VOICE RECOGNITION AND TWITTER SENTIMENT ANALYSIS

Nitin Prince Reuben,
Masters of Science in Information Systems,
Northeastern University, Boston(MA)

ABSTRACT

With the advancement of web technology and its growth, there is a huge volume of data present in the web for internet users and a lot of data is generated too. Internet has become a platform for online learning, exchanging ideas and sharing opinions. Social networking sites like Twitter, Facebook, Google+ are rapidly gaining popularity as they allow people to share and express their views about topics, have discussion with different communities, or post messages across the world. Twitter is a platform widely used by people to express their opinions and display sentiments on various occasions. There has been lot of work in the field of sentiment analysis of twitter data. This project focuses mainly on sentiment analysis of twitter data by extracting key words from an audio file, say from a speech using Google's speech recognition library to help analyze the information in the tweets where opinions are highly unstructured, heterogeneous and are either positive or negative, or neutral. Then we will deploy the whole model online using FLASK API.

KEYWORDS

Sentiment analysis, Machine Learning, Natural Language Processing, Python, Flask, GUI, Speech Recognition, Tableau.

INTRODUCTION

A lot of factors like stock prices, relationship between countries, etc. depends upon the political speeches. It is important for us to understand how general people react to it. Though we can miss some facts while we are noting things down but if we capture the speech in audio file and convert, there are very little chances we will miss any. The topics on which these speeches are delivered can create a huge debate in social media and public forums. One such social media where we can find general people and leaders who run the country is TWITTER.

In the past few years, there has been a huge growth in the use of microblogging platforms such as Twitter. Spurred by that growth, companies and media

Organizations are increasingly seeking ways to mine Twitter for information about what people think and feel about their products and services. Companies such as Twitrratr (twitrratr.com), tweetfeel (www.tweetfeel.com), and Social Mention(www.socialmention.com) are just a few who advertise Twitter sentiment analysis as one of their services.

On Twitter, users can share their opinions in the form of tweets, using only 280 characters. This leads to people compacting their statements by using slang, abbreviations, emoticons, short forms etc. Along with this, people convey their opinions by using sarcasm and polysemy.

To extract sentiment from tweets, sentiment analysis is used. The results from this can be used in many areas like analyzing and monitoring changes of sentiment with an event, sentiments regarding a brand or release of a product, analyzing public view of government policies etc.

A lot of research has been done on Twitter data to classify the tweets and analyze the results. In this paper we aim to review of some researches in this domain and study how to perform sentiment analysis on Twitter data using Python. But before that we are aiming to convert audio file to text. This is done with a vision that we can convert audio recordings of political speeches and perform twitter analysis on them which will help us understand how these speeches effects life of normal people.

In a time where everything is getting automated, we are trying to run our model online. We will build a web interface through which we can take input from people online and give them the output through web page. Though you can run twitter sentiment analysis on local desktop but our vision was to provide an application which can help people especially those who have less knowledge about data science, be able to perform twitter sentiment analysis.

VOICE RECOGNITION

So, we started with converting audio to text by using SpeechRecognition, a library provided by GOOGLE. This library takes audio file as input and gives text as output. There isn't much scope for tweaking this library as this is provided by google and it is difficult to understand the back-end code. There are few drawbacks as well. We can convert only 'English' audio as the system is not very well equipped to convert other language audio files. There are libraries like 'pocketsphinx' which does the same work but the conversion happens in our local desktop. This makes things easier for us as we can clearly understand the back-end code and also get a chance to tweak it. Neural network is also an option to make this possible. But because of time constraint, we proceeded with SpeechRecignition only.

FILE TYPE CONVERSION

On analyzing, we found that Google speech library accepts only '.wav' and it is difficult to make sure every time to provide audio file in that format.So, we used another library called 'pydub' which converts the file format of audio file from any one format to another. We used this to convert

audio files which are in other format to '.wav' file which helps our voice conversion library.

SENTIMENT ANALYSIS

Sentiment analysis is a process of deriving sentiment of a statement or sentence. It's a classification technique which derives opinion from the tweets and formulates a sentiment and based on which, sentiment classification is performed. Sentiments are subjective to the topic of interest. We are required to formulate that what kind of features will decide for the sentiment it embodies.

In the programming model, sentiment we refer to, is class of entities that the person performing sentiment analysis wants to find in the tweets. The dimension of the sentiment class is crucial factor in deciding the efficiency of the model.

For example, we can have two-class tweet sentiment classification (positive and negative) or three class tweet sentiment classification (positive, negative and neutral).

Sentiment analysis approaches can be broadly categorized in two classes – lexicon based, and machine learning based.

Lexicon based approach is unsupervised as it proposes to perform analysis using lexicons and a scoring method to evaluate opinions. Whereas machine learning approach involves use of feature extraction and training the model using feature set and some dataset.

The basic steps for performing sentiment analysis includes data collection, pre-processing of data, feature extraction, selecting baseline features, sentiment detection and performing classification either using simple computation or else machine learning approaches

NATURAL LANGUAGE PROCESSING (NLTK)

Natural Language toolkit (NLTK) is a library in python, which provides the base for text processing and classification. Operations such as tokenization, tagging, filtering, text manipulation can be performed with the use of NLTK.

The NLTK library also embodies various trainable classifiers (example – Naïve Bayes Classifier).

NLTK library is used for creating a bag-of words model, which is a type of unigram model for text. In this model, the number of occurrences of each word is counted. The data acquired can be used for training classifier models. The sentiment of the entire tweets is computed by assigning subjectivity score to each word using a sentiment lexicon.

FLASK

Flask is a micro web framework written in Python and based on the Werkzeug toolkit and Jinja2 template engine. It is BSD licensed.

Flask is called a micro framework because it does not require tools or libraries. It has no database abstraction layer, form validation, or any other components where pre-existing third-party libraries provide common functions. However, Flask supports extensions that can add application features as if they were implemented in Flask itself. Extensions exist for object-relational mappers, form validation, upload handling, various open authentication technologies and several common framework related tools. Extensions are updated far more regularly than the core Flask program
Using Flask, we can make a web page and run entire code through a browser. This makes our model more automated. We can deploy the application on web using either docker or command prompt.

DATA COLLECTION AND CLEANING

We are using Tweepy - client for Twitter Application Programming Interface (API) to scrape data from twitter and then since most of the tweets are in different languages as well we shall be using a dictionary to covert other languages to English as well to get a broader analysis.

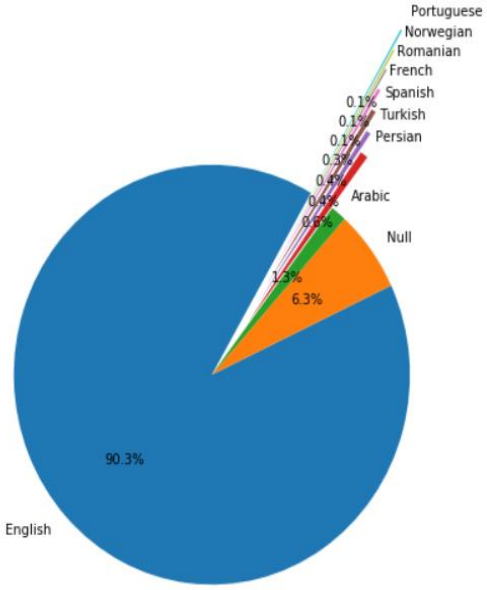

Fig 1

After scraping data on topic '#FirstDayOfSpring', we got 775 tweets from various languages. We converted the languages from their abbreviation to full form using a dictionary. The description of language is depicted in fig 1.

Along with tweet, location from where it's tweeted is also scrapped. Using 'tableau', it's plotted on a geological graph. Fig 2 describes the countries from where got the tweets. Fig 3 describes the states in USA from where people have tweeted about this topic.

Fig 2

Fig 3

Setting Up Environment for Sentiment Analysis Using Python

The following components are required to be downloaded and installed properly.
Download and install Python 2.6 or above in a desired location.

Download and install NumPy.

Download and install NLTK library.

Download and install Scikit-learn library.

APPLICATIONS

Commerce: Companies can make use of this research for gathering public opinion related to their brand and products. From the company's perspective the survey of target audience is imperative for making out the ratings of their products. Hence Twitter can serve as a good platform for data collection and analysis to determine customer satisfaction.

Politics: Majority of tweets on Twitter are related to politics. Due to Twitter's widespread use, many politicians are also aiming to connect to people through it. People post their support or disagreement towards government policies, actions, elections, debates etc. Hence analyzing data from it can help is in determining public view.

Sports Events: Sports involve many events, championships, gatherings and some controversies too. Many people are enthusiastic sports followers and follow their favorite players present on Twitter. These people frequently tweet about different sports related events. We can use the data to gather public view of a player's action, team's performance, official decisions etc.

FUTURE SCOPE

We can automate this application even more. We can create a docker image of this application and upload in AWS or google SDK. Once it's live, we can get a URL with which we can run this application in any device at any place. This makes this model even more dynamic and usable to local public.

CONCLUSION

Twitter sentiment analysis comes under the category of text and opinion mining. It focuses on analyzing the sentiments of the tweets and feeding the data to a machine learning model to train it and then check its accuracy, so that we can use this model for future use according to the results. It comprises of steps like data collection, text pre-processing, sentiment detection, sentiment classification, training and testing the model. This research topic has evolved during the last decade with models reaching the efficiency of almost 85%-90%. But it still lacks the dimension of diversity in the data. Along with this it has a lot of application issues with the slang used and the short forms of words. Many analyzers don't perform well when the number of classes are increased. With the growing and evolving technology, were everything is automated, we are trying to move a step ahead with the resources which we have. We have made progress in that direction and will continue moving forward.

REFERENCES

1. *David Zimbra, M. Ghiassi and Sean Lee, "Brand-Related Twitter Sentiment Analysis using Feature Engineering and the Dynamic Architecture for Artificial Neural Networks", IEEE 1530-1605, 2016.*

2. *Varsha Sahayak, Vijaya Shete and Apashabi Pathan, "Sentiment Analysis on Twitter Data", (IJIRAE) ISSN: 2349-2163, January 2015*.

3. *Peiman Barnaghi, John G. Breslin and Parsa Ghaffari, "Opinion Mining and Sentiment Polarity on Twitter and Correlation between Events and Sentiment", 2016 IEEE Second International Conference on Big Data Computing Service and Applications.*

4. *International Journal of Computer Applications (0975 – 8887) Volume 165 – No.9, May 2017*

5. Flask - https://en.wikipedia.org/wiki/Flask_(web_framework)