Finance Data Analysis: A Comparison Between Yearly Economy Change and Tone Management

Norlan Prudente

University of the Pacific

3601 Pacific Ave  
Stockton, CA 95211

n\_prudente@u.pacific.edu

1. **Abstract**

Companies are using tone management to gain recognition in press. How to use tone management effectively based on how the economy is doing in the US can be one of the many questions that comes up in many companies. A research to compare yearly GDP to the sentimental value of reports may benefit companies that have the same question above. Data will be downloaded two ways: GDP from an online database and forms, 10-K and 10-Q, using a web scraper. Web scraper deals with ip rotation to solve the DDOS issue by rotating through a pool of proxies. After gathering all of data required, data analysis is done through python and excel. Excel mainly does the graph. By performing this research, company may improve on using tone management based on how the economy is doing.

1. **Keywords**

Python, Web Crawler/Scraper, Data Analysis, Economy, and Tone Management.

# Introduction

Business, whether small or big, has their way to get their names out there. Using advertisement such as television, internet, radio, or in the street using building, bus, etc. Another way to get recognition is the use of tone management, the act of managing tone of words in order to earn press release. [1] This research paper will focus on tone management and United States yearly economy change. The goal is to determine whether the change in the economy, GDP, correlates on how tone management is utilized. used. The research will consist of data gathering and data analyzation.

Data gathering can be done either by manually downloading the data from a database or using a web scraper to gather them from websites that does not provide them. Both have been used by researchers. This research will use both method due to the reasons mentioned above. As for data analysis, it will be performed through the conversion of words to sentimental score.

Business may benefit from this research because companies will have a higher awareness of how tone management are used base on the economy change. By having this type of information in hand, business will know how to use tone management no matter how the economy changes.

This research paper will focus on determining whether tone management and change in economy has any correlation at all. The next part of the paper is structured as follows: section 4 will cover background to help readers understand some of the terminology used; Section 5 will cover literature review; Section 6 will cover data gathering technique, manually downloading and using a web scraper, then cover the procedures of data through sentimental analysis, and comparison with GDP. Section 7 will conclude the whole paper.

1. **Background**
   1. *Python*

Python [2] is a high-level programming language that is free. It uses an interpreter instead of a compiler. An interpreter [3] can execute a line of code without going through the whole program. A compiler [3] needs to go through the whole code, compile it to machine code, and create an executable before running it. Python follows an object-oriented [4] paradigm, which mean programming style will be based on objects, which are called classes. A class will have its own data and methods. Classes and objects are treated the same as real world objects, so no two objects should be the same. An example would be having a car with its make, model, and year as data, while turning the engine on will be one of its methods. Python uses dynamic semantics such as dynamic typing and binding. [5] Dynamic semantics makes variable assignment easier by not making the programmer initialize the type of the variable: integer, float, double, Boolean, etc. By using dynamic typing and binding, programmers can use the same variable name, and if they are done with it, assign a new type without the need to specify it. The purpose of dynamic typic and binding is to make coding faster by focusing more on the logic rather than on each variable. Python has a nice and easy structure for people who are just starting to program because it emphasizes in readability and indentation. Lastly, Python has huge libraries and packages that will assist programmers because difficult algorithms are already implemented to perform heavy calculations such as in statistics and calculus. With a wide variety of libraries and packages provided by python, solving DDOS would be easier than implementing it from scratch.

* 1. *Defense Mechanism of DDOS*

Distributed Denial of Service, DDOS, [6] is a defense mechanism that a website can have. DDOS can be triggered in many ways, such as when a repetitive task that a human would not likely perform occurs. Another way to trigger a defense mechanism is when a repetitive amount of request/download from a website occurs. Lastly, accessing links that are meant to catch a web crawler is called a honey pot. Honey pots are links that cannot be accessed by regular users and can only be seen if the page source is viewed, which makes it a great trap. There are good and bad reasons for having DDOS. The good part is getting rid of malicious attackers, but the bad part is not being able to automate certain tasks. Things such as automatically checking if a website is working properly through the use of web crawler and one of the biggest companies, google, uses a web crawler to make a website visible when searching through their search engine.

# Literature review

* 1. *Quantitative VS Qualitative Data*
     1. *Quantitative*

Quantitative data [7] are data that are in forms of numbers. They can be encounter in survey such as questions that requires a number as an answer, Likert scale, bullet choice, etc. “Quantity” is the key to its name. Meaning data that relates to, measurement, or measured by the quantity rather than the quality. Quantitative data is easier to analyze because it is already in the forms of number, where qualitative data are not.

* + 1. *Qualitative*

Qualitative data [8] are data that comes in forms of words. Some example are questions that are answered with sentences, paragraphs, even a whole essay, reports, etc. Similar with quantitative data, key words “quality” is in its name and it looks for quality rather than quantity. Quantitative and qualitative can also be mixed up when it comes to experiment.

* + 1. *Mixed Data*

Mixed data [9] ca be tackled in many ways, but triangulation was used in the paper called “*Mixing Quantitative and Qualitative Methods: Triangulation in Action*” written by Todd D. Jick. “They argued that more than one method should be used in the validation process to ensure that the variance reflected that of the trait and not of the method. Thus, the convergence or agreement between two methods." [9] The paper used it for social science, but the research would be similar for data science because the result will have both quantity and quality in it. For this research paper, we used mixed data: sentimental values as qualitative and yearly GDP of US as quantitative.

* 1. *Tone Management*

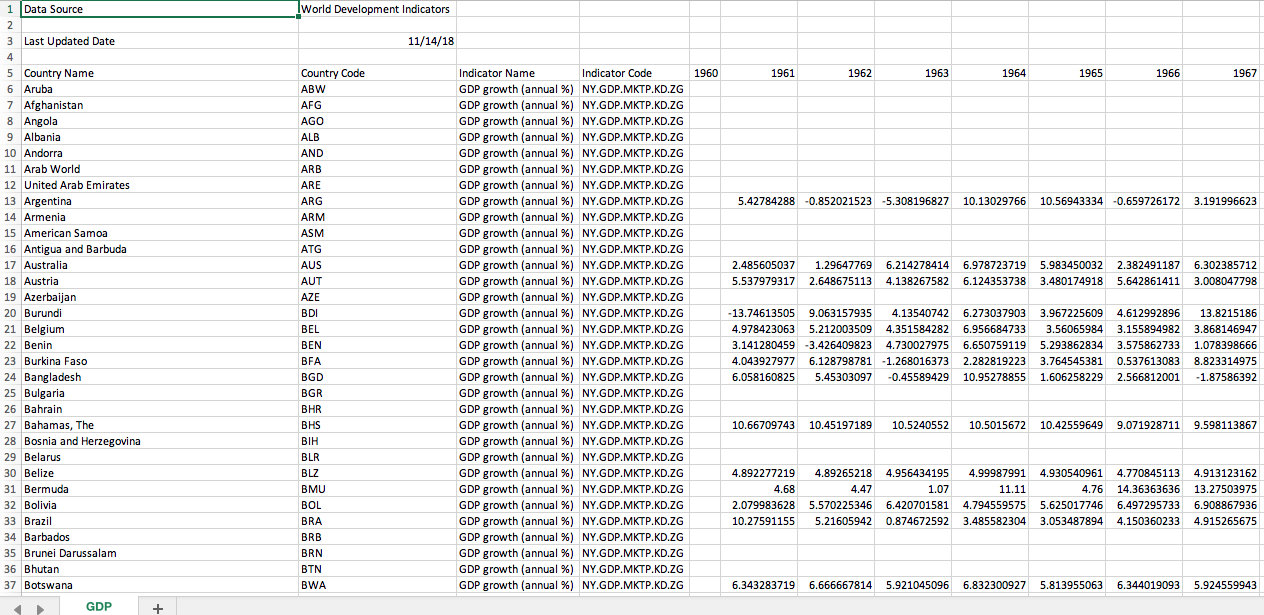
Sentimental analysis is widely used in research, but tone management are mainly used in business or financial analysis. This makes it a challenge to do because not a lot of study, or research paper, are done, therefore resources on how to perform such research is limited. The previous study was mainly concern about the frequency of abnormally positive financial sentiment. Researchers found that more positive tone are used when the company are doing good or have met their goals, but less positive when only the CEO benefits such as receiving stocks, which can impact the stocks and their price are reduced. Additional information that was mentioned in the study is that tone management can mislead investors on believing that the company is performing in a different manner than what they really are. As for this research paper, it will be studying, as mentioned in the intro, whether there is a correlation between the yearly GDP of US and the average sentimental values of each company in a yearly base.

# Procedures and analysis

* 1. *Data Gathering*
     1. *Manual Gathering Through Database*

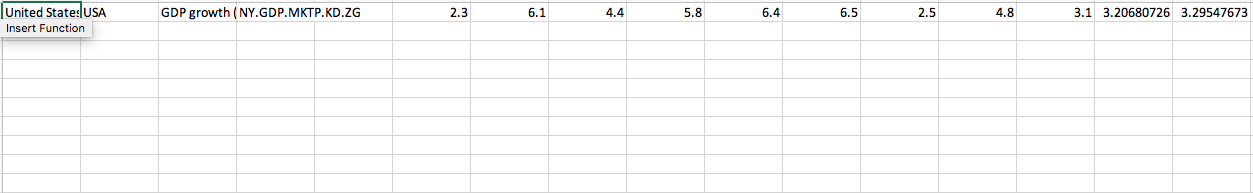
One of the many good things when it comes to data gathering is that some websites provides a data that are formatted nicely and can be downloaded. A web site called The Web Bank holds GDP data for the whole world. The downloaded data is nicely formatted to be separated in countries name and the GDP in different year, 1961 up to 2007. A sample can be seen in Table 1.

Table 1. GDP World



The data of interest is only for the GDP of the US, therefore cleaning the data will be required. Cleaning up is done by writing a script that grabs only the US data and saving it to another file. The script is a simple for loop that looks at each data that looks for “United States.” Since the position of the countries name is the first row, then comparing the first index of each data to the string “United States” is all that was needed. If a match is found, then the data in that index is grabbed. Table 2 shows the new file.

Table 2. US GDP



* + 1. *Web Scraper Using Python*
       1. *Gathering Procedure*

To start gathering the data, all the CIKs, a number representing companies, will be loaded and be used by the scrapers to find forms that will be gathered, 10-K and 10-Q. The scraper will then ask for the range of dates, beginning and end dates, of the forms that the user wants to gather. Leaving the two dates blank will gather all the forms available. Additionally, the user will be asked to enter the type of form they want to gather: 10-K, 10-Q, or both. Upon answering the prompt, the program will then get proxies to use from <https://free-proxy-list.net/>. The purpose of the proxies is to prevent DDOS from websites, fooling a website by letting it think that multiple computers are using it even though only one computer are accessing it. [10] The scraper will get ten https type of proxies and test it our site of interest, <https://www.sec.gov/>, to make sure it can access the website within 5 seconds. After getting ten proxies and testing them, the program will then load any data that was previously gathered.

The behavior of how the data are gathered is in a Breadth First Search Algorithm. [11] The scraper behaves in BFS algorithm because it will gather all the links at the current URL, for each CIK, before visiting all the links that was found. The first set of data to gather are the links that point to the links of forms. The second set of data will be the direct links to the actual forms. Lastly, the third set are individual sections of the forms. A section called item 7 is what being gathered for 10-K forms and item 2 for 10-Q forms. Those two sections are both titled “Management’s Discussion and Analysis of Financial Condition and Results of Operations.” For every 100 CIKs or failure to connect, the proxy will choose another one from the ten proxies that was picked. If all ten were to fail to connect, then a new set will be gathered.

* + - 1. *Data Structure*

To understand how the data are gather, different data structure will be shown. The first two sets, links, are in JSON [12] form, while the last one is in list [2]. The data structure for the first set, mentioned above, is shown on Data Structure 1, Data Structure 2 for the second set, and Data Structure 3 for the last set.

Data Structure 1: Links to the links of forms

Data Structure 2: Direct links to the forms

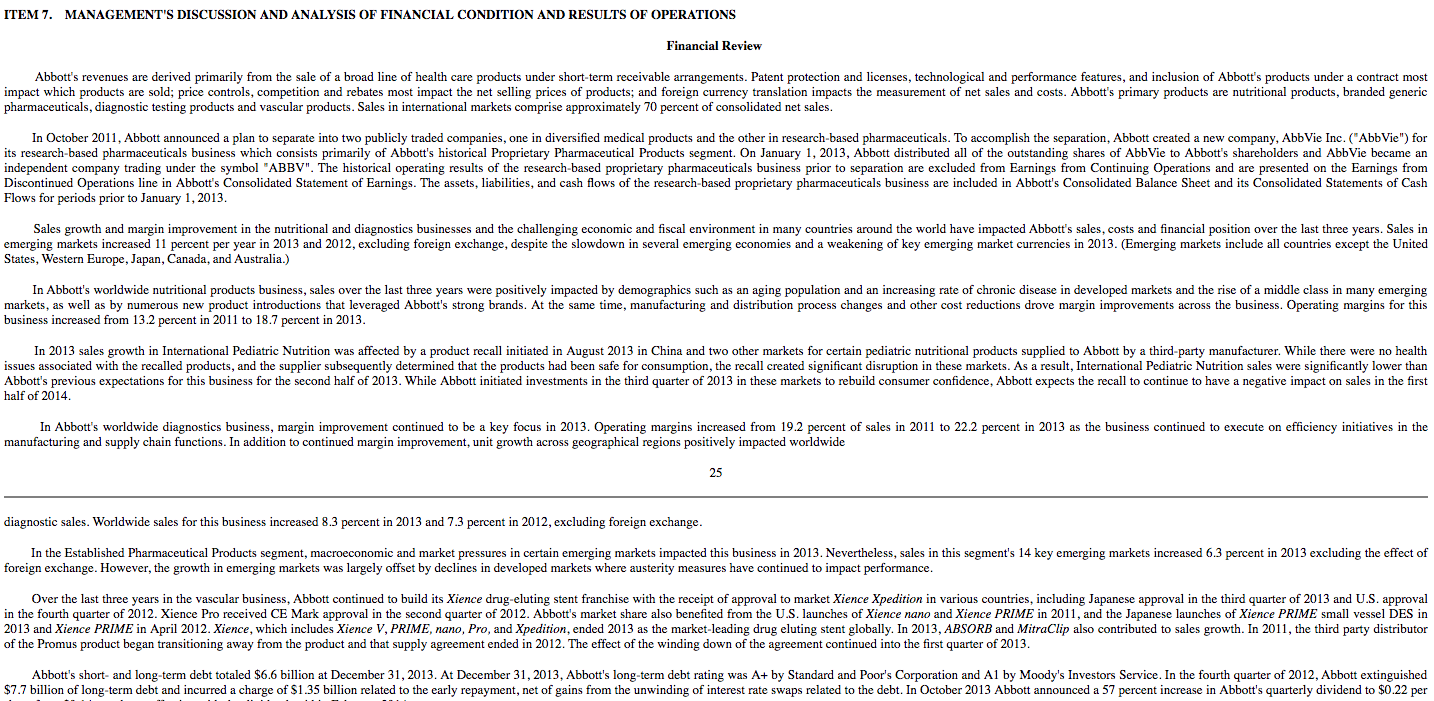
Data Structure 3: Section needed (converted to sentimental value)

The last data structured showed the converted section of that was gathered, but how did it get converted? The Next section will talk about how the data got converted.

* 1. *Data Analysis*
     1. *Data Conversion*

As mentioned earlier, item 7 and item 2 are the data that were gathered. A partial example of this is shown in Figure 1. The section can range from 4 whole page or more, also depending on the resolution of the screen and the font size that was being used by users, which then can be clarified as using more than 3000 words. With that in mind, the conversion started by using a predefined sentimental values using AFINN-111.txt that can be downloaded at <http://www2.imm.dtu.dk/pubdb/views/publication_details.php?id=6010>. The text file holds all the sentimental value, or score, of each, word that were available in the file, which are both positives or negatives. The range are from -5, worst, to 5, best. The section gathered earlier were read line by line and were split into individual words. To clean up the data, numbers and symbols were removed. After cleaning up the data, we check for each word for their sentimental value and they were added together. The purpose of adding them together is to see whether they would end up on either the positive side or negative side. Additionally, by adding the sentimental score, the score shows how strongly positive or negative the data was. After getting each individual form’s sentimental value, a final clean-up was done. The final clean-up was done by adding the sentimental score that goes together by CIK number and the years that they were reported. After that, a comparison is made with the yearly GDP.

Figure 1: Management’s Discussion and Analysis of Financial Condition and Results of Operations



* + 1. *GDP and Sentimental Comparison*

For the purpose of graphing the data, the data is fixed in the format that excel works on, horizontally with each column as year. Figure 2 shows the graph of GDP change in the United States from 1994 up to 2017. Since the 10-K and 10-Q data, that we currently have, only covers 1994 to 2019, the graph focused on those years. Which is why a zoomed in of the GDP graph is shown.

Figure 2: GDP Yearly Changes



* + - 1. *10-K Analysis*

Looking at the yearly changes in GDP graph, 2009 is the biggest downfall, worst economy, and this year would be the most interesting to compare the data into. To be clearer, looking at the data from the worst economy year and how tone management are used can give the research, and the reader, a view on how companies used tone management. Table 3 shows the final modification on 10-K data and Figure 3 shows the graph of the first ten data. Figure 3 looks messy, but as the graph get zoomed in, 2007 to 2011 that is shown in Figure 4, shows that an increase in sentimental value is used in tone management on majority of the companies. Another interesting finding is that looking at 1999 data, the economy is rising, but the sentimental values are falling. A reason for this is that companies probably think that if the economy is good then tone management is not as important as it is in the worst time. Additional reason might be that tone management was not practice as much as it is in the later years. Figure 5 shows 1999 zoomed in data.

Table 3: Final Data Modification, 10-K

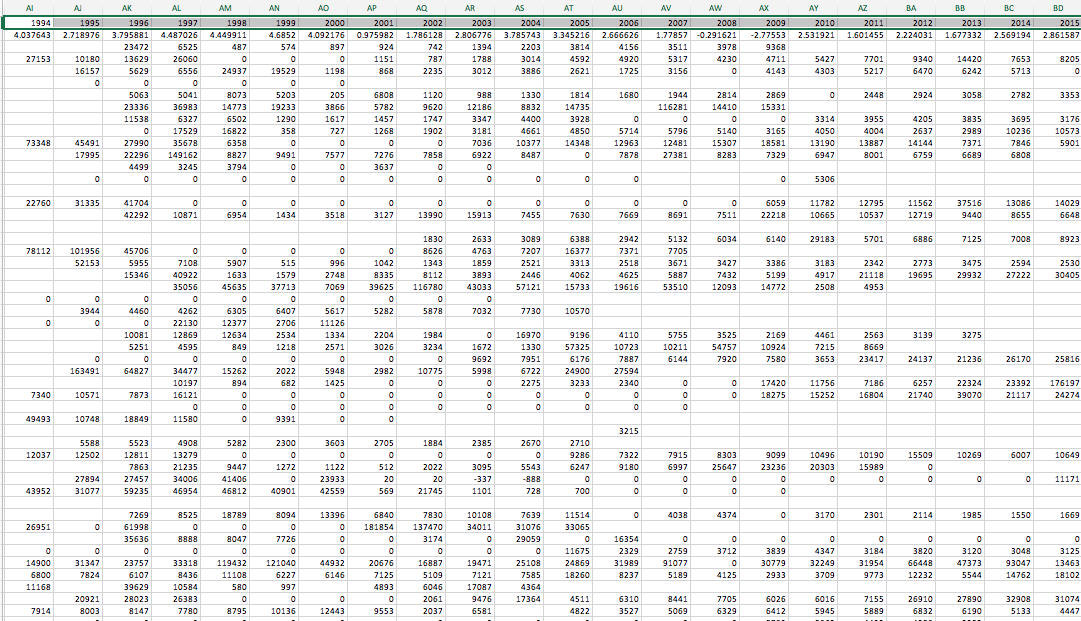


Figure 3: 10-K First Ten Data

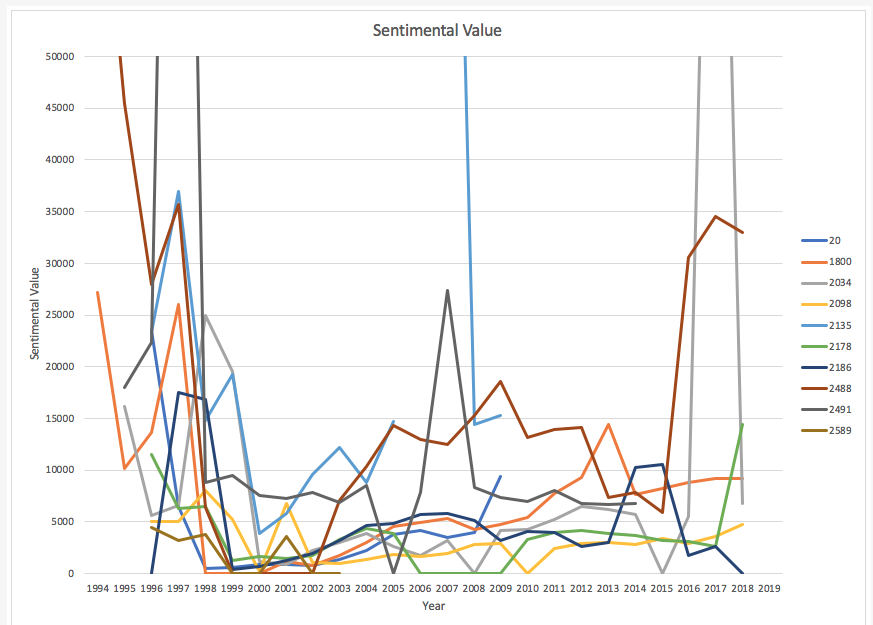


Figure 4: 10-K Zoomed-In Data, 2007-2011

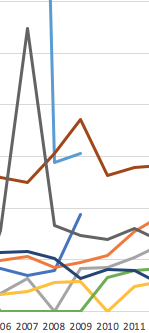
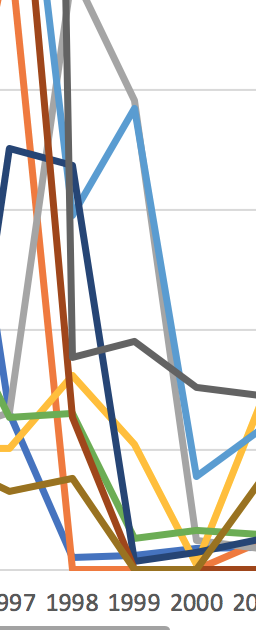


Figure5: 10-K Zoomed-In Data, 1999



* + - 1. *10-Q Analysis*

Similarly, Table 4 shows the final 10-Q data and Figure 6 shows the Graph of 10-Q. Looking at 10-Q data and comparing it to the yearly GDP changes, Figure 7 shows that 6 out of ten has a decreased sentimental value on the worst economy, 2009. All the while, looking at the similar year as 10-K, 1999, Figure 8 shows that shows that half of the data increased 4 decreased and one neutral. Compare to 10-K, 10-Q does not seem to have a pattern when it comes to using tone management because the graph shows mixed data.

Table 4: Final Data Modification, 10-Q

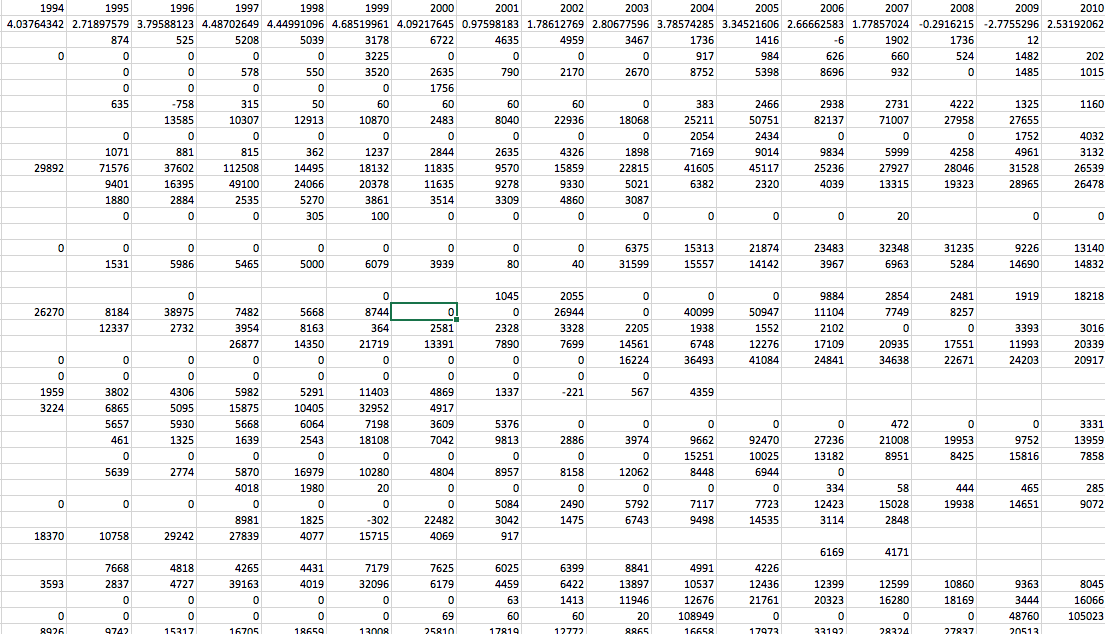


Figure 6: 10-Q First ten Data

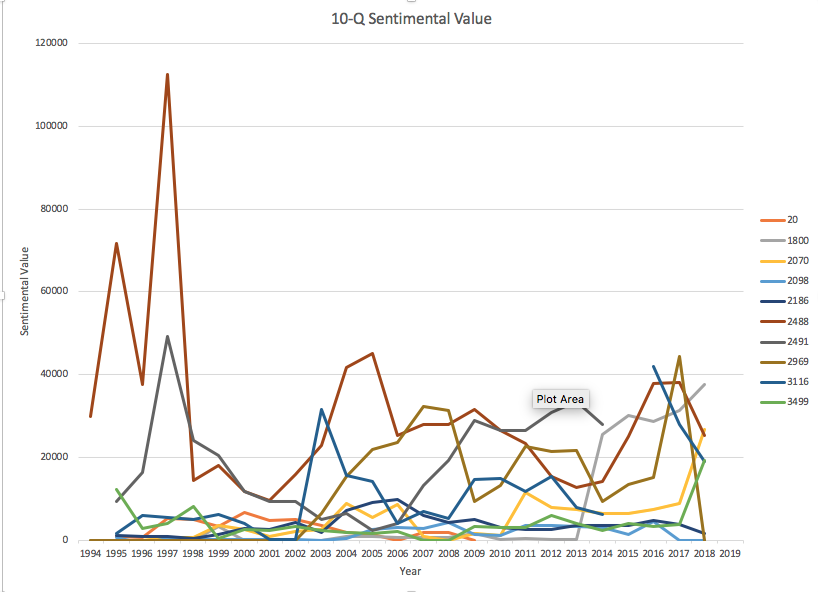


Figure 7: 10-Q Zoomed-In Data, 2006-2011

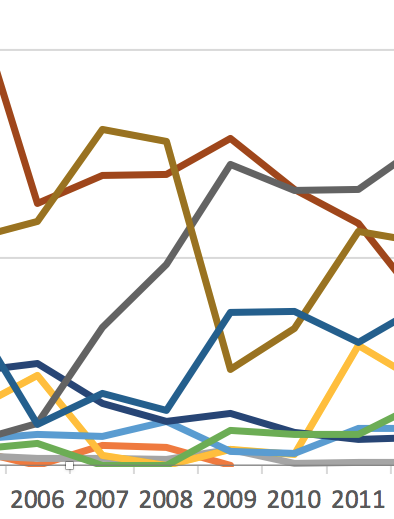
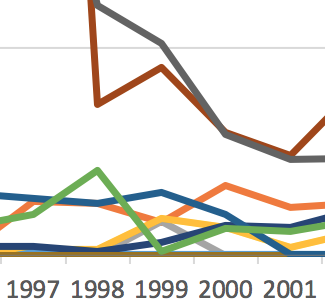


Figure 8: 10-Q Zoomed-In Data, 1999



1. **Conclusion**
   1. *Summary*

Companies has been using tone management to earn recognition to the press. Researching towards how tone management are used based on the economy standing may benefit companies on how to utilize tone management in their reports. Data gathering and data analyzation was performed using python and excel. Gathering data was done by downloading the yearly GPD changes from 1961 to 2017 was easier than using a data scraper. Getting data using web scraper was harder due to data inconsistency. Initially, data that was supposed to be collected is the reports, item 7 and item 2 sections, and the total net or total gross. The forms sometimes have tables or certain pattern, but as CIK change so as the form format. Each CIK is different and even within the same CIK, it changes the report format. Therefore, comparison went to yearly GDP instead. Analyzing the data, it seems that 10-K forms use tone management better on bad economy and relaxes during better years. As for 10-Q, there seem to be no pattern when using tone management. Of course, a lot of factors can also add up to how tone management are used such as how company are doing, total sales made, and employee performance.

* 1. *Future Work*
     1. *Web scraper*

Web scraper still needs improvement on speed. It takes a long time to gather everything: links to that contain the links to forms, links that holds the actual forms, and scraping html content to be converted into sentimental value. To give the actual speed for 1000+ CIK, the web scraper takes a week to gather each one of the data to finish, 3 weeks total. Another problem is that when it comes to collecting html content, some links break the program. The solution was, for now, to delete “broken” links.

* + 1. *Data Analysis*

Data analyzation comes as personal interpretation from the current researcher. Other researcher can find different finding when they look at the data gathered and do a different way to analyze them. This research only did a simple addition and get the total sentimental value of each year. There are other statistic techniques that can be applied to this research to make it have better findings.

1. **References**
2. X.Huang, S. H. Teoh, and Y. Zhang, “Tone Management,” The Accounting Review, vol. 89, no. 3, pp. 1083–1113, Dec. 2013.
3. Python Software Foundation Python Language Reference, version python 3.6, Available at [https://www.python.org](https://www.python.org/).
4. Programiz. *Interpreter Vs Compiler : Difference Between Interpreter and Compiler*. Available: <https://www.programiz.com/article/difference-compiler-interpreter>.
5. TutorialsPoint. OOAD Object Oriented Paradigm. Available: <https://www.tutorialspoint.com/object_oriented_analysis_design/ooad_object_oriented_paradigm.htm>.
6. S. Ferg, "Static vs. dynamic typing of programming languages," 2009. Available: <https://pythonconquerstheuniverse.wordpress.com/2009/10/03/static-vs-dynamic-typing-of-programming-languages/>.
7. M. Li, "An approach to reliably identifying signs of DDOS flood attacks based on LRD traffic pattern recognition," *Computers & Security,* vol. 23, *(7),* pp. 549-558, 2004. Available: <https://www.sciencedirect.com/science/article/pii/S0167404804001245>. DOI: 10.1016/j.cose.2004.04.005.
8. A. Bryman and D.Cramer, *Quantitative data analysis for social scientists*. Florence, KY, US: Taylor and Frances/Routledge, 1990.
9. A. Lacey and D. Luff, “Qualitative Data Analysis,” National Institute for Health Research, Available: <http://healthindisasters.com/images/Books/9_Qualitative_Data_Analysis_Revision_2009.pdf>
10. T. D. Jick, “Mixing Quantitative and Qualitative Methods: Triangulation in Action,” *Administrative Science Quarterly*, Vol. 24, No. 4, pp. 602-611, 1979. Available: <https://www.jstor.org/stable/pdf/2392366.pdf>
11. ScrapeHero, "How to prevent getting blacklisted while scraping," 2014. Available: <https://www.scrapehero.com/how-to-prevent-getting-blacklisted-while-scraping/>.
12. GeeksforGeeks. Breadth First Search or BFS for a Graph. Available: <https://www.geeksforgeeks.org/breadth-first-search-or-bfs-for-a-graph/>
13. JSON, Inroducing JSON. Available: <https://www.json.org/>