RESEARCH ARTICLE OPEN ACCESS

Sentimental Analysis of Movie Reviews in Kannada

Sai Darshan B.R, Shashank Bhatt.P, Karthik H.S, Sachin Kumar B.S
Department of CSE, DBIT
Merin Thomas
Asst. Prof, Department of CSE, Don Bosco Institute of Technology
Dr.Latha .C.A
Professor, Dept. of CSE, Global Academy of Technology
VTU

ABSTRACT

With the development of Internet technologies, there is an enormous amount of information that is getting accumulated in the World Wide Web, such as reviews, blogs, tweets, posts etc. In recent years sentimental analysis has gained momentum due to the increase in the size of information. Sentimental Analysis has been important in reviewing products, movies etc. Firstly, we use web scraping to scrape information about movie reviews from a website. Approach is to determine the sentiment polarity of Kannada movie reviews. This paper introduces a hybrid method including both Lexicon based and Machine Learning.

Keywords:– Sentimental Analysis, Machine Learning, Web Scraping.

I. INTRODUCTION

The Internet has revolutionized the computer and communications world like nothing before. The Initial concepts originated in several Computer Science laboratories in the United States, Great Britain, and France, but the internet today is a widespread information infrastructure. With the development of Internet technologies, there is a huge amount of information and raw data such as reviews, blogs, tweets, posts, and other such information on all kinds of websites. People have started to share information about entities such as products, movies etc., through different kinds of sites such as face book, Amazon etc. This information plays an important role in deciding whether an entity is good or bad. With the increase in the amount of such information, the analysis and categorization of these become extremely difficult. The large amount of the information also attracts researchers to make efforts to organize the information clearly. So the automatic text categorization is come up with. One of the ways in which the above problem can be tackled is through Sentiment Analysis.

II. SENTIMENT ANALYSIS

In recent years Sentiment Analysis has gained momentum by the increase of social networking sites. Sentiment Analysis is the process of determining whether a piece of writing is positive, negative or neutral. It's also known as opinion mining, deriving the opinion or attitude of a speaker. A common use-case for this technology is to discover how people feel about a particular topic. Sentiment Analysis can be of three levels -Document level (such as blog), Sentence level (such as comments) and Word level.

Sentiment classification looks for words or emotional states such as good, bad, nice, sweet, awful etc. in a piece of sentence in a review.

In this paper we present the implementation of Sentimental Analysis for Kannada Movie Reviews. We use a hybrid method of analysis where in we utilize both the lexicon based approach and the machine learning approach called Naive Bayes classifier.

III. INTRODUCTION TO WEB SCRAPING

Web scraping also known as Web Data Extraction, Web Harvesting is a technique employed to extract data from websites here the data is extracted and saved to a local file in your computer or to database. Data displayed by most websites can only be viewed using a web browser

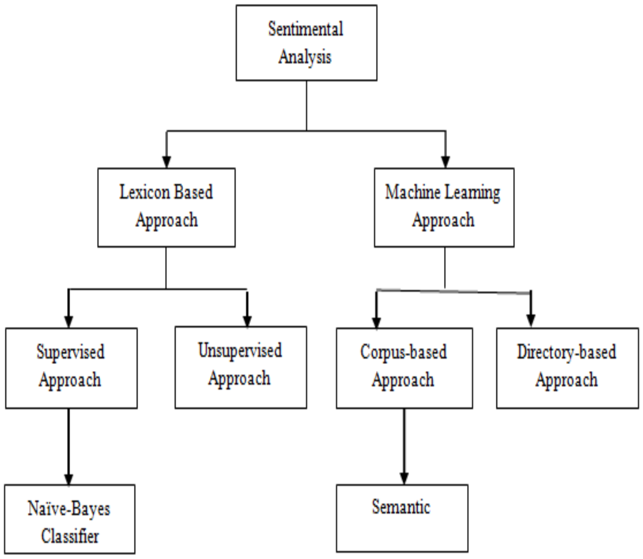


Fig.1: Sentimental Analysis Hierarchy

. Examples are data listings at yellow pages directories, real estate sites, social networks, online shopping sites, contact databases etc. Most websites do not offer the functionality to save a copy of the data which they display to your computer. The only option then is to manually copy and paste the data displayed by the website in your browser to a local file in your computer - a very tedious job which can take many hours or sometimes days to complete. Web Scraping is the technique of automating this process, so that instead of manually copying the data from websites, the Web Scraping software will perform the same task within a fraction of the time. Web scraping software will interact with websites in the same way as your web browser. But instead of displaying the data served by the website on screen, the Web Scraping software saves the required data from the web page to a local file or database. Below is a sample code to extract data from a test site using php :

```
<?php
$url = 'http://www.xyz.com';
$output = file_get_contents($url);
echo $output;
?>
```

The $output contains the source code of the website xyz.com. The php pregmatch() is performed on $output so as to extract only the required data from the source code.

The Data extracted is the Kannada movie review. This information is then translated to English using Google's translation API.

IV.ARCHITECTURE OF THE KANNADA MOVIE ANALYZER

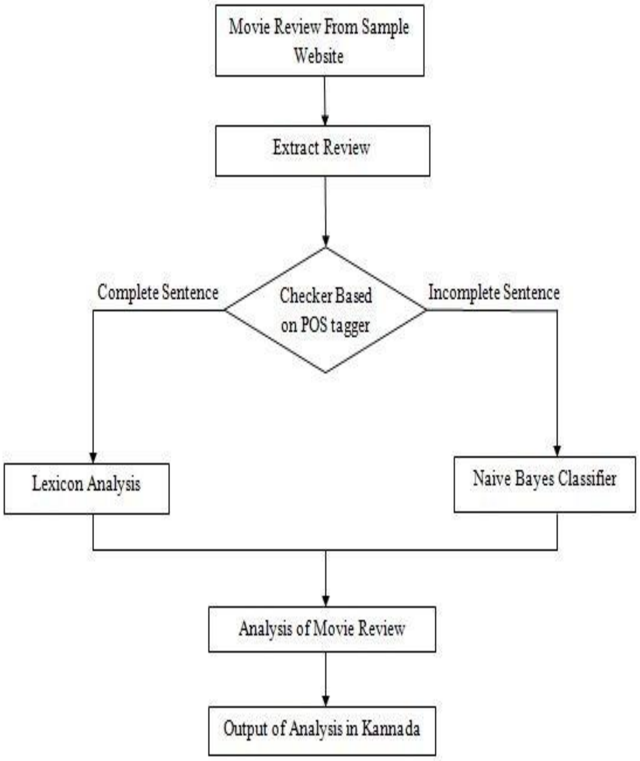


Fig.2: Architectural Diagram of Proposed System.

Updated at Wed,6th Apr, 2016 10:08AM IST

Your Default Edition | English Edition | ePaper

Udayavani.com

ಉದಯವಾಣಿ

Home

ಪ್ಲ್ಯಾನ್ ಕೇಳೋಕೆ ಚೆನ್ನಾಗಿದೆ!; ದಿ ಪ್ಲ್ಯಾನ್

Nov 28, 2015, 5:39 PM IST

V. LEXICON BASED APPROACH

Lexicon method finds co-occurrence seed sentiment words using semantic technique. This is done by deriving polarities using the co-occurrence of axioms as adjectives, adverbs in a corpus. It is also possible to use a document on the web as the corpus for the construction of seed sentiment words covering the maximum unavailable words, if the corpus with small list of seed words is used as training data set. The semantic approach gives sentiment values directly and gives similar sentiment values to the words that are close to the seed word semantically. Tool for computing the similarity between words is WordNet. It provides different kinds of semantic relationships between words that are used to calculate sentiment polarities. WordNet could be used too for obtaining a list of sentiment words by iteratively expanding the initial set with synonyms and anonyms and then determining the sentiment polarity for an unknown word by the relative count of positive and negative synonyms of this word.

Here, the following three conditions are handled – First, if blind negation is found, then the sentence is considered as negative review. Second, if positive sentiment (such as good) word is preceded by a negation then the polarity is reversed for negative sentiment word it remains unchanged. Third, if the only sentiment words then, their corresponding polarities are retrieved from training data set then, perform sum of polarities to get the final value for the review.

VI. NAÏVE BAYES CLASSIFIER (NB)

The Naïve Bayes classifier is the simplest and most commonly used classifier, it is also known as baseline algorithm. Naive Bayes classifier technique is based on the Bayesian theorem.

This method computes the posterior probability of a class, based on the distribution of the words in the sentence. This model ignores the position of the word in the sentence. The words of the sentence are collected together called Bag of words. Each words positive and negative polarity is calculated by they equation given below.

$$P(label|features) = \frac{P(label) * P(features|label)}{P(features)}$$

Using a set of sample reviews as training data, the words from the sample reviews collected together results the bag of words.

A posterior probability is the probability of the assigning the observations into groups of data after relevant evidence is taken into account. A prior probability is the probability that an observation will fall into a group before you collect the data. Prior probability would express the belief before collecting the evidence. P(label|feature)is the posterior probability of label. P(features|label)is the likelihood that a random feature. P(features) is the prior probability that a given feature set is occurred.

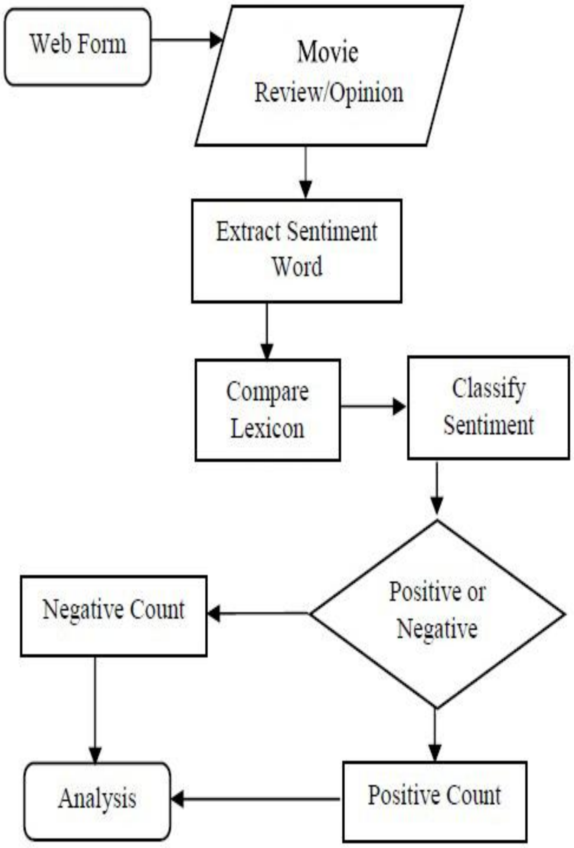


Fig.4: Flow Diagram of Lexicon Analysis.

Given the Naïve assumption which states that all features are independent, the equation could be rewritten as follows:

$$P(label|features) = \frac{P(label) * P(f1|label) * \ldots\ldots * P(fn|label)}{P(features)}$$

VII. REQUIREMENT

A set of sample reviews ranging from 100 to 200 reviews based on certain domain such as movie, product etc and its corresponding value. The values are either 0 or 1.The value one represents positive and zero as negative.

Algorithm: Naive Bayes Classifier

Input: Data M={m1,m2,m3……..}
Output: Positive, Negative
Step 1: Divide a message into words Mi={w1,w2,w3……}
Step 2: If wi not belongs to NTRetrieve +ve and –ve polarity.
Step 3: Calculate probability of +ve polarity of wi with total +ve polarity of M

$$P(Ck|x) = \frac{p(Ck)\,p(x|Ck)}{p(x)}$$

Step 4: Calculate overall polarity of Word= (+ve polarity) + ve polarity)
Step 5: Repeat step 2 until end of words
Step 6: Add the polarities of all words of a message
Step 7: Message can be positive or negative. Repeat step until M NULL

VIII. IMPLEMENTATION

The Implementation of lexicon based algorithm produces the following output, the words in database for the input and its scores are shown in the seed table (2) below. The steps of processing and tagging are shown in processing table (1).

Seed words here are few sentiment words from the database of the sentiword.net , along with their corresponding positive and negative score. Input is preprocessed and this processed data is input to lexicon-based algorithm known as sentiment calculation.

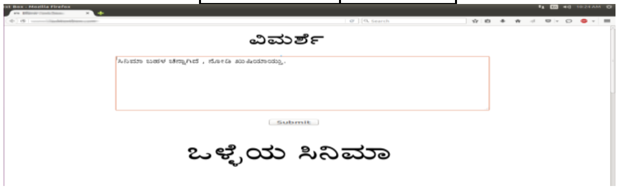
In the preprocessing step, stop word, punctuations and stemming words removed. Tagging of the sentence is done by POS tagger it transform the sentence into tokens by detecting axioms such as adjectives, nouns, adverb, which are the key essentials to find sentiment words in the text.

Table: 1 processing table

Sentence	ಸಿನಿಮಾ ಅದ್ಭುತವಾಗಿದೆ ನನಗೆ ಇಷ್ಟವಾಯಿತು.	ನಿರ್ದೇಶನ ಕೆಟ್ಟದಾಗಿದೆ, ಚಿತ್ರಕಥೆ ಅಷ್ಟೇನು ಗಮನಾರ್ಹವಲ್ಲ.
Stop words	ಸಿನಿಮಾ ಅದ್ಭುತವಾಗಿದೆ ಇಷ್ಟವಾಯಿತು.	ನಿರ್ದೇಶನ ಕೆಟ್ಟದಾಗಿದೆ, ಚಿತ್ರಕಥೆ ಗಮನಾರ್ಹವಲ್ಲ.
Score	3+2=5	-2+(-2)=-4
Type	ಸಕಾರಾತ್ಮಕ ವಿಮರ್ಶೆ	ನಕಾರಾತ್ಮಕ ವಿಮರ್ಶೆ

Table 2:Seed table

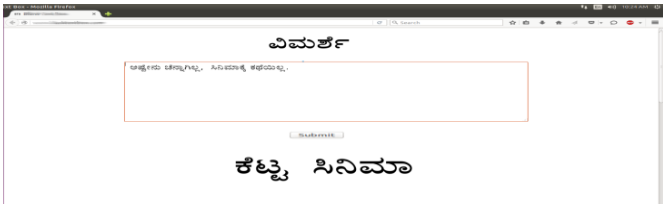
Word	Score
ಸಾಹಸ	1
ಅದ್ಭುತ	3
ಅಸಹ್ಯ	-3
ವಿಚಿತ್ರವಾಗಿ	-1
ಕೆಟ್ಟ	-2
ಆಕರ್ಷಕ	2
ಉಥಮ	4
ಇಷ್ಟ	2


ವಿಮರ್ಶೆ

ಸಿನಿಮಾ ಬಹಳ ಚೆನ್ನಾಗಿದೆ , ನೋಡಿ ಖುಷಿಯಾಯಿತು.

ಒಳ್ಳೆಯ ಸಿನಿಮಾ

Fig.5: Analysis of a Good Review.


ವಿಮರ್ಶೆ

ಅಷ್ಟೇನು ಚೆನ್ನಾಗಿಲ್ಲ, ಸಿನಿಮಾಕ್ಕೆ ಕಥೆಯಿಲ್ಲ.

ಕೆಟ್ಟ ಸಿನಿಮಾ

Fig.6: Analysis of a Bad Review.

In the Implementation of the Naïve Bayes classifier (NB) Words in the review form the bag of words. Each word is compared with the words of sample review collected in the database and retrieves its rating. A word such as movie is present in 4 of 12 sample positive reviews then likelihood P(x/c) of positive review for the word movie is 4/12 similarly likelihood of negative review is 2/8.

Class prior probability P(c) is the probability of total positive review to the total number of review in the database. Class prior probability for negative review is calculated in the similar way. Class prior probability P(c) is the probability of total positive review to the total number of review in the database. Class prior probability for negative review is calculated in the similar way. Predictor prior probability P(x) is probability of total review for a word such as movie to the total sample review in the database. The Posterior probability of each word is calculated, if the value is positive then it is positive review else negative.

Table 3: Sample class prior probability Table

Bag of Words	POS	Neg	P(X)
ನಿರ್ದೇಶನ	4/18	2/20	6/38
ಕೆಟ್ಟದಾಗಿದೆ	0/18	6/20	6/38
ಚೆಲನಚಿತ್ರ	4/18	2/20	6/38
ಚಿತ್ರಕಥೆ	3/18	2/20	5/38
ಅಲ್ಲ	0/18	4/20	4/38
ಸಾಲು	4/18	3/20	7/38
ಅದ್ಭುತ	4/18	0/20	4/20
ನೋಡುತಿದ್ದೇನೆ	4/18	4/20	8/20
ಇಷ್ಟ	3/18	5/20	8/20
ಆಕರ್ಷಕ	3/18	0/20	3/38
P(C)	29/38	28/38	

IX. RESULTS

Table 5: Precision and Recall of hybrid model

	Positive	Negative
Precision	0.9361	0.9245
Recall	0.9351	0.8215
Accuracy	94%	86%

X. CONCLUSION

Lexicon-based method is accurate than Naive Bayes classifier when sentence is processed completely with training set data and retrieve their respective scores. On the other hand Naïve Bayes classifier is inefficient than Lexicon-based method algorithm in accuracy but gives better results in cases where data is incomplete or uncertain and has a wide application. Since both the above methods have their respective disadvantages, our hybrid method of analysis combines the advantages of both the Lexicon based approach and the Naive Bayes classifier method and produces a more accurate analysis. This method of implementation can be further enhanced to movie reviews, not just in Kannada but in other languages also.It can also be extended to produce an analysis of a wide range of other products such as electronics, clothing and other such accessories.

REFERENCES

[1] B. Pang and L. Lee "Opinion mining and sentiment analysis" Foundations and Trends in Information Retrieval, 2(1-2):1 { 135,2008.}

[2] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up? Sentiment classification using machine learning techniques.[2002]

[3] I.Hemalatha1, Dr. G.P Saradhi Varma, Dr.A.Govardhan "Sentiment Analysis Tool Using Machine Learning Algorithms".

[4] Peter Turney. 2002. "Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews". In Proc of the ACL.

[5] Bas Heerschop, Frank Goossen, Alexander Hogenboom, "Polarity Analysis of Texts using Discourse Structure", Erasmus University.

[6] Comparative Study of Classification Algorithms used in Sentiment Analysis, Amit Gupte, Sourabh Joshi, Pratik Gadgul, Akshay Kadam.