IDS 561 - Analytics for Big Data

Twitter Sentiment Analysis - "Pepsi Vs Coke"

Submitted by -
Dhananjay Singh (668437546)
Shivani Narahari (675954089)

1. INTRODUCTION

Coca-Cola and PepsiCo are well established names in the soft drink industry and the ‘Cola wars’ refers to the long-time rivalry between their products Coke and Pepsi. Current market share over the last decade for Coke is 17.8%, and while for Pepsi it is 8.4%. Both these companies have engaged in mutually targeted marketing campaigns throughout history in order to win this war. The customer sentiments regarding both Coke and Pepsi are also divided. There are various conflicting articles on the internet, some in favor of Coke and others in favor of Pepsi. The goal of this project is to gauge the customer sentiments by gathering tweets posted on Twitter related to Pepsi and Coke. We will then analyze the sentiments of the customers of these two soft drinks and present a score to conclude who’s currently winning this Cola War.

Sentiment Analysis is a process to determine whether some textual data has a positive, or negative connotation. People share a lot of information on social media in the form of posts, or comments. Twitter sentiment analysis can be done to gather and analyze people’s tweets to gauge their feelings about any brand. Many surveys/research have been conducted to figure out if people like Pepsi more or Coke. In this project, we will perform sentiment analysis on twitter data to understand people’s feelings towards Pepsi and Coke as either Positive or Negative.

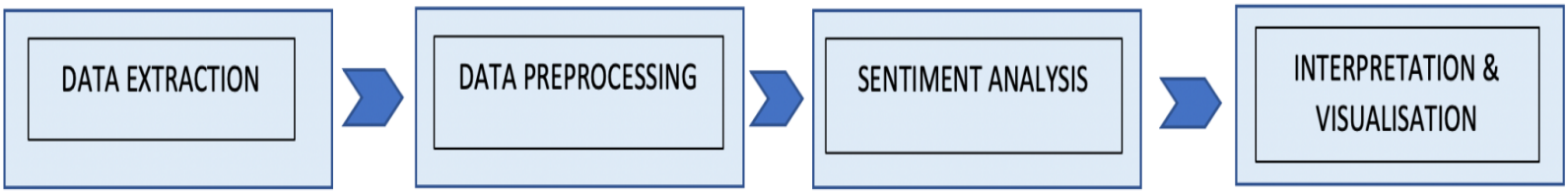
2. DATA DESCRIPTION

We have utilized the Tweepy Python library and built a twitter dataset which contains 55,438 rows(Tweets) and 10 columns. The features include username, account description, location, following count, followers count, total count of tweets, user created, tweet created, retweet count, tweet text, and hashtags. We have used the tweets ‘text’ column to do sentiment analysis.

A snippet of our dataset is as follows -

username	acctdesc	location	following	followers	totaltweets	usercreatedts	tweetcreatedts	retweetcount	text	hashtag
ts_RT_Today	Aussie Realist. Defend #Taiwan🇹🇼.\nSupport the...	Australia / Taiwan	463	916	3363	2009-07-08 13:50:11+00:00	2022-03-18 05:20:28+00:00	0	@Amaralee23 Haven't bought @CocaCola in years,...	[{'text': 'DrinkPepsi 'indices': [93, 104]},..
greyisde4d	that mf grey	Melbourne, Victoria	139	33	2067	2018-05-11 21:04:50+00:00	2022-04-23 02:38:32+00:00	0	you ordered diet coke, that's a joke right?	
razy_wolf16	Hi I'm Razor, Razy For short\nPAW Patrol Rocky...	NaN	101	20	53	2022-04-21 15:10:28+00:00	2022-04-23 08:58:30+00:00	0	I voted for #Pepsi, and you?! This is the fin...	[{'text': 'Pepsi 'indices': [12, 19]}
AimanXlucky	Hoping to have the best year ever 🍀\n#BIGWINSOON	Dhaka, Bangladesh	3975	128	14009	2022-02-18 17:05:05+00:00	2022-04-23 08:39:44+00:00	0	I don't respect anybody who can't tell the dif...	
rencors	bad at journalling so I tweet instead • 20 • (...	fatphobes dni	279	467	4879	2021-02-14 07:58:26+00:00	2022-04-23 07:45:28+00:00	0	This vanilla coke zero is like electricity on ...	

3. METHODOLOGY



3.1 Data Scraping and Cleaning

We created a Twitter Developer account with Elevated Access and got the Consumer Key, Consumer Secret, Access Token, and Access Token Secret. Then, we scraped Twitter to get Tweets using Python's Tweepy library. We utilized "Search API" under Tweepy. The search query that was passed to the API to get information on tweets were "#coke", "#cokevspepsi", "pepsi", and "#pepsivscoke". We also included the parameter "-filter:retweets" to avoid getting retweets so that we do not get duplicate records in our dataset. We collected 55,438 tweets in total.

We performed data cleaning on the "text" column using the NLTK and re- libraries to remove the following:

- Urls
- Emojis
- Stop-words
- Punctuations
- Non-English words
- Numbers
- Multiple full-stops
- Very long or very short words

In addition to this, we converted the text to lowercase and dropped the following columns which serve no purpose in our analysis - 'acctdesc', 'location', 'following', 'followers', 'totaltweets', 'usercreatedts', 'tweetcreatedts', 'retweetcount', 'hashtags'. The only features that we made use of were ‘username’ and ‘text’.

3.2 Mapper

Input:
We converted our dataframe to a text file with 2 columns, username and cleaned tweets, separated by a tab character.
This was used as the input to our mapper function along with the sentiment dictionaries. We got a list of positive and negative sentiment words in 2 text files, positive-words.txt and negative-words.txt, which were obtained by using nltk.corpus.reader.opinion_lexicon.

Output:
The output of our mapper function gave us a (key, value) pair in the following format - (username, “coke”/ “pepsi” with ratio score). We compared each word of the tweet with the sentiment dictionary and counted the total number of positive and negative words present in that tweet. Two ratio scores were then obtained by dividing the number of positive or negative words in a tweet and the length of that particular tweet. The final ratio score was calculated by subtracting the negative ratio score from the positive ratio score.

We stored the (key, value) output of the mapper in a MapperOutput.txt file which is displayed below -

nattycakes	coke	0.2
ife0luwa	coke	0.3333333333333333
chellelembo	coke	0.1
xcoileray	coke	-0.25
theakshay18	coke	0.1
Paul_Hawkins83	coke	0.25
leejonathan	coke	0.0
trent_the_tiger	coke	0.25
bareribs	coke	0.2
mkerobert	pepsi	0.2
mario_gelinas	coke	0.0625
AyoAquaire	coke	0.0

3.3 Reducer

Input:
The MapperOutput.txt file was the input to the Reducer function.

Output:
The output of our reducer function gave us two (key, value) pairs in the following format - (coke, average score), and (pepsi, average score). The reducer calculated the sum of the ratio scores by matching the key with coke and also for pepsi. It also counted the total number of tweets for each of the two keys. The value of the output was calculated by dividing the total sum of ratio score by the total count of the tweets for each of the two keys.

4. RESULTS

The final score that the Reducer gave to the two brands Pepsi and Coke were -

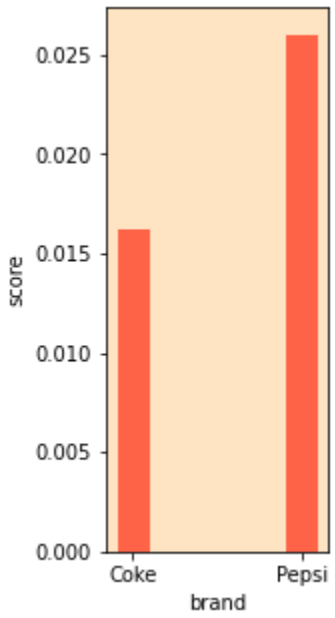
```
Total avg score of products

coke    0.0166
pepsi   0.0266
```

This shows that people favor Pepsi over Coca Cola, as Pepsi's overall score was higher.
The reasons for this could be:

1. The number of tweets for Coke may be higher than for Pepsi, hence decreasing the ratio value.
2. Sentiment analysis is not adept at capturing sarcasm in tweets, hence negative sentiments expressed by positive connotation words are taken as positive.

Visualizing the average score for Coke and Pepsi -



Coke and Pepsi both have a positive sentiment score. However, twitter sentiment analysis shows that Pepsi wins the "Cola wars".

5. ROLE OF TEAM MEMBERS

Both the team members contributed fully and equally for this entire project.

Member Name	**Contribution**
Dhananjay Singh	100%
Shivani Narahari	100%