A Natural Language Analysis of Warren Buffett's Annual Letter to Shareholders

Vatsal Nahata
vatsal.nahata@yale.edu

4 May 2020

Abstract

This study uses Python's various natural language processing libraries, including Sentiment Analysis to analyse Warren Buffett's annual shareholders letter in order to uncover insights about his investing strategy. The study finds that his outlook has been fairly neutral, that it has a good balance of subjectivity facts and is highly stable despite fluctuations in market prices and sentiment over the period 1977-2019. It also uncovers his strategy as being driven by value investing and buying companies based on their intrinsic valuations. It finds that long term investing and the investment hypothesis of "buying successful businesses" rather than taking high risk-reward bets has been a big contributor to his financial success.

1 Introduction

In the tumultuous and unnerving world of investing, it is rare to find someone who an aspiring investor can look up to. This is precisely where the motivation for my project comes from. Ever since my teenage days, I have carefully followed Warren Buffett, who is arguably the most successful investor the stock markets have ever seen.

His net worth currently stands at a staggering $73.5 Billion and continues to grow healthily. What is most astounding is the fact that he uses simple, easy to understand and time tested strategies of investing and never invests in companies he does not understand. He also mostly deals in equities. This is a far cry from what most hedge funds and private equity investors do today; they employ the best of mathematicians, build the most intricate of models and deal in the most exotic of derivatives. Buffett has consistently beaten all of them over the past three decades.

In this study, I use Python's various libraries to do a textual and graphical exposition of how Buffett thinks. His annual letter to Berkshire Hathaway's shareholders is hailed as the Bible for investors. This paper takes text data from his letters spanning 1977-2019 and applies Python's natural language processing techniques in order to get investment insights and to ultimately understand the man himself[1].

2 Data & Methodology

The data for this study has been web-scraped from Berkshire Hathaway's website comprising 43 letters spanning 1977-2019 using Selenium's Chrome web driver. In total, I had the opportunity to analyse 2.96 Million words[2]. I first created a dictionary with each year as the key and the corresponding letter as its value. Subsequently, I cleaned the letter (the primary data) by getting rid of all numbers, special characters, punctuation's and line spaces using Python's regular expressions library (re library). I then tokenized the text and removed a long list of stopwords such as prepositions, conjunctions, simple nouns and words with no meaning etc. I then converted all data into title case. All the above were done using Python's NLTK library.

Once the text was cleaned, I created a Pandas DataFrame which had each year as the index and the letter variable as a column (which included all of Buffett's letters from 1977-2019). This provided me with a lot of flexibility in doing Basic sentiment analysis as well as finding out the frequency distribution of common words per decade. At times, I also found that storing all 2.96 Million words (the entirety of his letters) in a single string was helpful. All libraries used in this study have been mentioned in the Reference section.

The Results section lay out the various ways that I have presented my findings. To generate the findings, I crunched my data in three forms: (i) I found out common words by using NLTK's (Natural Language Toolkit) FreqDist() function and used Matplotlib to plot these; (ii) in order to do Sentiment Analysis, I ran a loop over my DataFrame to separately get sentiment scores for each year's letter in terms of polarity and subjectivity; (iii) Finally for lexical dispersion, I used inbuilt NLTK functions which allowed me to plot lexical graphs. I also wanted to compare common words and bigrams by decade but that would have extended the length of this study to far too long without adding immense value to the overall study.

[1] I would like to thank Professor Casey King, Flynn Chen and Nick Marwell for their inputs and support on this project.

[2] Data has been scraped from https://www.berkshirehathaway.com/letters/letters.html

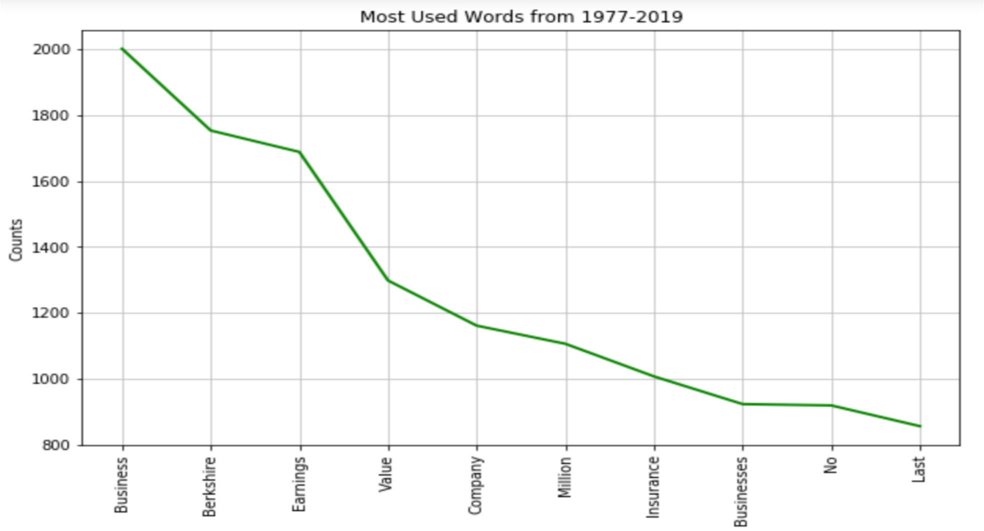
3 Results

Results are presented in the following formats: (i) A simple frequency plot and pie chart of the most common words and bi-grams across the 1997-2019 period is undertaken. A word cloud of the most frequent thirty words is also shown; (ii) A basic sentiment analysis using TextBlob for polarity and subjectivity scores for every year in the sample, plotted against the annual return provided by the Dow Jones Index to show how Buffett's behaviour has changed compared to volatility in Market prices and sentiment; (iii) A Lexical Dispersion Plot which tells us the intensity with which a particular word has been used over time.

3.1 Common Words, Bigrams and Word Clouds

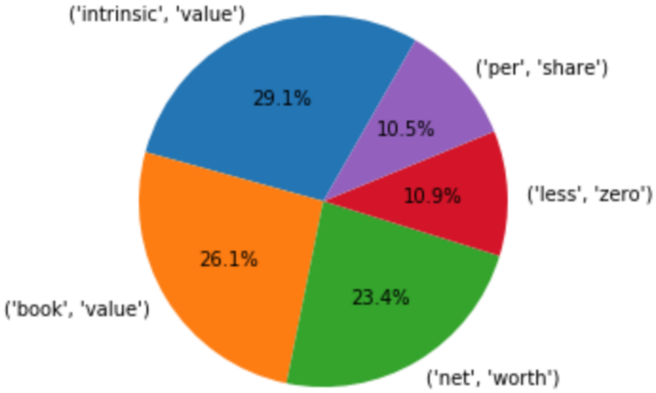
In this section, I give a very brief and introductory overview of the most frequent words and bigrams used in the form of a frequency plot, a pie chart and a word cloud.

Figure A



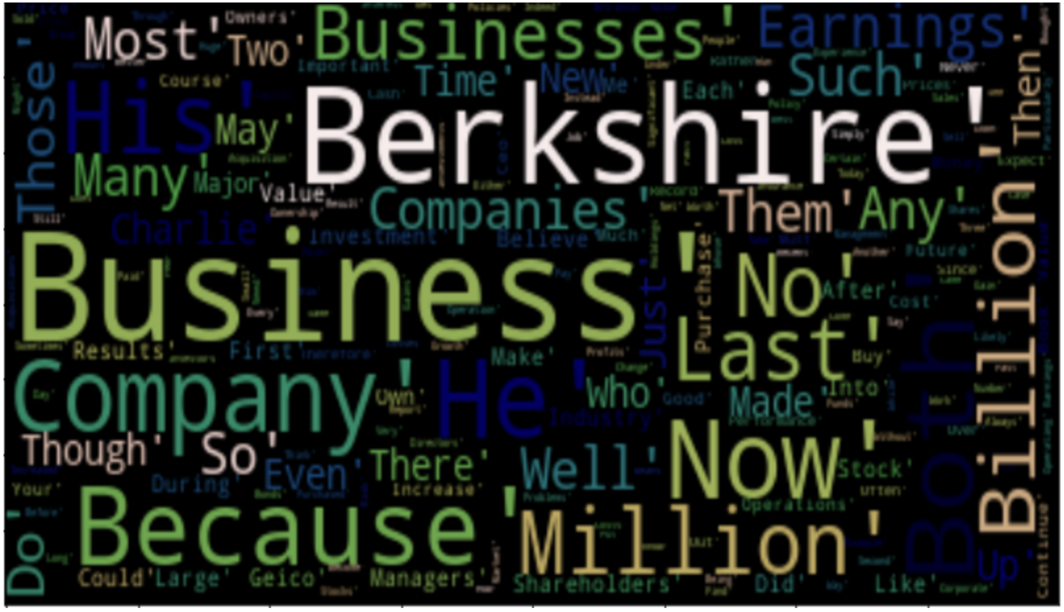
I find that words such as 'Business', 'Berkshire', 'Value' and 'Earnings' are the most frequently used words implying that Buffett's investment strategy is fairly simple, straightforward and pillared around very fundamental concepts. He goes after companies that have high earnings growth and uses metrics such as the Price to Book Ratio to assess the Intrinsic Value of a company. When it comes to his use of bigram words, we find that intrinsic-value' features heavily over 43 years.

Figure B



This reflects the core of his investment strategy. Because of his focus on Intrinsic Value, other bigrams like net-worth, per-share, book-value and less-zero also tie into how Buffett thinks about Intrinsic Value. Intrinsic Value is the discounted value of the cash that can be taken out of a business during its remaining life. Buffett follows the Benjamin Graham school of value investing, which looks for securities whose prices are unjustifiably low based on their intrinsic worth. Rather than focus supply and demand intricacies of the stock market, Buffett looks at companies as a whole. Value investors look for securities with prices that are unjustifiably low based on their intrinsic worth. There isn't a universally accepted way to determine intrinsic worth, but it's most often estimated by analyzing a company's fundamentals. Like bargain hunters, the value investor searches for stocks believed to be undervalued by the market, or stocks that are valuable but not recognized by the majority of other buyers.

Figure C



The Word Cloud tells us in an eye-catching manner about the relative importance of each word based on intensity, consistent use and frequency. While there are simple and common words such as 'Though', 'Because', 'So' etc. (because there are 30 words in the word cloud), a deeper look closely mirrors what we say in Figure's A & B. Warren Buffett's investment philosophy has evolved over the last 43 years to focus almost exclusively on buying high quality companies with promising long-term opportunities for continued growth. Some investors might be surprised to learn that the name Berkshire Hathaway comes from one of Buffett's worst investments. Two of Buffett's famous quotes sum it all up for the Word Cloud: "If you aren't thinking about owning a stock for ten years, don't even think about owning it for ten minutes.";"Our favorite holding period is forever." – Warren Buffett.

3.2 Sentiment Analysis

We can see from the graphs below that remarkably, Warren Buffett's sentiment has largely been stable and constant over a period of 43 years.

Figure 1

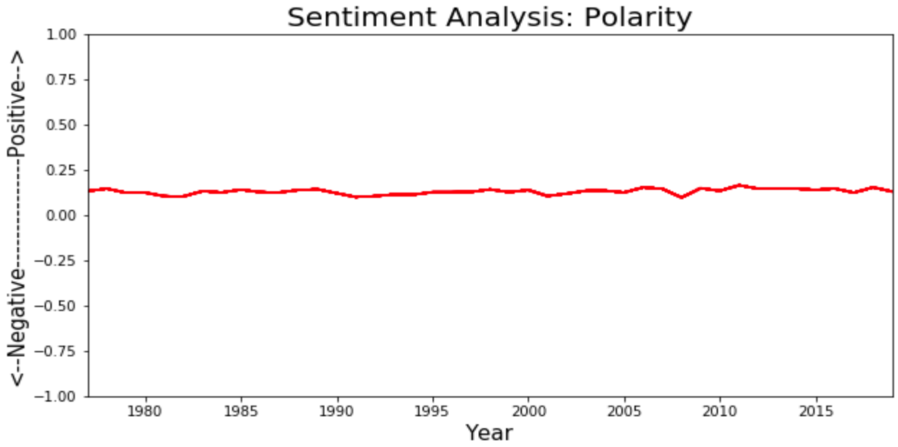


Figure 2

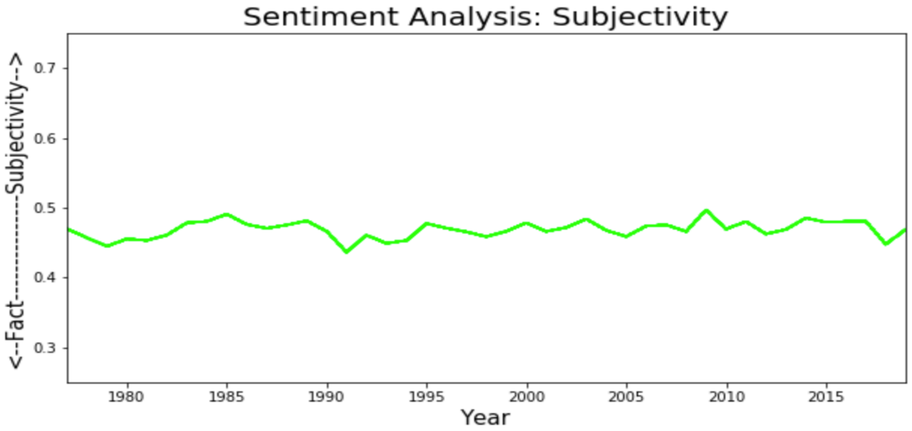
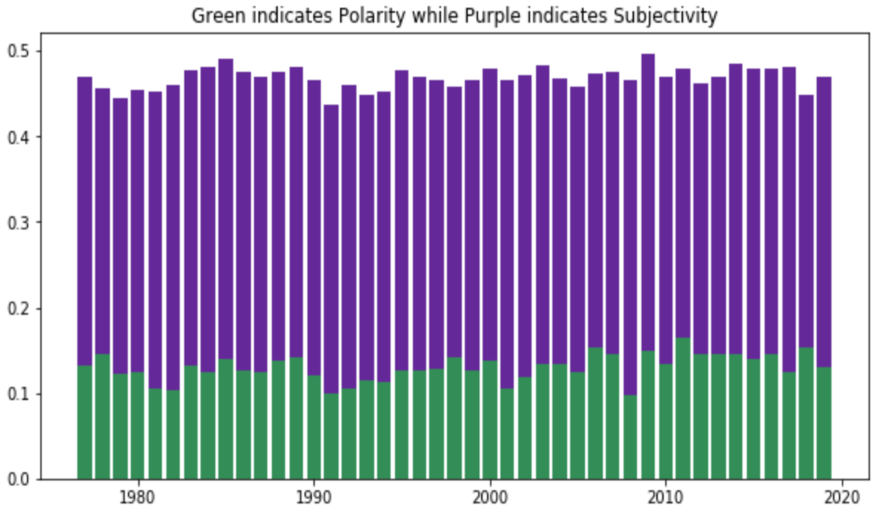


Figure 3



His polarity score (defined as how positive or negative his sentiment is: -1 being the most negative and +1 being the most positive) has hovered between between 0.09-0.16 while his subjectivity score (defined as how opinionated or fact based his sentiments are: 0 being purely fact based while 1 being very subjective or opinionated) has been within 0.43-0.49.

This indicates a very high degree of stability in his investment perspective. It also tells us that his sentiment does not change depending on how the market performs which is extremely remarkable. This is the very first insight we can draw from his mind and indicates the mental characteristics one should possess in order to succeed in the market. Equanimity is Warren Buffett's prime virtue, not just in terms of being positive or negative, but also in terms of having a healthy mix of facts and subjectivity in his opinions. His Polarity has always been neutral to slightly positive while his thinking is grounded in a good mix of both facts and subjectivity, based on how he has scored. These score were derived from TextBlob but for the sake of Robustness, I also ran sentiment scores on Python's VaderSentiment module and found that his neutrality score was always between 0.77-0.80.

I next analyse whether there has been any correlation between how the market has performed over the years versus how his sentiments have behaved. I take the annual return in the Dow Jones as a proxy for market sentiment rather than taking the level of the Dow Jones since that has tended to rise over time. Again, I find that he has been remarkably stable compared to the market and has never let the market get the better of him. The Pearson correlation coefficient between the annual return on the Dow Jones and his polarity scores is 0.005 while the same for the annual return on the Dow Jones and his subjectivity scores is 0.035. To further check for robustness and to further validate my results, I calculate the correlation between the rate of change in his polarity score and the annual return given by the Dow Jones in order to perhaps find how his sentiment has changed *over time* vis-a-vis the market.

Here too, the Pearson correlation coefficient for polarity is -0.098 while for subjectivity it is -0.092. These results clearly suggest that he has hardly changed his attitude and sentiment with respect to the market and speaks volumes about his temperament. The following charts also highlight this.

Figure 4

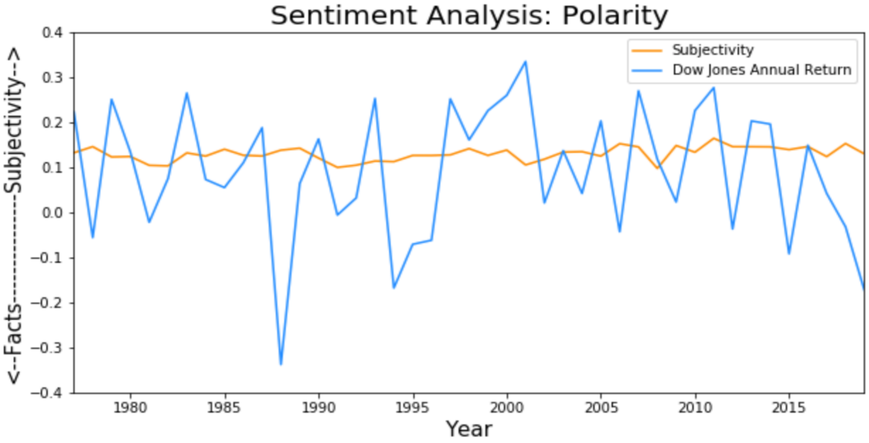


Figure 5

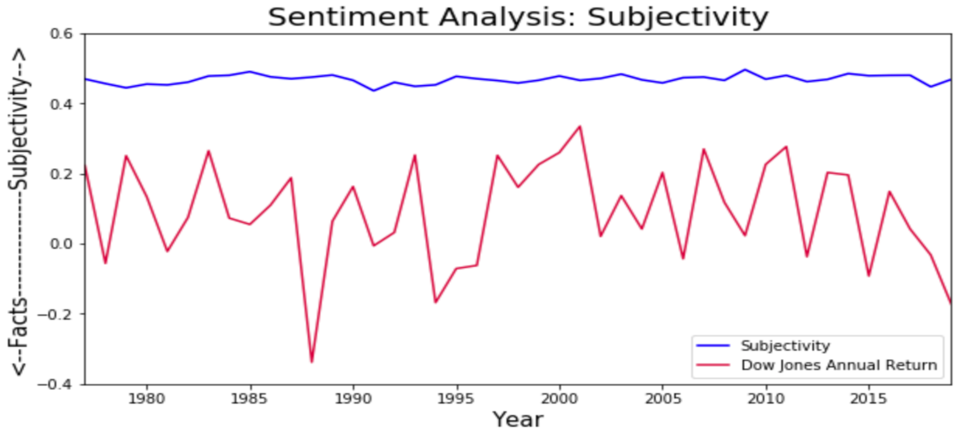


Figure 6

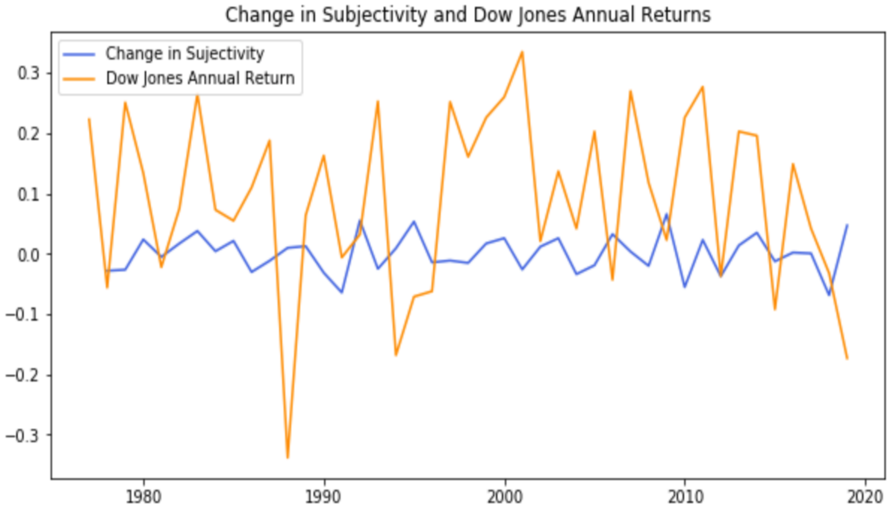
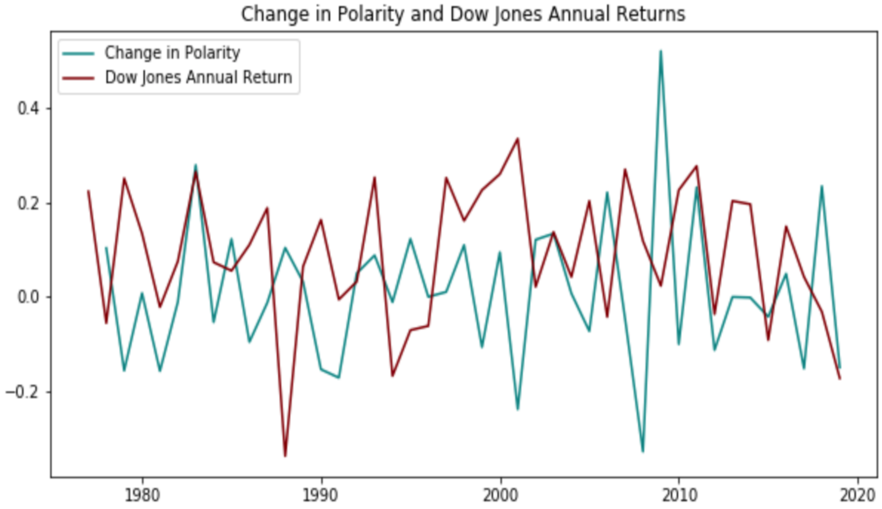
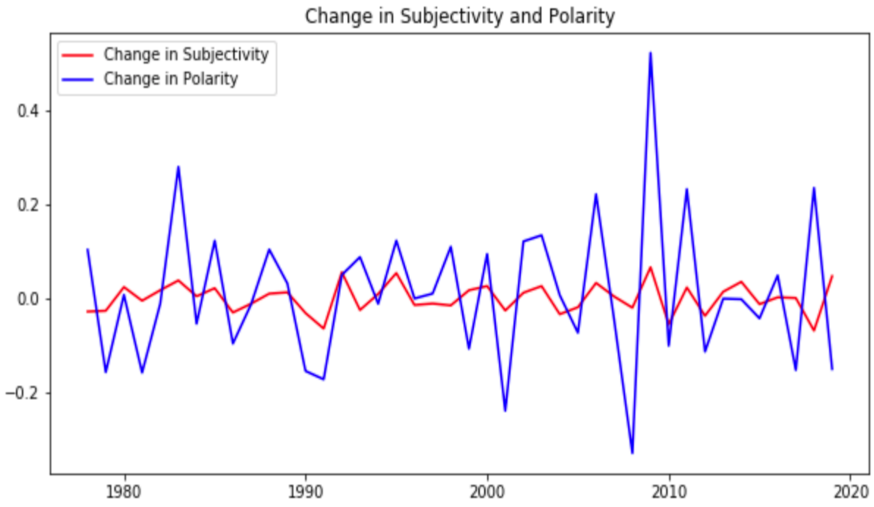


Figure 7



Another interesting finding is that his positivity has been more stable than his subjectivity. This is understandable because subjective traits are generally harder to preserve over time, especially in an industry that is very numbers driven and where information is key. This is highlighted by Figure 8.

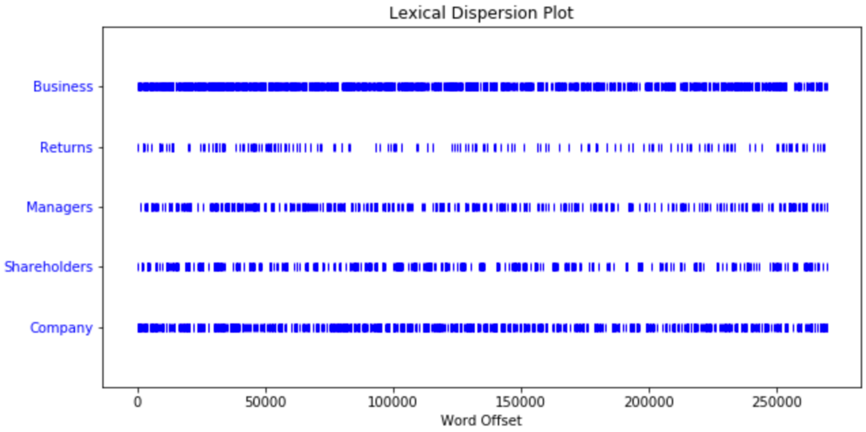
Figure 8



3.3 Lexical Dispersion

In this section we try to check for the constancy and intensity of certain key words. These are meant to tell us the importance of that word over a period of time and indicates how frequently that word was used over the 43 year period of this study. I divide my words into four categories: (i) Words that signal his priorities from the business or firm side such as 'Manager', 'Shareholders', 'Business' etc. (ii) Words that give out his fundamental investment philosophy such as 'Value','Intrinsic','Capital' etc. (iii) Words that are emotional inputs or reactions to his experience in the Stock Market such as 'Happy', 'Rational','Bomb' etc. and (iv) Miscellaneous but important words like 'Million','Billion','Charlie' etc.

Figure 9



In Figure 9, we find that he has been (since 1977) using the word 'Business' and 'Company' very frequently and consistently. This tells us that he has always thought of buying shares in the market as essentially buying a business. This has a deeper meaning than what one concludes from first impression. His philosophy has been to think of buying a share exactly in the same way one would think when buying a property or a car.

There is a lot of due diligence done when buying a property because the asset stays with a person for a much longer period of time and the consequences of any foolish decision would stick around for a very long time. He thinks of buying equities as exactly an exercise in that. Doing thorough background checks and investing in high quality and sustainable businesses has been a key feature of his investment strategy. Surprisingly, the word 'Returns' has not been as intensely used indicating perhaps that he does not like to boast about or think too much about the returns his investments are making.

Figure 10

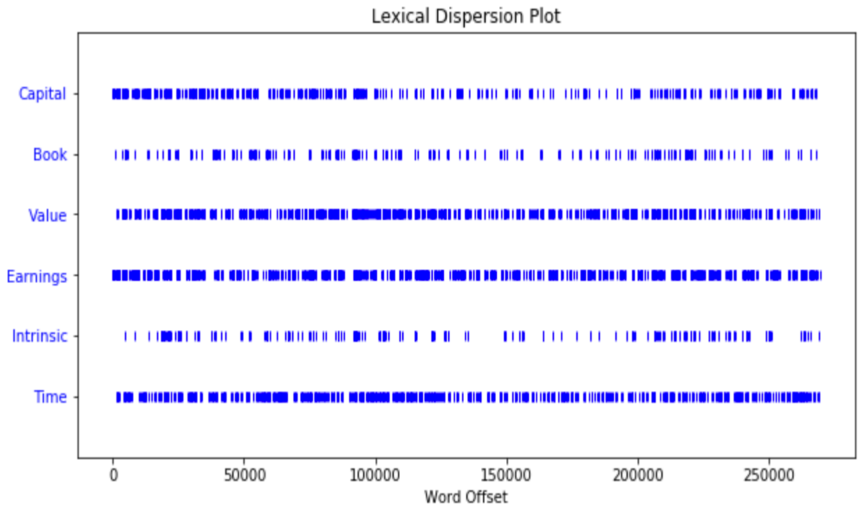


Figure 10 tells us that Words like 'Value', 'Earnings' and 'Time' have been the cornerstone of his investment strategy. He is a strong believer in the power of compound interest which is how 'Time' has shaped his choice of businesses. There is no point in buying a business if it is not going to last for at least 3 decades into the future. Similarly, there is no point in investing if it is not for the long term. The extremely dense and consistent use of the word 'Value' since 1977 indicates how in a world where investors frequently change themselves to adapt to the economy or in finding new ways to make money, Buffett has never divorced himself from the belief in Value Investing which essentially means that one buys companies at the correct price and valuations and below their 'Intrinsic' value in order to generate handsome returns.

Figure 11

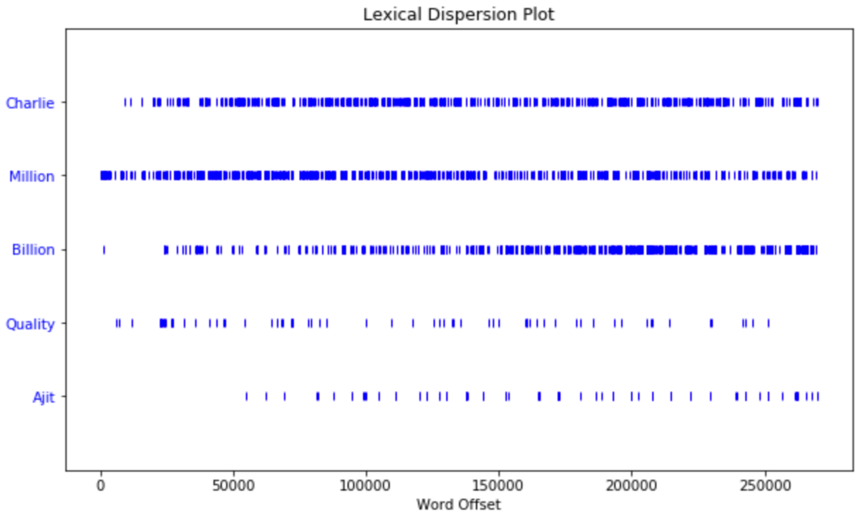
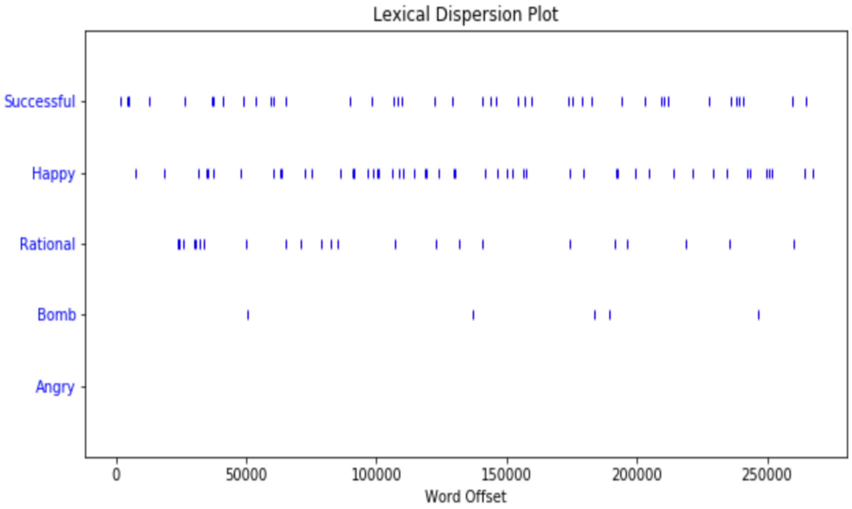


Figure 11 tells about Charlie Munger who has been his long time partner and confidant. The Berkshire Hathaway conglomerate has been built together by Munger and Buffett. Mention must also be made of Ajit Jain who currently leads the insurance business and joined Berkshire in the mid-1980s. His importance though is very secondary to the importance of Munger in Buffett's investing career. Another interesting trend is to see how the word 'Million' and 'Billion' have behaved. Over time, the usage of the latter has dominated the usage of the former showing how rich Buffett got over time. Million was used very frequently in the 1980s and 1990s but Billion has overtaken that, for obvious reasons.

Figure 12



A final lexical dispersion (Figure 12) tells us about perhaps Buffett's emotional side but we find that word like 'Successful', 'Happy' and 'Rational' are not as frequently used perhaps showing equanimity, humility and wisdom in his thought process.

4 Conclusion

We have found, through various tools and charts that Warren Buffett's key message, throughout his investing career, has been one of keeping it simple and focusing on value. He has also displayed exemplary temperament in his sentiment and this shows how important it is to not get swayed by what the overall market sentiment is: when the Market booms or busts, Buffett is more or less the same level-headed individual. We also find that certain words have held a key prominence in his letters since 1977 such as 'Charlie', 'Value', 'Earnings', 'Time' and 'Business'.

5 References

5.1 Python Libraries Used

NLTK, regex, Selenium, Textract, pickle, TextBlob, Vader, Pandas, Matplotlib and WordCloud