ISSN:0975-9646

Sentiment Classification of News Articles

Kiran Shriniwas Doddi[1], Dr. Mrs. Y. V. Haribhakta[2], Dr. Parag Kulkarni[3]

[1,2] *Department of Computer Engineering,*
College of Engineering Pune, India
[3]*EkLaT Research Center,*

Abstract **– The advent use of new online social media such as articles, blogs, message boards, news channels, and in general Web content has dramatically changed the way people look at various things around them. Today, it's a daily practice for many people to 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. corruption, rapes, thefts etc. Reading such news is spreading negativity amongst the people. Positive news seems to have gone into a hiding. The positivity surrounding the good news has been drastically reduced by the number of bad news.**

This has made a great practical use of Sentiment Analysis and there has been more innovation in this area in recent days. Sentiment analysis refers to a broad range of fields of text mining, natural language processing and computational linguistics. It traditionally emphasizes on classification of text document into positive and negative categories. Sentiment analysis of any text document has emerged as the most useful application in the area of sentiment analysis.

The objective of this project is to provide a platform for serving good news and create a positive environment. This is achieved by finding the sentiments of the news articles and filtering out the negative articles. This would enable us to focus only on the good news which will help spread positivity around and would allow people to think positively.

Keywords **- Document classification, Sentiment Analysis, Support vector machine (SVM).**

I. INTRODUCTION

With the arrival of internet, there has been a drastic change in the social life, lifestyle and decisions of common people. Today it's everyday activity and regular practice for each person to read news online and watch advertisements regarding a movie, a product or a book before actually putting the money into it. As it has changed their lifestyle, it also has impact on the social life of an individual. The exposure to new online social media such as articles, blogs, message boards, news channels such as Web content is influencing their social life and the way people look at various things around them. People's perspective tends to undergo a change as per the content they read.

The social media has now occupied the major space on the Web. The new user-centric Web hosts a large volume of data every day. Users are now co-creators of content on web, rather than just passive consumers. The user is now contributing to social media ranging from articles, blog posts, news, tweets, reviews, photo/video upload, etc. This is creating a large amount of the data on the Web as unstructured text.

The majority of the content that we read today is on the negative aspects of various things e.g. corruption, rapes, thefts etc. Reading such news is spreading negativity amongst the people. Positive news has been dominated and getting less attention. The positivity surrounding the good news has been drastically reduced by the number of bad news.

The objective of this project is to provide a platform for serving good news and create a positive environment. The new challenging task here is to analyse large volume of unstructured text to be more specific news articles and devise suitable algorithms to understand the opinion of others and find positive and negative aspect of it. This would enable us to focus only on the good news which will help spread positivity around and would allow people to think positively.

II. LITERATURE SURVEY

Today researchers have worked on finding sentiments of movie reviews, product reviews, tweets, prediction on rise or drop in stock price etc. The aspect of such dataset is that it includes short and rich structured information about individuals involved in communication. Finding sentiments of structured data is easier than finding it from unstructured data.

A. Classification methods and Feature extraction

For automation of sentiment analysis, different approaches have been invented to predicting sentiment from words, expressions and from documents. There are many natural language processing based and pattern based algorithms like Naïve Bayes(NB), Support Vector Machine(SVM), Maximum Entropy (ME) etc.

However, some research has investigated more complex algorithms in few years back. Martineau and Finin invented Delta TFIDF in 2009, an intuitive general purpose technique, so that words can be efficiently weighted before applying to classification [12]. In most of the papers, researchers have used dictionary based document classification technique. Some have used Bag of words technique along with some machine learning algorithms most likely support vector machine. SVM algorithm was mostly used classification algorithm because it is highly

generalized and its performance is different for various ranges of applications. It is also considered as one of the most efficient classification algorithm which provides a comprehensive comparison for text classification in supervise machine learning approach.

B. Model building:

In old papers, the model was built once using training data and used to classify the new dataset based on built model. While using so, if new dataset has new features (with the advent of Web and social media), the model fails to classify the document correctly and its accuracy falls. To avoid this, they had to rebuild the new model with new training dataset considering new features and calculate the accuracy. But this is not feasible solution and approach for this problem. News corpus from different sources consists of news articles and editorial content with a broad range of discussions and topics. So challenging task here is to design a generic heuristic algorithm that will correctly extract sentiments from these mixed news corpus for document classification and feature extraction.

III. PROPOSED SOLUTION

Analyzing Sentiment of the text is itself a challenging task in Natural Language Processing. There are many approaches studied by me while finding solution for sentiment of news articles. Initially, we collected different words which carry sentiments (like positive, negative and neutral) from DAL, GI and WordNet [3]. Later we needed test data for analyzing the structure of the text. We have read RSS feeds from various sources based on category [4]. With this mixed news corpus on broad range of topics, we want our model to run correctly by extracting new features from dynamic data. So one way to achieve this is, we need to dynamically prepare training dataset after every interval of time, so that system will build the new model with this new corpus at that time and classify correct new test instance. Below is the heuristic approach that I have planned to use:

A. Tag sentence using POS tagger:

We can use the POS tagger to tag the words in each sentence. We can find the verb, adverbs and adjectives from each sentence using standard POS tagger. It tags the different words in a sentence and Stanford NLP group has provided different trained models for different languages [5]. Here we are interested in grammatical rules for English language only.

Major challenge here is to choose the set of words which defines our feature vector. Here we will try with two different set of attributes, like

a. Adjective + Adverb
b. Adverb + Verb

We also need to find the negation word from the sentence.

B. Find sentiscore using sentiwordnet

Sentiwordnet is an open source library which is a lexical resource for opinion mining. It gives a three sentiment scores (sentiscore) for a word: positivity, negativity and objectivity [3].

We will use this library to get the sentiscore of attributes mentioned above i.e. for adjective, adverb and verb. If library doesn't give you the sentiscore, we will convert that word to its base form using wordnet libraries and then lookup again to find the sentiscore.

For example, we have passed a word "extreme" as "a" means Adjective to find the sentiment of it. So this library will return sentiment value and its classification as below:

Sentiment value: -0.0883838383838384
Sentiment Classification: weak_negative

C. Feature Extraction and document classification

It is very difficult task to process text which contains millions of different unique words, so it makes text analytics process difficult. Therefore, feature-extraction is one of the approaches used when applying machine learning algorithms like Support Vector Machine (SVM) for text categorization. With the survey, it has been found that a feature is a combination of keywords (attributes), which captures essential characteristics and sentiment of the text. A feature extraction method detects and filters only important features which a far smaller set than actual number of attributes and make them a new set of features by decomposition of the original data. Therefore this process enhances the speed of supervised learning algorithms [10]. For our purpose, we will focus on the adverb, adjective and verb as our features. We will also apply some weighting scheme and some of the most popular ones are Binary, Term-Frequency (tf), Term-Frequency Inverse Document Frequency ($tf - idf$)

For document classification to prepare training data set, we will apply some algorithm to aggregate the sentiment of sentiwords (L). So the sentiment of overall document (global sentiment) can be calculated by applying some aggregation function as follows:

$$G = F(L_i)$$

This global sentiment will define sentiment of overall news article.

D. Apply Support Vector Machine

SVM is a linear/non-linear classifier used for classification of the text data. SVM algorithm was mostly used classification algorithm because it is highly generalized and its performance is different for various ranges of applications. It is also considered as one of the most efficient classification algorithm which provides a comprehensive comparison for text classification in supervise machine learning approach.

In SVM, we want a give set of points to be classified into two classes such that hyper plane between two classes will maximize the distance between the two classes. This will ensure the better classification of the points which are unseen, i.e. better generalization. With the classified dataset from document classification, SVM will prepare the model and classifier. We will use Weka for this purpose.

E. Data set and overall system

For our analysis, we have gathered news articles from different RSS feeds mentioned at [4]. We have pre-defined some news channels in the database and we have cron job that runs every hour and fetches news articles. This includes all categories of articles like sports, Bollywood, Hollywood, economy, cricket, nation, etc.

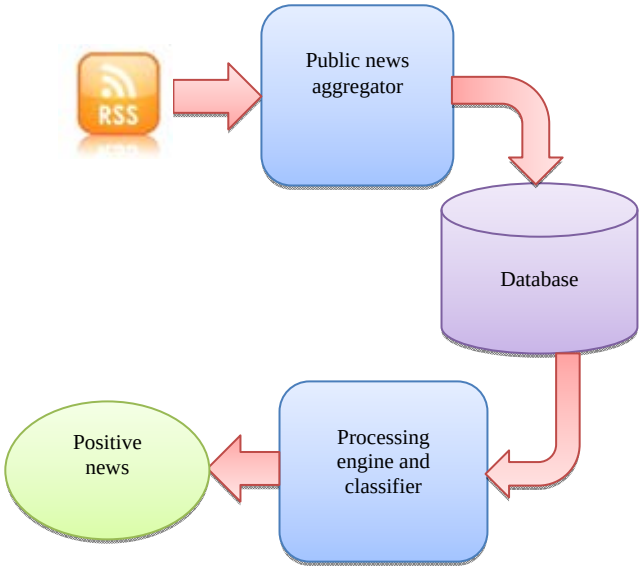

Figure1- System Diagram

V. CONCLUSIONS

Our aim is to provide a platform that filter outs negative articles and serve only good news after classifying news articles using sentiment analysis and create a positive environment among the society. This will help spread positivity around and would allow people to think positively.

REFERENCES

[1] Sentiment Analysis of movie review by V.K. Singh, R. Piryani, A. Uddin & P. Waila
[2] Probabilistic Model-based Sentiment Analysis of Twitter Messages by Asli Celikyilmaz, Dilek Hakkani-T, Junlan Feng
[3] WordNet: An Electronic Lexical Database http://wordnet.princeton.edu
[4] RSS feeds from feedburner.com and feedporter.com
[5] POS Tagger (The Stanford Natural Language Processing Group)
[6] Sentiment Classification for Stock News
[7] Polarity Detection in Reviews (Sentiment Analysis) by Manish Agarwal and Sudeep Sinha
[8] Sentiment Analyzer: Extracting Sentiments about a Given Topic using Natural Language Processing Techniques
[9] Polarity detection in movie reviews by Hitesh Khandelwal (Y5202)
[10] Text Mining with Support Vector machine and Non-negative Matrix Factorization algorithms by Neelima Guduru
[11] Polarity detection in movie reviews by Ramnik Arora
[12] J. Martineau and T. Finin, "Delta TFIDF: An Improved Feature Space for Sentiment Analysis," In Proceedings of the 3rd AAAI International Conference on Weblogs and Social Media, 2009.
[13] Wikipedia http://en.wikipedia.org/wiki/Unstructured_data
[14] World Wide Web Size http://www.worldwidewebsize.com/