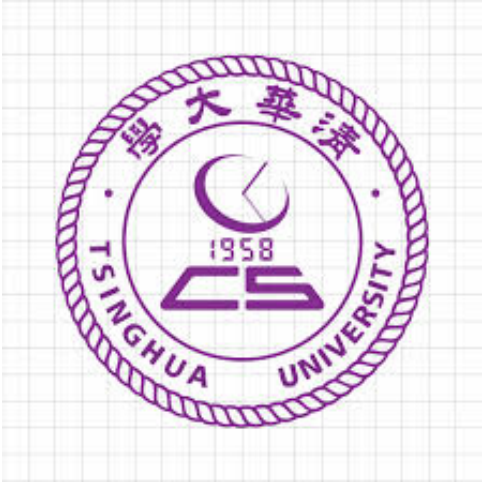
Final Project Report

# Sentiment Analysis Search Engine

Isingizwe Didier Frank 2019280803

Course: Web Information and Retrieval



June 7th, 2020

# Contents

# Chapter 1

# Introduction

Sentiment analysis (Or opinion mining) refers to us as the process of determining the emotional tone behind a series of words, used to gain an understanding of the attitudes, opinions and emotions expressed, while text classification is the process of assigning tags or categories to text according to its content. In this work, i integrated a smart sentiment detector in a search engine that retrieves Positive, Negative  Neutral tagged data/results relevant to the query searched with a filter feature that let the user see positive results (good) only, negative results (Bad) only or neutral results (Not bad nor good).

This experiment can be applied to other broad applications such as topic labeling, spam detection, and intent detection. A search engine refers to us as a software system that is designed to carry out web search, which means to search the World Wide Web in a systematic way for particular information specified in a textual web search query. In my experiment, it is just a sentiment analysis oriented search engine which means that the user can choose to see positive (good) or negative (bad) results considering the searched query.

# Chapter 2

# Motivation

Social Media:
-Sentiment analysis is extremely useful in social media monitoring as it allows us to gain an overview of the wider public opinion behind certain topics. Shifts in sentiment on social media have been shown to correlate with shifts in the stock market.

Politics:
-The Obama administration used sentiment analysis to gauge public opinion to policy announcements and campaign messages ahead of 2012 presidential election. Being able to quickly see the sentiment behind everything from forum posts to news articles means being better able to strategise and plan for the future.

Business:
-It is also able to be an essential part of your market research and customer service approach. Not only you can see what people think of your own products or services, you can see what they think about your competitors too. The overall customer experience of users can be revealed quickly with sentiment analysis, but it can get far more granular too.

-The ability to quickly understand consumer attitudes and react accordingly is something that Expedia Canada took advantage of when they noticed that there was a steady increase in negative feedback to the music used in one of their television adverts.

The novelty of my system is that every search results will be quantified for large scale companies to have it in terms of analytics to understand what people think about their products/services and it can also be applied to all businesses or politicians.

# Chapter 3

# Related Works

There are several tools for sentiment analysis and most of them are special purpose oriented like sentiment analysis search for twitter comments (Social media related comments) and other systems that is accredited or integrated in companies but my system is general purpose oriented which means that it can be used by anyone including businesses, celebrities and other users who would want to check good or bad comments about products before buying them for free and easy of search instead of going on each every commercial website because our source of information is on google so, the search engine gathers everything from google which is more convenient compared to existing system etc...

Some of similar existing system (Special purpose):

Quick Search — Social media search engine:
Quick Search gives you an instant overview of your brand online. It's a social media search engine that offers extensive coverage of social networks - including news sites, blogs, and forums. The social media search engine allows you to dig deeper and find the source of this negative sentiment.

RapidMiner — New areas for business expansion:
A data science software platform that provides text mining to help brands perform sentiment analysis. Online reviews and social media posts can be analyzed, plus official publications and documents. Brands can identify trending topics that are buzzing with consumers and customers, collect feedback on product launches, and find new areas for business expansion.

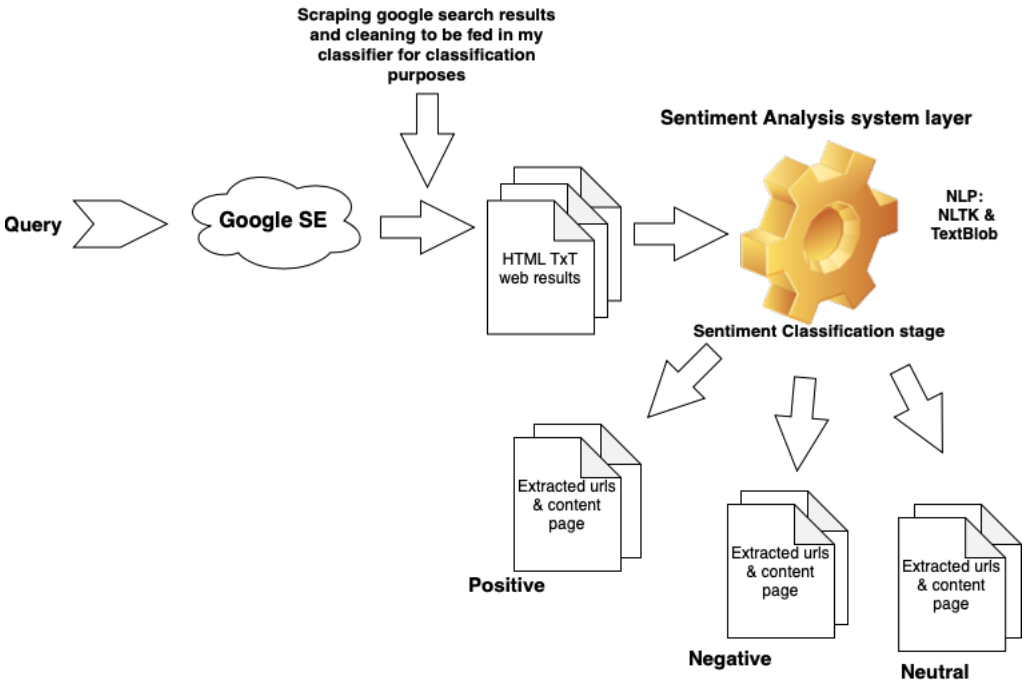
NCSU Tweet Visualizer — Sentiment Viz:
This is a cool freebie for Twitter sentiment analysis. Type in your keyword and the Tweet Visualizer pulls out recent tweets for the past week. Note that the time range is shorter for more popular subjects.

Social Searcher — Real-time search engine:
Real-time search engine for Twitter, and Google+. Includes filtering options that include post type, social channel, and sentiment. Each result has a color-coded button indicating positive, negative or neutral sentiment.

# Chapter 4

# System Architecture

From the figure below, the query goes to google with the use of API and perform a given query and then search results get cleaned for being feed into my classifier and then the output will depend on which filter the user chose to use depending on what the user wants to see either positive, negative or neutral.




Design Frank.png

Figure 4.1: Architectural design,designed from scratch using draw.io tool

# Chapter 5

# Methodology: Implementation

1. Searching from the web:
On this stage, the function takes input as any news topic and search for it in web and return website source and URL
@Parameter: topic(str)- News topic user input
@Return: webresultlist(list)- list of URLs.
Used Google Search: A python library that runs a Google search and fetch the individual results (full HTML and text contents)

2. Extracting content:
On this stage, the function takes URL as input and parse it to beautiful soup library to extract content and also clean it at the primary level with basic blacklist filter
@Parameter: URL(str)- needs to be extracted.
@Return: text(str)- extracted text.
Beautiful Soup is a Python package for parsing HTML and XML documents. It creates a parse tree for parsed pages that can be used to extract data from HTML, which is useful for web scraping. It is available for Python 2.7 and Python 3.

3. Cleaning text:
On this stage, the function is used to clean text for sentimental analysis
@Parameters: text(str): text to be cleaned.
@return: finaloutput(dict): url and search result.

4. Sentiment analysis:
On this stage, the function is being used to know the sentiment of text with the use of NLTK and TextBlob libraries.
@Parameter: text(list): list of words.
@return: sentiment(string): Positive or Negative.
NLTK is a leading platform for building Python programs to work with human language data. It provides easy-to-use interfaces to over 50 corpora and lexical resources such as WordNet, along with a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning, wrappers for industrial-strength NLP libraries.
TextBlob is a Python (2 and 3) library for processing textual data. It provides a simple API for diving into common natural language processing (NLP) tasks such as part-of-speech tagging, noun phrase extraction, sentiment analysis, classification, translation, and more.

5. Thread function:
On this last stage,the function takes the topic to be searched on web. First step, i will be getting the list of URLs and second is thread function to extract content.
@Parameter: topic(str)- to be searched on web.

The whole project is implemented in Python 3 ¿ Dgango library for web development.

# Chapter 6

# Experimental Results

The initial Homepage.

It contains a search bar for inputting the query and the below filter options as positive, negative or neutral. You can only choose one filter option.

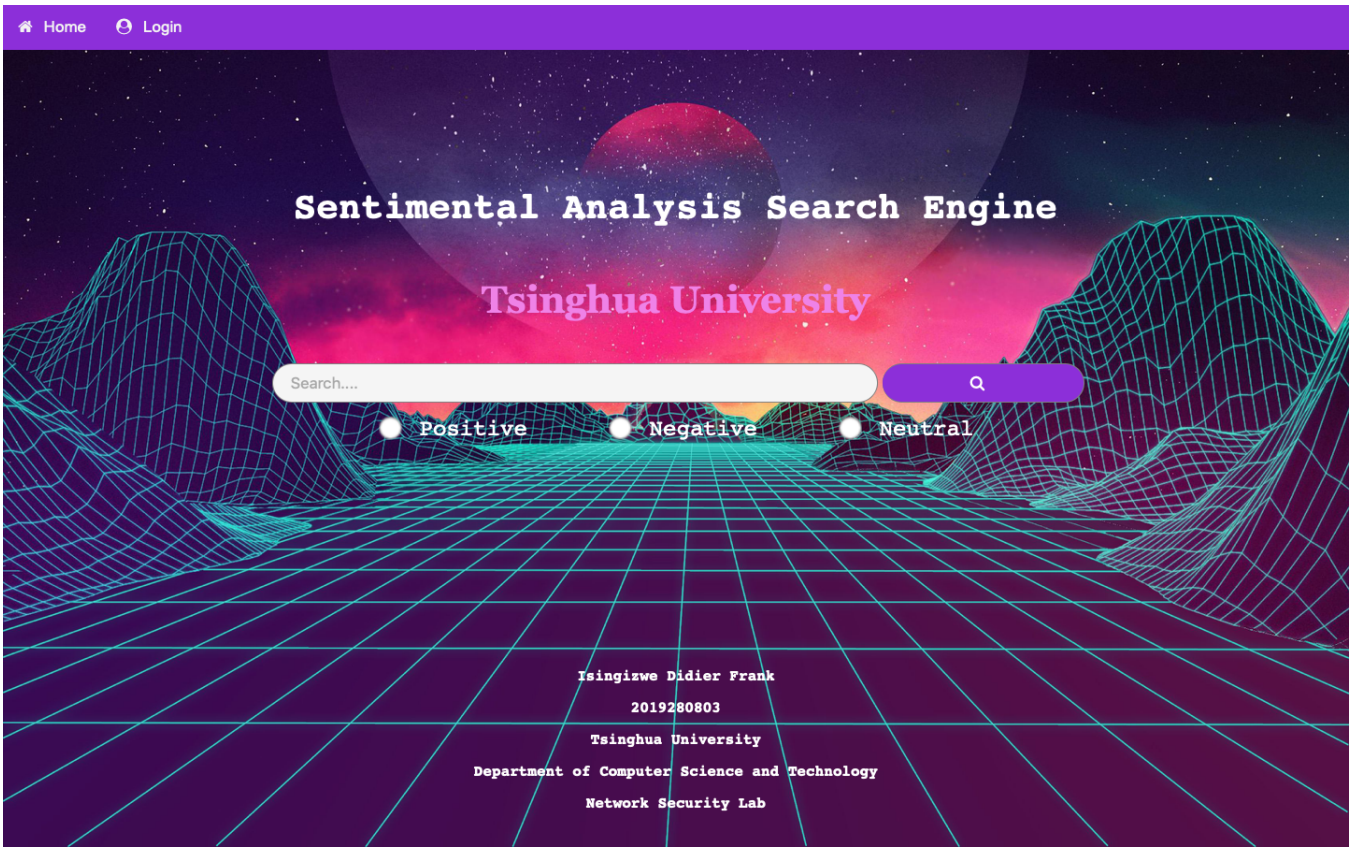



Figure 6.1: Homepage

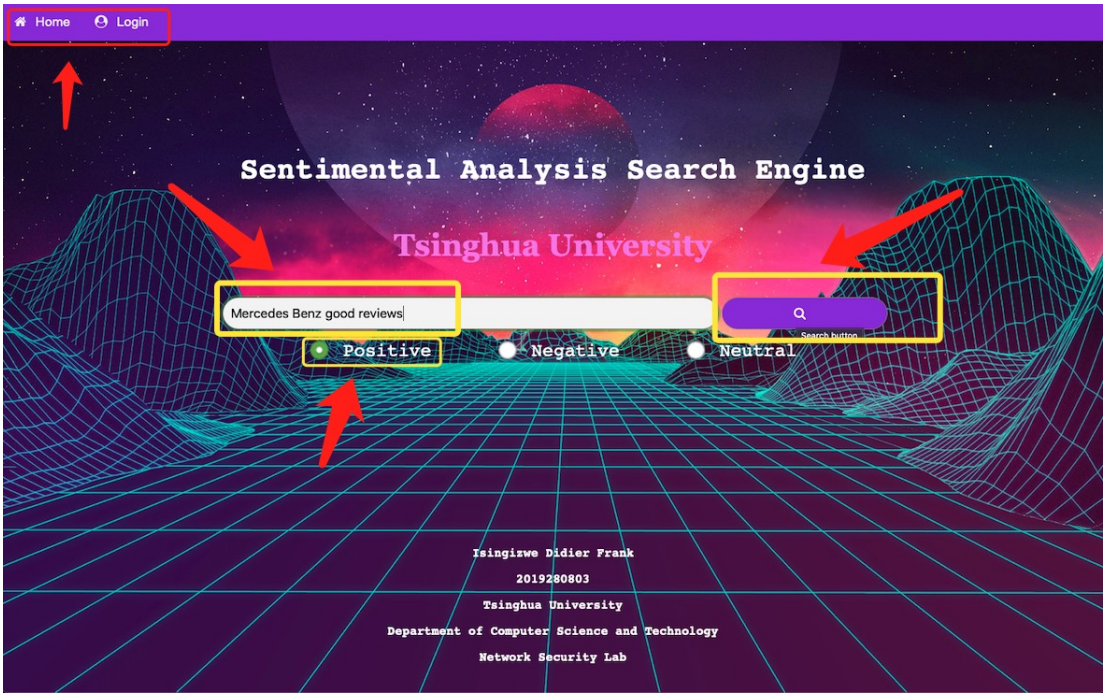



Figure 6.2: Positive scenario

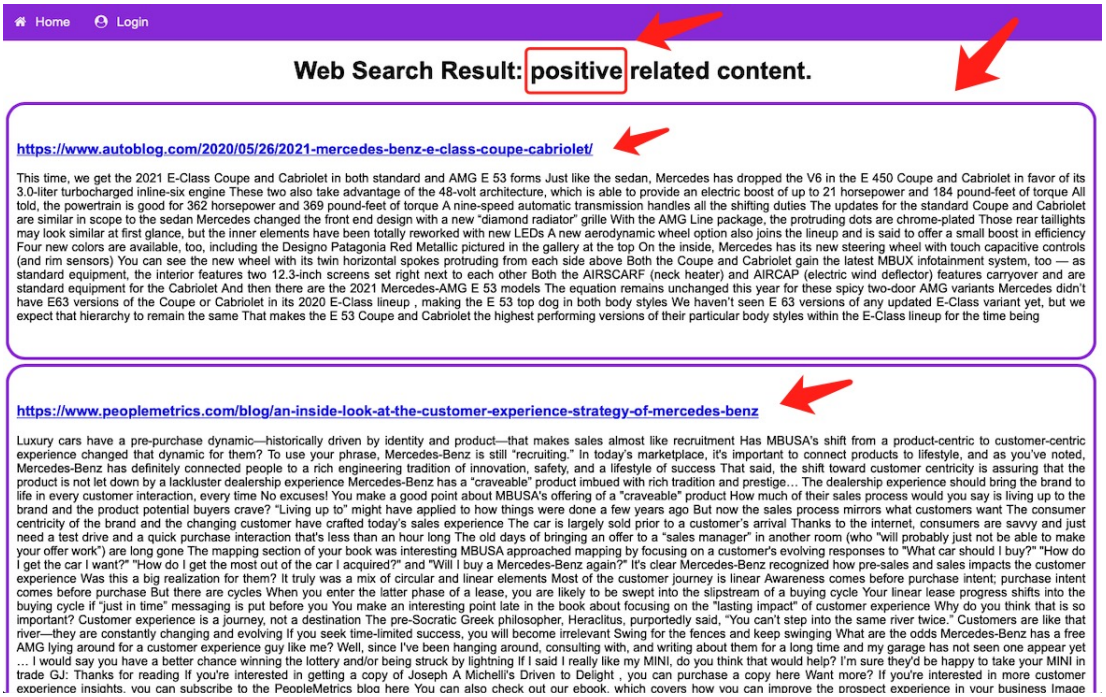



Figure 6.3: Positive scenario output

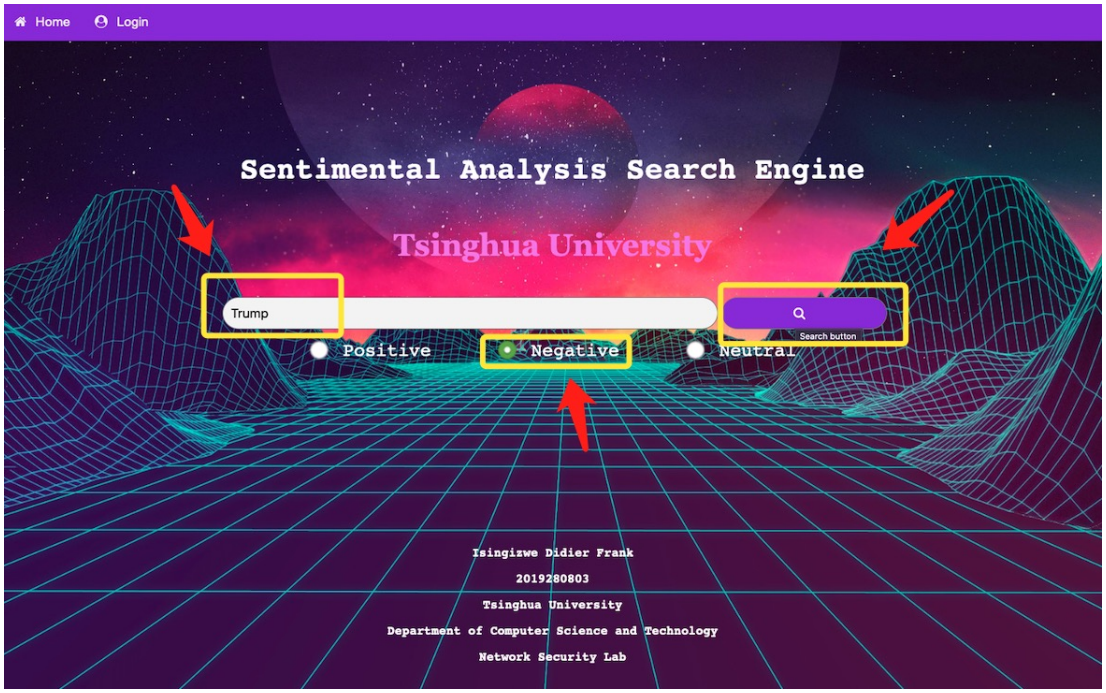



Figure 6.4: Negative scenario

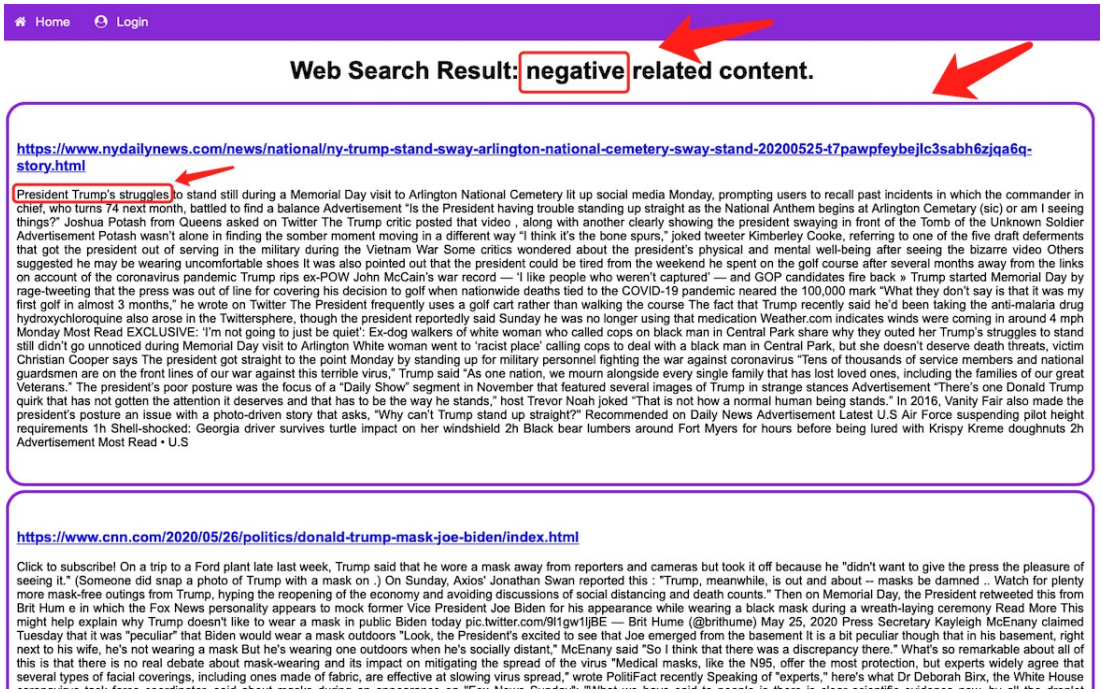



Figure 6.5: Negative scenario output

# Chapter 7

# Challenges

I faced issues with using Google,bing  baidu custom search APIs because most of them are paid. So, I manually did queries and fed them directly to my classifier.

The existing builtin library TextBlob accuracy is over 0.80 and it's not really great and reliable, it's still need further improvement.

Shortage GPU resources while using neural network approach in real-time in Django framework.

# Chapter 8

# Future Works

In the future contribution, the approach can be trying to use Convolutional Neural Networks or other approaches in classifying the data to compare accuracies with both approaches and come up with a great way to improve the accuracy.

The positive or Negative results shown will label essential keywords that affected the classifier to classify the whole paragraph as positive or negative. Preferably, all these keywords will be quantified to be considered in analytics (visualization purposes) for large scale companies.

Login Option for future personalizations and profiling.

# Chapter 9

# References

A. https://www.talkwalker.com/blog/best-sentiment-analysis-tools

B. https://textblob.readthedocs.io/en/dev/quickstart.html

C. https://pypi.org/project/beautifulsoup4/

D. https://www.geeksforgeeks.org/performing-google-search-using-python-code/

E. https://pypi.org/project/google-search/