Sentiment Analysis of News Headlines using Naïve Bayes Classifier

Harpreet kaur
Assistant Professor
Department of Information Technology
Jalandhar, India
sonia29590@gmail.com

Dr. Vinay Chopra
Assistant Professor
Department of computer science
Jalandhar, India
vinaychopra222@yahoo.co.in

Abstract—The amount of user generated content is increasing day by day and it involves detection of opinions about particular topic or an object. Sentiment analysis is used to extract sentiments of people about products, moviews, political events etc. It identifies the viewpoint of opinion holder and polarity of the content i.e. positive negative or neutral. Given the large amount of news being generated these days through various websites, it is possible to mine the general sentiment of particular news. Nowadays many people read news online. People's perspective tends to undergo a change as per the news content they read. The majority of the content that we read today is on the negative aspects of various things e.g. rapes, thefts, corruption etc. Reading such negative news is spreading negativity amongst the people. The positivity surrounding the good news has been drastically reduced by the number of bad news. The objective of this work is to provide a platform for serving good news and create a positive environment.

Keywords—Sentiment analysis, text mining, machine learning, naive bayes.

1. INTRODUCTION
Sentiment Analysis focuses on the extraction of the sentiments of people from the text available on internet. The information presented in different formats on internet such as posts, news articles, comments, and reviews is available in huge volume. If someone wants to buy any product then they can look up the reviews of other people and take decision regarding the product. This user generated is too large to analyze by normal user. To automate this process various Machine Learning and Natural language Processing techniques are used. By collecting and analyzing these reviews people will be able to compare the features of the products and take their own decisions. Moreover manufacturers will be able to find out the strengths and weaknesses of their products. By analyzing the weaknesses they will be able to improve their products and use the business intelligence for future investments.

Text documents can be classified into two categories, namely facts and opinions. Facts are the objective statements about some entity or event or product, and opinions describe attitudes, sentiments, feelings and emotions regarding a product, service, topic or an issue.

Earlier businesses and organizations hire people for the task of reading newspapers and manually compiling lists of positive, negative and neutral references to the organization.

Today newspapers are published online. There are a wide range of opinionated articles posted online in blogs and other social media. So there is a possibility of automatically detecting positive or negative mentions of an organization in articles published online. The effort required to collect this type of information gets dramatically reduced.

Research in such an area was difficult in the past, due to the scarcity of large manually-annotated corpora. This problem has been solved as a result of the recent increase in the number of news websites. News Articles can be collected daily by manually copying the news or using RSS feeds that are mostly available in every news website like The Times of India, BBC News etc. By serving as data sources, these websites facilitate the research on identifying the polarity of the news articles that can be helpful to readers as this can be beneficial to them.

In this paper, we classify online news articles on the basis of polarity like positive, negative or neutral. More formally, we have a set N of news articles and a set P of polarity. The goal is to find a function f: N → E such that f sends n ∈ N to the correct e ∈ E. To achieve this goal, two approaches can be used, machine learning approach and Lexicon based approach.

Machine Learning approach is classified into two categories. Supervised machine learning and unsupervised machine learning. In supervised machine learning the program is “trained” on a pre-defined set of training examples, which then facilitate its ability to reach an accurate conclusion when given new data. The training data is labeled with the correct answers. In Unsupervised machine learning, the program is given a bunch of data and must find patterns and relationships therein. A collection of unlabeled data is given, which is to be analyze and discover patterns within. Further supervised learning includes various techniques like Naïve Bayes, support vector machines, nearest neighbor, decision tress etc. Unsupervised learning techniques include Lexicon based
techniques. There are three methods to construct a sentiment lexicon: manual construction, corpus-based methods and dictionary-based methods. Each concept used in the application (e.g. one word or the entire article) has three associated values positive, negative and neutral.

In this paper, we examine classification performance by using naïve bayes technique on news articles. A lot work has been done on classification of product reviews, movie reviews, e-commerce etc. We are going to analyze reader’s emotional state while reading news articles.

The paper is organized as follows: Section II describes some related work on sentiment analysis in news articles; Section III describes our process for analysis of news on the basis of polarity, which is the main focus of this paper; Section IV describes our experiments and the results we obtained finally, Section V outlines future directions for research emerging from our work.

II. RELATED WORK

Arora, Rajeev, and Srinath Srinivasa have performed research on Sentiment Analysis in News Articles Using Sentic Computing. In this paper opinions are classified into four types positive, negative, neutral and constructive and various facets of opinion mining are discussed like opinion structure and opinion mining tools and techniques. Entity discovery, aspect identification, Lexical acquisition. An integrated opinion mining flow for an opinion mining engine has been proposed [3].

Mostafa Karamibekr, Ali A. Ghorbani, have proposed differences between products and services and Social issues. The author has collected a dataset consisting of more than 1000 comments, which express public opinions regarding “abortion” from CNN, ProCon.org, Yahoo Answers, and Women’s Issues at About.com. Author proposed a method which focuses on the role of verbs, adjectives, adverbs and nouns as opinion terms. The proposed method is compared with BOW approach and the accuracy achieved is 65%.

Ms. K. Mouthami, Ms. K. Nirmala Devi, Dr. V. Murali Bhaskaran, have proposed system that uses fuzzy set theory. Sentiment classes are refined as three fuzzy sets positive sentiment, neutral sentiment and negative sentiment. Evaluation measures like F-measure, precision, recall and accuracy are discussed by the author.

Prashant Raina, presents opinion mining engine which performs fine grained sentiment analysis to classify sentences as positive, negative or neutral. The parameters for the evaluation are precision, recall and F-measure. Experiments have been conducted on a sample of 3,181 sentences from the MPQA opinion corpus. The corpus is a collection of over 500 articles from news sources. The accuracy obtained is 71%.

S Padmaja, Sasidhar Bandu, Prof. S Sameen Fatima, Pooja Kosa, M C Abhinaya did comparison between sentiwordnet and two machine learning based algorithms Naïve Bayes and Support vector machines are shown using weka tool. Datasets are created from three different news databases. The parameters for the evaluation are precision, recall, F-measure and accuracy. SentiWordNet has the highest coverage with 74.2% for UPA (congress) whereas Naïve bayes with 72.5% is the next highest coverage score for TDP and SVM accounts for a fairer score of coverage with 66.7% for TRS.

K. Saraswathi, A. Tamilarasi. Phd proposed a system to investigate the efficiency of Bagging to predict opinions as positive or negative for online movie reviews from IMBD dataset. 300 instances (150 positive and 150 negative) were used for evaluation. The classification accuracy of Bagging was compared against Naïve Bayes and CART. Results demonstrate the efficiency of Bagging. Bagging achieves 14 to 15.34% better classification accuracy than the other classifiers.

Wael M.S. Yafooz, Site Z.Z. Abidin, Nasiroh Omar, investigate the challenges and issues that relate to online news domain that includes news extraction, news clustering, news topic detection and tracking, multilingual news, news summarization and aggregation. The name entity technique is found to be very useful in order to handle the huge amount of information from the heterogeneous websites and languages.

<table>
<thead>
<tr>
<th>Title of the paper</th>
<th>Author</th>
<th>Year</th>
<th>SA Methods (MLB / LB)</th>
<th>Dataset</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>[4]</td>
<td>Peasant Riana</td>
<td>2013</td>
<td>MLB</td>
<td>MPQA opinion corpus</td>
<td>71%</td>
</tr>
<tr>
<td>[10]</td>
<td>Side Raquel Hider, Rehab Mehrotra</td>
<td>2011</td>
<td>MLB</td>
<td>News Corpo r</td>
<td>79.93%</td>
</tr>
<tr>
<td>[19]</td>
<td>Alexandr a Balahur et al.</td>
<td>2009</td>
<td>LB</td>
<td>EMM news data</td>
<td>50%</td>
</tr>
</tbody>
</table>
III. PROPOSED FRAMEWORK

The framework comprises of four stages: Data collection and cleaning; preprocessing; sentiment analysis and finally experiments and results. Following describes the brief description of the stages. The architecture of the proposed framework for sentiment analysis is presented in Fig.3.1

A. Text Collection and Cleaning Stage

For the analysis of news headlines, dataset has to be collected. For our analysis, we have gathered news articles from RSS feed. For our study a sample of 103 news, dating from January 1, 2015 to February 28, 2015, was obtained from online Indian newspaper namely The Times of India. The collected text is noisy and methods for cleaning and parsing of the data to form a corpus for further processing.

B. Pre-processing Stage

At this stage, the corpus is transformed in to feature vectors; for our purpose of conducting this work, strings are converted into words using some filtration techniques. We adapted a simple feature selection or pre-processing method to transform or tokenize the text stream to words; these methods constitute a sequence of the following tasks; removing delimiters, removing numbers and stop words.

C. Sentiment Analysis Stage

This stage handles the polarity measurement and sentiment. We approach these tasks by employing machine learning methods.

I. Machine learning-based sentiment classification

In the machine learning based approach, beside the corpus, the fundamental requirement is a dataset, already coded with sentiment classes. As described, the classifier is modeled with the labeled data. For the purpose of this report, multinomial naive Bayes is used as a baseline classifier because of its efficiency. We assume the feature words are independent and then use each occurrence to classify headlines into its appropriate sentiment class. Naive Bayes is used because this is easy to implement and requires small amount of training data to estimate the parameters. It follows from [21] that our classifier which utilizes the maximum a posteriori decision rule can be represented as:

$$c = \arg\max_{c \in C} \left( \frac{P(c|d)}{d} \right)$$

$$= \arg\max_{c \in C} \left( P(c) \prod_{1 \leq k \leq n_d} P(t_k|c) \right)$$

(1)

Where $t_k$ denotes the words in each headline and $C$ is the set of classes used in the classification, $P(c|d)$ denotes the conditional probability of class $c$, $P(c)$ is the prior probability of a document occurring in class $c$ and $P(t_k|c)$ denotes the conditional probability of word $t_k$ given class $c$. To estimate the prior parameters, equation (1) is then reduced to:

$$C = \arg\max_{c \in C} \left( \log P(c) + \sum_{1 \leq k \leq n_d} \log P(t_k|c) \right)$$

(2)

IV. EXPERIMENTS AND RESULTS

In this section we discuss the dataset used in our experiments and the classification results obtained with our model.

A. Datasets

The work described in this paper is part of a larger research to improve the accuracy of sentiment analysis in the daily news present in online news articles. For our study a sample of 103 news articles, dating from January 1, 2015 to February 28, 2015, were obtained from online Indian news website namely The Times of India.
B. Parameters

Following are the parameters computed in sentiment analysis of news articles using naïve bayes algorithm.

i. Precision and recall

Precision and recall are two widely used metrics for evaluating performance in text mining, and in text analysis field like information retrieval. Precision is used to measure exactness, whereas recall is a measure of completeness.

\[
\text{Precision} = \frac{TP}{TP + FP} \quad (3)
\]

\[
\text{Recall} = \frac{TP}{TP + FN} \quad (4)
\]

ii. F-measure. F-Measure is the harmonic mean of precision and recall. This gives a score that is a balance between precision and recall.

\[
\text{F measure} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \quad (5)
\]

iii. Accuracy: Accuracy is the common measure for classification performance. Accuracy is the proportion of correctly classified examples to the total number of examples, while error rate uses incorrectly classified instead of correctly.

\[
\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (6)
\]

V. RESULTS

Following screenshot shows the implementation of naïve bayes on new headlines for sentiment analysis that measures the accuracy, precision, recall, F-measure, TP rate, FP rate. Accuracy obtained is 53.3%.

CONCLUSION

In this paper we have discussed about the polarity of news articles in terms of positive, negative and neutral using Naïve Bayes algorithm. There are many machine learning techniques that can be used for sentiment analysis in any area like social media, medical etc to determine the polarity of news. News articles present an even greater challenge, as they usually avoid overt indicators of emotion or attitudes. First a framework is developed for harnessing and tracking these headlines and experiences using text mining and sentiment analysis, we then conducted sentiment analysis of news headlines using naïve Bayes classifier was used to obtain a baseline result for assessing other classifiers. Future work can be done on news headlines by using other machine learning techniques or by combining different techniques to get good results.

References


