Knowing Your Twitter Sentiment

Ansari Nusrat Fatima
Kasali Heena Ikbal
Khan Iqra Bashir
ansarinusratfatima@gmail.com
heena.kasali.10@gmail.com
iqrakhan2515@gmail.com

Abstract:

The World Wide Web has intensely evolved a novel way for people to express their views and opinions about different topics, trends and issues. Online Micro blogging on social networks have been used for indicating opinions about certain entity in very short messages. Existing some popular micro blogs like twitter, face book etc ,in which twitter attains maximum amount of attention in the field of research areas related to product, movie reviews, stock exchange etc.Twitter is one of the widely used social media platform to express the thoughts. Sentiment analysis relates to the problem of mining the sentiments from online available data and categorizing the opinion expressed by an author towards a particular entity into at most three preset categories: positive, negative and neutral. By doing sentiment analysis in a specific domain, it is possible to identify the effect of domain information in sentiment classification.

Keywords-

Introduction:

Ongoing increase in wide-area network connectivity promise vastly augmented opportunities for collaboration and resource sharing. Now-a-days, various social networking sites like Twitter, Face book, MySpace, YouTube have gained so much popularity and we cannot ignore them. They allow people to build connection networks with other people in an easy and timely way and allow them to share various kinds of information and to use a set of services like picture sharing, blogs, wikis etc. With the rise of social networking epoch, there has been a surge of user generated content. Micro blogging sites have millions of people sharing their thoughts daily because of its characteristic short and simple manner of expression. Although Twitter may provision for an excellent channel for opinion creation and presentation, it poses newer and different challenges and the process is incomplete without adept tools for analyzing those opinions to expedite their consumption. Instigated by this the research carried out to use sentiment analysis to gauge the public mood and detect any rising antagonistic or negative feeling on social medias. The area of Sentiment Analysis intends to comprehend these opinions and distribute them into the categories like positive, negative, neutral. Sentiment analysis refers to the use of natural language processing, text analysis, computational linguistics, and biometrics to systematically identify, extract, quantify, and study affective states and subjective information. Sentiment analysis on Twitter posts is the next step in the field of sentiment analysis, as tweets give us a richer and more varied resource of opinions and sentiments that can be about anything from the latest phone they bought, movie they watched, political issues, religious views or the individuals state of mind. It is identifying the emotional tone in the series of words that make one tweet.

We use manually annotated Twitter data for our experiments. One advantage of this data, over previously used data-sets, is that the tweets are collected in a streaming fashion and therefore represent a true sample of actual tweets in terms of language use and content.

Literature Survey:

Sentiment analysis has been handled as a Natural Language Processing task at many levels of granularity. Starting from being a document level classification task (Turney, 2002; Pang and Lee, 2004), it has been handled at the sentence level (Hu and Liu, 2004; Kim and Hovy, 2004) and more recently at the phrase level (Wilson et al., 2005; Agarwal et al., 2009).

Microblog data like Twitter, on which users post real time reactions to and opinions about "everything", poses newer and different challenges. Some of the early and recent results on sentiment analysis of Twitter data are by Go et al. (2009), (Bermingham and Smeaton, 2010) and Pak and Paroubek (2010). Go et al. (2009) use distant learning to acquire sentiment data. They use tweets ending in positive emoticons like ":)" ":-)" as positive and negative emoticons like ":(" ":-(" as negative. They build models using Naive Bayes, Max Ent and Support Vector Machines (SVM), and they report SVM outperforms other classifiers. In terms of feature space, they try a Unigram, Bigram model in conjunction with parts-of-speech (POS) features. They note that the unigram model outperforms all other models. Specifically, bigrams and POS features do not help. Pak and Paroubek (2010) collect data following a similar distant learning paradigm. They perform a different classification task though: subjective versus objective. For subjective data they collect the tweets ending with emoticons in the same manner as Go et al. (2009). For objective data they crawl twitter accounts of popular newspapers like "New York Times", "Washington Posts" etc. They report that POS and bigrams both help (contrary to results presented by Go et al. (2009)). Both these approaches, however, are primarily based on n-gram models. Moreover, the data they use for training and testing is collected by search queries and is therefore biased. In contrast, we present features that achieve a significant gain over a unigram baseline. In addition we explore a different method of data representation and report significant improvement over the unigram models. Another contribution of this paper is that we report results on manually annotated data that does not suffer from any known biases. Our data is a random sample of streaming tweets unlike data collected by using specific queries. The size of our hand-labeled data allows us to perform cross validation experiments and check for the variance in performance of the classifier across folds.

Another significant effort for sentiment classification on Twitter data is by Barbosa and Feng (2010). They use polarity predictions from three websites as noisy labels to train a model and use 1000 manually labeled tweets for tuning and another 1000 manually labeled tweets for testing. They however do not mention how they collect their test data. They propose the use of syntax features of tweets like re tweet, hash tags, link, punctuation and exclamation marks in conjunction with features like prior polarity of words and POS of words. We extend their approach by using real valued prior polarity, and by combining prior polarity with POS. Our results show that the features that enhance the performance of our classifiers the most are features that combine prior polarity of words with their parts of speech. The tweet syntax features help but only marginally.

Gamon (2004) perform sentiment analysis on feedback data from Global Support Services survey. One aim of their paper is to analyze the role of linguistic features like POS tags. They perform extensive feature analysis and feature

selection and demonstrate that abstract linguistic analysis features contributes to the classifier accuracy. In this paper we perform extensive feature analysis and show that the use of only 100 abstract linguistic features performs as well as a hard unigram baseline.

Software Design Approach:

Software process model deals with the model which we are going to use for the development of the project. There are many software process models available but while choosing it we should choose it according to the project size that is whether it is industry scale project or big scale project or medium scale project.

Accordingly the model which we choose should be suitable for the project as the software process model changes, the cost of the project also changes because the steps in each software process model varies. This software is build using the waterfall model. This model suggests work cascading from step to step like a series of waterfalls. It consists of the following steps in the following manner:

1. Analysis Phase: We gather the Requirements of Twitter Sentiment Analysis like BRS that is "Business Requirement Specifications" and SRS that is "System Requirement specifications".
2. Design Phase: The objective of design is to determine how the problem will be solved. During design the analyst's focus shifts from the logical to the physical. Data elements are grouped to form physical data structures, screens, reports, files and databases.
3. Coding Phase: The system is created during this phase. Programs are coded, debugged, documented and tested. New hardware is selected and ordered. Procedures are written and tested. End-user documentation is prepared.
4. Databases and files are initialized. Users are trained.
5. Testing Phase: Once the system is developed, it is tested to ensure that it does what it was designed to do. After the system passes its final test and any remaining problems are corrected, the system is implemented and released to the user .All these phases are described with respect to the project in the rest of the document.

Cost Estimation:

For a given set of requirements it is desirable to know how much it will cost to develop the software to satisfy the given requirements, and how much time development will take. These estimates are needed before development is initiated. The primary reason for cost and schedule estimation is to enable the client or developer to perform a cost-benefit analysis and for project monitoring and control. Cost in a project is due to the requirements for software, hardware and human resources. Most cost estimates are determined in terms of Person month (PM). We have used COCOMO (Constructive Cost Model). The Intermediate COCOMO model computes software development effort as a function of program size and a set of cost drivers that include subjective assessments of product, hardware, personnel and project attributes. This model estimates the total effort in terms of person-months of the technical project staff.

Cost estimation is usually measured in terms of effort. Cost estimation can be defined as the approximate judgement of the costs for a project. Cost estimation will never be an exact science because there are too many variables involved in the calculation for a cost estimate, such as human, technical, environmental, and political.

- Obtain an initial estimate of the development effort from the estimate of thousand of delivered

lines of source code (KLoC)38

- The initial estimate (also called as nominal estimate) is determined by an equation of the form used in the static single-variable models, using KLoC as the measure of size.
- To determine the initial effort Ei in person- months the equation used is,Ei = a * (KLoC) b
- Where, a and b are constants which are determined depending on the type of the project. Since,this project is of Windows based type, therefore the values of a = 1.40 and the value of b =0.6 and KLoC is the number of lines of source code which is .874 KLoC. Thus the value of Ei is:

  Ei = 1.40 * (0.874) * 0.60 = 0.73416 PM
- Adjust the effort estimate by multiplying the initial estimate with the entire multiplying factor.We have taken the factors:
  1. Reliability
  2. Complexity
  3. Time Constraints
  4. Turnaround time
  5. Analyst capability
  6. Programmer capability
  7. Programming language experience
  8. Modern Programming practices
  9. Use of SW tools
  10. Development Schedule
- Based on these factors we have calculated, Effort Adjustment Factor (EAF) as follows:

EAF = 1.15 * 0.85 * 1.00 * 0.87 * 1.00 * 1.00 * 1.07 * 1.10 * 0.91 * 1.00= 0.91087

- The final effort estimate, E is determined by multiplying the initial estimate by the EAF:

E=EAF * Ei= 0.91087 * 0.73416= 0.6687 Person Month

- We take the assumption charges are 40 rupees per day.

Total estimation = 191 * 0.6687 * 40= 5100/- Rupees.

Implementation:

1. Fetching Twitter Data using Twitter API: Develop a twitter API [10] for downloading the tweets. The Twitter API directly communicates with the Source and Sink. The Authentication keys and tokens are established that helps in communication over Twitter Server. The source is twitter account and the sink is HDFS (Hadoop Distributed File System) where all the tweets are stored.
2. Pre-processing of tweets: A tweet contains a lot of opinions about the data which are expressed in different ways by different users .The twitter dataset used in this survey work is already labeled into two classes viz. negative and positive polarity and thus the sentiment analysis of the data becomes easy to observe the effect of various features. The raw data having polarity is highly susceptible to inconsistency and redundancy. Preprocessing of tweet include following points,
   - Remove all URLs (e.g. www.xyz.com), hash tags (e.g. #topic), targets (@username)
   - Correct the spellings: sequence of repeated characters is to be handled Replace all the emoticons with their sentiment.

- Remove all punctuations, symbols, numbers
- Remove Stop Words Expand Acronyms(we can use a acronym dictionary)
- Remove Non-English Tweets

3. Feature Extraction: In sentiment analysis the training data are classified into features (attributes) based on the content, and then weights are assigned to the features to distinguish their importance. The preprocessed dataset has many distinctive properties. In the feature extraction method, we extract the aspects from the processed dataset. Later this aspect are used to compute the positive and negative polarity in a sentence which is useful for determining the opinion of the individuals using models like unigram, bigram.
4. POS tagging: The part of speech (POS) tagging is a method of splitting the sentences into words and attaching a proper tag such as noun, verb, adjective and adverb to each word based on the POS tagging rules. POS tagging has been widely used in various tasks including text classification, speech recognition, automatic machine translation, and so on. We can generate syntactic dependency patterns by parsing or dependency trees.
5. Classifier: The Naive Bayesian Classification represents a supervised learning method as well as a statistical method for classification. It is probabilistic model and it permit us to capture uncertainty about the model in a principled way by determining probabilities. It helps to solve diagnostic and predictive problems. This Classification is named as Naïve Bayes after Thomas Bayes, who proposed the Bayes Theorem of determining probability. Bayesian classification provides useful learning algorithms and past knowledge and observed data can be combined. It helps to provide a useful perspective for understanding and also evaluating many learning algorithms. This helps to determine exact probabilities for hypothesis and also it is robust to noise in input data.

$$P(C \mid X) = \frac{P(C \mid X).\ P(C)}{P(X)}$$

Where,

- P(C | X) is posterior probability
- P(X | C) is likelihood
- P(C) is class prior probability
- P(X) is predictor prior probability

6. Result: As a result, program will be categorized sentiment into positive and negative, which is represented in a pie chart and html page. 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.

Applications:

1) Reviews from Websites: Today Internet has a large collection of reviews and feedbacks on almost everything. This includes product reviews, feedbacks on political issues, comments about services, etc. Thus there is a need for a sentiment analysis system that can extract sentiments about a particular product or services. It will help us to automate in provision of feedback or rating for the given product, item, etc.
2) Business Intelligence: It has been observed that people nowadays tend to look upon reviews of products which are available online before they buy them. And for many businesses, the online

opinion decides the success or failure of their product. Thus, Sentiment Analysis plays an important role in businesses.

3) Applications across Domains: Recent researches in sociology and other fields like medical, sports have also been benefitted by Sentiment Analysis that show trends in human emotions especially on social media.
4) Smart Homes: Smart homes are supposed to be the technology of the future. Based on the current sentiment or emotion of the user, the home could alter its ambiance to create a soothing and peaceful environment.

Advantages:

1)Develop product quality: Sentiment analysis helps you complete your market research by getting to know what your customers' opinions are about your products/services and how you can align your products/services' quality and features with their tastes.

2) **Improve customer service:** There are many factors that contribute to great customer service, such as on-time delivery, being responsive in social media, and adequate compensation for product's errors. Sentiment analysis can pick up negative discussions, and give you real-time alerts so that you can respond quickly.

3) Crises management: Constant monitoring of what is currently happening in social media conversations also helps you to prevent or at least mitigate the damage of online communication crisis. If you have sentiment analysis in place, you can detect potential manifestations.

4) Sales Revenue: The biggest benefit of doing sentiment analysis is to boost sales revenue. Increase in sales revenue is the final outcome of successful marketing campaigns, improved products/service quality, and customer service, which can be achieved with sentiment analysis.

5) **Adjust marketing strategy:** Most companies, if not all, are active in social media, and use the public forum to promote their brands and services. The information you get from sentiment analysis provides you with means to optimize your marketing strategy.

Limitations:

- While fetching the real time data from twitter it judge our patience, that is it requires a huge amount of time.
- The another one is the languages that is being supported is hindi and English. The country where we live in people often speak hindi combined with English, most popularly known as Hinglish, so it gets difficult even for the application to get it what the people are trying to convey.

Conclusion And Future Work:

Twitter Data in the form of opinion, feedback, reviews, remarks and complaint are treated as big data and it cannot be used directly. These data first convert as per requirement. In this paper, we discussed pre-processing of data to remove noise from the data. We have implemented sentiment analysis for movie data set, on Hadoop framework and analyzed with large number of tweets. This type analysis will definitely help any organization to improve their business productivity. The analysis of twitter data are done on various perspective like Positive, Negative and Neutral sentiments on tweets. It also provide the fast downloading approach for efficient Twitter

Trend Analysis. Tweets can also be useful in prediction of product sales, quality of services offered by company, feedback

of users etc. Hence, the future scope in the sentiment analysis for the other social networking websites like Facebook, Google

Plus etc.

References:

https://ieeexplore.ieee.org/abstract/document/7066632/118

https://dl.acm.org/citation.cfm?id=318505

https://www.researchgate.net/publication/301408174_Twitter_sentiment_analysis

https://keio.pure.elsevier.com/.../sentimet-analysis-in-twitter-from-classification-to-qu.

https://www.analyticsvidhya.com/blog/2018/07/hands-on-sentiment-analysis-dataset-python/

https://www.google.com/search?q=twitter+sentiment+analysis&rlz=1C1CHBF_enIN816IN816&oq=twitter+sentiment+analysis+&aqs=chrome..69i57j0j69i61l2j35i39l2.13190j0j7&sourceid=chrome&ie=UTF-8#

https://www.google.com/search?q=twitter+sentiment+analysis&rlz=1C1CHBF_enIN816IN816&oq=twitter+sentiment+analy..69i57j0j69i61l2j