International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 03 Issue: 04 | Apr-2016 www.irjet.net p-ISSN: 2395-0072

SENTIMENT ANALYSIS: CLASSIFICATION AND SEARCHING TECHNIQUES

Dr. G V Garje[1], Apoorva Inamdar[2], Apeksha Bhansali[2], Saif Ali Khan[2], Harsha Mahajan[2]

[1]HOD of Computer Engineering and IT, PVG's College of Engineering and Technology, Pune, Maharashtra, India
[2]Department of Computer Engineering and IT, PVG's College of Engineering and Technology Pune, Maharashtra, India

---

Abstract - *Sentiment analysis is now widely used to achieve greater business precision. It is implemented using multiple machine learning algorithms such as Naïve Bayes, SVM, Neural Network, Regression etc. Usually, sentiment analysis results into sentiments being classified in three polarities: Positive, Negative and Neutral. None the less, the use of the neutral label is very scarce. However, a human perspective is required in sentiment analysis, as automated systems are not able to analyze historical tendencies of the individual commenter, or the platform and are often classified inappropriately with respect to the expressed sentiment. In case of classification, often sentiment words are required, for which, multiple searching techniques are applied, but the most commonly used searching technique is the sequential search. This paper discusses some of the sentiment classification and searching approaches, and presents an alternative approach for human sentiment analysis.*

Key Words: **Sentiment analysis, classification, dictionary**

1. INTRODUCTION

Existing approaches to sentiment analysis can be grouped into three main categories: knowledge-based techniques, statistical methods, and hybrid approaches. Knowledge-based techniques classify text by categories based on the presence of unambiguous sentiment words such as happy, sad, afraid, etc. Statistical methods leverage on elements from machine learning such as latent semantic analysis, support vector machines, "bag of words" and many more. More sophisticated methods try to analyse the expressers' sentiment along with its effect. In this case, deep parsing is required in order to obtain grammatical dependency. Hybrid approaches apply both machine learning and elements from knowledge representation through the analysis of concepts that do not explicitly convey relevant information, but which are implicitly linked to other concepts that do so. Sentiment analysis generally progresses in three major steps- Identification, Classification and prediction. Out of this, identification part is divided into information retrieval, data cleaning, parsing, tokenization etc. Classification involves creation of labels/categories to classify sentiments and classification algorithm. The classification methods can be chosen from statistical, linear, probabilistic and kernel models. Each classification model has different features and different supporting algorithms. Prediction, though an optional part of sentiment analysis, is very essential. Many classification models also support prediction.

Classification is often based on sentiment polarities such as Positive– happy, good, better, nice, etc., Negative -sad, angry, frustrated, etc. and Neutral. Neutral polarity is created due to the need of a buffer for uncertain scenarios.

When it comes to searching techniques applied in sentiment analysis, most of the algorithms apply sequential search, while some algorithms apply hash technique for storage and search.

2. POPULAR SENTIMENT CLASSIFICATION METHODS

Table -1: An overview of classification methods

Approach	Definition	Features
Bag of Words	The bag-of-words model is a model in which a text is represented as the bag of its words, disregarding grammar and even word order but keeping multiplicity.	Simple. Easy to implement.

Approach	Definition	Features
Naive Bayes	Naive Bayes Classifiers are a family of simple probabilistic classifiers based on applying Bayes' theorem with strong independence assumptions between the features.	Applies Bayes theorem. Based on probabilistic model.
Logistic Regression	Logistic regression measures the relationship between the categorical dependent variable and one or more independent variables by estimating probabilities using a logistic function, which is the cumulative logistic distribution.	Performs regression analysis. Largely used for prediction along with classification.
Support Vector Machine	A Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyper plane.	It is non probabilistic binary linear classifier, but can also perform non-linear classification
Decision tree	A decision tree is a decision support tool that uses a tree-like graph or model of decisions and their possible outcomes.	Decision tree can be linearized into decision rules.
Neural Network	A neural network usually involves a large number of processors operating in parallel, each with its own small sphere of knowledge and access to data in its local memory.	It is used for hard tasks such as speech recognition, computer vision etc.
Clustering	Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar to each other.	It is an iterative process of similarity recognition.

3. NOTABLE SHORTCOMINGS OF POPULAR CLASSIFICATION METHODS

3.1 No. of classes

Usually, in sentiment analysis, the classification is binary. Most of the times, classification revolves around only two classes- positive sentiment and negative sentiment. Methods such as clustering and neural network offer more than two classes for classification, but it faces issues of ambiguity. Binary classification has highest precision rate than n-ary classification in any method.

3.2 Unreliable dictionary

Many sentiment classification methods use automatically generated dictionary of good and bad words. Some classification algorithms have algorithms to generate automatic dictionary from given dataset. The guiding rules for this dictionary are ambiguous. Most of the sentiment dictionary algorithms are open source, thereby allowing developers to add their own constraints. This leads to high degree of versatility and ambiguity. There is no formal, globally approved standard sentiment dictionary. Also, many such dictionaries are data-specific, i.e. they are specially created for a specific dataset. This makes sentiment dictionaries somewhat unreliable.

3.3 Use of probabilistic classification approach

The probabilistic classification approach is used in some classification methods, where classification of next item is dependent upon the probability of previously classified items. Class containing most items is favored and most likely; the next item is classified into that class. When it comes to sentiment classification, probabilistic approach may be convenient, but it is not entirely preferable. With respect to sentiment classification, probabilistic model may cause many errors, leading to incorrect classification of some words. This decreases accuracy of classification.

3.4 Need to understand difference between human emotions

Sentiment analysis finds most of its application in product review, trend analysis, advertisement survey etc. But, recently, sentiment analysis is been applied to analyze emotions of people. This application involves study of human emotions. Human sentiments are much more complicated and multi-leveled. They are mostly situational, in context of a subject and people's reaction, and fickle. Sentiment analysis in this case needs to be detailed and highly unambiguous. The analysis that a "person is happy" is incomplete. It requires details such as - the extent of person's happiness, cause of happiness, likely situations to change this sentiment along

with its impact, and so on. Such sentiment analysis is the need of time where many people are facing psychological issues because of social media and are going to the extent of suicides and committing serious crimes through this social media platform. The spectrum of human emotions needs to be considered with more detail in sentiment analysis.

4. AN ALTERNATIVE FOR CLASSIFICTION: BINARY SEARCH TREE

Binary search tree is one of the basic data structure. It applies fundamental concepts of binary search which is one of the simplest and most efficient searching algorithms. This classification implementation suggests on manually creating a sentiment dictionary, customized if necessary as per the required number of sentiment classes. Use of N-grams is highly recommended in this classification for better results. For N-Grams, a two-word prefix window is advised. For each class, create a BST of sentiment words. Then scan any test dataset. This will require only single scan of dataset, and with use of binary search technique, the searching time is reduced to half. This method provides better and faster classification. It also provides a way to convert unstructured sentiment data into structured data.

Binary search is efficient for larger data. In this we check the middle element. If the value is bigger than what we are looking for, then look in the first half; otherwise, look in the second half. Repeat this until the desired item is found. The input data must be sorted for binary search. It eliminates half the data to be searched per iteration. It is logarithmic. If we have 1000 elements to search, binary search takes about 10 steps, linear search1000 steps. Time complexity of Linear Search is O (N), whereas of Binary Search is O (log2 N).

For our test implementation, we used Twitter dataset for stress detection. We created 6 sentiment classes (Stress levels) – No stress, Very weak stress, Weak Stress, Moderate stress, Severe Stress, Very Severe stress. The dictionary was made for all these classes and then BST was applied. It gave us classification accuracy of 87%.

3. CONCLUSIONS

The suggested alternative implementation of BST is easy to implement, but it requires a bit of manual work. It reduces search time by half, providing faster classification. Also, by fitting a large unstructured data into a basic data structures, many operations can be performed on the data easily.

REFERENCES

[1] Yiping Li, Jing Huang, Hao Wang, Ling Feng 'Predicting Teens' Future Stress Level from Microblog', 2015

[2] Dr. G.V.Garje, Apoorva Inamdar, Harsha Mahajan, Apeksha Bhansali, Saif Ali Khan 'STRESS DETECTION AND SENTIMENT PREDICTION: A SURVEY',Jan 2016

[3] Yuanyuan Xue, Qi Li , Li Jin, Ling Feng, David A. Clifton,Gari D. Clifford ,"Detecting Adolescent Psychological Pressures from Micro-Blog", 2013.K. Elissa, "Title of paper if known," unpublished.

[4] Aditya Mogadala, "Twitter User Behavior Understanding with Mood Transition Prediction" ,2015