International Journal of Research in Advent Technology (E-ISSN: 2321-9637) Special Issue
National Conference "NCPCI-2016", 19 March 2016
Available online at www.ijrat.org

Negative News No More

Classifying news articles on the basis of social impact

Ashish Mokalkar, Aishwarya Deshmukh, Rahul Tiwari, Suresh Akudari
ashishmokalkar79@gmail.com,aishwaryadeshmukh8@gmail.com, rahultiwari0812@gmail.com, sureshakudari315@gmail.com

Abstract—Most news we see are negative, and consists of terrorism, disasters, crime and corruption. News Media shows mostly negative news, which makes users think the world is a dangerous place to live. The media just tries to make money with sensationalism. The project focuses to develop an algorithm that will classify a news headlines according to how positive or negative or neutral that news story is. The algorithm should be able to distinguish between positive story "Efforts to use bamboo to rebuild post-quake Nepal" and negative headline "2 policemen die in road mishap in Surat". The algorithm will also identify neutral news i.e., that have neither strongly positive or negative impact on society (e.g. "iPad changing how college textbooks are used").For this we collected news RSS(Rich Site Summary) feed from national newspapers like, "The Times Of India". We also provide user a scale of negativity so that user can see news according to his choice. The focus is on classifying the headlines of the articles as positive/negative/neutral on the basis of their social impact.

***Keywords*—**Text Mining, Machine Learning, Natural Language Processing, News Classification

I. INTRODUCTION

Automatic text classification has always been an important application and research topic since the inception of digital documents. Today, owing to the very large amount of text data it is necessary to classify text data.

Intuitively Text Classification is the task of classifying a document under a predefined category. More formally, if a_i is a text articles of the entire set of articles A and $\{c_1,c_2,c_3,\ldots.c_n\}$ is the set of all the categories, then text classification assigns one category c_j to a text article a_i.

As in every supervised machine learning task, an initial dataset is needed. A text article may be classified in more than one category (Ranking Classification), but in this paper only researches on Hard Categorization (assigning a single category to each document) are taken into consideration. Moreover, approaches, that take into consideration other information besides the pure text, such as image related to texts or published date, are not presented.

In Figure 1 is given the graphical representation of the Text Classification process.

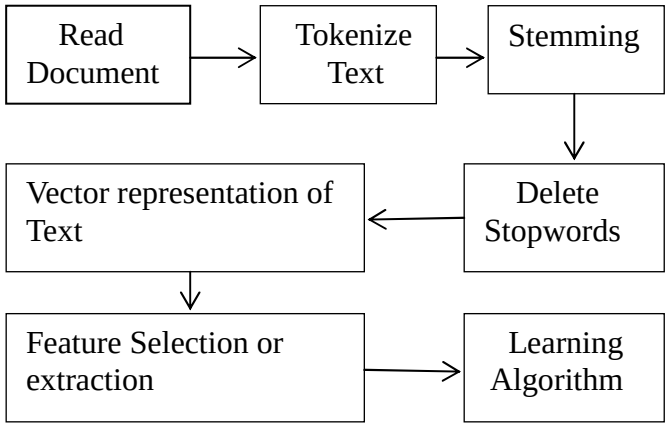


Figure 1: Text Classification Process

II. DATA COLLECTION AND DATA SETS

For this task we collected two data sets. The first is a set of 2000 news article headlines. The second is a set of 1000 news article headlines with short text excerpts from the article (typically the first 2-4 sentences of article text).the data samples were extracted from RSS feeds over several weeks.The sources of the news feeds include Times of India, CNN, Indian Express and Zee News.

Each data sample, corresponding to a single news article, was assigned to one of three classes, "positive," "negative," or "neutral." The data samples were classified by the two project team members. Articles were classified as "positive" if they featured a happy, inspiring, motivating, or funny topic. Articles classified as "negative" typically includes violence, crime, and loss of life or property. Articles that did not strongly fall into either category, including polarizing articles (e.g. on controversial or political subjects), were classified as "neutral".

Each of these classes are further classified into 3 sub-classes as "More", "Medium" and "Less" on the basis of their social impact.

III. NEWS ARTICLE PRE-PROCESSING

Text pre-processing is a primary step in news headlines classification process. Text is pre-processed effectively where the unstructured text data is initially obtained, which mostly is

the combination of both garbage and useful data. All of the data comes from variety of data gathering sources and is to be cleaned. Firstly, the text data is made free from all noisy and useless information, which include punctuation marks, semicolons, irrelevant texts, quotes, exclamation marks, dates etc.

A. Tokenization

Breaking huge text into small tokens or segments is said to be text tokenization. Each word in the Headline and content is treated as a string and these are broken in small tokens. Final output obtain is then served as input for further processing of text mining. All documents are combined and a set of words is obtained. This process is called dictionary creation for each news headline document.

B. Diacritics Removal

Diacritics are defined according to the language. For English, all irrelevant words that include commas, semi-colons, quotes, double quotes, full stops, underscores, special characters, dashes, and brackets etc. are removed. Simple way to implement this is to replace all those words, which are diacritics with simple space.

C. Stop Words Removal

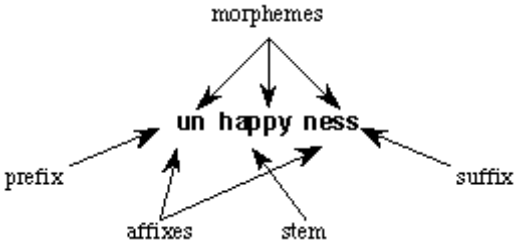
All those words, which appear frequently in text are said stop words. They are considered of low worth and are removed eventually. Those words, which are frequent in many news headlines, are defined useless with respect to frequency. On the other hand, if words deliver very less conceptual knowledge, they are said to be of low worth either their frequency lies in highest or lowest domains.

D. HTML Tags Removal

As all the data is gathered from website by parsing, the unstructured data contains large number of unnecessary html tags. So it is necessary to remove html tags from the data to do efficient analysis.

E. Morphological Analysis

Morphology is the study of the smallest meaning bearing units constituting a root wordstructure and formation of words. Its most important unit is the *morpheme*, which is defined as the "minimal unit of meaning". Consider a word like "unhappiness". This has three parts:



There are three morphemes, carrying a certain amount of meaning. *un* means "not", while *ness* means "a headland or promontory ". *Happy* is a *free morpheme* because it can appear on its own (as a "word" in its own right). *Bound morphemes* need to be attached to a free word, and so cannot be words in their own right.

F. N-gram analysis

N-gram is a sequence of words in a given text. For example, consider the news "India beats Australia". Without using n-gram analysis, this news will be classified as "Negative" because of the word "beats" in it. So, we applied bi-gram analysis that contains two consecutive words at a time which will be useful to avoid ambiguity.

IV. Feature Selection

When a huge amount of features are given and each of the features is a well known descriptive word for each class, a lot of time may be consumed in headlines classification and it may be possible that expected classification accuracy may not be achieved, and data may get over-trained. Hence for overcoming such issues, a process named as feature selection is adopted, where only those relevant and highly effective features are chosen, which may prove more noticeable as well as emphasizable for better headlines classification.

. We use a bag-of-words model for headline classification. One set of features was generated for each of the two data sets. Each feature set was based on a word dictionary extracted from the headline articles. Features represent individual word fragments, with an additional feature indicating whether a numeric value appears in the headlines. The feature set excludes stop-words (e.g. "a", "an" ,"the", "about," "over", "with"). Suffixes were removed, both automatically and manually, from dictionary words to create the list of tokens appearing in each data set. The feature set for the headline-only data includes 3654 features (3653 tokens, 1 numeric) and the feature set for the headline-plus-text data includes 5945 features (5944 tokens, 1 numeric).

V. News Classification

After feature selection, classification phase is important where unseen news headlines are classified into their respective classes. Major aim of this classification is to assign headlines to their respective classes. Classification has been performed for various purposes i.e. Short text classification, News headlines classification, Short messages classification, Social blogs messages classification, Email classification, Positive or negative news classification, Financial news classification on basis of news headlines, and much more. Our main focus here is to classify news articles as "positive", "negative" or "neutral".

The news headlines classification methods used are enlisted briefly.

A. Naive Bayes

We used two different naive Bayes models; one models the three classes, "positive," "neutral" and "negative" while the other models two classes, "positive" and "negative," and uses a threshold parameter to define the "neutral" class from this model. Both the two-class and three-class naive Bayes classifiers use a multinomial event model with Laplace smoothing.

Our successful use of the two-class naive Bayes classifier attempts to exploit the fact that the neutral class represents an intermediate class between positive and negative. In our nine-class model, the data set was trained on "positive" and "negative" and "neutral" examples. The prediction for the "neutral" class is based on a thresholding scheme; if the difference in posterior probabilities for the positive or negative classes is below a specific threshold for a given test sample, it will be classified as "neutral." The best thresholding method (absolute difference of log-probabilities) and the threshold value were selected usingcross-validation. We got 71% accuracy in naïve bayes algorithm.

B. Support Vector Machine(SVM):

SVM has been used a lot for news text classification. The unique fact regarding SVM is that it incorporates both positive and negative training sets, which is not preferred in other classification algorithms. SVM has helped researchers a lot for performing short text news classifications as compared to full text and have shown considerable results.

Initially, the SVM did not include regularization. As a result the classifier tended to over-fit the training set and we have 63% accuracy in this algorithm.

VI. PERFORMANCE METRICS

One of the difficulties of categorizing news stories as positive, negative or neutral. For this application, we are most concerned with our algorithm correctly differentiating between positive, negative and neutral stories.

We also did a survey in our institute for assessing the performance of our algorithm. Ten news were given to ten students and they were asked to rate the positivity for each news and accordingly average was calculated as sum of positivity score for each news by 10 students divided by number of students. It is mentioned as "Positivity by user" in figure 1.Then these news were rated by our model.It is denoted as "Positivity by program" in figure 1.

VII. RESULTS

In our milestone, diagnostics showed that our algorithm performance may be improved with more data (either more headline samples or more text per article).Figure 1 shows that the accuracy of the three-class naive Bayes levels out and no longer improves significantly for larger training sets. Therefore, while we observe that the results still show moderate bias, we do not believe that the current performance is limited by the size of our data set.

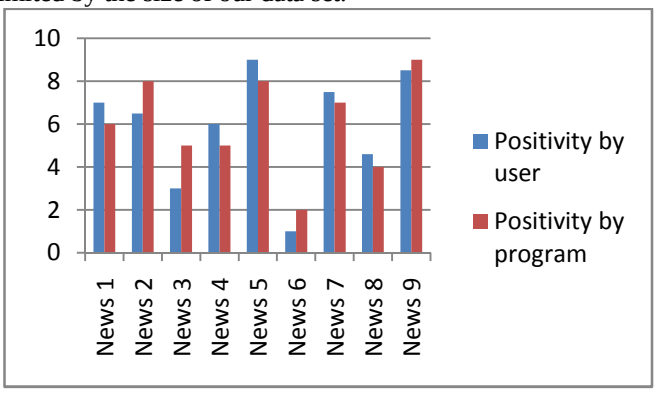


Figure 1: Analysis chart

VIII. CONCLUSION AND FUTURE WORK

Given the subjective nature of our classes, future work may involve creating more personalized recommendations of positive and negative stories based on the user's preferences.

We plan to make our data set publicly available for the machine learning community.

REFERENCES

[1] Xin Chen, Mihaela Vorvoreanu, and Krishna Madhavan,2013,Mining Social Media Data For Understanding Students Learning Experiences, IEEE Transactions on Learning Technologies, DOI :- 10.1109/TLT.2013.2296520,14.

[2] LU YE, RUI-FENG XU, JUN xu,2012,Emotion Prediction of News Articles from Reader's Perspective Based On Multi-label Classification, International Conference on Machine Learning and Cybernetics,15-17 July,6.

[3] Prashant Raina,2013,Sentiment Analysis in News Articles Using Sentic Computing, IEEE 13th International Conference on Data Mining Workshops,978-0-7695-5109-8/13,DOI 10.1109/ICDMW.2013.27,4.