SENTIMENT ANALYSIS OF REAL TIME GEO TAGGED TWEETS

Prepared by

QAIS MAKANI…..71101110041

YASH RATHOD…..71101110009

CHINMAY PARIKH…..71101128001

Under the guidance of

Prof. KETAN SHAH

in partial fulfillment of the requirements of PROJECT

for the award of the degree of

Bachelor of Technology

IN

INFORMATION TECHNOLOGY

At

Department of Information Technology
Mukesh Patel School of Technology Management & Engineering
NMIMS (Deemed –to-be University)
JVPD Scheme Bhaktivedanta Swami Marg,
Ville Parle (W), Mumbai-400 056.

CERTIFICATE

This is to certify that the project entitled "SENTIMENT ANALYSIS OF REAL TIME GEO-TAGGED TWEETS" is the bonafide work carried out by QaisMakani, YashRathod, Chinmay Parikh…of B.Tech (IT), MPSTME, Mumbai, during the VII Semester of the academic year 2014-2015, in partial fulfillment of the requirements for the award of the degree of Bachelors of Technology as per norms prescribed by NMIMS. The project work has been assessed and found to be satisfactory.

Signature of the Mentor
Prof. Ketan Shah
Associate Professor
Date: 20-11-2014
(Internal Mentor)

HOD (IT)
(Prof. Vijay Raisinghani)

Dean
(Dr.SharadY.Mhaiskar)

DECLARATION

We, QaisMakani, YashRathod, Chinmay Parikh roll numbers: A029, A042, A059 respectively, understand that plagiarism is defined as any one or the combination of the following:

1. Uncredited verbatim copying of individual sentences, paragraphs or illustrations (such as graphs, diagrams, etc.) from any source, published or unpublished, including the internet.
2. Uncredited improper paraphrasing of pages or paragraphs (changing a few words or phrases, or rearranging the original sentence order)
3. Credited verbatim copying of a major portion of a paper (or thesis chapter) without clear delineation of who did or wrote what. (Source: IEEE, The Institute, Dec. 2004)

We have made sure that all the ideas, expressions, graphs, diagrams, etc., that are not a result of our work, are properly credited. Long phrases or sentences that had to be used verbatim from published literature have been clearly identified using quotation marks.

We affirm that no portion of our work can be considered as plagiarism and we take full responsibility if such a complaint occurs. We understand fully well that the guide of the seminar/project report may not be in a position to check for the possibility of such incidences of plagiarism in this body of work.

Signature:
Name: QaisMakani
Roll No.: A029
Date: 20-11-2014

Signature:
Name: YashRathod
Roll No.: A042
Date: 20-11-2014

Signature:
Name: Chinmay Parikh
Roll No.: A059
Date:20-11-2014

Contents

List of Figures

List of Table

Abstract

Due to leaps in social media and technology, almost everyone is connected via a social network. Twitter is the most popular of the social media sites. It uses a simple follow model (you only receive updates from the people you follow). Users update their wall by posting “tweets”. “Tweets” are one short (140 character limit) updates.

Companies which also have a twitter account are followed by their customers. Such companies desire to know the how well their product is doing on the market by analyzing customer tweets. This analysis for positive and negative feedback is known as “Sentiment Analysis”.

It’s easy for humans to decipher the tweets left by the customers and sort them into the three categories. But computers don’t understand human language or the context of the tweet or concepts such as sarcasm. The application of Natural Language processing together with Machine Learning helps close the gap between computers and humans. Sentiment Analysis uses Machine Learning to train a “Classifier”. The “Classifier” is trained using training data, in which tweets are fed with “labels” to the “Classifier”. Example of the Training Data set:

"I love Holidays", "Positive" "I hate when spiders appear out of thin air", "Negative" "Chocolate is the BEST", "Positive" "The Rains are depressing", "Negative".

Then Extraction which is done using Natural Language processing to extract “Features” from the tweet. We get:

’Love’ , Positive probability 97.43% ’hate’ , Negative probability 94.64% ’BEST’ , Positive probability 88.41% ’depressing’ , Negative probability 84.43%.

After Training the “Classifier” it can now Use it on test data. One may automate this so as to allow for self-learning.

Chapter 1:

Overview

Sentiment Analysis in twitter has been used previously on various English words used by people word wide. Well not only the words also the emoticons which are used have a specific meaning. For example, "☺" means happy, "☹" means sad, etc.
According to the literature survey Sentiment Analysis is has been handled as a Natural Language Processing task at various levels of granularity. It has been performed on various levels: document level, sentence level and recently on phrase level.
Since twitter is a microblogging site, users post real time opinions and reactions for almost everything they have done or seen. For example a person posts about a movie or comments about his or her day at the office, etc. Newer typing styles have made it difficult for any classifier which is designed to perform analysis, like using incomplete word or just an initial.

Previous analysis performed have done using different techniques like Naïve Bayes, Support Vector Machine or any other have shown a great variation in performance and efficiency.
The classifier used shows a better efficiency but there are some problems which still need to be countered, like sarcasm in sentence or named entity recognition, spelling mistakes takes or using short forms which increase the complexity of the classifier.

Alex Davis research paper shows how he has implemented his own method and technique. We have shown our mentor the demo version of the project. But since it was a prototype and the classifier needs to be trained with millions of tweets it will be appended at a later stage.

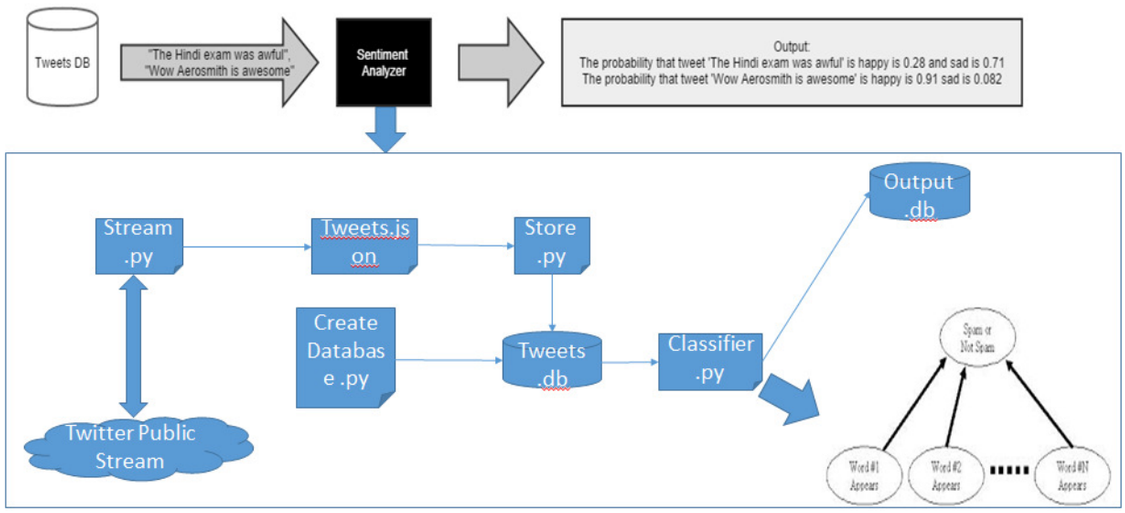
Chapters 2:

Proposal

Objectives:

Listed According to priority

1. Prototype Which Runs on a pre-compiled “Corpus”. Prototype Deliverable.
2. Enhance the Tweet Gathering System to make it Easy to use. GUI Version Deliverable.
3. Implement Either the Naive Bayes Classifier or the Support Vector Machine. Alpha Stage Deliverable.
4. Enhance the Feature Extraction and add Stemming or other methods to improve Accuracy. Alpha Stage Deliverable.
5. Implement a World map to display the real power use of the Geo-tagged property of the tweets. This allows companies to track users of their product and the regional areas in which their product is facing problems. Beta and Final Product Deliverable.



The data flow diagram above represents the overview of the entire project from the beginning till the end.

Starting with the collection of tweets and storing it in a database using SQLite. The tweets are stored into a database using a python API, “tweepy”. It provides with the consumer key which helps in collecting all the tweets.

After storing the tweets in a database we perform a sentiment analysis using a classifier from Alex Davis research paper.

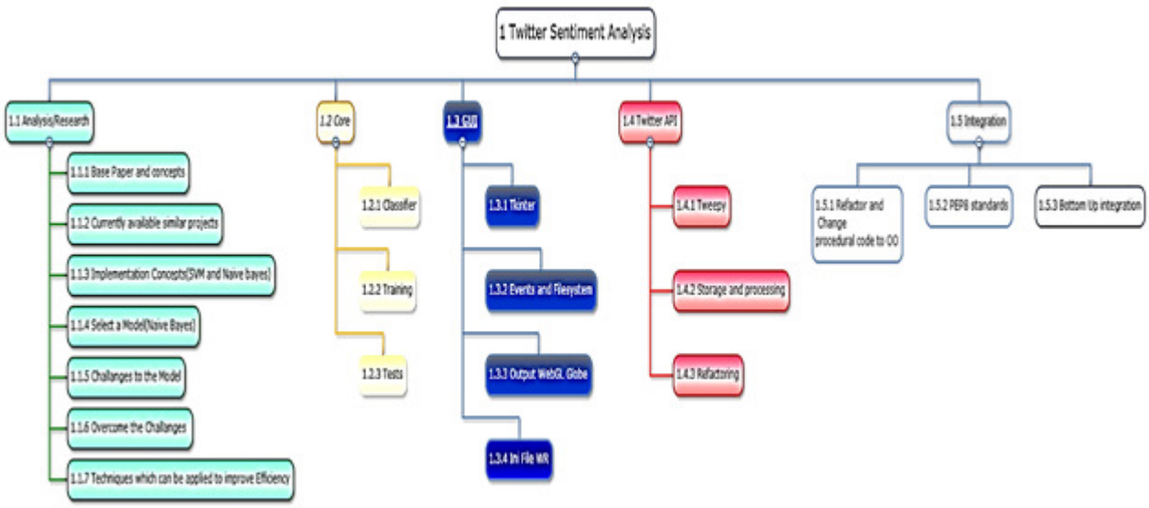
The Sentiment Analysis is performed using a naïve bayes classifier. We came to this conclusion of selecting naïve bayes classifier after comparing it with support vector machine. The result and performance of naïve bayes were better than svm. Streaming.py is a python file which helps in streaming with the help of API. It stores the tweet using JSON file parser which helps us to individually classify tweets via GET request. The tweets are filtered and then the analysis is performed sorting it according to positive and negative probability of the tweets and then storing back into the database.

The result of the classifier is stored into the database can now be connected to a Web Globe where it can display the origin of the tweet and the along with that also the happy and sad probability of the tweet indicating it with colors.

Chapter 3:

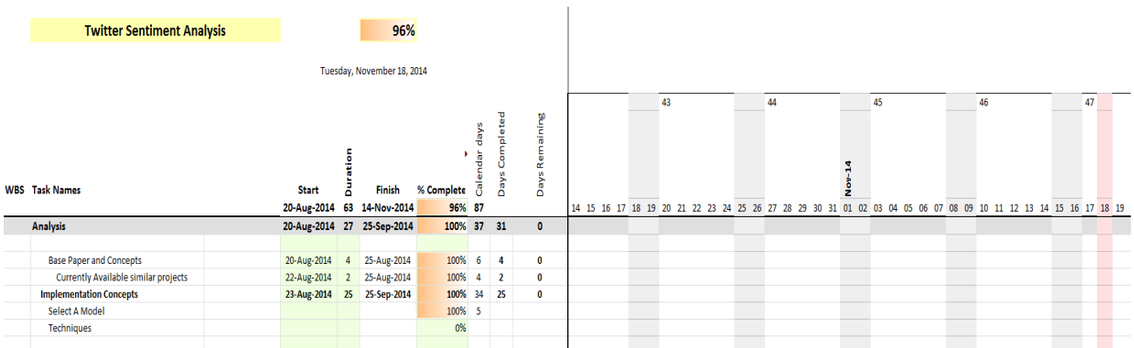
Planning

The whole project of sentiment analysis has been broken down into, separate task and each task into subtask. The Work Breakdown Structure is shown below:

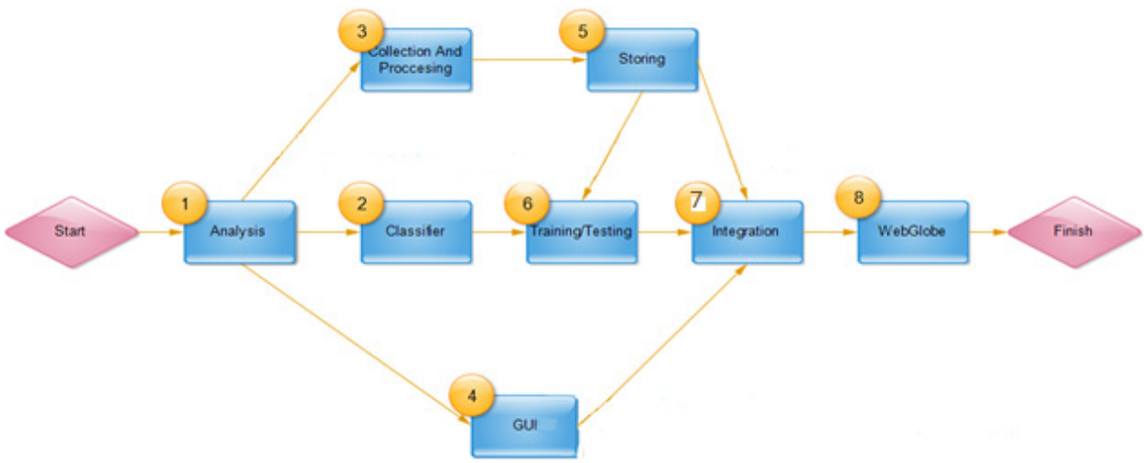


The work breakdown shown above is used in a Gantt chart to display the estimated schedule of the project.

Twitter Sentiment Analysis				96%			
				Tuesday, November 18, 2014			
WBS	Task Names	Start	Duration	Finish	% Complete	Calendar days	Days Completed	Days Remaining
		20-Aug-2014	63	14-Nov-2014	96%	87		
	Analysis	**20-Aug-2014**	**27**	**25-Sep-2014**	**100%**	**37**	**31**	**0**
	Base Paper and Concepts	20-Aug-2014	4	25-Aug-2014	100%	6	4	0
	Currently Available similar projects	22-Aug-2014	2	25-Aug-2014	100%	4	2	0
	Implementation Concepts	**23-Aug-2014**	**25**	**25-Sep-2014**	**100%**	34	25	0
	Select A Model				100%	5		
	Techniques				0%			
	GUI	**25-Aug-2014**	**60**	**14-Nov-2014**	**100%**	**82**	**80**	**0**
	Tkinter	**25-Aug-2014**	**60**	**14-Nov-2014**	**100%**	82	60	0
	Events And FileSystem	28-Aug-2014	10	10-Sep-2014	100%	14	10	0
	Ini File R/W	03-Sep-2014	10	16-Sep-2014	100%	14	10	0
	WebGlobe	03-Sep-2014	25	07-Oct-2014	50%	35	12	13
	Twitter API	**25-Aug-2014**	**30**	**03-Oct-2014**	**100%**	**40**	**30**	**0**
	Tweepy	**25-Aug-2014**	30	03-Oct-2014	**100%**	40	30	0
	Storage and Processing				100%	10		
	Refactoring				0%			



The next diagram shows the PERT chart of the project. The tasks have been numbered accordingly.



First we analyze our project, what all previous work has been performed in this topic which we have selected and what modifications can we implement. We analyze the resources and the risk which may be involved.

After analysis we select the classifier for tweets which will be gathered.

The next step is collection of tweets and processing them.

After the tweets have been stored if the user wants to filter the tweets he can do it by entering the keyword in GUI. The GUI helps in filtering the tweets and using location and after all the processing is done it is integrated with all the necessary components.

After the tweets are stored it is trained and tested for processing and for decision making.

All the components are integrated together and testing and refactoring of codes.

The result is displayed using 3D Web Globe which a graphical presentation of the tweets along with the probability.

Chapter 4:

Formulation

After thorough analysis of existing works and technologies, the following technology action plan was decided.

Table 4(a) – Activity Table

Sr. no.	Module	Technology
1	Input	Python TKinter GUI
2	Output	Google WebGl Globe
3	Storage	SQLite
4	Tweet gathering	Python tweepy (Twitter API)
5	Classifier	Python numpy
6	Tweet parsing	Python simpleJSON
7	INI R/W	Python ConfigParser

1. Python: Many advantages of python like platform independence, access to vast libraries, optional object orientation, and seamless integration with other programming languages made python the prime candidate while choosing a programming language for the project, but the main aspect of python that led to its selection was ease of implementation and maintenance. Python will be used to implement the entire project; the GUI, gathering and storing tweets and output, and the classifier.

2. Google Web Globe: It's a 3-D interactive globe which supports actions like rotating and zooming. This globe will be most useful to highlight the geo-tagged aspect of the tweets. It will display the classifier output for each tweet in a geographical accurate manner on the globe. The output will be color code; red for negative and green for positive. This will make it easier

for users to check how each part of the world is reacting to their product, etc.

3. SQLite: This will be used to store everything in the project; the tweets along with other relevant fields and the output as well, in separate tables. This was mainly chosen because it is not implemented as a separate process that a client program running in another process accesses. Rather, it is part of the using program.

4. JSON: This file does the pre-processing of the contents or fields which are necessary for the content filtering and the globe. There are around 40 attributes or fields which can be used for storing data.

   The main contents which are necessary for pre-processing are:

   a) User ID
   b) User Tag
   c) Location like city name, country name
   d) Location via coordinates (latitudes and longitudes)
   e) Retweet from and retweet count
   f) Exact time of the tweet
   g) Tweeted text
   h) Optional – hashtags included

5. Classifier: The classifier is the main part of the project. It'll use Naïve Bayes Algorithm to classify the gathered tweets as either positive or negative using a corpus. It will be implemented in Python and will require library named numpy.

Chapter 5:

Progress

The three parts of the projects are:

1. Get The Tweets - Twitter Streaming API.
2. Parse and Store the Tweets - JSON and SQLITE 3.
3. Analyze the Tweets - Sentiment Analysis.

A basic prototype of the project has been implemented in python which includes the three parts that have been mentioned above. The tweet gathering, parsing and storing are working as intended. The analyzing part has been implemented using a simple corpus based algorithm which reads from a dictionary of words, each assigned with a pre-defined happy and sad probability. The dictionary of words is stored in a csv file containing all the possible words which are most commonly used. The tweets are then broken down into words and each word, based on the corpus is assigned their happy and sad probability. Then the probability of the sentence or tweet is determined by adding its constituent words. It's a simple algorithm that has been implemented as a place holder for the real deal, which will be implemented in the near future.

The GUI part is still under development, and only a rough sketch of it has been developed. It will undergo many more changes until the design is finalized and implemented.

The output of the project is to be displayed on a globe with happy and sad tweets lighting up with green and red colors respectively at geographically accurate positions using the geo-tagged property of tweets. The colors can be changed as per the requirements. The code for the globe is available but it is undergoing some modification to display the colors for the tweets and connecting it to the database.

Chapter 6:

Conclusion

At this point in time, major chunk of analysis part has been complete. The technology implementing the modules has been chosen and a basic prototype of the project has been implemented. This was done to experiment and checked if the project is feasible; to rule out a chunk of the issues that might occur and get a picture as to what the project would be like given our current vision. This will be helpful in coming up with a better final product.

The GUI part will be developed using the Tkinter library for python.

The results of the classifier will be stored in the database and will be displayed on the WebGL globe which still needs to be developed.

References:

Corpus:

- http://www.sananalytics.com/lab/twitter-sentiment/
- http://thinknook.com/twitter-sentiment-analysis-training-corpus-dataset-2012-09-22/help.sentiment140.com/for-students/

Papers:

- http://kaikuehne.github.io/war/Pak%20and%20Paroubek%20%282010%29.%20Twitter%20as%20a%20Corpus%20for%20Sentiment%20Analysis%20and%20Opinion%20Mining.pdf
- http://s3.eddieoz.com/docs/sentiment_analysis/Twitter_Sentiment_Classification_using_Distant_Supervision.pdf
- http://www.sussex.ac.uk/Users/christ/crs/ml/lec02b.html
- http://eprints.pascal-network.org/archive/00008437/01/Language_Indepedent_Bayesian_Sentiment_Mining_of_Twitter.pdf
- Fast and accurate sentiment classification using an enhanced Naive Bayes model. Vivek Narayanan1, Ishan Arora2, Arjun Bhatia3. Department of Electronics Engineering, Indian Institute of Technology (BHU), Varanasi, India. http://arxiv.org/ftp/arxiv/papers/1305/1305.6143.pdf
- TweePy Library – (www.tweepy.org)
- B. Pang and L. Lee. "Opinion Mining and Sentiment Analysis" in Foundations and Trends in Information Retrieval, 2008.