A Study on Twitter 4j Libraries for Data Acquisition from Tweets

Shilpy Singh
Dept of ISE
BMSIT&M, Bangalore
Karnataka, India

Manjunath T N
Dept of ISE
BMSIT&M, Bangalore
Karnataka, India

Aswini N
Dept of ISE
BMSIT&M, Bangalore
Karnataka, India

ABSTRACT
Due to various programming languages & its compatibility issues with databases and utilization some lead to develop libraries for reusable patterns. In this paper, we explore the utilization of Twitter4J libraries which are reliable Twitter APIs and that can be integrated to any applications for data acquisition in any format. It is a cross-platform tool and can be used on several operating systems, with the latest versions of Java Runtime Environment. The utility can be used as it is without any customizations it has no dependencies to any other system on which it runs. The usage of Twitter4J is simple, as all you need to do is copy the JAR file to the preferred classpath and use it. Here we explore the method of using twitter4J libraries for data acquisition for data analytics. This work will help data scientist, data quality analyst and business users.

Keywords
Twitter 4j library, twitter sentimental analysis.

1. INTRODUCTION
Twitter4J is an official Java library for the Twitter API. Twitter4J one can easily integrate any application with the Twitter service. Twitter4J has features such as, 100% runs on Java Platform version 5 or later, Android platform and Google App Engine ready, Zero dependency, No additional jars required, Built-in OAuthsupport, Out-of-the-box gzip support, 100% Twitter API 1.1 compatible. By adding twitter4j-core-4.0.4.jar to any application class path. If you are familiar with Java language, looking into the JavaDoc should be the shortest way for you to get started twitter4j. Twitter interface is the one you may want to look at first.

1.1 Sentimental Analysis
Sentiment analysis is a popular research area in computer science. It aims to determine the attitude of a person with respect to some topic, such as his mood or opinion from textual documents generated by the person. With the proliferation of social micro-blogging sites, opinion text has become available in digital forms, thus enabling research on sentiment analysis to both deepen and broaden in different sociological fields. Twitter is a popular microblogging service where users create status messages (called "tweets"). These tweets sometimes express opinions about different topics. The purpose of this project is to build an algorithm that can accurately classify Twitter messages as positive or negative, with respect to a query term. Our hypothesis is that we can obtain high accuracy on classifying sentiment in Twitter messages using machine learning techniques. Generally, this type of sentiment analysis is useful for consumers who are trying to research product or service, or marketers researching public opinion of their company [3][4].

1.2 Twitter Sentiment Analysis
The sentiment can be found in the comments or tweet to provide useful indicators for many different purposes .Also, and stated that a sentiment can be categorized into two groups, which are negative and positive words. Sentiment analysis is a natural language processing techniques to quantify an expressed opinion or sentiment within a selection of tweets. Sentiment analysis refers to the general method to extract polarity and subjectivity from semantic orientation which refers to the strength of words and polarity text or phrases .There has two main approaches for extracting sentiment automatically which are the lexicon-based approach and machine-learning-based approach. Lexicon-based Approach Lexicon-based methods make use of predefined list of words where each word is associated with a specific sentiment .The lexicon methods vary according to the context in which they were created and involve calculating orientation for a document from the semantic orientation of texts or phrases in the documents. Techniques of Sentiment Analysis, The semantic concepts of entities extracted from tweets can be used to measure the overall correlation of a group of entities with a given sentiment polarity. Polarity refers to the most basic form, which is if a text or sentence is positive or negative.

2. RELATED WORK
There have been many papers written on sentiment analysis for the domain of blogs and product reviews. (Pang and Lee 2008) gives a survey of sentiment analysis. Researchers have also analyzed the brand impact of microblogging (Jansen). We could not find any papers that analyze machine learning techniques in the specific domain of microblogs, probably because the popularity of Twitter is very recent. Overall, text classification using machine learning is a well studied field (Manning and Schuetze 1999). (Pang and Lee 2002) researched the effects of various machine learning techniques (Naive Bayes (NB), Maximum Entropy (ME), and Support Vector Machines (SVM)in the specific domain of movie reviews. They were able to achieve an accuracy of 82.9%using SVM and a unigram model. Researchers have also worked on detecting sentiment in text. (Turney 2002) presents a simple algorithm, called semantic orientation, for detecting sentiment. (Pang and Lee 2004) present a hierarchical relevant to twitter because many users have emoticons in their tweets. Twitter messages have many unique attributes, which differentiates our research from previous research:

1. Length. The maximum length of a Twitter message is 140 characters. From our training set, we calculated that the average length of a tweet is 14 words, and the average length of a sentence is 78 characters. This is very different from the domains of previous research, which was scheme in which

text is first classified as containing sentiment, and then classified as positive or negative. Work (Read, 2005) has been done in using emoticons as labels for positive and sentiment. This is very mostly focused on reviews which consisted of multiple sentences.

2. Language model. Twitter users post messages from many different mediums, including their cell phones. The frequency of misspellings and slang in tweets is much higher than other domains [2][3].

3. J. Spencer, and G. Uchyigit in 2012 highlighted the usage of twitter4j for sentimental analysis [2].

4. Akshi Kumar, Prakhar Dogra and Vikrant highlighted the importance of Emotion Analysis of Twitter using Opinion Mining" during 2015.

3. PROCESSING METHOD & ANALYSIS

3.1 Natural Language Processing (NLP)

NLP techniques are based on machine learning and especially statistical learning which uses a general learning algorithm combined with a large sample, a corpus, of data to learn the rules .Sentiment analysis has been handled as a Natural Language Processing denoted NLP, at many levels of granularity. Starting from being a document level classification task, it has been handled at the sentence level and more recently at the phrase level .NLP is a field in computer science which involves making computers derive meaning from human language and input as a way of interacting with the real world.

3.2 Support Vector Machine (SVM)

Support Vector Machine is to detect the sentiments of tweets .together with stated SVM is able to extract and analyze to obtain upto70%-81.3% of accuracy on the test set. Collected training data from three different twitter sentiment detection websites which mainly use some pre-built sentiment lexicons to label each tweet as positive or negative. Using SVM trained from these noisy labeled data, they obtained 81.3% in sentiment classification accuracy.

3.3 Defining Sentiment Analysis

Here we have defined Sentiment to be "a personal positive or negative feeling."Here are some examples: For tweets that were not clear-cut, we use the following litmus test: If the tweet could ever appear as a newspaper headline or as a sentence in Wikipedia, then it belongs in the neutral class. For example, the following tweet would be marked as neutral because it is fact froma newspaper headline, even though it projects an overall negative feeling about the Tweet under consideration[12][11]:

Table-1: Sentimental Analysis Definition

Sentiment	Query	Tweet
positive	Jquery	dcostalis: Jquery is my new best friend.
Neutral	San Francisco	schuyler: just landed at San Francisco
Negative	exam	jvicious: History exam studying.

4. DATA ACQUISITION - REPRESENTATION

4.1 Sentiment Analysis

Tweets from JSON file will be assigned the value of each word by matching with the lexicon dictionary. As a limitation of words in the lexicon dictionary which is not able to assign value to every single word from tweets. However, as scientific language of python, which is able to analyze a sense of each tweet into positive or negative for getting a result.

4.2 Information Presented

The result will be shown in a pie chart which is representing a percentage of positive, negative and null sentiment hash tags. For null hash tag is representing the hash tags that were assigned zero value. However, this program is able to list a top ten positive and negative hash tags.

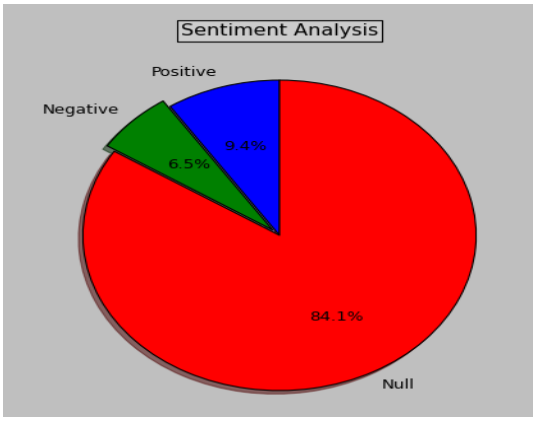

Fig-1: Pie Chart- Sentimental Analysis

As shown in Fig-1, the pie chart is representing of each percentage positive, negative and null sentiment hash tags in different color.

5. HIGH-LEVEL OVERVIEW

Here we stress on the high level overview of twitter streaming with different options:

Option 1: we can download thousands of tweets from Twitter. We will compare the various types of geographic information in these tweets.

Option 2: This is the creative option. Other common sources of geo referenced social media include Foursquare, Facebook, and Flickr, but there are many others as well. If we choose this option, our job will be to read through this assignment which is written for Twitter and do something of equal or greater depth on your social network of choice. The below steps highlights the twitter streaming:

Step 1: Setting Things Up: Sign up

Go to https://dev.twitter.com/ and sign up as a Twitter developer. Make a new application called whatever want (e.g. "CSCI5980Tweets_<YOURNAME>").Choose a language and a library for working with the Twitter API. A list of common libraries can be found here: https://dev.twitter.com/docs/twitterlibraries.The API we choose has "streaming API" support (Googling the API name and "Streaming API" should do the trick).we'll also want to choose a library in a language with which we are quite familiar.

Step 2: Analyze tweets from the Twin Cities (50%)

Using Twitter library, access the Streaming API and download all geo tagged tweets from the Twin Cities area as they are posted.

Step 3: Analyze tweets by keyword (40%)

Pick three keywords from the current events of the day. For instance, as of the week of Jan. 27, I might pick: "super bowl", "nsa", "Motorola" Use your Streaming API library to download all tweets that contain each keyword for at least 10 minutes each. Examine the geo Location field of these tweets. This is where tweet geotags are located. Also examine the location field in the profiles of the users who posted the tweets (You should not have to use an additional API for this).

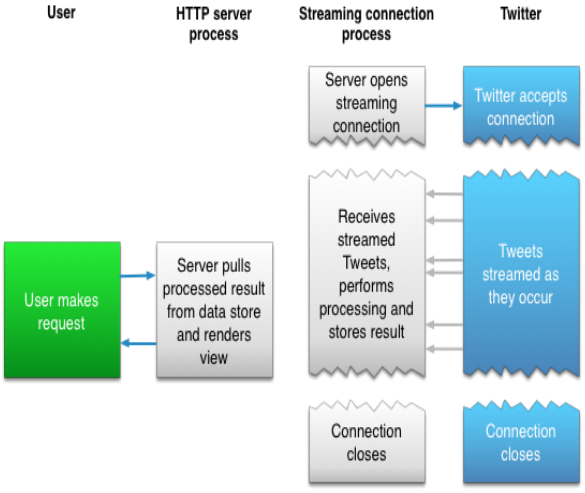

Fig-2: Twitter Streaming introduction

6. EXPERIMENTAL STUDY AND DISCUSSIONS

There are not any existing data sets of Twitter sentiment messages. We collected our own set of data. For the training data, we collected messages that contained the emoticons and Twitter API. The test data was manually. A set of 75 negative tweets and 108 positive tweets were manually marked. A web interface tool was built to aid in the manual classification task.

6.1 Feature Extractors

1. Unigram:

Building the unigram model took special care because the Twitter language model is very different from other domains from past research. The unigram feature extractor addressed the following issues:

a. Tweets contain very casual language. For example, you can search "hungry" with a random number of u's in the middle of the word on http://search.twitter.com to understand this. Here is an example sampling: huuuungry: 17 results in the last day huuuuuuungry:

4 results in the last day huuuuuuuuungry: 1 result in the last day besides showing that people are hungry, this emphasizes the casual nature of Twitter and the disregard for correct spelling.

b. Usage of links. Users very often include links in their tweets. An equivalence class was created for all URLs. That is, a URL like "http://tinyurl.com/cvvg9a" was converted to the symbol "URL."

c. Usernames. Users often include usernames in their tweets, in order to address messages to particular users. A de facto standard is to include the @ symbol before the username (e.g. @alecmgo). An equivalence class was made for all words that started with the @ symbol. The query term affect the classification.

2. Bigrams

d. Removing the query term. Query terms were stripped out from Tweets, to avoid having the reason we experimented with bigrams was we wanted to smooth out instances like 'notgood' or 'not bad'. When negation as an explicit feature didn't help, we thought of experimenting with bigrams. However, they happened to be too sparse in the data and the overall accuracy dropped in the case of both NB and MaxEnt. Even collapsing the individual words to equivalence classes did not help. Bigrams however happened to be a very sparse feature which can be seen in the outputs with a lot of probabilities reported as 0.5:0.5.For context: @stellargirl I looooooooovvvvvveee my Kindle2. Not that the DX is cool, but the 2 is fantastic in its own right. Positive [0.5000] Negative [0.5000]

3. Negate as a features

Using the Stanford Classifier and the base SVM classifiers we observed that identifying NEGclass seemed to be tougher than the POS class, merely by looking at the precision, recall and F1 measures for these classes. This is why we decided to add NEGATE as a specific feature which is added when "not" or 'n't" are observed in the dataset. However we only observed a increase in overall accuracy in the order of 2% in the Stanford Classifier and when used in conjunction with some of the other features, it brought the overall accuracy down and so we removed it. Overlapping features could get the NB accuracy down, so we were not very concerned about the drop with NB. However it didn't provide any drastic change with OpenNLP either.

4. Part of Speech (POS) features

We felt like POS tags would be a useful feature since how you made use of a particular word. For example, 'over' as a verb has a negative connotation whereas 'over' as the noun, would refer to the cricket over which by itself doesn't carry any negative or positive connotation. On the Stanford Classifier it did bring our accuracy up by almost 6%. The training required a few hours however and we observed that it only got the accuracy down in case of NB Handling the Neutral Class In the previous sections, neutral sentiment was disregarded. The training and test data only had text with positive and negative sentiments. In this section, we explore what happens when neutral sentiment is introduced.

5. Naive Bayes with Three Classes

We extended the Naive Bayes Classifier to handle 3 classes: positive, neutral, and negative. Collecting a large amount of neutral tweets is very challenging. For the training data, we simply considered any tweet without an emoticon to be part of the neutral class. This is obviously a very flawed assumption, but we wanted to see what the test results would be for the test data, we manually classified 33 tweets as neutral. The results were terrible. The classifier only obtained 40% accuracy. This is probably due to the noisy training data for the neutral class.

6. Subjective vs. Objective Classifier

Another way to handle the neutral class is to have a two phased approach:

1. Given a sentence, classify the sentence as objective or subjective.

2. If the sentence is subjective, classify it as positive or negative. We modified our Naive Bayes classifier to handle a subjective class and a objective class. Unfortunately, the results were terrible again, with an accuracy of only 44.9%. Again, this is probably due to the noisy training data of the neutral class.

7. CONCLUSION AND FUTURE IMPROVEMENTS

Here we emphasize the usage of machine learning techniques to perform for classifying sentiment in tweets. We also demonstrated how you can use twitter4j for streaming. It also supports programmatic access to the actions that any Twitter user can take, including posting messages, retweeting, following, and more.

As the future work, we plan to improve the accuracy of classification for such negative emotions as sadness and fear by gathering and analyzing more training data from at least 5 more participants. For the improvement of overall classification performance, we will investigate new features associated with user behavior. In this work, we have demonstrated a system to extract knowledge from tweets and then classify tweets based on the semantics of knowledge contained in them. For avoiding information loss, knowledge enhancer is applied that enhances the knowledge extraction process from the collected tweets. The maturity of knowledge gained using knowledge enhancer module has helped to filter tweet more precisely avoiding information loss. We have also measured missing information during specific keyword-based search and then proposed a method to collect more precise information about specific topic or domain. Sentiment analysis shows people attitude towards different topics. This data can also help to generate richer user profile and generate valuable recommendations. In future we are planning to integrate the proposed system with personalized profile management, sentiment analysis, and recommender system.

8. REFERENCES

[1] Sanket Sahu Mining Engineering (B. Tech)IIT-Kharagpur, Midnapore West Bengal, India, Suraj Kumar Rout Instrumentation & Engineering (B.Tech)College of Engineering & Technology, Debasmit Mohanty CEO StratLytics Consulting Private Limited," Twitter Sentiment AnalysisA more enhanced way of classification and scoring",2015.

[2] Onifade O.F.W Department of Computer ScienceUniversity of Ibadan Ibadan, Nigeria, Malik M.A.Department of Computer Science University of IbadanIbadan, Nigeria," SASM: A Tool for Sentiment Analysis on Twitter",2015.

[3] Akshi Kumar, Prakhar Dogra and Vikrant Dabas Dept. of Computer EngineeringDelhi Technological UniversityNew Delhi, India," Emotion Analysis of Twitter using Opinion Mining",2015.

[4] Lisa Madlberger Information & Software Engineering Group Vienna University of Technology Vienna, Austria," Predictions based on Twitter -A Critical View on the Research Process" 2014.

[5] Geetika Gautam Department of Computer Science &Engg .Jaypee Institute of Information technology Noida, India, Divakaryadav Department of Computer Science &Engg. Jaypee Institute of Information technology Noida, India," Sentiment Analysis of Twitter Data Using Machine Learning Approaches and Semantic Analysis", 2014

[6] Aliza Sarlan1, Chayanit Nadam2, Shuib Basri3Computer Information Science University Teknologi PETRONASPerak, Malaysia," Twitter Sentiment Analysis",2014.

[7] MalmazRoshanaei and Shivakant MishraDepartment of Computer ScienceUniversity of Colorado, BoulderBoulder, USA, "An Analysis of Positivity and Negativity Attributes of Users in Twitter", 2014.

[8] Hsiang HuiLek and Danny C.C. Poo Department of Information Systems School of Computing, National University of Singapore," Aspect-based Twitter Sentiment Classification",2013.

[9] Seyed-Ali Bahrainian, Andreas Dengel Computer Science Dept., University Of Kaiserslautern, Germany Knowledge Management Dept., DFKI, Kaiserslautern, Germany,"Sentiment Analysis and Summarization of TwitterData",2013

[10] J. Spencer, and G. Uchyigit, "Sentimentor: Sentiment Analysis of Twitter Data," Proc. of the 1st Int. Workshop on Sentiment Discovery from Affective Data, pp. 56–66, 2012.

[11] Balakrishnan Gokulakrishnan , Pavalanathan Priyanthan Thiruchittampalam Ragavan, Nadarajah Prasath, AShehan Perera Department of Computer Science and Engineering, University of Moratuwa, Moratuwa, Sri Lanka ," Opinion Mining and Sentiment Analysis on a Twitter Data Stream",2012

[12]]Alec Go, et.al, Twitter Sentiment Analysis, CS224N - Final Project Report, June 6, 2009.

[13] Twitter4J, Unofficial java library for the Twitter API, http://twitter4j.org/, 2007 Yusuke Yamamoto