THE DATAs CHALLENGE 2022

By Anshul Khairari, M.S., Business Analytics (NETID – AXK210031)

1. **How you proceeded with the data. Your thought process before you started working!**

Before start working on any data, I skim through the dataset, look at the size of the data and analyze the problem statement. After analyzing, it is critical to summarize the problem statement and come to a type of solution needed. For example, after reading the problem statement, –

''Many companies share their financial results for a quarter or a year through earning calls and reports. It is a way to communicate with shareholders about the company's performance and strategy for the future. Although, the full report with all the numbers and plans is a good way to look at company's present and future, the important part of this earning call is a speech by the CEO/CFO where they highlight company's progress and answer some questions by investors.

Your mission, should you choose to accept, is to analyze earning call speech by a CEO of a company. ***There is no right or wrong way to complete the task.*** *Please keep the following points in mind before working on the task."*

It was evident for me to use natural language processing to complete the task. My secondary thought was to display the entire data in the form of a word-cloud, using Tableau Public. Hence, **after my initial thought-process, following were my findings:**

a. The data is completely in string format, with few punctuations and random sentence split.
b. The text is a CEO speech, that needs to be analyzed. Hence, NLP is required.
c. For Sentimental Analysis of the speech, we need to check the frequency of the words, without stop words or filler words.
d. As this is a raw data, we need to access library of NLTK in python, to use pre-trained analyzer.

An overall flow of the task is as follows:

i. Importing Required Libraries
ii. Importing data into Python Jupyter Notebooks
iii. Converting data into list or arrays of strings
iv. Clean the Speech to remove miscellaneous unwanted text, like '\n'
v. Word Tokenization of speech sentences
vi. Removing Stop words and Punctuations
vii. Checking the Frequency of the Words
viii. Stemming and Lemmatization of remaining Words
ix. Part-Of-Speech (POS) Tagging and Name Entity Recognition (NER)
x. Chunking and Chinking, if necessary
xi. Transforming Initial speech into dataframe, line by line in rows
xii. Applying Valence Aware Dictionary and sEntiment Reasoner (VADER) to the dataframe
xiii. Finding Overall Speech Sentiment with most common sentiment sentence by sentence

The detailed flow of work is mentioned in the data files in the GitHub link, provided later in the document.

2. Tools, languages, software's used

Following are the tools, languages and software used:

a. Software:
   i. OS – MacOS (Monterey Version - 12.0.1)
   ii. IDE Launcher – Anaconda Navigator (Version 2.0.3)
   iii. IDE Platform – Jupyter Notebooks (Version 6.0.3)
b. Languages:
   i. Python – Version 3.8.12
c. Tools/Libraries:
   i. Pandas
   ii. Numpy
   iii. String (punctuation)
   iv. NLTK Tokenize (word tokenizer)
   v. NLTK Corpus (stopwords)
   vi. NLTK Probability (Frequency Distribution)
   vii. NLTK Stem (Snowball Stemmer and WordNet Lemmatizer)
   viii. NLTK (POS Tag, NER Chunking, Regular Expression Parser)
   ix. NLTK Chunk (tree2conlltags)
   x. NLTK VADER (Sentiment Intensity Analyzer)
   xi. Statistics (Mode)

3. Answers to these 2 simple questions:

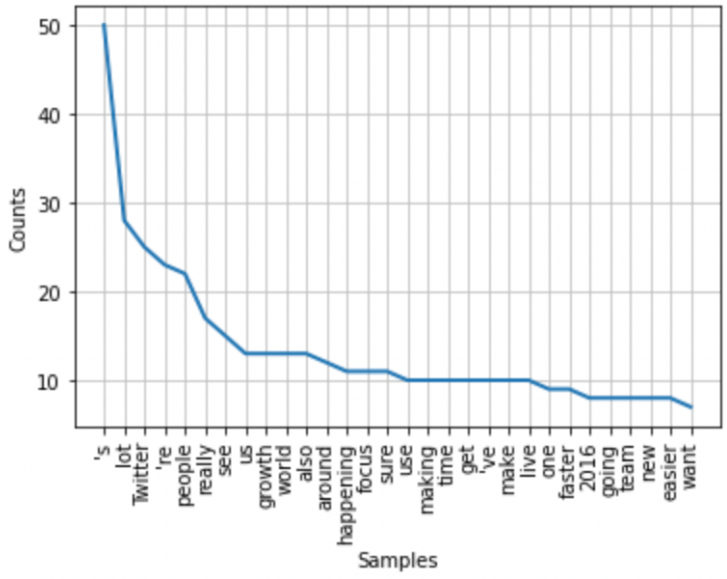
a. The company you think this is related to.

Using the NLTK probability library for frequency distribution, one of the top words used was “Twitter”. Thus, either the CEO is of the company Twitter or Competitor of Twitter. Upon reading the first and last few lines, it became evident that indeed the speech was given by the CEO of Twitter.

b. Financial year CEO is discussing?

Lines “2016 was a transformative year for us” and “We've hit the 10-year mark. We're about to be 11…”, and Twitter was formed in the year 2006, suggests that the financial year discussed is 2016-2017.

4. **The visualizations, insights you extracted from data – images, thoughts, anything!**



There are few visualizations and insights from the data. These are as follows:

i. The above graph represents the frequency of words within the speech. We can observe that, apart from error terms in cleaning, “lot”, “Twitter” and “people” are the most common words used.

ii. There are a total of 113 sentences in the speech. Out of these 113 sentences, 79 sentences are having positive sentiment, according to valence aware dictionary and sentiment analyzer, a.k.a., VADER. Thus, having approximately 70% positive sentiment, it is safe to say the entire speech is positive and motivating.

iii. As there is a reference to 2016, the speech is a reflection on the company’s performance and innovation in the year 2016 and thus, the speech is being delivered after 2016.

iv. Only 4 statements have negative sentiment, which may represent contrast to enhance the implication of the positive statements.

5. **What more you think can be done in future on this data!**

With this data, we can predict the structure of speeches and further use to provide prediction of sentimental analysis to other similar scenario speeches. With a combination of more speeches, we can even analyze the personality of the person and even create a bot mimicking the CEO for scheduled mails. The data of speech content can show the growth or fall of the company and hence, in future, we can use this data and such data to predict the stock market of the said company.

6. **Any code (if you have written) should be uploaded in a public GitHub repository. The link should be included in the report.**

Following is the link to the code:

https://github.com/anshulkhairari/ceo-speech-sentimental-analysis