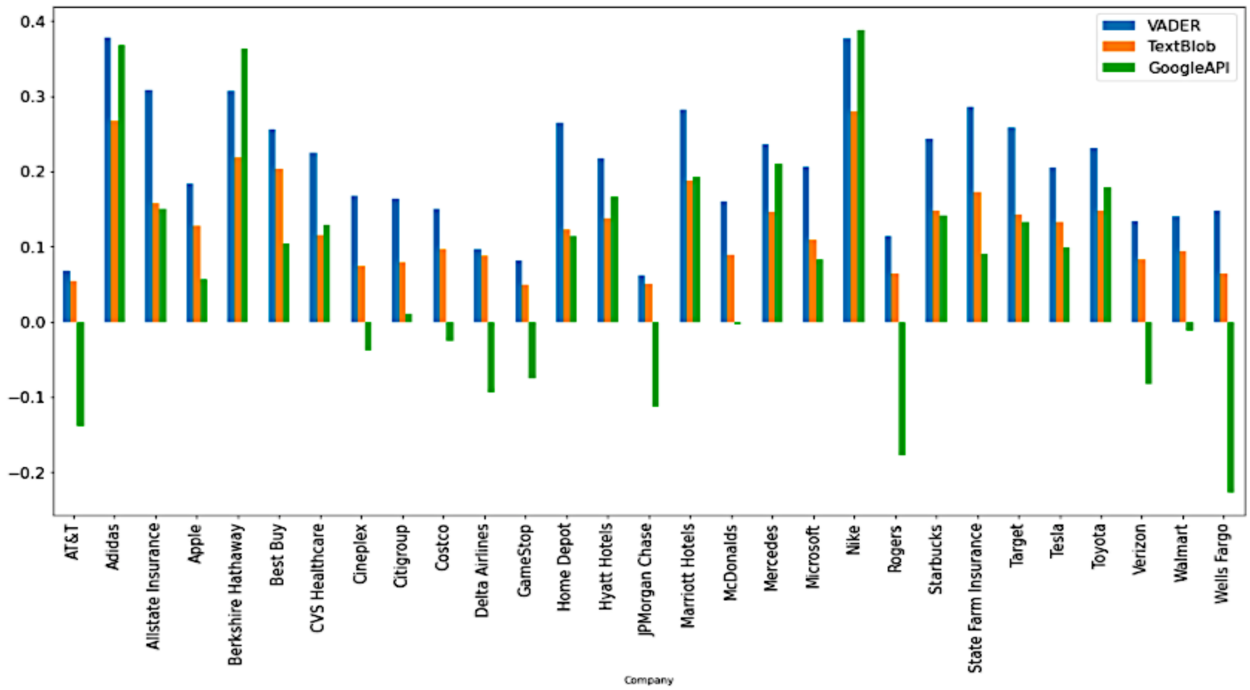
News Article Wordcloud

Blog Wordcloud

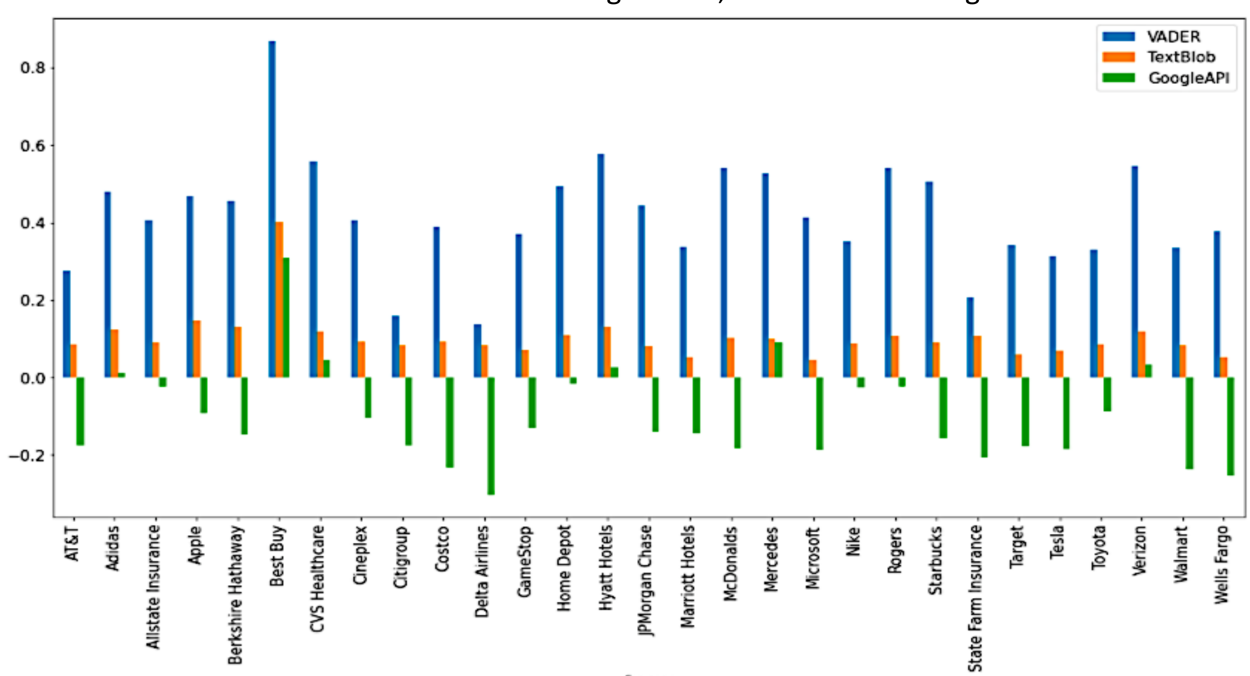
Review Wordcloud

After data cleaning was completed, sentiments were extracted using the three stated algorithms: VADER, TextBlob and Google Natural Language API. Below, I am reviewing the average sentiment extracted by the algorithms for each corporation for each source, as well as the scaled star ratings from Yelp.

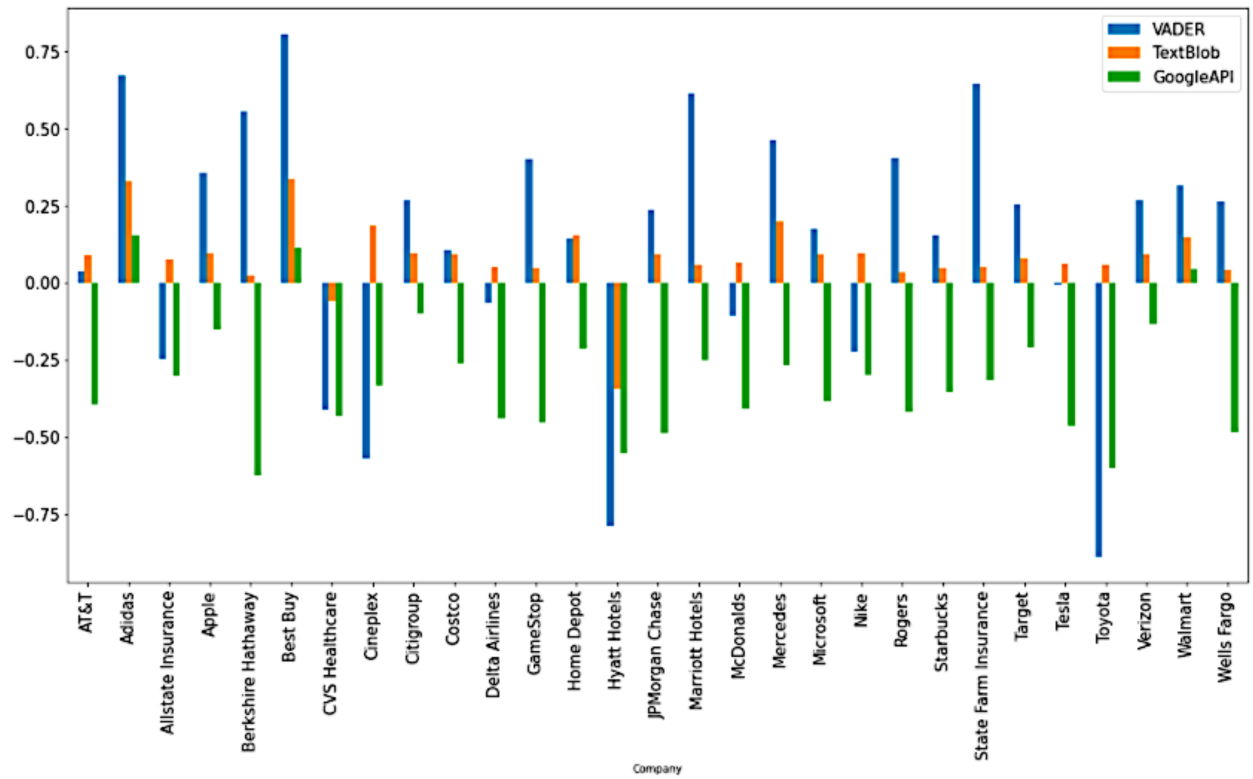
Tweets Extracted Sentiment using VADER, TextBlob and GoogleAPI



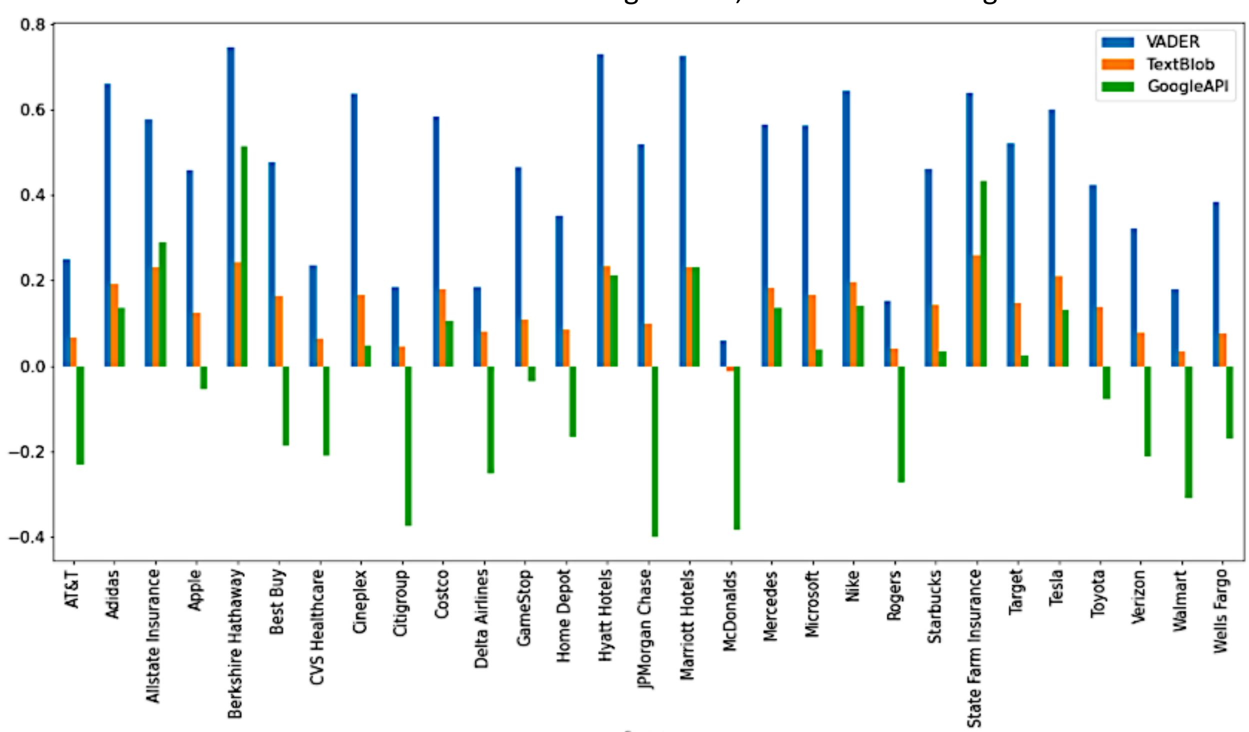
News Extracted Sentiment using VADER, TextBlob and GoogleAPI



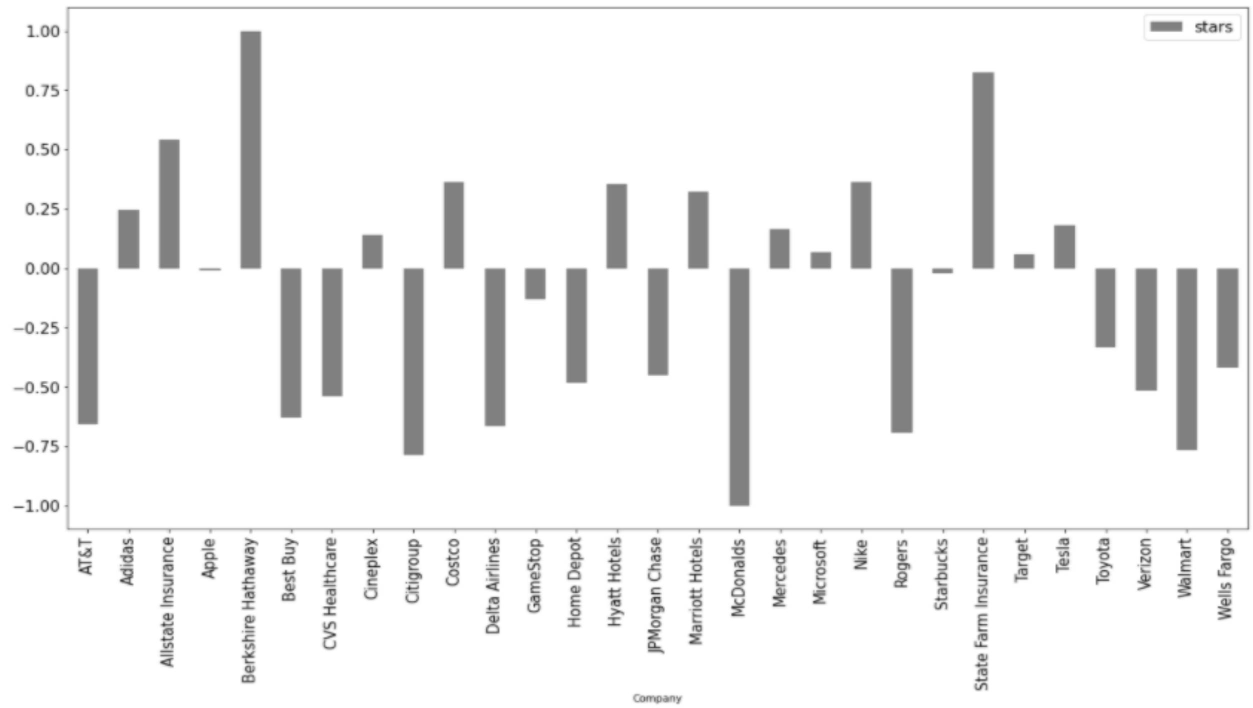
Blogs Extracted Sentiment using VADER, TextBlob and GoogleAPI



Reviews Extracted Sentiment using VADER, TextBlob and GoogleAPI



Scaled Yelp Star Ratings using Min-Max Scaler



The extracted sentiments using the three different algorithms does showcase that the GoogleAPI algorithm captured more negative sentiment than the rule-based VADER and TextBlob algorithms. Taking the average sentiment across the three algorithms may prove to be useful to capture the true corporate sentiment. Furthermore, the star ratings do seem to coincide with the extracted sentiments, specifically the sentiments extracted from reviews.

4.0 Methods and Experiments

4.1 Aim of Study

The aim of this study is to use the extracted sentiments from each of the sources to determine the optimal combination of sentiments that define corporate sentiment. The optimality will be determined by the magnitude and significance of the correlation with popularity and financial performance. By determining the optimal combination, I will identify the sources that efficiently define corporate sentiment and focusing on these platforms will help guide a corporation's future with a certain level of confidence.

4.2 Variables

In this experiment, there are only independent variables as this problem only requires unsupervised learning. The independent variables are:

1- Corporation name
2- Average Yelp star ratings
3- Average Yelp review sentiment

4- Average microblog (Twitter) sentiment
5- Average news summary sentiment
6- Average blog summary sentiment

4.3 Factors and Levels

In this experiment, the factors are the different algorithms being tested, i.e., VADER, TextBlob, and the Google Natural Language API. Since I am using pre-trained algorithms, the algorithms are already optimized using rules or a general corpus.

4.4 Experiments

1- Starting with just the sentiments extracted using the VADER algorithm from all sources, I took an average across the sources to determine the average corporate sentiment. Using this average, determine the correlation with the popularity of the corporation to identify if the generated corporate sentiment has a relationship with popularity.
2- Repeat experiment 1, but instead of determining the correlation with the popularity, determine the correlation with the financial performance indicators. This will help in identifying if the generated corporate sentiment has a relationship with financial performance.
3- Repeat experiments 1 and 2 using sentiments extracted using other algorithm combinations to identify the best available calculation for corporate sentiment by finding the combination that has the strongest correlation with popularity and financial performance.
4- Repeat experiments 1 through 3, isolating different sources using the combination of algorithms with the strongest correlation as determined after conducting experiment 3, e.g., use only microblogs and news summary instead of using all four sources. This will help determine if all sources are important in determining corporate sentiment or if any other combinations of sources are more effective in determining corporate sentiment.
5- After the above experiments for correlation, using a linear regression on the average corporate sentiment extracted from the four sources for all algorithms, R-squared will be derived to understand the degree to which the variability in the popularity, revenue growth and profit growth can be explained.

4.5 Measuring Performance

Performance for all available combinations of sentiments and star ratings will be measured using the correlation coefficient. If the variables are normally distributed, I will use the Pearson correlation and test the correlation for significance (parametric). If the variables are not normally distributed, I will use the Spearman correlation and test for significance (non-parametric).

5.0 Results

5.1 Experiment 1 to 3

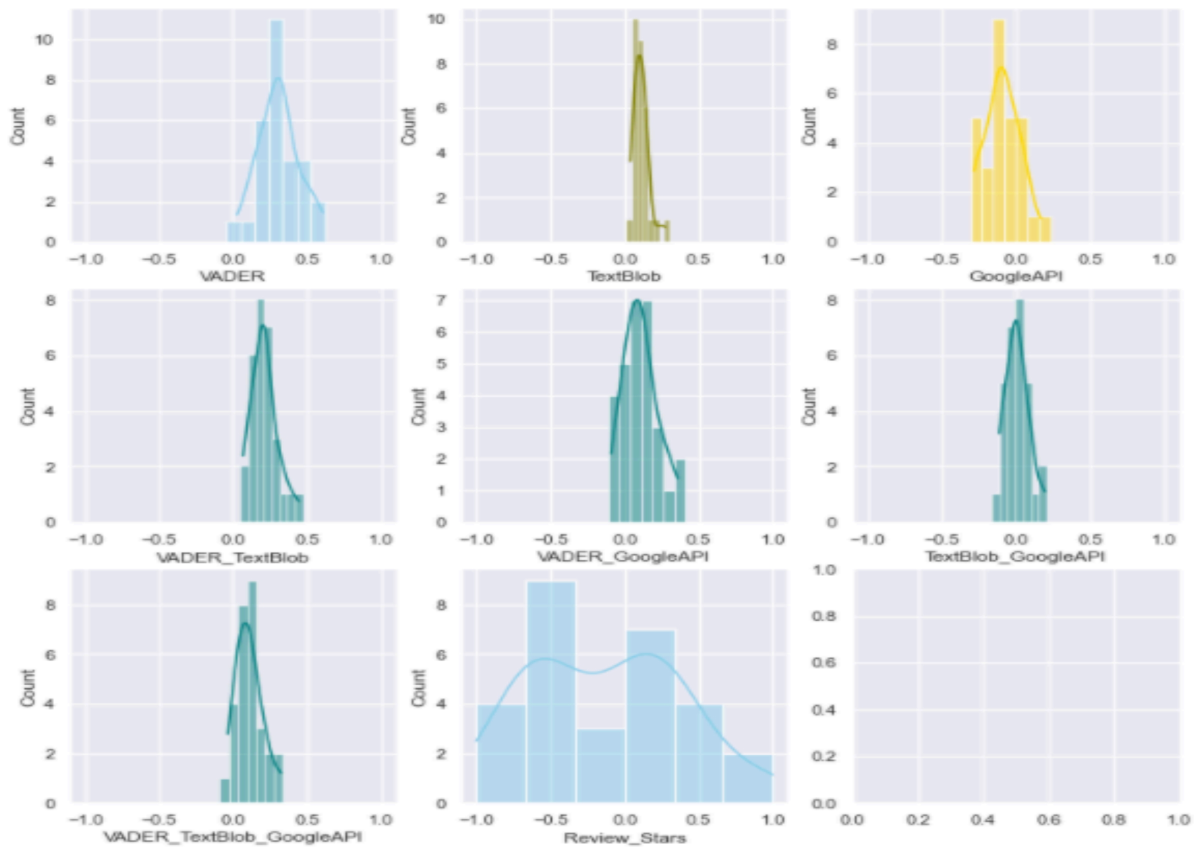
To run these experiments, I required seven combinations of the extracted sentiments (review star ratings were tested separately), as well as the star ratings for each source:

1- VADER only
2- TextBlob only
3- GoogleAPI only
4- VADER and TextBlob
5- VADER and GoogleAPI
6- TextBlob and Google API
7- VADER, TextBlob, and GoogleAPI

Once these combinations were generated, the combinations were averaged across each of the four sources to derive the corporate sentiment, i.e., the average of blog, news, reviews, and microblog sentiments were taken:

1- Corporate Sentiment using VADER
2- Corporate Sentiment using TextBlob
3- Corporate Sentiment using GoogleAPI
4- Corporate Sentiment using VADER and TextBlob
5- Corporate Sentiment using VADER and GoogleAPI
6- Corporate Sentiment using TextBlob and Google API
7- Corporate Sentiment using VADER, TextBlob, and GoogleAPI
8- Corporate Sentiment using scaled star ratings

Then, I created histograms to check the distribution of the eight possible corporate sentiments combinations, which I utilized to determine that every combination except the scaled star ratings were normally distributed including the metrics that will be used to compute the correlations (popularity, revenue growth, and profit growth). Therefore, I utilized Pearson correlation to generate the correlation between the normally distributed variables, and Spearman correlation for the scaled star ratings.
The results produced are shown below:



H0: The variables are uncorrelated, and the correlation shown is by chance

H1: The correlation between the variable is not 0

Popularity

	Algorithm	Metric	Correlation Method	Correlation	P-value
0	VADER	Popularity	Pearson	0.062536	0.747248
1	TextBlob	Popularity	Pearson	0.130708	0.499150
2	GoogleAPI	Popularity	Pearson	0.054575	0.778575
3	VADER_TextBlob	Popularity	Pearson	0.086241	0.656447
4	VADER_GoogleAPI	Popularity	Pearson	0.064572	0.739299
5	TextBlob_GoogleAPI	Popularity	Pearson	0.081434	0.674526
6	VADER_TextBlob_GoogleAPI	Popularity	Pearson	0.078690	0.684926
7	Review_Stars	Popularity	Spearman	0.030705	0.874366

Revenue Growth

	Algorithm	Metric	Correlation Method	Correlation	P-value
8	VADER	Revenue Growth	Pearson	0.202327	0.292536
9	TextBlob	Revenue Growth	Pearson	0.124694	0.519262
10	GoogleAPI	Revenue Growth	Pearson	0.048107	0.804281
11	VADER_TextBlob	Revenue Growth	Pearson	0.192440	0.317250
12	VADER_GoogleAPI	Revenue Growth	Pearson	0.144302	0.455167
13	TextBlob_GoogleAPI	Revenue Growth	Pearson	0.074825	0.699672
14	VADER_TextBlob_GoogleAPI	Revenue Growth	Pearson	0.144488	0.454579
15	Review_Stars	Revenue Growth	Spearman	-0.115067	0.552259

Profit Growth

	Algorithm	Metric	Correlation Method	Correlation	P-value
16	VADER	Profit Growth	Pearson	0.275915	0.147397
17	TextBlob	Profit Growth	Pearson	0.026778	0.890331
18	GoogleAPI	Profit Growth	Pearson	0.100464	0.604086
19	VADER_TextBlob	Profit Growth	Pearson	0.220834	0.249651
20	VADER_GoogleAPI	Profit Growth	Pearson	0.214319	0.264247
21	TextBlob_GoogleAPI	Profit Growth	Pearson	0.081265	0.675164
22	VADER_TextBlob_GoogleAPI	Profit Growth	Pearson	0.184914	0.336900
23	Review_Stars	Profit Growth	Spearman	0.131067	0.497961

From experiment 1 through 3, I noticed some correlation but there was no significance using the p-value, i.e., H0 could not be rejected. But, using the results, I decided to utilize the strongest correlation and lowest p-value algorithm for each metric to perform further testing to determine if the generated corporate sentiment values are capturing too much noise due to the combining step, i.e., not all sources are important. This might be due to the lack of data for some of the sources, hence the next experiment attempts to tackle that issue.

5.2 Experiment 4

To run this experiment (star ratings were ignored), I utilized the sentiment extracted using the TextBlob algorithm for the popularity metric, and the VADER algorithm for the revenue and profit growth metric. Using this, the following combinations of sources were generated to determine the corporate sentiment:

1- Blogs only
2- News only
3- Reviews only
4- Microblogs only
5- Blogs and News
6- Blogs and Reviews
7- Blogs and Microblogs

8- News and Reviews
9- News and Microblogs
10- Reviews and Microblogs
11- Blogs, News, and Reviews
12- Blogs, News, and Microblogs
13- Blogs, Reviews, and Microblogs
14- News, Reviews, and Microblogs
15- Blogs, News, Reviews, and Microblogs

Then, distributions were checked again, as done in experiments 1 through 3, and all generated corporate sentiments were normally distributed. Therefore, only Pearson correlation was utilized to perform the testing.

The results produced are shown below:

H0: The variables are uncorrelated, and the correlation shown is by chance
H1: The correlation between the variable is not 0

Popularity

	Algorithm	Metric	Correlation Method	Correlation	P-value
0	Blog_TextBlob	Popularity	Pearson	0.191469	0.319744
1	News_TextBlob	Popularity	Pearson	0.100638	0.603459
2	Review_TextBlob	Popularity	Pearson	-0.017870	0.926693
3	Tweet_TextBlob	Popularity	Pearson	-0.004906	0.979849
4	BN_TextBlob	Popularity	Pearson	0.190971	0.321029
5	BR_TextBlob	Popularity	Pearson	0.157324	0.415047
6	BT_TextBlob	Popularity	Pearson	0.153632	0.426218
7	NR_TextBlob	Popularity	Pearson	0.050008	0.796703
8	NT_TextBlob	Popularity	Pearson	0.060189	0.756446
9	RT_TextBlob	Popularity	Pearson	-0.013144	0.946046
10	BNR_TextBlob	Popularity	Pearson	0.164753	0.393074
11	BNT_TextBlob	Popularity	Pearson	0.158839	0.410513
12	BRT_TextBlob	Popularity	Pearson	0.118172	0.541513
13	NRT_TextBlob	Popularity	Pearson	0.031828	0.869807
14	TextBlob	Popularity	Pearson	0.130708	0.499150

Revenue Growth

	Algorithm	Metric	Correlation Method	Correlation	P-value
0	Blog_VADER	Revenue Growth	Pearson	0.289066	0.128292
1	News_VADER	Revenue Growth	Pearson	0.115057	0.552294
2	Review_VADER	Revenue Growth	Pearson	-0.176615	0.359405
3	Tweet_VADER	Revenue Growth	Pearson	0.082795	0.669389
4	BN_VADER	Revenue Growth	Pearson	0.300574	0.113119
5	BR_VADER	Revenue Growth	Pearson	0.183488	0.340705
6	BT_VADER	Revenue Growth	Pearson	0.295089	0.120174
7	NR_VADER	Revenue Growth	Pearson	-0.070987	0.714423
8	NT_VADER	Revenue Growth	Pearson	0.131920	0.495141
9	RT_VADER	Revenue Growth	Pearson	-0.108004	0.577071
10	BNR_VADER	Revenue Growth	Pearson	0.202162	0.292939
11	BNT_VADER	Revenue Growth	Pearson	0.303791	0.109127
12	BRT_VADER	Revenue Growth	Pearson	0.185439	0.335507
13	NRT_VADER	Revenue Growth	Pearson	-0.035190	0.856192
14	VADER	Revenue Growth	Pearson	0.202327	0.292536

Profit Growth

	Algorithm	Metric	Correlation Method	Correlation	P-value
15	Blog_VADER	Profit Growth	Pearson	0.361057	0.054317
16	News_VADER	Profit Growth	Pearson	0.052393	0.787222
17	Review_VADER	Profit Growth	Pearson	-0.103365	0.593628
18	Tweet_VADER	Profit Growth	Pearson	0.139222	0.471357
19	BN_VADER	Profit Growth	Pearson	0.346576	0.065500
20	BR_VADER	Profit Growth	Pearson	0.277537	0.144935
21	BT_VADER	Profit Growth	Pearson	0.375575	0.044670
22	NR_VADER	Profit Growth	Pearson	-0.050201	0.795934
23	NT_VADER	Profit Growth	Pearson	0.108029	0.576980
24	RT_VADER	Profit Growth	Pearson	-0.032856	0.865640
25	BNR_VADER	Profit Growth	Pearson	0.271325	0.154526
26	BNT_VADER	Profit Growth	Pearson	0.358072	0.056489
27	BRT_VADER	Profit Growth	Pearson	0.282836	0.137103
28	NRT_VADER	Profit Growth	Pearson	-0.002444	0.989959
29	VADER	Profit Growth	Pearson	0.275915	0.147397

From experiment 4, the following conclusions were derived:

- Popularity had a weak positive correlation, the strongest was produced using corporate sentiment derived by blogs (Blog_TextBlob) but H0 could not be rejected. Hence, there is noise in the data and the correlation occurs by chance.
- Revenue Growth had a weak positive correlation, the strongest was produced using the corporate sentiment derived by blogs and news (BN_VADER) but H0 could not be rejected. Hence, there is noise in the data and the correlation occurs was by chance.
- Profit Growth had a weak positive correlation (0.38), the strongest was produced using the corporate sentiment derived by blogs and microblogs (BT_VADER) and H0 was rejected as the p-value was lower than 0.05 (0.045). Hence, corporate sentiment

derived using blogs and microblogs has a weak positive correlation with profit growth, i.e., as sentiment in blogs and microblogs goes up, so does profit growth.

5.3 Experiment 5

This experiment was run using a simple ordinary least squares linear regression model from the scikit-learn library with the 29 corporations and the following combinations of corporate sentiment, averaging the extracted sentiments across the four sources as the independent variables, and the popularity, revenue growth and profit growth as the dependent variable:

1- Corporate Sentiment using VADER
2- Corporate Sentiment using TextBlob
3- Corporate Sentiment using GoogleAPI
4- Corporate Sentiment using VADER and TextBlob
5- Corporate Sentiment using VADER and GoogleAPI
6- Corporate Sentiment using TextBlob and Google API
7- Corporate Sentiment using VADER, TextBlob, and GoogleAPI

The following result table was derived using the OLS Linear Regression:

	Metric	Algorithm	Coefficients_1	Coefficients_2	Coefficients_3	Coefficients_4	Intercept	R-squared
0	Popularity	VADER	4.782414	-3.049429	-1.215380	-13.592469	14.751868	0.031155
1	Popularity	TextBlob	22.940080	13.269045	11.705868	-24.573296	9.025641	0.044024
2	Popularity	GoogleAPI	30.456155	-29.376375	4.050239	-5.374802	17.769775	0.189167
3	Popularity	VADER_TextBlob	9.621692	-1.329914	-1.060985	-13.735002	12.675843	0.033847
4	Popularity	VADER_GoogleAPI	12.138603	-12.675155	-1.830976	-4.056277	14.812047	0.071497
5	Popularity	TextBlob_GoogleAPI	38.046872	-28.709035	8.903879	-14.554812	16.105303	0.137853
6	Popularity	VADER_TextBlob_GoogleAPI	16.603908	-12.951892	-0.463057	-7.786412	14.244795	0.069539
7	Revenue Growth	VADER	0.167427	0.110472	-0.434470	0.722613	-0.107714	0.169418
8	Revenue Growth	TextBlob	0.451678	0.017622	-0.692116	0.449564	-0.092092	0.084600
9	Revenue Growth	GoogleAPI	0.228692	-0.108279	-0.239982	0.288660	-0.057248	0.076686
10	Revenue Growth	VADER_TextBlob	0.285718	0.085883	-0.577176	0.698080	-0.087756	0.154004
11	Revenue Growth	VADER_GoogleAPI	0.259090	-0.019981	-0.401594	0.543399	-0.052430	0.154558
12	Revenue Growth	TextBlob_GoogleAPI	0.364941	-0.135007	-0.352637	0.371084	-0.066118	0.086746
13	Revenue Growth	VADER_TextBlob_GoogleAPI	0.334214	-0.038721	-0.469643	0.567656	-0.061688	0.146042
14	Profit Growth	VADER	2.613503	-0.327295	-4.478272	10.035134	-0.943716	0.196151
15	Profit Growth	TextBlob	-4.948603	0.467844	-11.912539	20.551587	-1.365752	0.087286
16	Profit Growth	GoogleAPI	0.011166	-1.020656	-2.190781	5.661825	-1.285102	0.042549
17	Profit Growth	VADER_TextBlob	3.338909	-1.440986	-6.202869	12.619795	-0.965714	0.142465
18	Profit Growth	VADER_GoogleAPI	3.189643	-1.665090	-3.701886	8.338282	-0.529978	0.141156
19	Profit Growth	TextBlob_GoogleAPI	-1.690439	-0.655055	-4.520598	9.850247	-1.649012	0.054462
20	Profit Growth	VADER_TextBlob_GoogleAPI	3.208294	-2.125069	-4.538292	9.936454	-0.804004	0.109912

From experiment 5, the following conclusions were derived:

- 18.9% of the variability in popularity can be explained by sentiments extracted from the four sources using the GoogleAPI algorithm.
- 16.9% of the variability in revenue growth can be explained by sentiments extracted from the four sources using the VADER algorithm.

- 19.6% of the variability in profit growth can be explained by sentiments extracted from the four sources using the VADER algorithm.

5.4 Discussion

This project aimed to generate corporate sentiment using various online sources. To run consistent experiments, sampling was conducted using three goods and services categories: goods, services, and credence goods. In addition, for consistency, it was required for each selected corporation to have presence within each of the selected sources: news, blogs, microblogs, and reviews. This restriction implied that corporations with only an online presence, e.g., Shopify and Instagram, were not included as corporations withing the dataset. Furthermore, it was difficult to generate an even sample across all three good and services categories, especially the credence goods category. Also, due to a lack of presence of a few of the selected corporations within certain sources, the amount of data available varied. Lastly, due to the resources required to extract data from the selected sources, limited data was utilized to generate the corporate sentiment.

These restrictions on the sample may imply that the results don't apply to all types of corporations. To solve for these issues, more resources would be required to extract a larger dataset for more corporations from more data sources, e.g., Google Reviews, across the three goods and services categories. Also, from the experiments run in the project, it was shown that the blogs and microblogs had significant correlation with profit growth, which can now be further tested using corporations that have a presence in only these two sources, e.g., Shopify and Instagram.

To apply these experiments to other corporations, it would be beneficial to reduce the restriction of presence within sources to increase diversity in the selected corporations, extract a larger dataset, and test the hypotheses outside of the pandemic scenario. Since the sentiment algorithms were unsupervised, it would be useful to include an algorithm that is trained on a social media corpus to extract more accurate sentiment from social media (microblogs) platforms.

6.0 Conclusion and Future Work

For further research, it would be essential to test the significant correlation against another sample of corporations and compare the results with this project. Using additional sources for consumer opinions might require different algorithms to be utilized to extract the sentiments. The addition of sources and corporations from different industries with different presence within sources might not produce optimal results, but this can only be confirmed with further work.

Also, the experiment performed to find the degree to which the variability in popularity and the financial performance indicators can be utilized to train machine learning model to predict these metrics. Due to the lack of data available in this project, this was not possible as the models were susceptible to overfitting. Therefore, with additional corporations included in the sample, this would be a great idea to pursue. Since approximately 20% of the variability is explained with the corporate sentiment, this might require additional features to be included in the dataset to generate a robust model.

One of the introductory points mentioned in this project was that the recent pandemic might have an impact on how online consumer opinions impact a corporation’s performance. But this provided financial performance metrics that did not represent corporations correctly, e.g., profit growth of approximately -200%. It would be interesting to test these hypotheses with data outside of the pandemic, i.e., post-2021 or pre-2019.

Lastly, if a financial performance indicator that is updated at specific points in time over the year was included as a dependant variable, there would be the possibility of testing if corporate sentiment has a delayed or an instant effect on the financial performance indicator. This can be setup by isolating data from the four sources into quarters of the year and then compare the corporate sentiment of each quarter against the financial performance indicator at different quarters over the year. For example, the corporate sentiment for Q1 of 2020 would be compared to the financial performance of Q1, Q2, Q3 and Q4 of 2020 to test if the calculated sentiment has an instant or delayed reaction.

7.0 References

Bulbul, C., Gross, N., Shin, S., & Katz, J. (2014). When the path to purchase becomes the path to purpose. Think with google.

Chen, Y., & Xie, J. (2008). Online consumer review: Word-of-mouth as a new element of marketing communication mix. Management Science, 54(3), 477–491

Cheung, C. M., & Lee, M. K. (2012). What drives consumers to spread electronic word of mouth in online consumer-opinion platforms. Decision Support Systems, 53(1), 218–225

Dang, Y., Zhang, Y., & Chen, H. (2010). A lexicon-enhanced method for sentiment classification: An experiment on online product reviews. IEEE Intelligent Systems, 25(4), 46–53

Hu, Y.-H., & Chen, K. (2016). Predicting hotel review helpfulness: The impact of review visibility, and interaction between hotel stars and review ratings. International Journal of Information Management, 36(6), 929–944

Pavlou, P. A., & Dimoka, A. (2006). The nature and role of feedback text comments in online marketplaces: Implications for trust building, price premiums, and seller differentiation. Information Systems Research, 17(4), 392–414

Greco, F., & Polli, A. (2019). Emotional text mining: Customer profiling in brand management. International Journal of Information Management

Jeong, B., Yoon, J., & Lee, J.-M. (2017). Social media mining for product planning: A product opportunity mining approach based on topic modeling and sentiment analysis. International Journal of Information Management, 48, 280–290

Rathore, A. K., & Ilavarasan, P. V. (2020). Pre-and post-launch emotions in new product development: Insights from twitter analytics of three products. International Journal of Information Management, 50, 111–127

Maqsood, H., Mehmood, I., Maqsood, M., Yasir, M., Afzal, S., Aadil, F., Selim, M., & Muhammad, K. (2020). A local and global event sentiment based efficient stock exchange forecasting using deep learning. International Journal of Information Management, 50, 432–451

Al-Natour, S., & Turetken, O. (2020, October 1). A comparative assessment of sentiment analysis and star ratings for consumer reviews. ScienceDirect. https://www.sciencedirect.com/science/article/abs/pii/S0268401219311181

Campbell, K. (2021, March 6). Measuring Corporate Sentiment. Reputation X. https://blog.reputationx.com/measuring-corporate-sentiment

About: SEC classification of goods and services. (n.d.). DBpedia. Retrieved June 17, 2021, from https://dbpedia.org/page/SEC_classification_of_goods_and_services

Adams, R. L. (2017, March 2). 10 Wildly-Successful Blogs That Earn Outlandish Incomes. Forbes. https://www.forbes.com/sites/robertadams/2017/03/02/top-income-earning-blogs/?sh=70f123372377

Medical, B. (2019, November 26). 11 Yelp Statistics For 2020 You Need To Know As A Business Owner. Boost Medical. https://boostmedical.com/yelp-statistics/#:%7E:text=A%20study%20from%202017%20found,influential%2C%20and%20most%20used.%E2%80%9D

Densmore, J. (2019, July 23). Ethics in Web Scraping - Towards Data Science. Medium. https://towardsdatascience.com/ethics-in-web-scraping-b96b18136f01

Paul, T. (2020, December 29). Is web scraping legal? A guide for 2021. LinkedIn. https://www.linkedin.com/pulse/web-scraping-legal-guide-2021-tony-paul/

Seni, G. (2010). Ensemble Methods in Data Mining: Improving Accuracy Through Combining Predictions. Morgan and Claypool Publishers. https://www.morganclaypool.com/doi/abs/10.2200/S00240ED1V01Y200912DMK002

Intellica.AI, I. (2019, September 19). VADER, IBM Watson or TextBlob: Which is Better for Unsupervised Sentiment Analysis? Medium. https://intellica-ai.medium.com/vader-ibm-watson-or-textblob-which-is-better-for-unsupervised-sentiment-analysis-db4143a39445

Terry-Jack, M. (2019, May 2). NLP: Pre-trained Sentiment Analysis - Mohammed Terry-Jack. Medium. https://medium.com/@b.terryjack/nlp-pre-trained-sentiment-analysis-1eb52a9d742c