Sentiment Analysis on Product Review

Mrinal B K ,1MS12CS060,
Rahul Sherikar,1MS12CS082,
Rajesh S M,1MS12CS087,
ManuKumar K J,1MS13CS413,
Department of Computer Science and Engineering,
M.S.R.I.T, Bangalore, India

Abstract-Sentiments are expressed in sentences that are put on the social website product reviews, these reviews dataset we used for our analysis. User opinions are important to analyse the product. People post their thoughts and feedbacks about the products that they are using. A simple sentiment analysis algorithm to classify a document as 'positive' or 'negative' based on the opinion expressed in it. New opportunities and challenges still arise in the field of natural language processing. Sentiment analysis system designed using Naive Bayes algorithm.That saves running time and reduces computational complexity.

There are three types of approaches for sentiment classification (i) Machine learning based text classifier -such as Naïve Bayes (ii) Unsupervised semantic orientation scheme of extracting relevant n-grams of the text (iii) SentiWordNet based publicly available library that provides positive, negative and neutral scores for words.

This analysis is done on product reviews from the Amazon product dataset. This helps the user to better decisions whilst purchasing product an appropriate product(s).

I. INTRODUCTION

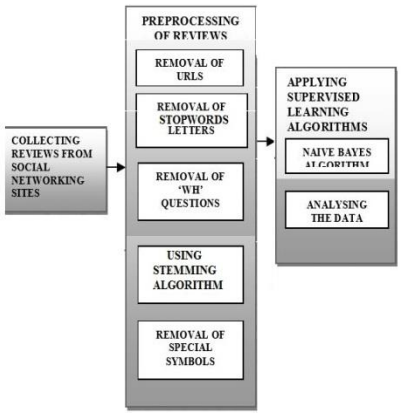
This project is a product review mining using machine learning and semantic orientation. Supervised classification and text classification techniques are used in the proposed machine learning approach to classify the product review. A corpus is formed to represent the data in the documents and all the classifiers are trained using this corpus.

Thus, the proposed technique is more efficient. Though, the machine learning approach uses supervised learning, the proposed semantic orientation approach uses "unsupervised learning" because it does not require prior training in order to mine the data.

Thus, the study concludes that the supervised machine learning is more efficient but requires a considerable amount of time to train the model. On the other hand, the semantic orientation approach is slightly less accurate but is more efficient to use in real time applications. The results confirm that it is practicable to automatically mine opinions from unstructured data.

The project used machine learning techniques to investigate the effectiveness of classification of documents by overall sentiment. Experiments demonstrated that the machine learning techniques are better than human produced baseline for sentiment analysis on product review data. The experimental setup consists of product-review corpus with randomly selected 1 lakh sentiment reviews. Learning methods Naive Bayes, was employed. The machine learning techniques are better than human baselines for sentiment classification. Whereas the accuracy achieved in sentiment classification is much lower when compared to topic based categorization. Sentiment Analyser to extract opinions about a subject from online data documents. Sentiment analyser uses natural language processing techniques. The Sentiment analyser finds out all the references on the subject and sentiment polarity of each reference is determined. The sentiment analysis conducted by the researchers utilized the sentiment lexicon and sentiment pattern database for extraction and association purposes. Online product review articles for digital camera and music were analysed using the system with good results.

II. **DESIGN**



A. Naïve Bayes Classifier

Naive Bayes is a simple technique for constructing classifiers: models that assign class labels to problem instances, represented as vectors of feature values, where the class labels are drawn from some finite set. It is not a single algorithm for training such classifiers, but a family of algorithms based on a common principle: all naive Bayes classifiers assume that the value of a particular feature is independent of the value of any other feature, given the class variable.

With a multinomial event model, samples (feature vectors) represent the frequencies with which certain events have been generated by a multinomial $(p_1, \ldots, p_n)$where p_i is the probability that event i occurs (or K such multinomial in the multiclass case). A feature vector $\mathbf{x} = (x_1, \ldots, x_n)$ is then a histogram, with x_i counting the number of times event i was observed in a particular instance. This is the event model typically used for document classification, with events representing the occurrence of a word in a single document (see bag of words assumption). The likelihood of observing a histogram x is given by

$$p(\mathbf{x}|C_k) = \frac{(\sum_i x_i)!}{\prod_i x_i!} \prod_i {p_{ki}}^{x_i}$$

The multinomial naive Bayes classifier becomes a linear classifier when expressed in log-space

$$\begin{aligned} \log p(C_k|\mathbf{x}) &\propto \log \left(p(C_k) \prod_{i=1}^{n} {p_{ki}}^{x_i} \right) \\ &= \log p(C_k) + \sum_{i=1}^{n} x_i \cdot \log p_{ki} \\ &= b + \mathbf{w}_k^\top \mathbf{x} \end{aligned}$$

Where $b = \log p(C_k)$ and $w_{ki} = \log p_{ki}$

If a given class and feature value never occurs together in the training data, then the frequency-based probability estimate will be zero. This is problematic because it will wipe out all information in the other probabilities when they are multiplied. Therefore, it is often desirable to incorporate a small-sample correction, called pseudo count, in all probability estimates such that no probability is ever set to be exactly zero. This way of regularizing naive Bayes is called Laplace smoothing when the pseudo count is one, and Lid stone smoothing in the general case.

Here is a worked example of naive Bayesian classification to the document classification problem. Consider the problem of classifying documents by their content, for example into spam and non-spam e-mails. Imagine that documents are drawn from a number of classes of documents which can be modelled as sets of words where the (independent) probability that the i-th word of a given document occurs in a document from class C can be written as $p(w_i|C)$

Then the probability that a given document D contains all of the words w_i, given a class C, is

$$p(D|C) = \prod_i p(w_i|C)$$

The question that we desire to answer is: "what is the probability that a given document D belongs to a given class C?" In other words, what is $p(C|D)$?

Now by definition

$$p(D|C) = \frac{p(D \cap C)}{p(C)} \text{ and } p(C|D) = \frac{p(D \cap C)}{p(D)}$$

Bayes' theorem manipulates these into a statement of probability in terms of likelihood.

$$p(C|D) = \frac{p(C)}{p(D)} p(D|C)$$

Assume for the moment that there are only two mutually exclusive classes, S and ¬S (e.g. positive and negative review), such that every element is in either one or the other;

$$p(D|S) = \prod_i p(w_i|S) \text{ and } p(D|\neg S) = \prod_i p(w_i|\neg S)$$

Using the Bayesian result above, we can write:

$$p(S|D) = \frac{p(S)}{p(D)} \prod_i p(w_i|S) \text{ and } p(\neg S|D) = \frac{p(\neg S)}{p(D)} \prod_i p(w_i|\neg S)$$

Dividing one by the other gives:

$$\frac{p(S|D)}{p(\neg S|D)} = \frac{p(S)}{p(\neg S)} \prod_i \frac{p(w_i|S)}{p(w_i|\neg S)}$$

Thus, the probability ratio p(S | D) / p(¬S | D) can be expressed in terms of a series of likelihood ratios. The actual probability p(S | D) can be easily computed from log (p(S | D) / p(¬S | D)) based on the observation that p(S | D) + p(¬S | D) = 1.

Taking the logarithm of all these ratios, we have:

$$\ln \frac{p(S|D)}{p(\neg S|D)} = \ln \frac{p(S)}{p(\neg S)} + \sum_i \ln \frac{p(w_i|S)}{p(w_i|\neg S)}$$

Finally, the document can be classified as follows. It is positive review if $p(S|D) > p(\neg S|D)$(i.e., $\ln \frac{p(S|D)}{p(\neg S|D)} > 0$), otherwise it is a negative review.

III. IMPLEMENTATION

There are 4 modules used in this project for the analysis

- Module for retrieving the datasets
- Module for preprocessing the data
- Sentiment analysis module
- Filtering or classification module

A. Implementation of the Modules.

The 4 modules can be described as follows:

- Retrieving the datasets: The datasets that has been fetched is in the raw format, we have to retrieve the datasets in the proper format. The attributes required for the analysis should extracted so that so they can be further used.
- Preprocessing of the data: The datasets or the reviews that are extracted have to be processing before being used for analysis. Preprocessing can done by
    1. Removing the stop words.
    2. Stemming of the reviews.
- Sentiment analysis module: The reviews are categorized and word count is done using the HPC tool like Hadoop and categorizing is done on the reviews and divided into positive or negative words.
- Classifying and filtering module: The classifications of the reviews are done using the Naïve Bayes classifier and the final output is retrieved.

B. Algorithm design

1) Porter's Algorithm (stemming)

Step 1 : Gets rid of plurals and -ed or -ing suffixes.
Step 2 : Turns terminal y to i when there is another vowel in the stem.
Step 3 : Maps double suffixes to single ones: -ization, -ational, etc.
Step 4 : Deals with suffixes, -full, -ness etc.
Step 5 : Takes off -ant, -ence, etc.
Step 6 : Removes a final –e.

2) Naive bayes algorithm (classification)

```
TRAINMULTINOMIALNB(ℂ, 𝔻)
1  V ← EXTRACTVOCABULARY(𝔻)
2  N ← COUNTDOCS(𝔻)
3  for each c ∈ ℂ
4  do N_c ← COUNTDOCSINCLASS(𝔻, c)
5     prior[c] ← N_c/N
6     text_c ← CONCATENATETEXTOFALLDOCSINCLASS(𝔻, c
7     for each t ∈ V
8     do T_ct ← COUNTTOKENSOFTERM(text_c, t)
9     for each t ∈ V
10    do condprob[t][c] ← (T_ct+1) / (Σ_t'(T_ct'+1))
11 return V, prior, condprob
```

IV. SCOPE AND FUTURE WORK

The scope of this project aims to satisfy the user community to make their own analysis of this classified data and also can be used for further research topic.

Creating more and new path of exploration in the field of entertainment in this case products. This expands the horizons of these current fields which were not possible before this.

V. CONCLUSION

Our reseach aims to determine the attitude of a speaker or a writer with respect to some topic or the overall contextual polarity of a document. The attitude may be his or her judgment or evaluation , affective state , or the intended emotional communication. So, that the user community can have these somewhat processed raw data for this dataset to be expanded to produce different products in this topic.

VI. REFERENCE

1. Chen Mosha,"Combining Dependency Parsing with Shallow Semantic Analysis ".
2. Yuanbin Wu, Qi Zhang, Xuanjing Huang, Lide Wu,"Phrase Dependency Parsing forOpinion Mining".
3. Shailendra Singh Raghuwanshi, PremNarayan Arya "Comparison of K-means and Modified K-mean algorithms for Large Data-set".