Sangrok Lee
Professor Johannes Hachmann
August 14, 2020

RESEARCH ASSISTANTSHIP FINAL REPORT

1. INTRODUCTION

The world is over-saturated with data, expected to reach 44 zettabytes by 2020. The multinational businesses invested millions of dollars in artificial intelligence (AI) and data technologies to improve business operations more effectively and efficiently and increase the competitive advantages over competitors. For example, with historical financial statements such as 10K, an annual cumulative financial statement required by the Securities and Exchange Commission, machine learning can use in portfolio management, fraud detection with or without sentiment analysis.

In the same manner, the university researchers use machine learning to derive meaningful insights. For example, the accounting researchers analyze the large textual data manually, which they spend an extreme amount of time, and often it is not accurate. The use of machine learning and natural language processing (NLP) enables researchers to perform more effectively and efficiently.

The researchers use and develop machine learning and natural language processing such as sentiment analysis to study the relationship between social media and financial statements or financial market, stock, or predict the business bankruptcy, which feeds in textual communication of employees with financial accounts as the balance sheet.

Professor Inho Suk is an Associate Professor in the School of Management, Accounting, and Law Department at University at Buffalo. His research focuses on voluntary disclosure, earnings management around earning thresholds; analyst forecast revision; the interaction between product market and capital market; information costs and asset pricing; accounting information and market microstructural behavioral finance theory and managerial behavior, and corporate governance and corporate legal risk management.

The research assistantship focuses on sentiment analysis based on business' social media such as Twitter. He needs a student with a business such as accounting or finance, and data management background, who can collect the textual data from the social media or the news,

and perform the sentiment analysis. He also wants to know how to use machine learning libraries such as Scikit-Learn, SpaCy, and NLTK, and the basic Python syntaxes.

The research assistantship started on June 22, 2020; however, there were communication problems due to the pandemic, COVID-19. It began on July 1, 2020, with understanding natural language processing: tokenization, stemming, lemmatization, stopwords, phrase matching, part-of-speech, named entity recognition, sentence segmentation, semantics, and sentiment analysis. In late July, we worked on sentiment analysis based on Twitter. The research assistantship takes six to ten hours per week (Appendix A), *three credit hours*, and it ended on August 10, 2020.

2. BACKGROUND

Natural Language Processing (NLP) interacts between machine and human language, where it breaks sentences into words to understand the relationship and perform the machine translation, automatic summarization, sentiment analysis, and etcetera.

Sentiment analysis determines the emotion, positive, neutral, or negative, within textual data through the sentiment scores, which positive words count as +1 and negative words count as -1. The data preprocessing is an essential initial step to reduce the noise and dimensionality of textual data that improves the sentiment classification. It needs tokenization, stemming, lemmatization, removes stopwords, and etcetera techniques to perform preprocessing the textual data.

Sentiment analysis is indispensable in business activities and university researches. In businesses, it monitors the brand through customer reviews but also improves products or services and customer support. In university researches, it used in business, economics, engineering, and medical area.

In this research assistantship, the textual data is from Jeff Benzos, the CEO and president of Amazon. It is a multinational e-commerce business and focuses on logistics, hardware, data storage, payment, and media. Nowadays, businesses are in a slump due to the current situation, but Amazon increases the business operation cash flow. It could be an interesting topic to understand his business strategies.

3. METHOD

Python libraries: re, Pandas, NumPy, Matplotlib, Seaborn, NLTK (natural language toolkit), and TextBlob used to perform the data preprocessing and sentiment analysis. NLP requires textual data to must be preprocessed to perform analysis. NLTK and TextBlob provide the easy-to-use to preprocess: tokenization, stemming and stopwords, and analysis.

3.1. TEXTUAL DATA PREPROCESSING

Twitter API (Application Programming Interface) enables programmatic access to extract the textual data. However, it requires to create a developer account, https://developer.twitter.com/en. To extract the textual data to the Python, the use of *Tweepy* is well suited with tweets with more than 140 characters.

The use of NLTK: tokenization, stemming, lemmatization and stopwords, and removal of punctuations, numbers, and special characters in data preprocessing. The stemming has two methods: Porter Stemmer and Snowball Stemmer from NLTK. Both methods reduce the word from affixes to suffixes, but Snowball Stemmer performs more accurately and more aggressively than the Porter Stemmer. Therefore, Snowball Stemmer is more favorable to the preprocessing the textual data. Lemmatization reduces the word from prefixes to lemma, the roots of the words, but needs to know the part of speech. Both stemming and lemmatization reduce the size of the dictionary, but also these normalize the textual data.

3.2. SENTIMENT ANALYSIS

Among the NLP libraries: SpaCy, NLTK, and TextBlob, NLTK is an excellent tool for education and research. The VADER (Valence Aware Dictionary and sEntiment Reasoner) module will fit into the analysis because it is a lexicon and rule-based sentiment analysis tool. TextBlob returns two properties: polarity and subjectivity in sentiment analysis. In this research assistantship, these two approaches will perform sentiment analysis.

4. RESULT

4.1. TEXTUAL DATA PREPROCESSING

In this research assistantship report, the textual data is from Jeff Benzos, the CEO, and president of Amazon; a resent two hundred tweets from Twitter. However, it needs data preprocessing such as cleaning and transformation before performing the sentiment analysis. Here are the lists of the observation from the raw textual data:

- It uses the *hash-tags (#)* to categorize the tweets and uses the *at (@)* to incorporate the other users on Twitter. For example, Jeff Benzos uses #BenzosEarthFund, #ClimatePledge, and @Emmanuel.
- It contains the URL (Uniform Resource Locator) address in each tweet. For example, https://.../.

- It contains the punctuations, special characters, and numbers that do not add value to the textual analysis; instead, these could create the noise.
- It contains the contractions, with an apostrophe, and non-lexical vocabularies or short words.
- All the words need to convert into lowercase due to case sensitivity. For example, the word Apple and apple will count separately, although it is the same word.

After the data preprocessing, the textual data needs to convert text into the numerical form, which machine can understand and perform the sentiment analysis.

```
0    [discussing, climate, sustainability, preservi...
1            [just, took, test, turns, biggest, sbliv]
2    [alexa, show, everyone, upcoming, super, bowl,...
3                                              [jamal]
4    [india, rolling, fleet, electric, delivery, ri...
Name: tweets_adj, dtype: object
```

From the *NLTK Tokenizer* module, the use of *TweetTokenizer* divides (tokenizes) the tweets into words (tokens) to perform stemming, lemmatization, and remove stopwords. The lists above show the tokenization of the textual data that preprocessed.

```
0    [discuss, climat, sustain, preserv, natur, wor...
1             [just, took, test, turn, biggest, sbliv]
2    [alexa, show, everyon, upcom, super, bowl, thank]
3                                              [jamal]
4    [india, roll, fleet, electr, deliveri, ricksha...
Name: tweets_adj, dtype: object
```

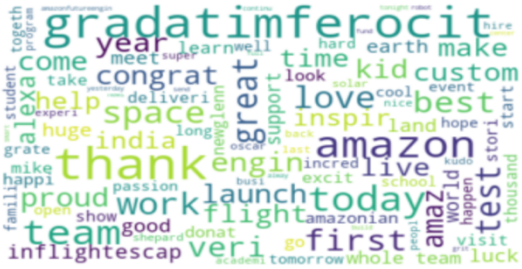
The *NLTK Snowball Stemmer* module uses *SnowballStemmer* to reduce inflected words to the word stem (affixes to suffixes). It cut off the “-s,” “-es,” “-ed,” “-ing,” “-ly,” and etcetera from the end of the words to standardize it. For example, the word “discussing” changed into the “discuss,” as shown above.

From the *NLTK Stem* module, the use of *WordNetLemmatize* to reduce inflected words to the lemma (affixes to prefixes). It removes the inflected words and returns the dictionary form of the word. However, the stemming and the lemmatization results almost identical, which is better to use stemming because lemmatization is expensive.

	tweets	tweets_adj
0	discussing climate, sustainability, and preser...	discuss climat sustain preserv natur world pre...
1	i just took a dna test, turns out i’m 100% @li...	take test turn biggest sbliv
2	hey, alexa — show everyone our upcoming super ...	alexa show everyon upcom super bowl thank
3	#jamal https://t.co/8ej1rubxvb	jamal
4	hey, india. we’re rolling out our new fleet of...	india roll fleet electr deliveri rickshaw full...

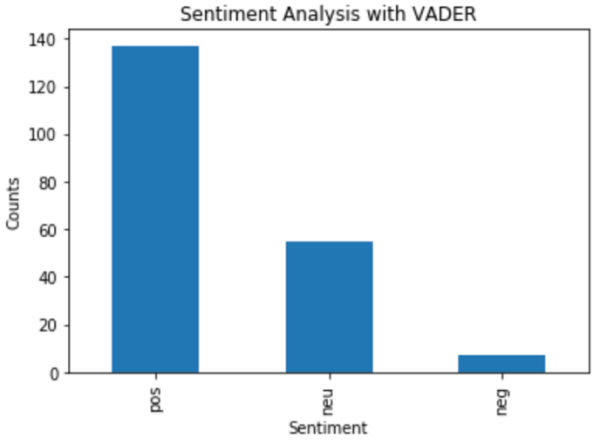
From the *NLTK Corpus* module, the use of *stopwords* to remove the common words such as “i,” “we,” “you,” “they,” and etcetera (*stopwords.words(“english”)* to see list of the stop words) to make machine learning to focus on words that define the meaning. The table above shows the raw textual data called *tweets*, and standardized textual data called *adj_tweets*.

4.2. WORD CLOUD AND SENTIMENT ANALYSIS



The use of *Wordcloud* shows frequent words from the preprocessed textual data. The word cloud, also known as a tag cloud, above indicates that Jeff Benzos uses the words: “Gradatim Ferociter,” which is Latin for step by step ferociously, “thank,” “best,” “love,” “team,” “amazon,” “india,” and etcetera often in the Twitter which most of the words are positive or neutral. These positive words from the CEO could increase business performance and attract investment.

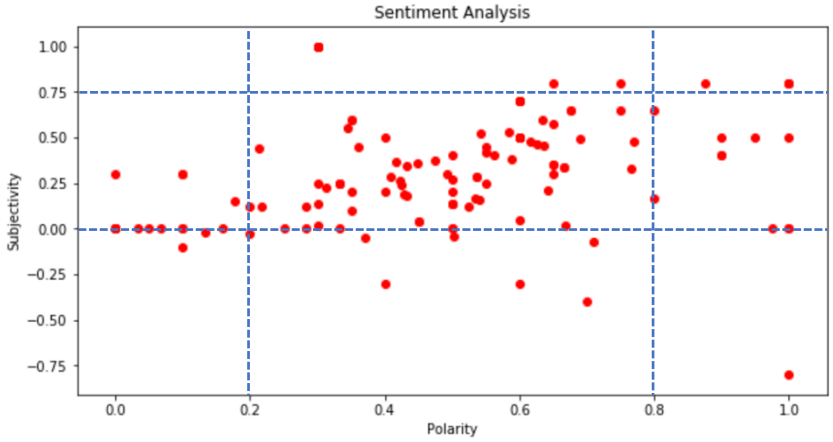
From the *NLTK Sentiment Vader* module, the use of *SentimentIntensityAnalyzer* performs the sentiment analysis based on the preprocessed textual data. It works through the sum of the scores, positive words get a +1, and negative words get a -1, in each tweet, and results in either positive, neutral, or negative sentiment.



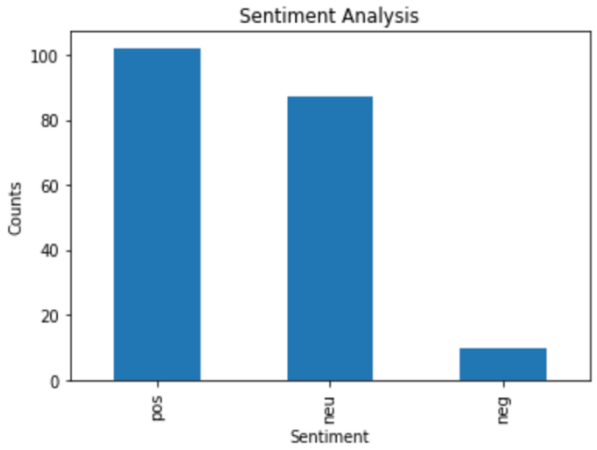
As shown above, the most significant number of tweets result in positive and followed by neutral sentiment, which positive has 137 tweets, and neutral has 55 tweets. However, seven tweets

resulted in negative sentiment, but sometimes it misclassifies the tweets when users use negative words to express their opinions. For example, Jeff Benzos tweeted, *"grateful for the journalists at the @washingtonpost and around the world who do the work, no matter the risk or dangers they face."* The respect of the journalist working under the pandemic, COVID-19, treated as a negative sentiment due to the word selection.

4.3. SENTIMENT ANALYSIS USING TEXTBLOB



As shown above, it shows the scatter plot based on the polarity and subjectivity. Polarity lies in the range between -1 and 1, where 1 indicates positive and -1 indicates negative sentiment. Subjectivity lies the range between 0 and 1, where 0 indicates objective and 1 indicates subjective. Most of the points are between 0.2 and 0.8 in polarity and between 0.00 and 0.75 in subjectivity.



As shown above, the most significant number of tweets result in positive and neutral sentiment, which positive has 102 tweets, and neutral has 87 tweets. As compared with VADER sentiment analysis, the number of positive tweets decreased, but neutral and negative tweets increased.

However, ten tweets resulted in negative sentiment, but it misclassified some tweets. For example, Jeff Benzos tweeted, *“long wait is over… Clarkson, Hammond, may are back… #thegrandtour premieres tonight in the UK, Iceland, Germany, Japan, and globally in Dec.”*

5. CONCLUSION AND IMPROVEMENT

Multinational businesses use social media to show the achievements and communication to the board of directors, shareholders, and customers. Jeff Benzos, the CEO, and president of Amazon use affirmative words: "thank," "best," "love," "team," "amazon," "india," and etcetera often to tweet in Twitter. Frequently business owners use positive words because it creates trust between businesses and customers, which increases purchase intent and brand loyalties, and increases the business' influence. The MIT Sloan Management Review states, *"by tweeting, CEO has an opportunity to initiate and influence online conversations. Rather than waiting for impression driven by the media."*

Both VADER (Valence Aware Dictionary and sEntiment Reasoner) and TextBlob perform a lexicon-based, also known as a knowledge-based approach for sentiment analysis. However, when the linguistic rules considered, VADER and TextBlob tend to perform poorly on recognize the sentiments. To improve the sentiment analysis, the use of the statistical methods includes text categorization such as the support vector machine and information extraction such as sequential stochastic model, convolutional neural network (CNN), and Word2vec.

6. RESEARCH ASSISTANTSHIP

In this research assistantship, the main objectives are: to perform sentiment analysis with preprocessed textual data from social media, Twitter, and to assist the research professor with Python 3 syntaxes and machine learning libraries: Scikit-Learn, SpaCy, NLTK, and TextBlob and simple libraries: Pandas, NumPy, Matplotlib and Seaborn, but also the foundation of the natural language process (NLP) through the communication service such as Zoom.

Zoom is an excellent communication tool with a screen-share function and access to the research professor's computer, which settled the problem under the pandemic social distancing. However, online communication has limitations: a lack of interaction and concentration, and low quality of materials includes connection of the internet.

During the research assistantship, the use of e-book: *"Hands-On Machine Learning with Scikit-Learn and TensorFlow"* written by Aurelien Geron and *"Natural Language Processing with Python"* written by Steven Bird, Ewan Klein, and Edward Loper, online-course such as Udemy

and Coursera helped to understand the foundation of NLP and use of the SpaCy, NLTK, and TextBlob.

Overall, the research assistantship achieved the objectives without implementing the SpaCy library for the sentiment analysis. The research professor, Inho Suk, has a basic foundation of the machine learning knowledge and experiences syntaxes such as MATLAB and R that speed up the learning.

6.1. LEARN FROM RESEARCH ASSISTANTSHIP

The research assistantship provided opportunities to learn new machine learning, natural processing language, and improve the Python syntaxes. However, it took time to learn the NLP foundation with the machine learning libraries: SpaCy and NLTK.

The e-books and online-courses provided NLP theories: tokenization, stemming, lemmatization, stopwords, part-of-speech, tagging, named entity recognition, and etcetera, and NLP analysis: semantics and sentiment analysis with applications. The importance of the basic NLP is to understand the logic and mathematics behind the machine learning libraries.

It was a challenge for a student who graduated from the School of Management at University at Buffalo with less experience with the programming language. However, throughout the research assistantship, it forced to improve the Python syntaxes but also techniques.

Also, it forced to improve professionalism and time management throughout working on the time-pressured environment. However, it made motivate to learn about natural language processing and always be ready for questions.

Overall, it was a worthwhile opportunity. However, it could be better if the communication was face-to-face instead of online. If there are research assistantship opportunities in the future focus on sentiment analysis, the objective will be to create the algorithm that fits into the research will perform better instead of using the machine learning libraries.

WORKS CITED

Bird, Steven, et al. Natural Language Processing with Python: O'Reilly, 2009.

Chadwick, Richard. “Tweepy for Beginners.” Medium, Towards Data Science, 1 July 2019, towardsdatascience.com/tweepy-for-beginners-24baf21f2c25.

Géron Aurélien. Hands-on Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems. O'Reilly, 2019.

Heidenreich, Hunter. “Stemming? Lemmatization? What?” Medium, Towards Data Science, 21 Dec. 2018, towardsdatascience.com/stemming-lemmatization-what-ba782b7c0bd8.

Hou, Zontee, et al. “How to Use Your CEO's Twitter Account to Build Brand Loyalty.” Content Marketing Consulting and Social Media Strategy, 26 Oct. 2015, www.convinceandconvert.com/social-media-strategy/ceo-twitter-account/.

Jain, Shubham. “NLP For Beginners: Text Classification Using TextBlob.” Analytics Vidhya, 11 Feb. 2018, www.analyticsvidhya.com/blog/2018/02/natural-language-processing-for-beginners-using-textblob/.

Karani, Dhruvil. “Introduction to Word Embedding and Word2Vec.” Medium, Towards Data Science, 1 Sept. 2018, towardsdatascience.com/introduction-to-word-embedding-and-word2vec-652d0c2060fa.

Malhotra, Claudia Kubowicz, and Arvind Malhotra. “How CEOs Can Leverage Twitter.” MIT Sloan Management Review, 14 Dec. 2015, sloanreview.mit.edu/article/how-ceos-can-leverage-twitter/.

Mayo, Matthew. “Natural Language Processing Key Terms, Explained.” KDnuggets, 2017, www.kdnuggets.com/2017/02/natural-language-processing-key-terms-explained.html.

Mourri, Younes Bensouda, et al. “Natural Language Processing Specialization.” Coursera. 2020, www.coursera.org/specializations/natural-language-processing.

Pandey, Parul. “Simplifying Sentiment Analysis Using VADER in Python (on Social Media Text).” Medium, Analytics Vidhya, 23 Sept. 2018, medium.com/analytics-vidhya/simplifying-social-media-sentiment-analysis-using-vader-in-python-f9e6ec6fc52f.

Portilla, Jose. “2020 Complete Python Bootcamp: From Zero to Hero in Python.” Udemy. 2020, www.udemy.com/course/complete-python-bootcamp/.

Portilla, Jose. “NLP - Natural Language Processing with Python.” Udemy. 2020, www.udemy.com/course/nlp-natural-language-processing-with-python/.

“Tweepy Documentation¶.” Tweepy Documentation - Tweepy 3.9.0 Documentation, docs.tweepy.org/en/latest/.

Vaish, Rahul. “TextBlob and Sentiment Analysis - Python.” Medium, Medium, 14 June 2018, medium.com/@rahulvaish/textblob-and-sentiment-analysis-python-a687e9fabe96.

APPENDIX A: TIMESHEET

Date	In	Out	In	Out	Hours
06 / 29 / 2020	03:00 pm	03:30 pm	05:00 pm	07:30 pm	03:00
					03:00
07 / 01 / 2020	10:00 am	12:30 pm			02:30
07 / 03 / 2020	10:00 am	12:00 pm	01:30 pm	02:30 pm	03:00
					05:30
07 / 06 / 2020	03:00 pm	06:45 pm			03:45
07 / 08 / 2020	10:00 am	02:00 pm			04:00
07 / 10 / 2020	11:30 am	12:00 pm	03:00 pm	05:00 pm	02:30
					10:15
07 / 13 / 2020	11:00 am	02:00 pm			03:00
07 / 15 / 2020	01:00 pm	04:15 pm			03:15
07 / 17 / 2020	01:30 pm	02:00 pm			00:30
					06:45
07 / 20 / 2020	07:00 pm	08:30 pm			01:30
07 / 21 / 2020	02:00 pm	03:30 pm			01:30
07 / 22 / 2020	07:00 pm	08:30 pm			01:30
07 / 23 / 2020	12:00 pm	01:00 pm			01:00
07 / 24 / 2020	07:00 pm	08:30 pm			01:30
					07:00
07 / 27 / 2020	01:00 pm	04:30 pm			03:30
07 / 28 / 2020	10:30 am	12:00 pm			01:30
07 / 30 / 2020	01:00 pm	04:30 pm			03:30
07 / 31 / 2020	10:30 am	12:00 pm	01:00 pm	03:00 pm	03:30
					12:00
08 / 04 / 2020	03:00 pm	06:00 pm			03:00
08 / 05 / 2020	12:30 pm	01:30 pm	05:00 pm	07:30 pm	03:30
08 / 06 / 2020	02:30 pm	06:00 pm			03:30
08 / 07 / 2020	01:30 pm	03:00 pm			01:30
					11:30
08 / 10 / 2020	01:00 pm	05:30 pm			04:30
Total					**59:50**