DESIGN AND IMPLEMENTATION OF HYBRID CLASSIFICATION ALGORITHM FOR SENTIMENT ANALYSIS ON NEWSPAPER ARTICLES

Harpreet Kaur
Department of Information Technology, Punjab Technical University Jalandhar, India

Dr. Vinay Chopra
Department of Computer Science, Jalandhar, India

ABSTRACT
A significant growth in the volume of research in sentiment analysis has been seen from last few years, mostly on highly subjective text types. The basic aim of sentiment analysis is to determine the sentiments or emotions of the writer with respect to some topic. In social media, sentiment analysis is useful to automatically characterize the emotions or mood of consumers toward a specific product or company and determine whether they reflect positive attitude or negative. In our work, we focus on news articles. The main tasks identified for news opinion mining consists of extracting news articles from online news websites and identifying positive and negative sentiments that exist in that article. A large number of companies use news analysis to help them make better business decisions. The news that we read today contains mostly negative content e.g. corruption, theft, rapes etc. Positive news seems to be hiding and negative news is highlighting. The objective of this research is to provide a platform for creating positive environment. This is achieved by finding the sentiments of the news articles and filtering out the positive as well as negative articles. This would help us to spread positivity around and would allow people to think positively.

KEYWORDS—sentiment analysis, machinelearning, classification, News analysis.

I. INTRODUCTION
Sentiment analysis uses a computational approach to identify opinionated content as it is language processing task and categorize it as positive or negative. It is a combination of natural language processing and text mining. Sentiment analysis tries to identify the expressions of opinion and mood of writers about an object. Today every person reads news online and watches advertisements regarding a product, a movie or a book before actually putting the money into it. Many newspapers are published online like The Times of India, BBC News etc. In addition to this, there are a wide range of opinionated articles posted online in blogs and other social media thus stimulating the possibility of automatically detecting positive or negative mentions of an organization in articles published online and thereby dramatically reducing the effort required to collect this type of information[4]. The automatic processing of texts to detect opinion expressed has been demonstrated as opinion mining or sentiment analysis.

Sentiment analysis contains two types of information, namely, facts and opinions. Facts are the statements which describe the nature of a product or event and are objective. Opinions describe the appraisals, attitude and emotions regarding to that entity but the recent trend is to focus on opinions. Opinion mining consists of number of challenges. The first one is word based challenge. Sometimes it is difficult to understand the emotion of a given word because same word can have positive as well as negative meaning in one statement or the other. Human beings can understand this change and interpret the statements but it’s difficult for a machine to interpret. Earlier the organizations had only own way to track their reputation in the media and that was only surveys on paper. The surveys were conducted by the surveyors who were appointed in the organization but now there is no need to hire people to take surveys. Today newspapers are published online and people can give their views online. The earlier was an expensive one. The automatic detection of polarity reduces the effort of collecting such type of information.

There are two approaches to achieve sentiment analysis, machine learning approach and Lexicon based approach [6]. 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 that facilitate the ability to reach an accurate conclusion when new data is given. The training data is labeled with the correct answers. In Unsupervised machine learning, bunch of data is input to the program and it finds 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 Naive Bayes, support
In this paper, we examine classification performance by implementing naïve bayes classifier and our proposed hybrid approach on news articles. A lot of 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 [3] 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.

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.Mouthani, 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 fuzzy set, neutral sentiment fuzzy set and negative sentiment fuzzy set. Evaluation measures like F-measure, precision, recall and accuracy are discussed by the author.

Prashant Raina, [4] 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%.

The authors of [5] 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 [8] 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, Siti 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.

UBALE SWATI, CHILEKAR PRANALI, SONKAMBLE PRAGATI [22] implemented naïve bayes classifier on news articles. They proposed a conceptual framework for analyzing the polarity of news articles.

III. PROPOSED FRAMEWORK

The News Sentiment Analysis automatically analyses news articles. It identifies the positive, negative or neutral opinions. The conceptual 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 [26]. The
architecture of the proposed framework for sentiment analysis is presented[2] in Fig.3.1

Text Collection and Cleaning Stage

For the analysis of news headlines, data is collected from online Indian newspaper namely The Times of India. News. The first task is crawling of news using web crawler. After that the news is stored as .arff file. For our analysis, we have gathered news articles from RSS feed. For our study a sample of 112 news, dating from January 1, 2015 to February 28, 2015, was obtained. The collected text is noisy and methods for cleaning and parsing of the data to form a corpus for further processing.

Pre-processing Stage

At this stage, the corpus is transformed into 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. For stop words removal, a list of stop words is provided in the filtration process.

Fig. 3.1 Architecture of the proposed framework

C. Sentiment Analysis Stage

This stage of framework handles the polarity measurement and sentiments. We approach these tasks by employing following machine learning methods.

I. Machine learning-based sentiment classification using Naïve bayes classifier

The fundamental requirement in machine learning based approach the dataset, already coded with sentiment classes. As described, the classifier is modeled with the labeled data. For this purpose, multinomial naïve 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. Naïve 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:

\[
C = \arg \max_{c\in C} (P(c/d))
\]

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

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 [26]:

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

We have implemented Sentiment analysis for newspaper articles using naïve bayes in [27]. In this paper hybrid approach is being implemented and compared with naïve bayes classifier on newspaper articles.

II. Machine learning-based sentiment classification using Hybrid approach

In Hybrid approach, we have combined the functionality of naïve bayes and decision table classifiers and this has been proved that this combination gives improved performance of classification. The polarity of news articles has been identified using naïve bayes as well as using hybrid approach and the results of the hybrid approach are better that the naïve bayes classifier. The Decision table algorithm stores the input data in condensed form based on a selected set of attributes and uses that
table when making predictions for the data [25]. Based on the observed frequencies each record in the table is associated with class probability estimates. The approach is to choose a set by maximizing cross-validated performance.

In our approach we used forward selection to select attributes in decision table. At each point in the search it evaluates the merit associated with splitting the attributes into two disjoint subsets: the Decision table, the other for Naïve Bayes. We use a forward selection, where, at each step, selected attributes are classified and modeled by Naïve Bayes and the remaining attributes are modeled by Decision table. Initially all attributes are modeled by Decision Table.

The class probability estimates of the Decision Table and Naïve Bayes must be combined to generate overall class probability estimates.

IV. EXPERIMENTS AND RESULTS

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

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.

Parameters

Following are the parameters [5] 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.

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

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

We conducted experiments on 112 news headlines that have been collected from The Times of India news website. We implemented Multinomial Naïve Bayes Classifier and Our proposed hybrid approach. The accuracy obtained from naïve bayes classifier is 53.5%. In comparison to other datasets. News data contains less emotional words. The social media datasets contains more emotional words rather than other datasets. So it’s very difficult to classify news articles. Then we implemented our proposed approach on same dataset and the results obtained are 71.5%. As we can see the performance has been increased using our proposed approach. Table 1. Contains the comparison of Naïve bayes classifier and hybrid approach. Various parameters like precision, recall, F-measure are compared.

<table>
<thead>
<tr>
<th>S.no</th>
<th>Parameters</th>
<th>Naïve Bayes</th>
<th>Hybrid approach</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Accuracy</td>
<td>53.5%</td>
<td>71.4%</td>
</tr>
<tr>
<td>2.</td>
<td>Precision</td>
<td>0.54</td>
<td>0.79</td>
</tr>
<tr>
<td>3.</td>
<td>Recall</td>
<td>0.536</td>
<td>0.714</td>
</tr>
<tr>
<td>4.</td>
<td>F-measure</td>
<td>0.537</td>
<td>0.661</td>
</tr>
</tbody>
</table>

Table 2. Contains the confusion matrix. Each column of the confusion matrix represents the instances in a predicted class, while each row represents the instances in an actual class. The table reports reports the number of false positives, false negatives, true positives, and true negatives.pos refers to positive, neg refers to negative and neu refers to neutral.
TABLE 5.2. Confusion matrix

<table>
<thead>
<tr>
<th>Predicted</th>
<th>Actual</th>
<th>Pos</th>
<th>Neg</th>
<th>Neu</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pos</td>
<td>24</td>
<td>9</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Neg</td>
<td>0</td>
<td>22</td>
<td>12</td>
<td></td>
</tr>
<tr>
<td>Neu</td>
<td>2</td>
<td>8</td>
<td>34</td>
<td></td>
</tr>
</tbody>
</table>

VI. CONCLUSION

In this paper we have discussed about the polarity of news articles in terms of positive, negative and neutral using Naïve Bayes algorithm and our proposed hybrid approach. 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. Identifying polarity in news articles is great challenge. In this paper 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 that can be used to obtain a baseline result for assessing other classifiers. Then our proposed hybrid approach is used to detect the polarity of news headlines. This approach is working over complete dataset. Future work can be done on news headlines by using other machine learning techniques or by combining different techniques to get good results. The work can also be done over single news headline by using different technique to detect the polarity of single news headline.

REFERENCES


