International Journal of Applied Engineering Research
ISSN 0973-4562 Volume 10, Number 5 (2015) pp. 13447-13457

http://www.ripublication.com

Big Data Analytics: A Supervised Approach for Sentiment Classification Using Mahout: An Illustration

Jaswant U and PN Kumar

Dept. of CSE, Amrita Vishwa Vidyapeetham, Coimbatore
jaswanth.uthukota@gmail.com
Dept. of CSE, Amrita Vishwa Vidyapeetham, Coimbatore
pn kumar@cb.amrita.edu

Abstract

With the growing popularity of social media channels, huge volumes of data are being produced. Analyzing such huge volumes of data so as to make meaningful and useful sense of it is the drive behind big data analytics. One approach to such analytics is sentiment analysis. Sentiment analysis of social networking content provides reliable feedback on public views about any theme of discussion. It has marked its relevance in various domains such as consumer analytics, political analytics, trend mining, brand reputation, etc. This paper aims at accessing social media content posts that may or may not be regarding specific products, and extracting useful tokens, features, or patterns, that would quantify the weight of the opinion for further classification of new incoming data. The characteristic of the posts found in social networking sites, especially Twitter, bear unstructured grammar and mis-spelt words. We propose to parse the data so available for context-specific opinion mining. This paper proposes a method for the same in a supervised learning approach of sentiment analysis on large volumes of Twitter data. We have shown that the Naïve Bayes method achieved higher accuracy, compared to the methods of Support Vector Machines (SVM), Maximum Entropy method (MaxENT) and K-Nearest Neighbor(KNN) methods.

Keywords: Sentiment analysis, social media, supervised learning, Opinion mining, Naïve Bayes Method, Mahout.

1. INTRODUCTION

Big Data is a massive volume of both structured and unstructured data which continues to grow so much that it becomes difficult to manage it using traditional

database management concepts and tools. The difficulty can be related to data storage, searching, transfer, analysis and visualization. Big Data spans across five dimensions: *Volume, Velocity, Variety, Variability and Veracity*. By volume, we refer to the large size of data- in terabytes and petabytes; velocity implies that for timely usage, it should be generated and processed fast; variety refers to the structured and unstructured data of all categories: text, video, social media, log files etc; variability refers to the inconsistency which can be shown by the data at times, thus hampering the process of being able to handle and manage the data effectively, and finally, veracity implies that the quality of the data being captured can vary greatly, thus accuracy of analysis depends on the veracity of the source data. Another characteristic is its *Complexity*, wherein the data management can become a very complex process, especially when large volumes of data come from multiple sources. These data need to be linked, connected and correlated in order to be able to grasp the information that is supposed to be conveyed by these data. This situation, is therefore, termed as the 'complexity' of Big Data.

The domain focused here is the data generated from a Social Networking Site (SNS), particularly Twitter. The internet has given rise to huge volumes of data. A very significant source is the social media and networking and the Internet of Things (IoT).

Gaining strategic advantage over competitors requires the innovative application of information technologies. Social Customer Relationship Management (CRM) is the use of social media services, techniques and technology to enable organizations to engage with their customers, to optimize the power of social interactions so as to get closer to the customers. Companies of every kind are beginning to focus more on consumers and engaging with them on Facebook, Twitter and other social media platforms. Analytics therefore is the killer application for social media. One such method is the use of sentiment analysis, which allows companies to find out what customers think about them in the market place.

According to the Harvard study, most companies do not know what their customers are saying about them and where they are saying [1].Customer Relationship Management (CRM) has become easier when social media analytics is integrated into the system. The increase in on-line collaboration has stressed a need to arrive at a central portal that bridges the social networking feeds with the product inventory statistics. Hence social media analytics focuses on retrieving the useful data from social media and automatically transforming this into something meaningful and quantifiable.

2. METHODOLOGY: SENTIMENT ANALYSIS

Sentiment analysis aims to detect the attitude of a text. A simple subtask of sentiment analysis is to determine the polarity of the text: positive, negative or neutral. This can be seen as a classification task and Naïve Bayes algorithm is suitable for this classification. The Naïve Bayes algorithm uses probabilities to decide which class best matches for a given input text. The classification decision is based on a model obtained after the training process. Model training is done by analysing the

relationship between the words in the training text and their classification categories. The algorithm is considered naive because it assumes that the value of a particular feature is independent of the value of any other feature, given the class variable. This can be seen as a classification task and Naïve Bayes is a suitable algorithm for this task.Each text that needs to be classified has words W_j (j=1..n) .

- Probabilities extracted from the training data set (for words W_j):

$$P(W_j \text{ given } +ve) = \frac{\text{The number of +ve Texts with } W_j}{\text{The number of +ve Texts}} \quad (1)$$

$$P(W_j \text{ given } -ve) = \frac{\text{The number of -ve Texts with } W_j}{\text{The number of -ve Texts}}$$

- Probabilities are extracted from the test set:

$$P(+ve)` = \frac{\text{Number of +ve Texts}}{\text{The total number of Texts}} \quad (2)$$

$$P(-ve) = \frac{\text{Number of -ve Texts}}{\text{The total number of Texts}}$$

- Applying Naïve Bayes Theorem (to classify a Text being +ve or -ve):

$$P(+ve \text{ given Text}) = \frac{P(\text{Text given } +ve) \times P(+ve)}{P(\text{Text})} \quad (3)$$

$$P(-ve \text{ given Text}) = \frac{P(\text{Text given } -ve) \times P(-ve)}{P(\text{Text})}$$

- $P(+ve \text{ given Text}) = P(\text{Text given } +ve) \times P(+ve) =$
  $[P(W_1 \text{ given } +ve) \times P(W_2 \text{ given } +ve) \times \ldots$
  $\times P(W_n \text{ given } +ve)] \times P(+ve)$ (4)

- $P(-ve \text{ given Text}) = P(\text{Text given } -ve) \times P(-ve)$
  $= [P(W_1 \text{ given } -ve) \times P(W_2 \text{ given } -ve) \times \ldots$
  $\times P(W_n \text{ given } -ve)] \times P(-ve)$ (5)

P(+ve given Text) and P(-ve given Text) can then be compared and the text sentiment is declared positive or negative based on the higher probability values. In this paper, we employ two phases for sentiment analysis:

- Preprocessing using Natural Language Tool Kit (NLTK).
- Data Mining using Mahout.

Appendix gives the basic definitions of the various terms used in this work.

3. BACK GROUND AND RELATED WORK

Sentiment analysis can be handled in two ways: Supervised and Unsupervised learning.Supervised learning takes an annotated training dataset that is subjected to a classifier for learning [2]. Upon learning, the classifier saves its model based on the algorithm being used. Subsequently, validation can be done on this classifier model to ascertain its accuracy when a new entry, not available in the training dataset comes in. This saved model is used for testing future input data which may or may not be annotated. An important task in supervised learning algorithms is feature extraction. In sentiment analysis, commonly used feature extraction techniques include: term presence and their frequency, part-of-speech information, negations, opinion words and phrases. Examples of supervised learning algorithms include Naïve Bayes, Maximum Entropy method (MaxENT), Support Vector Machines (SVM), Boosting Algorithm, etc. The effectiveness of the classification is dependent on the amount and quality of the training data used. In general, supervised learning is seen to have better performance and accuracy than unsupervised learning (especially in opinion mining). However, it may fail when training data is insufficient and acquisition of large amounts of training data is expensive.

Unsupervised learning takes unlabelled data and classifies by comparing the features of the given text against word lexicon whose sentiment values are determined prior to their use. Popular unsupervised algorithms are Clustering (K-means, Density-based, Hierarchical etc), Self- Organizing Maps (SOM), and Adaptive Resonance Theory (ART). Feature extraction by unsupervised learning is done by dimensionality reduction techniques such as Principal Component Analysis (PCA), Independent Component Analysis, and Singular-Value Decomposition (SVD) etc. However, despite better domain independency, unsupervised algorithms have poor performance in accuracy [3].

3.1 Commonly used NLP Tools

A variety of open-source text-analytics tools like natural- language processing for information extraction and classification can be applied for sentiment analysis. The tools listed below can work on textual sources only.

- *LingPipe:* LingPipe is a suite of java tools for linguistic processing of text including entity extraction, Parts-Of-Speech tagging (POS), clustering, classification, etc. It is one of the most mature and widely used open source NLP tool kits in industry. It is known for its speed, stability and scalability. It is not technically 'open-source'.

- *OpenNLP:* OpenNLP hosts a variety of java-based NLP tools which perform the most common NLP tasks, such as tokenization, sentence segmentation, part-of-speech tagging, named entity extraction, parsing and co-reference resolution. These tasks are usually required to build more advanced text processing services. OpenNLP also includes maximum entropy based machine learning.
- *Stanford Parser and Parts-Of-Speech (POS) Tagger*: Java packages for sentence parsing and POS tagging from the Stanford NLP group. It has implementations of probabilistic natural language parsers.
- *Natural Language Tool Kit (NLTK):* NLTK is a comprehensive python library for NLP and text analytics. Designed for teaching and researching, classification, clustering, tokenizing, removing stop words, speech tagging and parsing and more. It contains a set of tutorials and data sets for experimentation. It is written by Steven Bird, from the University of Melbourne.
- *Opinion Finder:* Opinion Finder, which was initially released in 2006, employs a multi-stage NLP process. It aims to identify subjective sentences and to mark various aspects of subjectivity in these sentences, including the source (holder) of the subjectivity and words that are included in phrases expressing positive or negative sentiments.
- *NLP Tool Suite:* A comprehensive NLP tool suite is used for the application purposes of semantic search, information extraction and text mining. Most of their continuously expanding tool suite is based on machine learning methods and thus is domain and language independent.
- *TweetNLP Tagger:* TweetNLP Tagger is POS tagger that is exclusively used for mis-spelt words and SMS slang tenses where the sentences have no structured grammar or spelling rules.

In the proposed system, NLTK has been used for data pre- processing because it has most of the functionalities needed to perform NLP tasks like tokenization, POS tagging, stop words removal and other filtering tasks. It comes with a stop words text corpora that contains word list for many languages. Moreover we can extend the stop word list. The main advantage lies in handling large dataset.

3.2. Mahout

Mahout is an open source machine learning library from Apache for big data analysis. The algorithms it implements fall under the broad umbrella of machine learning or collective intelligence. It's also scalable. It aims to be the machine learning tool of choice when the collection of data to be processed is very large, perhaps far too large for a single machine. In its current incarnation, these scalable machine learning implementations in Mahout are written in Java, and some portions are built upon Apache's Hadoop distributed computation project.

3.3. Discussion

An overview of the various studies in the past is categorized based on supervised and unsupervised learning.

3.3.1 Supervised Learning

The work "Twitter Sentiment Analysis: The Good the Bad and the OMG!"[4], investigated the utility of linguistic features for detecting the sentiment of Twitter messages. They evaluated the usefulness of existing lexical resources as well as features that capture information about the informal and creative language used in microblogging. A supervised approach has been used to tackle the problem, leveraging existing hashtags in the Twitter data for building training data. They used the hash-tagged data set (HASH) from Edinburgh Twitter corpus and the emoticon data set (EMOT) for training purpose. For evaluation, a manually annotated data set from iSieve Corporation is used. Their results showed 74% accuracy using only HASH, and 75% accuracy when a combination of HASH and EMOT were used as training data.

However, in the work done by Thomas Lake[5], dataset from three different sources was used to perform semantic sentiment analysis in the following steps: Extracting entities and Concepts (using five open API)/Feature incorporation to Naïve Bayes/feature reduction/semantic interpolation. Unigrams were trained in Naïve Bayes, POS tagging was provided by TweetNLP, and topics of similar sentiments were grouped together using weakly supervised Joint Sentiment Topic classifier.

SentiMentor[6], utilizes the Naïve Bayes classifier to classify the live Tweets into positive, negative or objective sets. Data collected by the Twitter API was pre-processed by tokenization, stop-word removal, POS tagging (using OpenNLP), extracting unigrams and bigrams. The classifier was trained with 219 annotated tweets and testing was done on the grams with and without POS tagging. Accuracy was 52.31%.

"Twitter Sentiment Classification using Distant Supervision"[7], made a benchmark among the works done in supervised classification of sentiments. It is a two step approach: feature extraction and classification. Feature extraction step classifies query, simplifies lengthy words, and removes most of the emoticons in the tweets. The feature set is subjected to Naïve Bayes, MaxENT and SVM to train the dataset and test on incoming tweet upon processing. Test data containing emoticons have no effect on the classifier. The results show that MaxENT and Naïve Bayes models have the best performance using word unigram and bigrams.

Agarwal[8], used a dataset of 5127 tweets and SVM was used to evaluate different cases of unigram models. The features set explored were: number of negation words, number of positive words, number of negative words, number of extremely positive, extremely negative, number of positive and negative emoticons, positive and negative hash tags, capitalized words and exclamation words. 3-fold and 5-fold cross validation was performed on binary classification problem (positive or negative). While 3 fold cross validation gave 60.83% accuracy, the 5 fold cross validation gave 75.39% accuracy.

The work done by Wilas[9] , approached the problem in two phases: Preparation and Analysis. The preparation phase is arranged as follows: Micro-blog retrieval, Opinion classification based on a classification algorithm, assigning polarity score from sentiWordNet dictionary, and retrieving feature set of a particular product. The analysis phase includes utilizing the twitter search Application Protocol Interface to collect tweets, training and testing the opinion filtering module using SVM with Information Gain feature selection, tokenization/stop-word removal/link-removal/term-normalization/slang handling at the polarity module, and finally outlining the features in each tweet at the feature module.

The assumptions made here includes retrieving tweets purely related to smart phone brands, duplicated content are ignored, non-English tweets are filtered out, tweets that contain more than one product are ignored, emoticons are removed, sarcasm was not handled and abbreviations were removed. The dataset includes tweets from March 2010 June 2010 and 100,000 tweets were retrieved from the crawler that was split into 10 categories of 10,000 each. Within each category, 1,000 were randomly selected for training the opinion classification module where 600 annotated dataset was subjected to SVM training.

In the work of Hassan Saif,[10], before classifying sentiment, they first did subjectivity v/s objectivity classification. Then the polarity classification was performed on the subjectivity set. They split the feature set into two categories: meta-features and tweet syntax features. Meta features consist of words, its POS (using POS dictionary). Tweet syntax features consist of re-tweets, hash-tags, reply, link, punctuations and emoticons (based on a dictionary of 11 entries). An error rate of only 18.7% was claimed.

The supervised learning approaches have predominantly used statistical based learning algorithms. Naïve Bayes classifier, Support Vector Machines and maximum entropy are the commonly found algorithms in the history of literature work. These algorithms are based on tokenizing and quantifying patterns that are crucial to the opinion mining context.

3.3.2 Unsupervised Learning

A prominent work on unsupervised sentiment analysis was done by Turney[11] where point- wise mutual information (PMI) was calculated between the tokens of the review versus the cluster centers of poor and excellent. The sentiment of a document is calculated by the average semantic orientation of all such phrases. Accuracy of 66% was achieved using this technique over the movie review domain.

3.3.4. Inferences

From the literature survey it can be inferred that most of the experiments have considered emoticons as a noise and there is no effect on the classifier when the testing data with emoticons are introduced. Another observation is, the sentiments that define the entire phrase are adjectives supported by the verbs and adverbs. POS tagger can be used that tags each word in a phrase with its corresponding part-of-speech (noun, verb, adjective, etc.). Through this, only adjectives, adverbs and verbs are retrieved and polarized for a given sentence.

4. PROPOSED EXPERIMENTAL SYSTEM AND RESULTS

The proposed model, shown in Figure1, is a semi-supervised approach that uses Naïve Bayes method to deduce a probability count of the tokens in the entire corpus. The Map Reduce for machine learning is given in Figure2. The learning proceeds in three phases: Training, Validation and Testing.

4.1 Classifier

In this work, the Naïve Bayes classifier model is used, since the works mentioned in the literature survey have established that Naïve Bayes method has provided good results for sentiment analysis than the other supervised learning algorithms [5][6][7]. Naïve Bayes classifier is based on the Naïve Bayes theorem from probability. The prerequisite to using Naïve Bayes is an annotated dataset bearing the set of data and its corresponding class. The data can be a document, a sentence, or an entity, depending on the context of the classification problem.

4.2 Training and Testing

The training and testing module flow involved in the supervised Naïve Bayes learning is given in Figure3.

4.3 Results

For training the classifier, an exhaustive training dataset having thousands of annotated review feeds is used. The validation data contains reviews about a specific product, scaling up to 5 varieties of products: books, electronics, DVD, movie, kitchen, By size it is 1% of the training data and accuracy, precision & recall are calculated over the confusion matrix hence generated. The dataset is equally distributed and contains approximately 352 positive and negative records in each case. The testing data is with unlabelled content. The maximum accuracy achieved was 84% of the cases when tested with bi-grams with POS tagger.

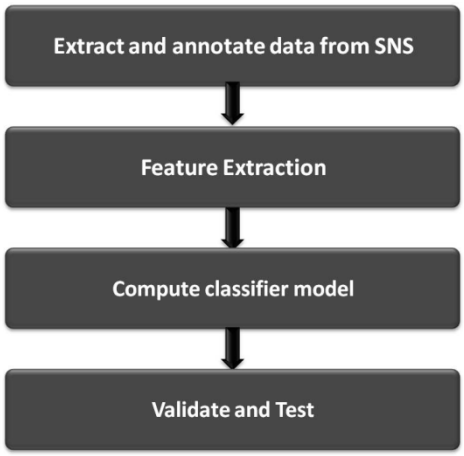

Figure1: Approach for classification

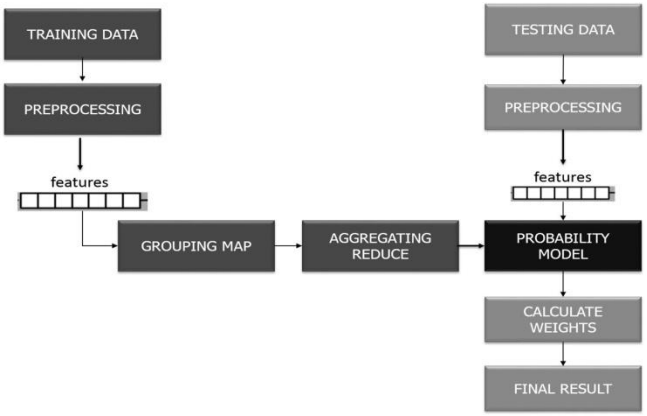

Figure2: Map Reduce for machine learning

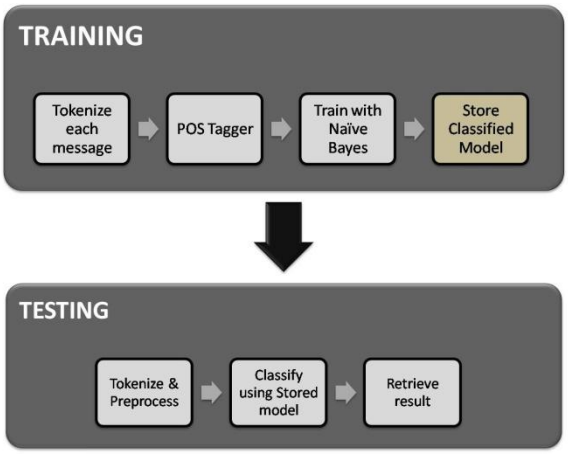

Figure3: Training and Testing Module

5. CONCLUSION

The sentiment mining of Twitter data was implemented using Mahout. The Naiive Bayes method adopted by us produced an accuracy of 84%. MapReduce framework was incorporated, so as to implement the work in a distributed environment using Mahout. The sentiment analysis done here is content-based filtering where every user is tabulated against every product with an opinion score. Polarizing the score is on a zero or one notation since only binary classification is considered. But it can be scaled to more number of classes to suit the requirement specifications.

REFERENCES

[1] http://www.cio.in/article/social-crm-enterprise-how-analytics-can-move-you-greater-success dated 15 Nov 2012.

[2] Bo Pang and Lillian Lee, "Opinion Mining and Sentiment Analysis", in Foundations and Trends in Information Retrieval, Vol. 2, pp 1-135,2008

[3] Tonghui li and Xixi Xiao, "An Unsupervised Approach for Sentiment Classification", in IEEE Symposium on Robotics and Applications (ISRA), pp 638-640, 2012.

[4] E. Kouloumpis, T. Wilson and J. Moore, "Twitter Sentiment Analysis: The Good the Bad and the OMG!", in Proceedings of the Fifth International Association for the Advancement of Artificial Intelligence (AAAI) Conference on Weblogs and Social Media, (ICWSM), 2011.

[5] Thomas Lake, "Twitter Sentiment Analysis", Western Michigan University, Kalamazoo, MI, For client William Fitzgerald, April, 2011.

[6] James Spencer and Gulden Uchyigit, "Sentimentor: Sentiment Analysis of Twitter Data", in First International Workshop on Sentiment Discovery from Affective Data (SDAD), pp 417-424, 2012.

[7] Alec Go, Lei Huang and Richa Bhayani, "Twitter Sentiment Classification using Distant Supervision", in CS224N Project Report, Stanford, pp 1-12, 2009.

[8] Apoorv Agarwal, Boyi Xie, IIia Vovsha, Owen Rambow and R Pas- sonneau, "Sentiment Analysis of Twitter Data", in Proceedings of the Workshop on Languages in Social Media (LSM), pp 30-38, 2011.

[9] Wilas Chamlertwat, Pattarasinee Bhattarakosol and Tippakorn Rungkasiri, "Discovering Consumer Insight from Twitter via Sentiment Analysis", in Journal of Universal Computer Science, Vol. 18, no. 8,2012.

[10] Hassan Saif, Yulan He and Harith Alani, "Semantic Sentiment Analysis of Twitter", in 11th International Semantic Web Conference (SWC), 2012.

[11] P. Turney, "Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews", in Proceedings of the 40th Annual Meeting of the Association of Computational Linguistics (ACL),pp 417-424, 2002.

Appendix

Basic Definitions

- Sentiment: A sentiment is the author's opinion or emotion expressed on a topic.
- Subjectivity Analysis: Subjectivity is used to express private states in the context of a text, generally in term of opinions, emotions etc. This term originated from the areas of information retrieval, artificial intelligence and natural language processing.
- Objective Sentiment: If there is no opinion conveyed on the source content, then it is called objective sentiment.
- Opinion: An Opinion is a belief or judgment of a large number or majority of people formed about a particular thing, not necessarily based on fact or knowledge. In general, opinion refers to what a person thinks about something. In other words, opinion is a subjective belief and is the result of emotion or interpretation of facts.
- Opinion Mining or Sentiment Analysis: Opinion mining is a technique to detect and extract subjective information in text documents. In general, sentiment analysis tries to determine the sentiment of a writer about something. The sentiment may be his or her judgment, mood or evaluation, while sentiment classification refers to evaluation of a target object (positive or negative viz film, book etc.)
- Tokenization: Is the process of breaking up a stream of text into words, phrases, or other meaningful entities called tokens. The list of tokens becomes input for further data pre-processing (eg stop words removal and tagging).
- StopWords: Stop words do not convey any information while attempting to classify sentiment. Some of the most common stop words are: the, is, at, which, etc. Stop words are filtered out in the pre-processing stage.
- Stemming: Stemming is a process of transforming variant forms of a word into a common form. For example, learn, learnt, learning can be transformed into learn. Deriving these different forms of a same word into root word on big data helps in improving the performance of informational retrieval.
- Tagging: Tagging is the process of reading the text in some language and assigning parts of speech to each word (tokens), such as noun, verb, adjective, etc. It is useful in retrieving informational words that gives sentiment of the text.