International Journal on Data Science and Technology
2016; 2(4): 41-45
http://www.sciencepublishinggroup.com/j/ijdst
doi: 10.11648/j.ijdst.20160204.11
ISSN: 2472-2200 (Print); ISSN: 2472-2235 (Online)

Comparative Twitter Sentiment Analysis Based on Linear and Probabilistic Models

Kiplagat Wilfred Kiprono, Elisha Odira Abade

School of computing and informatics, University of Nairobi, Nairobi, Kenya

Email address:

Wilkiprono@gmail.com (K. W. Kiprono), eabade@uonbi.ac.ke (E. O. Abade), elisha.abade@gmail.com (E. O. Abade)

To cite this article:

Kiplagat Wilfred Kiprono, Elisha Odira Abade. Comparative Twitter Sentiment Analysis Based on Linear and Probabilistic Models. *International Journal on Data Science and Technology.* Vol. 2, No. 4, 2016, pp. 41-45. doi: 10.11648/j.ijdst.20160204.11

Received: June 12, 2016; **Accepted**: June 23, 2016; **Published**: August 1, 2016

Abstract: The transition from web 1.0 to web 2.0 has enabled direct interaction between users and its various resources and services such as social media networks. In this research paper we have analyzed algorithms for sentiment analysis which can be used to utilize this huge information. The goals of this paper is to device a way of obtaining social network opinions and extracting features from unstructured text and assign for each feature its associated sentiment in a clear and efficient way. In this project we have applied naïve bayes, support vector machines and maximum entropy for analysis and produced an analytical report of the three qualitatively and quantitatively. We performed the project empirically and analyzed the resulting data using an excel tool so as to obtain comparative analysis of the three algorithms for classification.

Keywords: Pos, Svm, Maxent, Naive Bayes, Feature Selection, Sentiment Classification, N-grams, Bigrams, Unigrams, Trigrams

1. Introduction

Direct interaction in the web and the environment has led to the availability of huge information in the internet. Social media networks such as tweeter, facebook, linkeldn and what sup has enabled people to share opinions realtime. Companies and business organizations in the world and Kenya have taken advantage of the platform to advertise, make sales and product reviews. Amazon, e-bay, Google shopping and OLX are examples and the number of reviews especially for popular products grow rapidly. Thus, they make use of people's opinions to make decisions not only for individuals but also for government and commercial sectors. Having such mass volume of data from different information sources make it difficult to take useful and satisfactory decision due to three factors. People cannot read the mass amount of data available, data on the web is unstructured, semistructured and heterogeneous in nature and information about the same product is often spread over a large number of sites and user accounts. Furthermore, differential feature formats and some products using different names make the resulting output of opinion mining and sentiment analysis concerning that domain of the online products. The levels of classifying sentiments include document level, sentence level/phrase level and aspect /feature level. We use it according to the level interest. In our research project we have used feature level since we are collecting opinions about several aspects of the same product and within the same document. We are going to subject the data to the three algorithms naive bayes, support vector machines and maximum entropy.

1.1. Tweeter

This is a real time information network that connects individuals to the latest stories, ideas, opinions and news about what you find interesting. To follow conversations and most compelling information, you will simply search their accounts. Bursts of information called tweets will be seen in the tweeter accounts. A tweet has 140 characters long but it gives a lot of information to be discovered. You will find photos, videos, and conversations directly in the tweets to get the whole story at once. In this project we used raw tweeter data collected from several accounts using the tweeter API and preprocessed for the purpose of experimenting.

Tab. 1. Comparison of web 1.0 and web 2.0.

WEB 1.0	WEB 2.0
Application Based	Web Based
Isolated	Collaborative
Offline	Online
Licensed/Purchased	Free
Single Creator	Multiple Collaborators
Proprietary Code	Open Source
Copyrighted Content	Shared Content

1.2. Sentiment Analysis

Sentiment analysis or opinion mining is the computational study of people's opinions, attitudes and emotions towards and entity. This could be individuals, events or topics. The topics are most covered by reviews. A Company like Safaricom Kenya limited in Kenya which is basically doing business in voice and data can launch a tariff for calls and expect people to comment from the product. This review will make them want to improve and add value on their products. Opinion mining extracts and analyses people's opinions about a product while sentiment analysis identifies the sentiment expressed in a text then analyze it. The aim is to find opinions, identify the sentiments they express and then classify their polarity so as to be used for decision making. It is composed of machine learning and lexicon based approach. Our project will dwell on machine learning approaches bust specifically the supervised approaches.

1.3. Sentiment Analysis Techniques

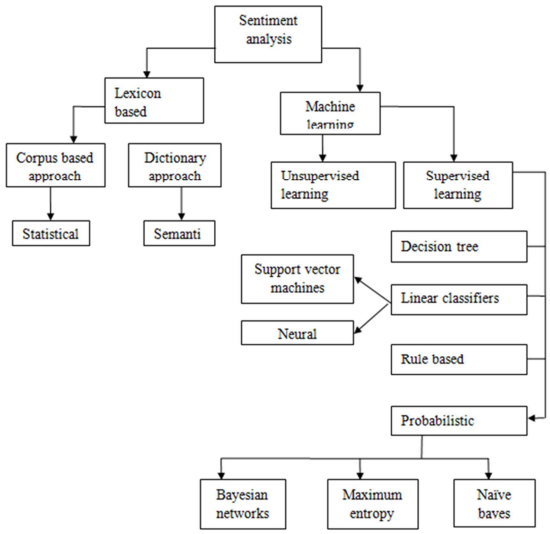

Fig. 1. Machine learning techniques.

1.4. Machine Learning Approach

Machine learning use algorithms to solve the sentiment analysis as a regular text classification problem that makes use of syntactic or linguistic features. The classification model is related to the features in the underlying record to one of the labels. The model is used to predict a class label for every instance of unknown class. It is hard to classify when only one is assigned to an instance.

2. Supervised Learning

The supervised learning methods depend on the existence of labeled training documents. There are many kinds of supervised classifiers in literature. The brief details some of the most frequently used classifiers in sentiment analysis.

2.1. Probabilistic Classifiers

Probabilistic classifiers use mixture models for classification. The mixture model assumes that each class is a component of the mixture. Each mixture component is a generative model that provides the probability of sampling a particular term for that component. These kinds of classifiers are also called generative classifiers.

2.2. Support Vector Machines

Support vector machines is a linear classifier which is effective and can achieve good performance. In high d dimensional feature set space. In our project it showed that the classifier proved the most reliable in terms of accuracy, precision and accuracy of the sentiment process. We trained with LIBSVM (Chang and Lin, 2011) a widely used tool in many researches.

2.3. Maximum Entropy

The idea behind MaxEnt classifiers is that we should prefer the most uniform models that satisfy any given constraint. MaxEnt models are feature based models. MaxEnt makes no independence assumptions for its features, unlike Naïve Bayes. This means we can add features like bigrams and phrases to MaxEnt without worrying about feature overlapping. The principle of maximum entropy is useful explicitly only when applied to testable information. A piece of information is testable if it can be determined whether a given distribution is consistent with it. The major advantages of using Max Ent or its variations are:

- Accuracy
- Consistency – This algorithm shows consistency in results and if priors are used results also improve over a period of time.
- Performance / Efficiency - Can handle huge amounts of data
- Flexibility - The algorithm is flexible of having many different typed of data in a unified platform and classify it accordingly.

3. Naive Bayes

Naïve Bayes is used as a classifier in various real world problems like Sentiment analysis, email Spam Detection, email Auto Grouping, email sorting by priority, Document Categorization and Sexually explicit content detection. Naive Bayes classification model computes the posterior probability of a class, based on the distribution of the words in the document. The model works with the BOWs feature extraction which ignores the position of the word in the

document. It uses Bayes Theorem to predict the probability that a given feature set belongs to a particular label. The major advantage of Naïve Bayes is it requires low processing memory and less time for execution. It's advised that this classifier should be used when Training time is a crucial factor in the system. Naïve Bayes is the baseline algorithm for researches in decision level classification problem. In presence of limited resources in terms of CPU and Memory Naïve Bayes is recommended classifier.

4. Methodology

We conducted our research empirically and the data results were conducted quantitatively and qualitatively. The analysis of the resulting data was done using an excel application since the data set was not large. The whole process can be summarized as follows:

4.1. Architectural Design

- Extracted tweets from social media using an extraction script. Twitter API was used to collect tweets and then stored in a MSQL database.
- Preprocessing and cleaning of the data.
- The data is then divided into 75% for training and 25% for test data set.
- Training the data so as to come up with a model that can be used to classify new and pure tweets.
- Using the model generated to classify posts which feature from the tweets collected and classifies them into the three polarities i.e. negative, positive and neutral.
- Results analysis is achieved from the classifiers developed and the conclusions drawn.

4.2. Preprocessing

Preprocessing the data is done by cleaning and preparing the text for classification. Online texts contain usually lots of noise and uninformative parts such as HTML tags, scripts and advertisements. In addition, on words level, many words in the text do not have an impact on the general orientation of it. Keeping those words makes the dimensionality of the problem high and hence the classification more difficult since each word in the text is treated as one dimension (Abassi etal 2011). To reduce the noise in the text should help improve the performance of the classifier and speed up the classification process, thus aiding in real time sentiment analysis.

The whole process involves several steps: online text cleaning, white space removal, expanding abbreviation, stemming, stop words removal, negation handling and finally feature selection. Features in the context of opinion mining are the words, terms or phrases that strongly express the opinion as positive or negative. This means that they have a higher impact on the orientation of the text than other words in the same text.

4.3. Filtering

Repeated words like good to show their intensity of expression are eliminated as they are not present in the sentiwordnet hence extra letters in the word must be eliminated. This elimination follows the rule that a letter can't repeat more than three times.

4.4. Questions

Questions such like which, how, what etc. are not going to contribute to polarity hence in order to reduce the complexity, such words are removed.

4.5. Removing Special Characters

Special characters like () {} [] etc. should be removed in order to eliminate discrepancies during assignment of polarity. For example "it's good" means if the characters are not removed may concatenate with the words and make those words unavailable in the dictionary.

4.6. Removing Retweets

Many people may copy another person's tweets and retweet using a different account. This happens if he likes another user's tweet.

4.7. Removing Urls

Generally Urls does not contribute to analysis of the sentiment in informal text e.g. "I have logged into www.ecstacy.com as I am bored". This is negative but may be neutral because of the presence of the word ecstasy.

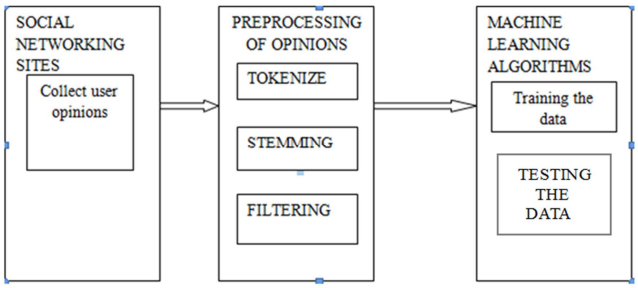

Fig. 2. *Preprocessing in sentiment analysis.*

4.8. Feature Selection in Sentiment Classification

Sentiment Analysis task is considered a sentiment classification problem. The first step in the SC problem is to extract and select text features. Some of the current features are:

Terms presence and frequency: These features are individual words or word n-grams and their frequency counts. It either gives the words binary weighting (zero if the word appears or one if otherwise) or uses term frequency weights to indicate the relative importance of features.

Parts of speech (POS): finding adjectives, as they are important indicators of opinions.

Opinion words and phrases: these are words commonly used to express opinions including good or bad, like or hate. On the other hand, some phrases express opinions without

using opinion words. For example: cost me an arm and a leg.

Negations: the appearance of negative words may change the opinion orientation like not good is equivalent to bad.

5. Comparative Analysis of the Algorithms

Tab. 2. *Results and findings from the tests.*

Training Data	Support Vector Machines			Maximum Entropy			Naïve Bayes		
	Bigram	Unigram	Trigram	Bigram	Unigram	Trigram	Bigram	Unigram	Trigram
5000	76	80	61	74	72	55	61	74	45
4000	73	78	61	72	72	54	58	71	43
3000	69	74	74	71	71	54	56	70	42
2000	67	78	56	70	71	71	55	68	43
1000	66	75	56	69	69	44	54	71	44
Mean%	70.2	77	61.6	71.2	71	55.6	56.8	70.8	43.4

Tab. 3. *Comparative analysis of the algorithms.*

FEATURE	SUPPORT VECTOR MACHINES	MAXIMUM ENTROPY	NAÏVE BAYES
Accuracy	High	Good	Good
Memory Requirement	High	High	Low
Simplicity	Hard	Hard	Very Simple
Performance	Best	Better	Good
Training Time	High	Moderate	Less
Consistency Of Accuracy	Consistent	Variable	Variable

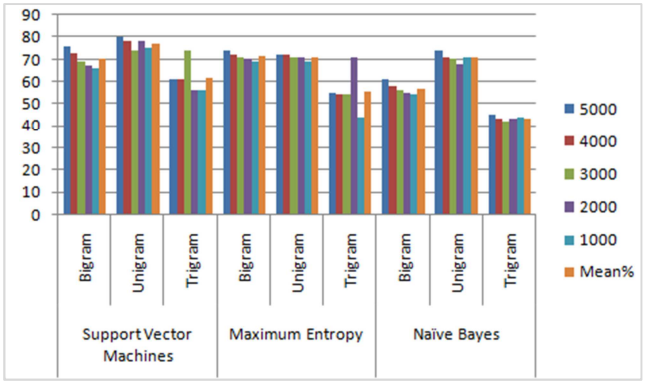

Fig. 3. *Graphical representations of the results.*

From our study it was evident that every kind of classification model had its own challenges. The selection of classification models can be decided on the basis of resources, accuracy requirement and training time available. Considering the support vector machines which showed that it was hard to implement, high memory requirements, consistent in data output and consumes more time in training, and the classifier was best fit for use in sentiment analysis. However it requires high training time and processing power this hence improved the accuracy of the classifier. If processing power is an issue and memory is an issue then the naïve bayes classifier is selected due to its low processing power and memory consumption less training is required time is required. If you powerful processing system and memory then maximum entropy proves to be a worthy alternative. Support vector machines proved to be average in all aspects and thus proved to be the best choice for sentiment analysis.

6. Conclusions and Future Work

In this project we presented a way in which machine learning techniques can be applied to large data sets to establish their performance when subjected to different features and classifiers in this case unigrams, bigrams and unigrams. We demonstrated how to collect original twitter posts for sentiment classification and the process of cleaning the data. We applied maximum entropy, support vector machines and naive bayes and found the process successful. The results analysis found that unigrams did best in all the classifiers followed by bigrams and trigrams. The best classifier was linear based, in this case support vector machines and probabilistic models did fairly well (naive bayes and maximum entropy). This results supported previous experiments done by pak etal, turney etal and (liu, 2013). The classification process gave an accuracy of 77% for support vector machines, 71% maximum entropy and 70.8% for naive bayes, however we feel this could be further improved. This was a near human accuracy which is argued by (ogneva, 2010) that humans may only agree on a polarity of a text 80% of the time meaning models with accuracy greater than 80% may be giving inconsistent results.

References

[1] Li Yung-Ming, Li Tsung-Ying. Deriving market intelligence from microblogs. Decis Support Syst 2013.

[2] Caro Luigi Di, Grella Matteo. Sentiment analysis via dependency parsing. Comput Stand Interfaces 2012.

[3] Liu B. Sentiment analysis and opinion mining. Synth Lect Human Lang Technol 2012.

[4] Pang B, Lee L. Opinion mining and sentiment analysis. Found Trends Inform Retriev 2008; 2: 1–135.

[5] Mohammad SM. From once upon a time to happily ever after: tracking emotions in mail and books. Decis Support Syst 2012 s; 53: 730–41.

[6] Fully Automatic Lexicon Expansion for Domain- oriented Sentiment Analysis by Hiroshi Kanayama Tetsuya Nasukawa, Tokyo Research Laboratory, IBM Japan, Ltd. 1623-14 Shimotsuruma, Yamato-shi, Kanagawa-ken, 242- 8502 Japan {hkananasukawa}@jp.ibm.com

[7] Text normalization in social media: progress, problems and applications for a pre-processing system of casual English - Eleanor Clarka* and Kenji Arakia Pre-processing very noisy text - Alexander Clark, ISSCO / TIM, University of Geneva, UNI-MAIL, Boulevard du Pont-d'Arve, CH-1211 Geneva 4, Switzerland.

[8] M. Almashraee, D. M. Diaz, and R. Unland, "Sentiment classification of online products based on machine learning techniques and multi-agent systems technologies," in Industrial Conference on Data Mining - Workshops, 2012.

[9] Mugenda, M., Mugenda, G. (1999). Research Methods. Quantitative and Qualitative Approaches. Nairobi, Kenya.

[10] Pauls, Adam, and Dan Klein. "Faster and smaller n-gram language models." *Proceedings of the 49th annual meeting of the Association for Computational Linguistics: Human Language Technologies*. Vol. 1. 2011.

[11] Socher, Richard, et al. "Semi-supervised recursive auto encoders for predicting sentiment distributions." *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2011.

[12] Kennedy, Alistair, and Diana Inkpen. "Sentiment classification of movie reviews using contextual valence shifters." *Computational Intelligence* 22.2 (2006): 110-125.