Emotion Detection through Tweets

Shivang Arora, Sumant Mishra

SRM University
Ms. R. Anita, Dept Of Computer Science Engineering, SRM University, Tamil Nadu, India

Abstract—Twitter is a popular microblogging service where users create status messages or small text-based Web posts called tweets. Twitter currently receives about 190 million tweets a day, in which people share their comments regarding a wide range of topics. A large number of tweets include opinions about products and services. Analyzing these tweets to extract opinions or sentiments help us determine the student behaviour. This project is aimed at building a sentiment analyzer tool for analyzing tweets which can be used to accomplish the above goal of determining the student behaviour. This project mainly focuses on classifying the tweets as to belong to student one of positive, negative or neutral category using pre-classified tweets as training data. We have used the Naïve Bayes algorithm for implementing the sentiment analyzer tool. The Sentiment analyzer tool developed can give an approximate estimation of the success of student. The tool has been developed using java programming language for the back end and JSP and HTML for the front end and the tool has been presented as a web application. The application retrieves tweets real time based on the user’s query, analyzes and classifies them as to belonging to one of positive, negative or neutral category, summarizes the result and presents the result in a format such as a pie chart, graph which in turn helps in determining the popularity of a student. The algorithm’s efficiency is mainly dependent on the quality of the training data, for the training data chosen for this project we obtained an accuracy of close to 42% with precision and recall standing out at 45.65% and 67.74% respectively.

1. INTRODUCTION

In the last couple of years the social medium Twitter has become more and more popular, since Twitter is the most used microblogging website with about 500 million active users and 340 million tweets a day; it is an interesting source of information. The messages, or in Twitter terms the tweets, are a way to share interests publicly or among a denoted group. Twitter distinguishes itself from other social media by the limited message size.

Because Twitter is widely adopted through all strata, it can be seen as a good reaction of what is happening around the world. Among all that happens, the latest trends are most interesting for companies. The latest trends can be analyzed and when indented, reacted to. From a marketing point of view, these latest trends can be used to respond...
with appropriate activities, like product advertisements. Analyzing tweets can therefore be a goldmine for companies to create an advantage over competitors.

Emotion Detection involves several research fields: natural language processing, computational linguistics and text analysis. It refers to the extraction of subjective information from raw data, often in text form. However, also other media types could contain subjective data, like images, sounds and videos but these types are less studied. In accordance, in all media types different kinds of sentiments exist. The sentiment can refer to opinions or emotions, even though these two types are related there is an evident difference. In sentiment analysis based on opinions, a distinction is made between positive, negative and neutral opinions. The sentiment analysis that is considered in this project is based on opinions and is often referred in literature as opinion mining. Sentiment analysis aims to determine the attitude of the opinion holder with respect to a subject. Other applications try to determine the overall sentiment of a document. Sentiment analysis can be difficult. For example, a text can contain more than one opinion about the same object or about several objects.

Opinion = (oj, fjk, ooiijkl, hi, tl)

Where oj is particular object, fjk is feature k of object oj, hi is an opinion holder, tl is the time and ooiijkl the actual opinion. Determining the actual polarity of some sentence is the most difficult task of the five properties of quintuple . This sentiment is subjective because different people have different mental scale for what they consider to be a strong or a weak opinion. Therefore it can occur that a sentence is labeled as positive to somebody and neutral as by somebody else.

Twitter offers a special orthography that includes special features, hash-tags, user mentions and retweets. The "#"-hash-tags are integrated in Twitter, and are used to categorize tweets. In basic usage this categorization is based on subject and topics. But these hash-tags can also be used to add an opposite direction of the tweet, like sarcasm or irony. Such tags can reverse the polarity of the message. The most used hash-tags at the moment are summarized in the trending topics.

With a "@" symbol tweets can be directed to another user. Normally, the tweets are posted in public, or to a restricted group. The prefix "@" with a username directs the message to a specific user. The other user is aware of this directed message, and can respond to it. Thus, conversations can arise by mentioning other user in tweets.

In this project, analysis of data requires the user keyword. Based on the query it retrieves data from Twitter and then it is analyzed by using machine learning technique. Naïve Bayes algorithm is used in this case. The result obtained from the algorithm along with the graphs which help the user understand the sentiment of the data. Hence, the representation of data becomes very important. Having understood this requirement in the early phase of the project, the adopted methodology will accomplish the objectives in a neat and intuitive way.

II. BACKGROUND

A. Emotion Theories

In this section, we will briefly mention Ekman’s emotion model and the OCC (Ortony/Clore/Collins) model. Emotion models stipulate needed knowledge to appraise events. Ekman’s emotion model consists of “sadness, happiness, anger, fear, disgust and surprise”. It has been used in systems that recognise facial expressions related to these emotional states. The OCC model presents emotions generally expressed by an agent. It includes 22 emotion categories designed to model humans in general. It is based on the premise that emotions “are not themselves linguistic things, but the most readily available non phenomenal access we have to them is through language”.

B. Information Retrieval

Information retrieval is the acquisition of specific information about a topic and is usually based on a query. Text classification consists of categorising text at the document, sentence or token level (e.g. document classification, opinion recognition, word sense
disambiguation). Classification is an information retrieval task used for analysing the content of unstructured data expressed in natural language on the web (i.e. email, scientific documents and government reports). Text classification is currently the main method for emotion detection in text. However, due to the highly subjective nature of what actually constitutes an emotion, this method faces many challenges which have generated interest among researchers. The major challenge identified is the inclusion of subjectivity detection mechanisms.

C. Computational approaches for emotion detection

Keyword based approach
This approach has a total reliance on the presence of keywords and may involve pre-processing with a parser and emotion dictionary. It is easy to implement, intuitive and straightforward since it involves identifying words to search for in text.

Learning based approach
This approach uses a trained classifier to categorise input text into emotion classes by using keywords as features. It is easier and faster to adapt to domain changes since it can quickly learn new features from corpora by supplying a large training set to a machine learning algorithm for building a classification model.

Hybrid based approach
This approach consists of a combination of the keyword based implementation and learning based implementation. The main advantage of this approach is that it can yield higher accuracy results from training a combination of classifiers and adding knowledge-rich linguistic information from dictionaries and thesauri.

1. Hidden phrase pattern detection: Subtle sentences expressing emotions should be identified based on syntactic and semantic details. sequence of keywords

2. Emotion prediction: Identify the emotion being expressed based on knowledge rich linguistic resources and trained classifiers. To address the requirements above, we have chosen the following design decisions.

1. Incorporate a mechanism to identify keywords.

2. Examples of hidden patterns should be provided by a human annotator to assist in training the classifier. Keyword based component

1. Web based blog data is collected in a corpus.

2. The text is then split into tokens and sentences are identified.

3. The tokens are then annotated with Part of Speech (POS) tags.

4. Syntactic and semantic data is used for Postprocessing tasks before machine learning takes place.
B. Learning based component

1. Supply corpus to be used as the training set.

2. Perform pre-processing tasks as listed in A above.

3. Select a classifier (algorithm) for classification.

4. Generate feature vectors from the training set.

5. Repeat steps 1-4 with the testing set.

6. Convert feature vectors to LibSVM format.

7. Run LibSVM to measure prediction accuracy of classifier.

4. IMPLEMENTATION

4.1. Programming Language Selection

The programming language chosen must reflect the necessities of the project to be completely expressed in terms of the analysis and the design documents. Therefore before choosing the language, features to be included in the project are decided. The Sentiment Analyzer project needs the following features in a language to be implemented. Some of the features required are stated as follows:

J2EE provides us with servlets and JSP which help in dynamically constructing web pages.

JSP and servlets make use of Java backend in a very optimal manner. They have special tags which help us exploit these features.

J2EE provides us with Java Beans which help in proper data manipulation.

Java's core classes are designed from scratch to meet the requirements of an object oriented system.

4.2. Platform

The Sentiment Analyzer tool was built and designed on Windows Operating system family. They were specifically tested on Windows 7 with Google Chrome and Mozilla Firefox browsers. Because the product is browser based, any user with the browsers mentioned above will be able to run the tool. The product is hence platform independent in the true sense.

4.3. Difficulties Encountered and the Strategies Used To Tackle

There were a number of challenges that were faced while implementing the Sentiment Analyzer tool. Some challenges were tricky and ended up in helping us think innovatively and come up with efficient solutions. Some major problems that were encountered have been stated in brief along with their solutions.

Problem 1

Initially we were using double data type to hold the intermediate and final computed probability values, but the probability values were so small that in some cases they ended being zero.

Solution

We used log function on the intermediate and final computed probability values, hence the values obtained were not close to zero.

Problem 2

In the initial stages of the project charts4j libraries were used to plot graphs and pi charts. There were some internal problems with the URL rendering.

Solution

This Problem was solved later by making use of Google's Charts API and Java script.

5. EXPERIMENT AND EVALUATION

In this section, the performance, experimental analysis of the algorithm implemented in the project and the results obtained are presented.

5.1 Evaluation Metric

The algorithm implemented is evaluated for the following four metrics.
Using the above table we can calculate the above mentioned measures by using the formulas below.

Accuracy = \( \frac{tp + tn}{tp + fp + fn + tn} \)

<table>
<thead>
<tr>
<th>predicted class (expectation)</th>
<th>actual class (observation)</th>
</tr>
</thead>
<tbody>
<tr>
<td>tp (true positive) Correct result</td>
<td>tp (true positive) Correct result</td>
</tr>
<tr>
<td>fp (false positive) Unexpected result</td>
<td>fn (false negative) Missing result</td>
</tr>
<tr>
<td>tn (true negative) Correct absence of result</td>
<td></td>
</tr>
</tbody>
</table>

Precision = \( \frac{tp}{tp + fp} \)

Recall = \( \frac{tp}{tp + fn} \)

In the above formula P is precision, R is recall and \( \alpha \) is factor that controls the trade off between precision and recall (\( 0 =< \alpha <= 1 \)).

If we substitute for \( \alpha = 0.5 \) (\( \beta = 1 \)) we get the F1 measure

\[
F = \frac{1}{\alpha \frac{1}{P} + (1-\alpha) \frac{1}{R}} = \frac{(\beta^2 + 1)PR}{\beta^2 P + R}
\]

\[
F1 = \frac{2 \cdot P \cdot R}{(P + R)}
\]

5.2 Performance Analysis

To compute the performance of the Naïve Bayes algorithm implemented the test data is extracted from the training data.

30 tweets from each of the positive and negative training data and 10 tweets from neutral training data was taken as the test data, hence the test data consisted of 70 tweets.

The algorithm was implemented using test data as the input

Now precision, recall and f measure is calculated as follows

Case1: Compute precision, recall and F1 measure for positive test data by comparing the program prediction and the actual true result.

Case2: Compute precision, recall and F1 measure for negative test data by comparing the program prediction and the actual true result.

Case3: Compute precision, recall and F1 measure for neutral test data by comparing the program prediction and the actual true result.

6. CONCLUSION

6.1 Summary

In this project we were able to successfully implement the sentiment analyzer tool for analyzing and classifying twitter data. The tool developed is capable of classifying the tweets as to belong to one of positive, negative or neutral sentiments.

Naïve Bayes algorithm has been used for the purpose of classifying the tweets, the training data was obtained from Stanford machine learning repository. The tweets to be analyzed are obtained real time based on the keyword (subject) entered by the user.

The training data consisted of pre classified tweets which in turn was used by Naïve Bayes algorithm for the purpose of classifying the retrieved tweets. The tweets classified are presented in the form of a pie chart which enables the user determine the overall sentiment of a subject.

The performance of Naïve Bayes algorithm implemented was satisfactory. The efficiency of the algorithm is mainly dependent upon the quality of the training data. By using the training data obtained from Stanford machine learning repository the algorithm implemented was able to reach an accuracy of close to 50%.
6.2 Limitations

Some of the shortcomings in our project are:

1. The application is not capable of analyzing very large number of tweets.
2. The application does not analyze tweets written in other languages accurately.
3. The accuracy of the algorithm implemented is highly dependent on the training data.

6.3 Future enhancements

Some of the future enhancements are:

- The number of tweets analyzed can be increased. For larger dataset (in terms of MBs, GBs) or big data, distributed computing technologies like Hadoop can be used.
- The application can be made to analyze tweets other than the one written in English language.
- The application could be made more responsive by using Threads and Parallel/Cloud Computing.
- Efficient algorithms such as Support Vector Machines (SVMs) for analysis can be used.
- This application can be converted into a mobile application (android, iPhone, iPad) where the users can analyze the tweets on the go and share the results on facebook.

7. REFERENCES


