Effective Sentiment Analysis of Social Media Datasets using Naive Bayesian Classification

Dhiraj Gurkhe
Fr. Conceicao Rodrigues College of Engineering Bandra (W), Mumbai-50

Niraj Pal
Fr. Conceicao Rodrigues College of Engineering Bandra (W), Mumbai-50

Rishit Bhatia
Fr. Conceicao Rodrigues College of Engineering Bandra (W), Mumbai-50

ABSTRACT

Effective Sentiment Analysis Of Social Media Datasets Using Naive Bayesian Classification involves extraction of subjective information from textual data. A normal human can easily understand the sentiment of a document written in natural language based on its knowledge of understanding the polarity of words (unigram, bigram and n-grams) and in some cases the general semantics used to describe the subject. The project aims to make the machine extract the polarity (positive, negative or neutral) of social media dataset with respect to the queried keyword. This project introduces an approach for automatically classifying the sentiment of social media data by using the following procedure: First the training data is fed to the Sentiment Analysis Engine for learning by using machine learning algorithm.After the learning is complete with qualified accuracy, the machine starts accepting individual social data with respect to keyword that it analyses and interprets, and then classifies it as positive, negative or neutral with respect to the query term.

General Terms:

Algorithm, Classification, Evaluation

Keywords:

Natural Language Processing, Machine Learning, Supervised Learning, Text Analysis

1. INTRODUCTION

With the explosion of internet there is an abundance of data available on-line, they can be numerical or text file and they can be structured, semi-structured or non-structured. Approaches and technique to apply and extract useful information from these data have been the major focuses of many researchers and practitioners lately. Advancement in computer technology along with many retrieval techniques and tools have been proposed according to different data types. In addition to data and text mining, there has seen a growing interest in non-topical text analysis in recent years. Sentiment analysis is one of them. Sentiment analysis,also known as opinion mining is to identify and extract subjective information in source materials which can be positive, neutral, or negative. Using appropriate mechanisms and techniques, this vast amount of data can be processed into information to support operational, managerial, and strategic decision making[8].

Sentiment analysis aims to identify and extract opinions and attitudes from a given piece of text towards a specific subject[11]. There has been much progress on sentiment analysis of conventional text which is usually found in open forums, blogs and the typical review channels. However,sentiment analysis of micro blogs like twitter is considered as a much harder problem due to the unique characteristics possessed by micro blogs (e.g. short length of status updates and language variations).

1.1 Social Media

In the past decade, social media has exploded with number of users reaching billions, a very good survey[7] shows Facebook has over 1 billion and Twitter has over 240 million active users on their respective sites. The survey[4] suggests that Facebook and Twitter make news a more participatory experience than before as people share news articles and comment on other people's posts. In 2010, according to CNN, 75% of people got their news forwarded through e-mail or social media posts, while 37% of people shared a news item via Facebook or Twitter.These honeycomb networks of social media users are slowly becoming the fastest way to spread news, reviews, opinions, comments, and other data throughout the world. These are ever increasing statistics and goes on to show why tapping into the data posted on these sites is ever so important and useful. The datasets of these sites are easily available through the respective API's like the Twitter API [1] which allows us to extract data based on query terms.

1.1.1 ***Challenges with social media data.*** The data available is not always ironclad, the general problems with social media data are

Grammar and Spellings With users being too causal when posting on the web they tend to make a lot of mistakes in the semantics of the language and even the spellings of words. These are generally checked in the pre processing stage of any application using these datasets.

Trustworthiness The most important property of social data is the views of different users on different subjects, but there are many fake accounts being made to give fake views and reviews to either push or pull an entity on the platform.

Format Every other social media site have its own style of posting data and also the way users post their data on these sites. Like

people using # to tag subjects or using @ to refer to different users. Hence, it is important to study and understand each site differently.

Language Social media sites provide options of using different languages to post views. There lies options to tackle this problem with either using translation mechanisms or building engines with respect to different languages.

2. RELATED WORK

Sentiment Analysis is in itself becoming a major area of study under Machine learning. The ideology used in this project is based on the underlying principles developed in [5] where the tweets were classified using unigram vectors and training was performed by distant supervision. The research in [5] elucidates that the use of emoticons as labels is effective in reducing dependencies in machine learning. The analysis in [5] is also on the basis of a query term and feature reduction using algorithms like Naive Bayes, Maximum Entropy and Support Vector Machines. The research and analysis conducted by Pang and Lee [3] was used to analyze the performance of different machine learning techniques in the movie review domain. It has also found implementations [11] as a sub component technology in augmentation with other systems like emails and online advertisements. With the help of improved Natural Language Processing capabilities and tools, this domain is gaining widespread importance and improved application in various other fields.

3. PROPOSED METHODOLOGY WORK

Figure 1 shows the overall architecture and process flow of various tasks for analyzing sentiments of social media dataset. Firstly, training data collected from various sources is subjected to preprocessing to eliminate features which do not contribute to polarity detection. This training data is fed into sentiment analysis engine for classifying test data. Secondly, the input query term is used to fetch data from social media for which polarity is to be detected. The sentiment analysis engine contains Naive Bayes classification algorithm which consults training data to calculate probabilities and predict the sentiment for given query term.

3.1 Pre-processing

Preprocessing eliminates the part which does not contribute significantly to the polarity detection .As Suggested by [5], there are many nooks and crooks of the social media datasets also known as tweets for Twitter. Tweets often contain usernames of account holder (@nirajp) which are replaced with the generic token USERNAME. Links(http://goo.gl/nirajp) are eliminated or replaced with the generic token URL. Additionally, [9] suggested further more preprocessing of tweets to reduce the feature which includes converting tweets to lower case characters to remove unevenness.'#' symbol used to denote hash tags are eliminated while keeping the succeeding hash tag word. Stop words such as *a, is, the* which do not contribute significantly to polarity detection are eliminated. Punctuation marks and additional white spaces are also eliminated. Two or more repetitive letters in a word are eliminated.e.g.: Happy is represented as *haaappy* or *haaaaaaappy* to stress emotion on social media platform is converted to 'happy'. Care is also taken that words must start with an alphabet. For the sake of simplicity, all those words which don't start with an alphabet are removed to reduce feature e.g. *21st, 7:30 pm.*

3.2 Feature Engineering

Feature Extraction is an extremely basic and essential task for Sentiment Analysis. Converting a piece of text to a feature vector is the basic step in any data driven approach to Sentiment Analysis

Unigram For text classification purpose, the unigram model was used which selects individual words from the data.*Apple is opening up the iPhone SDK. I'm stoked!*, for instance, contains following unigrams: Apple, opening, iPhone, SDK, stoked ,etc

Bigram In bigram model, a pair of words is extracted from data. The tweet, *Apple is opening up the iPhone SDK. I'm stoked!*, for instance, contains bigrams like: (Apple, opening), (opening, iPhone) etc

Unigram+Bigram In this model, a combination of unigram as well as bigram model is used to extract words from the data.

3.3 Model Building

3.3.1 ***Training***. For training purpose, the polarity labelled data from corpus is first parsed and relevant features are extracted from it to build the feature vector.This vector is used to create a Feature List which is a list of all the features of all the data items in dataset used for training, this list is stored in a text file on secondary memory for further use in both the training and classification

3.3.2 ***Naive Bayesian Classifier***. Naive Bayesian Text Classification algorithm is used for the purpose of classification of given trained model. It is the probabilistic approach to the text classification. Here the class labels are known and the goal is to create probabilistic models, which can be used to classify new texts. It is specifically formulated for text and makes use of text specific characteristics. The Naive Bayesian classifier treats each document as a "bag of words" and the generative model makes the following assumptions: firstly, words of a document are generated independently of context, and, secondly, the probability of the word is independent of its position. This is why the name naive was used for this algorithm. In real text documents the words often correlate with each other and the position of the word in text may play role.[10] Multinomial Naive Bayes model is shown in the equation 1.

$$P(c|d) := \frac{(P(c)\sum_{i=1}^{m} P(f|c)^{ni(d)})}{P(d)}. \quad (1)$$

In this formula, f represents a feature and ni(d) represents the count of feature fi found in tweet d. There are a total of m features. Parameters P(c) and P(f—c) are obtained through maximum likelihood estimates, and add -1 smoothing is utilized for unseen features.

3.3.3 ***Classification***. For classification purpose, the test data is preprocessed and feature vector of test data is formed. This test data is then fed into Naive Bayes algorithm along with the training data to calculate the probability using the Naive Bayes conditional probability formula to get polarity of the highest probability.

3.4 Data Extraction

Based on the input query term, the data is extracted from social media like Twitter using Twitter API [1].The retrieved data is sub-

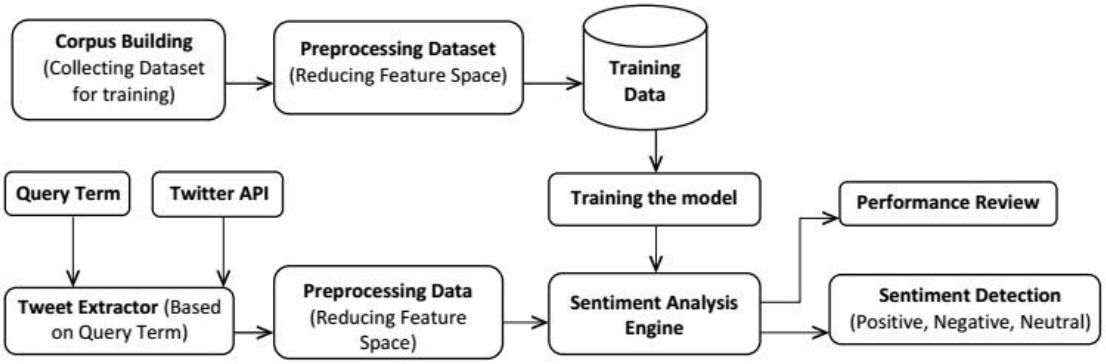

Fig. 1: Architecture and Process Work Flow

jected to preprocessing and is used as a test data for analysing sentiments using classification algorithm.

4. RESULTS

4.1 Experimental Setup

The implementation includes use of Python programming language along with Natural Language Toolkit (NLTK) libraries on Microsoft platform. Data used for training as well as for testing which includes social media data like tweets are represented using feature vector model. Results of experiments are used to calculate accuracy

4.1.1 ***Corpus Building*** . Machine learning classifiers like Naive Bayes use a large amount of training dataset for learning capabilities [2]. The training data set makes use of 113971 tweets which are classified as positive,negative and neutral. An amalgamation of training data from the following sources was done

Movie Review Datasets Each line in these two files corresponds to a single snippet (usually containing roughly one single sentence); all snippets are down-cased. The dataset contains 5331 positive snippets and contains 5331 negative snippets[3].

Sanders-Twitter Sentiment Corpus It consists of 5513 hand-classified tweets. These tweets were classified with respect to one of 4 different topics. It contains positive negative and neutral labeled data[12].

Twitter data based on emoticons Data collected by [5] based on emoticons

Sentiment Lexicon A lexicon which contains words classified as positive and negative[6]

For training, data collected from various sources was used to build corpus .These data sets were fed into training model and was appropriately preprocessed and trained to accurately classify the test data. Table 1 shows the numbers of training examples for individual labelled data.

Features	Total Training Data	Positive	Negative	Neutral
Unigrams	113971	56117	56117	1737
Bi-grams	13399	5831	5831	1737
Uni+Bigrams	14969	6616	6616	1737

Table 1. : Labelled training data statistics

4.1.2 ***Accuracy Testing***. Accuracy testing is done by measuring the true positive + true negative versus other possible results as shown in the equation 2.

$$Accuracy := \frac{TP+TN}{TP+TN+FP+FN} \quad (2)$$

Where TP is True Positive, TN is True Negative, FP is False Positive, and FN is False Negative. The cross validation technique was used to find these parameter. This technique involves partitioning the sample training data into complementary subsets, performing the analysis on one subset (called the training set), and validating the analysis on the other subset (called the validation set or testing set). In simple words, classifier was trained with bulk amount of traning data but some data was kept aside for calculating accuracy, this separated dataset is passed to the classifier for classification, the labels garnered are compared to the actual labels of the of the dataset to find accuracy testing parameters.

A total of 100 tweets were kept aside for accuracy testing distributed equally with respect to the three labels; this data is hand classified to make sure it is near gold standard. It is absolutely necessary to avoid negligible mistakes in the validation dataset as the errors in the generated parameters will be faulty and will lead to false analysis and evaluations.

As explained earlier, three types of feature extraction namely unigram, bigram, and unigram+bigram.The training for each file was done separately. Each training file is then subjected to check for accuracy using the cross validation technique.Evaluation of accuracy was kept in mind when considering only positive and negative data items in validation set. The results of the testing are given in following section.

Features	Training Data	With neutral	Without neutral
Unigram	113971	65	81.25
Bigram	13399	14	15
Uni+Bigram	14969	59	67.50

Table 2. : Accuracy statistics with different features

4.2 Comparision And Analysis

From the Table 2 it can be concluded that our sentiment analysis engine gives best results with Unigram detection without neutral labels. This result is a close match to the results found in [5] and [3] evaluations. The accuracy falls when testing with neutral labels this can be accounted to the fact that our training data had very less neutral data and also the inherent quality of neutral datasets being very difficult to classify.
A very low accuracy is observed in bi-gram features which can be simply put down to the fact that only bigrams are bound to have low accuracy as not all data items consist bi-grams indicating their sentiments.
The accuracy in [5] with unigram+bigrams was the best among all, even [3] showed a very high accuracy for the same. The accuracy is very decent when looked at the fact that we used only around 15,000 training data to train the classifier whereas both the other mentioned have used around 40,000 by [3] and around 1.6 million by [5]. This goes on to show that if the amount of training data is increased the accuracy is bound to increase.

5. CONCLUSION AND FUTURE WORK

The accuracy results produced by the engine using unigram feature extraction for negative and positive sentiment were highest. Other combination of grams has good potential especially the unigram+bigram combination. The project tries to label neutral data which has not been worked on significantly in the past, although the results are not satisfactory, but given higher amount of neutral data and a better quality of neutral datasets the results are bound to improve.
Machine Learning is an ever evolving field with newer and enhanced algorithms. These algorithms can be used to push the envelope even further with regards to speed, space and accuracy parameters. A few interesting works that can be done according to us are:

Adding other datasets We have only worked with twitter media dataset. This concept can be extended to other social media datasets like Facebook, LinkdedIn, Google+ etc.

Context Classification Language with its structure and words don't always easily juxtapose a statement as negative and positive. The word "kick" in the statement "I love kicking ball" denotes positive, but in the statement "I got kicked today" it denote negative. Similarly due to the structure, things like sarcasm can't be detected and lead to false classification.

Language Options We have worked only with the English language, but other world languages like Spanish, Hindi, Russian etc can be incorporated within the project.

6. REFERENCES

[1] Using the twitter search API, August 2013.

[2] Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow, and Rebecca Passonneau. Sentiment analysis of twitter data. In *Proceedings of the Workshop on Languages in Social Media*, pages 30–38. Association for Computational Linguistics, 2011.

[3] Lillian Lee Bo Pang and Shivakumar Vaithyanathan. Thumbs up?: sentiment classification using machine learning techniques. In *Proceeding EMNLP '02 Proceedings of the ACL-02 conference on Empirical methods in natural language processing - Volume 10 Pages 79-86*, pages 81–82, 2008.

[4] CNN Doug Gross. Survey: More americans get news from internet than newspapers or radio, 2010. [Online; accessed 9-July-2014].

[5] Alec Go, Richa Bhayani, and Lei Huang. Twitter sentiment classification using distant supervision. *CS224N Project Report,Stanford*, pages 1–12, 2009.

[6] Minqing Hu and Bing Liu. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177. ACM, 2004.

[7] Leveragenewagemedia. Social media comparison infographic, 2013. [Online; accessed 9-July-2014].

[8] Bing Liu. Sentiment analysis and subjectivity. *Handbook of natural language processing*, 2:568, 2010.

[9] Laurent Luce. Twitter sentiment analysis using python and nltk, 2014. [Online; accessed 9-July-2014].

[10] Christopher D Manning and Hinrich Schütze. *Foundations of statistical natural language processing*. MIT press, 1999.

[11] Bo Pang and Lillian Lee. Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, 2(1-2):1–135, 2008.

[12] Niek Sanders. Twitter sentiment corpus, 2014. [Online; accessed 9-July-2014].