Gary Chen

[Gchen09@syr.edu](mailto:Gchen09@syr.edu)

HW 5

API Naïve Bayes

IST-736 Text Mining

**Introduction:**

The Covid-19 pandemic and world conflicts have had an impact on everyone across the world and has caused some shifts in people’s world view. It disrupted supply chains, people’s lives, and jobs. Currently in 2023, the world is still going through a recovery but many still are facing the hardships and stress that was set off by the pandemic. The doom and gloom of the last couple of years has brought many topics such as banks, business, inflation, and layoffs to the forefront.

Politics have gotten even more polarizing in how different ideologies have clashed in maneuvering these tough times. The USA presidential election is coming up in 2024 and issues such as business, banks, inflation, and layoffs are huge talking points up for debate. These four topics are closely intertwined. The politics to raise interest rates caused huge ripples in massive white-collar layoffs with the intention of slowing down the economy to curb inflation. It is forcing businesses to rein in their spending by cutting off cheap capital and making it much more expensive to keep growing the company, so that means cutting down their work force.

On the flipside of the inflation coin, many businesses are recording record profits by increasing their prices. There were also recent major policy decisions by the Federal Reserve to protect some of the riches entities in America from regional bank runs such as Silicon Valley and First Republic Bank, which went against the principles of raising interest rates. These banks runs were largely due to the Federal Reserve heavily increasing interest rates. The high interest rates haven’t seemed to keep inflation in check. Many economic experts point out supply chain issues and high interest rates are only part of the equation to inflation. There hasn’t been much political policy to address why businesses are recording super high profits in these turbulent times. It will be interesting to see the News Headline patterns between these intertwined topics.

**Analysis:**

Business, Banking, Inflation, and Layoffs are current hot topics and are covered a lot in the news. To discover more insights and patterns between these topics, An API NewsAPI can search and gather the headlines for these topics. Application Programing Interface is a software interface for two or more computer programs to communicate. This NewsAPI searches through published articles from 80,000 news sources and blogs. This search can bring data about the articles into python within a JSON format. The parameters for the NewsAPI are “everything” as an endpoint, the four topics mentioned above as the keywords/phrases, and English as the language. The requests.get() function will make a request to a web page with the NewsAPI parameters and return the status code that makes it possible to be stored in a JSON format. The response objects are status, totalResults, articles, source, author, title, description, url, urlToImage, PublishedAt, and content. Then the “articles”, “source”, “publishAt”, “title”, and “description” are the response objects captured into a csv from the JSON status codes. Each of these objects are cleaned with proper date formatting (MM/DD/YYYY), removed of special characters, and removed of numbers.

A screenshot of a computer

Description automatically generated with medium confidence

After the JSON data with the selected response objects are put into a csv, the data will be imported back to Python as a dataframe with the objects as the columns. This dataset contains 400 rows of article documents with 100 articles from each of the Label topics (Business, Banking, Inflation, and Layoffs).

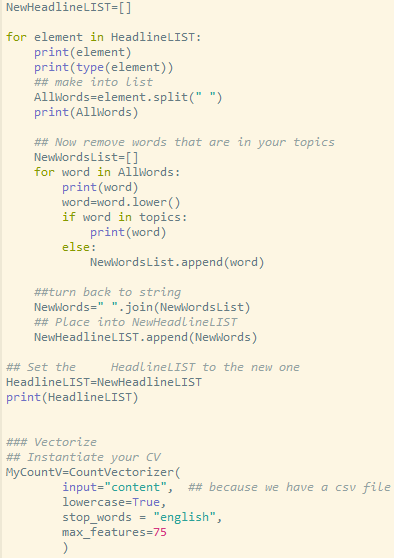
A screen shot of a computer code

Description automatically generated with low confidence

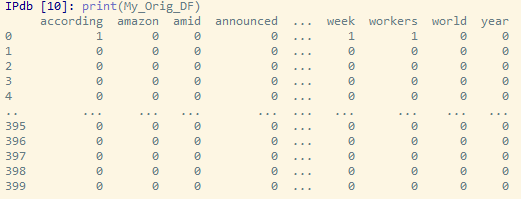
A screenshot of a computer

Description automatically generated with medium confidence

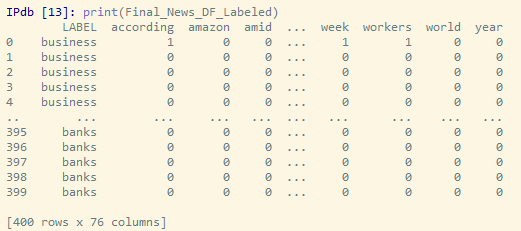
From this imported data frame, the Headline values and Label values are appended to empty lists to be vectorized. Before it goes into vectorization, all the headline phrases are split into individual lowercased words and the topic label words (Business, Banking, Inflation, and Layoffs) are removed before vectorization. The CountVectorizer is set to remove stop words and limit the word features to the top 75 most frequent words.



The cleaned Headline list is fitted into the CountVectorizer to get the top 75 most frequent word features into a data frame.



Next, the topic labels are concatenated into the Vectorized word features dataframe to create one dataset for modeling analysis.



Word Clouds were generated for each of the topic labels from the cleaned Vectorized data frame.

***Business Word Cloud***

******

***Layoffs Word Cloud***

******

***Inflation Word Cloud***

******

***Banks Word Cloud***

******

The Labeled Vectorized Dataframe was split into Training and Testing data sets for Naïve Bayes analysis. 70% of the data was split to the training set and 30% of the data was split to the testing set. The labels from both the sets were first stored into the TestLabels and TrainLabels variables to be used for model performance evaluations. Then the labels in the Training and Testing data sets are removed to run Multinomial Naïve bayes calculations.

A screenshot of a computer

Description automatically generated with medium confidence

The training set is used to train the MNB mode for predicting the labels in the testing set. The MNB modeler is created to fit the training data into a classifier. Then the predicted label will be compared to the actual labels in the testing set to see how accurate this process was. The confusion matrix will show how many predictions are correct and will be an easier way to view the accuracy.

**Results:**

The NewsAPI brought in the data needed to gather insights on the Business, Banking, Inflation, and Layoffs news topics. It produced 100 headline descriptions from each of the topics to be vectorized for Multinomial Naïve Bayes analysis. The CountVectorization was limited to the top 75 most frequent words and word clouds were produced from these frequent headline words.

In the Business word cloud, it gave very generic words that can mean a lot of things. It shows that many Business article headlines can be very generic with words like “financial”, “latest”, “rolling”, “coverage”, “company”, “morning”, “economic”, “world”, “growth”, & “markets”. It shows “Reuters” as a huge reporter in this space. Other words like “tech”, “china”, “apple”. “disney”, & “tech” give more context on what some of the recent news coverage on business.

The Layoffs word cloud gives a lot more context to the topic. It shows many big “tech” and “media” companies have been impacted like “amazon”, “twitter”, “ Disney”, “google”, “bloomberg”, “Microsoft”, & “apple”. Words like “company”, “workers”, “employees”, & “tech” have a big magnitude in identifying Layoff documents. “Company” also overlaps with the Business word cloud which can create some confusion for the MNB prediction model. “Elon” “Musks” also seems to appear frequently in layoff related articles which shows he’s been a huge part of this topic discussion. “bank” seems to have a very small magnitude in this layoffs word cloud.

The Inflation word cloud has very few words. “Reuters” is the highest magnitude word. It looks like “high”, “prices”, “rate”, “bank”, “data”, “says”, “despite”, “markets”, “investor”, “shares”, & “investor” were somewhat frequent words. It seems that “India” & “Canada” were heavily impacted or a huge part of this topic too.

Banks word cloud has many words that overlap the other word clouds. “Reuters”, “Bloomberg”, “financial”, “rates”, “shares”, “morning”, “billion”, “India”, “china”, & “tech” are all words found in the other word cloud topics. The unique words for the Banks are “regional”, and “republic”. This makes sense since many of the regional banks have been impacted by current financial market conditions and One Republic was one of them. The patterns in this word cloud show that the bank label will make it harder for MNB to predict the right topic for some of the documents.

The Multinomial Naïve Bayes model showed mixed results in its accuracy for each of the topics. The confusion matrix below shows that the model was able to predict inflation documents with perfect accuracy.

A picture containing text, screenshot, number, diagram

Description automatically generated

In the normalized confusion matrix, it shows that business had about 28% accuracy, layoffs had 27% accuracy, inflation had 100% accuracy, and banks had 50% accuracy. The model was most successful in predicting inflation, which makes sense from the word cloud analysis. It had the least number of words which can make it easier to identify inflation documents. It was also interesting that the model predicted most of the documents under the inflation label. The other topics had a bit too much noise and possible overlap. “Reuters” appeared in every single word cloud which could’ve caused some confusion in the modeling and should be removed from the word vectorization step in future analysis. “Banks” articles were probably a little easier to identify with words like “regional” and “republic”.

A picture containing text, screenshot, number, diagram

Description automatically generated

**Conclusion:**

Overall, this analysis showed how interrelated these topics are. Some of these document articles could’ve fit into multiple categories of these topics. Reuters appeared in all the of the topics and showed huge frequencies in the Banks, Inflation, and Business word clouds. This makes sense since it is a Financial Media outlet, and it is probably reporting on these topics. The only topic that was easily distinguishable from the headlines is inflation. This could be attributed to the small number of words associated with this topic. Words like “prices”, “high”, & “rate” helped identify Inflation related documents.

In the Business word cloud, it showed many of the headlines can be super generic. However, there were some nuggets to show some current business-related article. The “tech” sector has seen a huge downturn in its market cap. A lot of the “tech” stocks are down, and it is partly due to rising interest rates that impedes a lot of startups from growing. It costs a lot more money to try to expand when loans are more expensive from the high interest rates. Disney has been in a dispute with the Florida Governor from a policy perspective. Even with this dissection of some of the words to give insights on recent events, there is an overlap to the other topics.

Banks also showed a pattern of overlapping with the other topics. There were a couple of keywords like “regional” and “republic” that made it a little easier to classify these documents as banks. This proves the pattern of all these topics are similar and all matter to greater debate in the politics that America and the world is facing. Business, Banking, Inflation, and Layoffs should be discussed together in how the world recovers from the major events of the last couple of years.