Survey of sentimental classification algorithms and methods

1. Introduction

In this tech review, mainstream algorithms and various sentimental analysis algorithms and methods are presented briefly. There are many algorithms and applications of sentimental analysis were proposed every year. But most of them can be classified into several mainstream categories. This tech review aims to give a closer look at these algorithms and to summarize some sentimental analysis techniques. The tech view will mainly focus on the basic algorithms that could be applied in the final project, in order to get a better immersion of text analysis and sentimental classification.

2. Background review

Text sentimental analysis refers to the process of using natural language processing and text mining technology to analyze, process and extract subjective texts with some sentiment. At present, the research of text sentiment analysis includes natural language processing, text mining, information retrieval, information extraction and machine learning. According to the granularity of the analysis, sentimental analysis tasks can be divided into text level, sentence level, word or phrase level; according to the category of text processing, it can be divided into sentiment analysis based on product reviews and sentiment analysis based on news reviews.

Sentiment classification is to identify whether the subjective text is positive or negative for a given text, which is the most studied in the field of sentimental analysis. Generally, network text contains both subjective and objective texts. The objective text is the objective description of things, without sentiment/emotional tendency. The subjective text is the author's views or ideas on various things, with the author's likes and dislikes and other sentiment/emotional tendencies. The object of sentiment classification is the subjective text with emotional tendency.

3. Approaches in Sentiment Classification

According to Medhat et al.(2014)[1], Sentiment Classification techniques can be divided into three approaches. They are machine learning approach, lexicon based approach and hybrid approach.

3.1 machine learning approach

The machine learning approach contains two main parts: supervised and unsupervised learning methods. In most cases, supervised learning methods are more popular than unsupervised learning methods.

3.1.1 Supervised learning

For the supervised learning methods, labeled training documents will be one important thing for supervised classifiers.

3.1.1.1 Probabilistic classifiers

The probability classifier is based on probability or statistical model. The common probability classifiers are Naïve Bayes Classifier, Bayesian network and

Maximum Entropy Classifier.

Naïve Bayes Classifier is a method based on Bayes theorem and assuming that the feature conditions are independent. First, through the given training set and assuming the independence of feature words, the joint probability distribution from input to output is learned, and then based on the learned model, input x is used to obtain the output y with the maximum posterior probability.

Bayesian network, also known as belief network, or directed acyclic graphical model, is a probability graph model, which is composed of representative variable nodes and directed edges connecting these nodes. The nodes represent random variables, and the directed edges between nodes represent the relationship between nodes (from the parent node to its child node). The relationship strength is expressed by conditional probability, and the information is expressed by prior probability if there is no parent node. Bayesian network is considered to be a complete model for the variables and their relationships, which makes its computation complexity in text mining is very expensive.

Maximum Entropy Classifier, or conditional exponential classifier, is a probability classifier which belongs to the exponential model class. Unlike the Naive Bayes classifier discussed in the previous article, maximum entropy does not assume that these features are conditionally independent of each other. MaxEnt is based on the principle of maximum entropy and selects the model with maximum entropy from all models suitable for our training data. Maximum entropy classifier can be used to solve a large number of text classification problems, such as language detection, topic classification, sentiment analysis and so on.

3.1.2 Linear classifiers

In the field of machine learning, the goal of classification is to aggregate objects with similar features. A linear classifier clarifies data into labels through a linear combination of features. The feature of an object is usually described as an eigenvalue, while in a vector it is described as a feature vector.

Support Vector Machines Classifiers(SVM). Given a set of training instances, the simplest SVM algorithm will mark each training instance belongs to one or the other of the two categories. SVM algorithm creates a non-probabilistic binary linear classifier. In SVM model, instances are represented as points in the space, so that the instances of individual categories are separated by as wide and obvious intervals as possible. Text data are ideally suited for SVM classification because text data is sparse.

Neural Network consists of many neurons and the neuron is its basic unit. The connections between the neurons is synapses(or just call it edges). Neurons and edges typically have a weight that adjusts as learning proceeds. The weight increases or decreases the strength of the signal at a connection.

Decision tree classifier provides a hierarchical decomposition of the training data space in which a condition on the attribute value is used to divide the data[2]. Decision tree is a kind of tree structure machine learning algorithm. All samples start from the root node, and each parent node with child nodes has a judgment. According to the judgment results, the samples are distributed to the child nodes. The test samples flow down from the root node, and finally reach a leaf node without child nodes. This node is the category of the sample.

3.2 Lexicon-based approach

There are many methods using dictionaries of words annotated with their semantic orientation to classify text data.

3.2.1Dictionary-based approach

In the simplest case, sentiments can be viewed as binary: positive or negative. We can extend emotions to multiple dimensions, such as fear, sadness, anger, joy and so on, by extending dimensions or introducing emotion dictionaries. Based on the manually established adjective vocabulary, Hu and Liu[5] used the synonymy of words in WorldNet to judge the sentiment tendency of words, and to judge the sentiment polarity of views.

3.2.2 Semantic approach

Semantic approach is very similar to Dictionary-based approach. Both of them have a dictionary or vocabulary table to store words and their sentiments. Semantic approach always combined with some other methods such as mixed with some statistical methods. Zhang and Xu[6] used two different ways to find the weakness of the products from the customers' feedback. They not only utilized sentence based sentiment analysis method to determine the polarity of each aspect in sentences, but also identified the implicit features of products by using statistics-based selection method.

3.3 Hybrid approach

Except Machine learning approach and Lexicon-based approach, there are many other approaches that cannot easily be categorized. Formal Concept Analysis (FCA) is one of those techniques. FCA is a principled way of deriving a concept hierarchy or formal ontology from a collection of objects and their properties. Each concept in the hierarchy represents the objects sharing some set of properties; and each sub-concept in the hierarchy represents a subset of the objects (as well as a superset of the properties) in the concepts above it[3]. It was proposed by Wille[4] in 1981.

4. Conclusion

In addition to the algorithms introduced in this paper, there are many algorithms that can be used for sentiment analysis, such as unsupervised learning approaches, and some other methods based on FCA method. Similarly, the algorithms and applications introduced in this paper can also be used in other fields of text analysis and natural

language processing, such as sentiment extraction. Our final project for CS 410 is doing Kaggle competition "Toxic Comment Classification Challenge", which is basically trying to build a model to achieve sentiment classification. Through this survey of sentiment classification algorithms and methods, I hope I could get a better immersion of text analysis and sentimental classification for the project.

[1]Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams engineering journal*, *5*(4), 1093-1113.

[2]Quinlan JR. Induction of decision trees. Machine Learn 1986;1:81–106.

[3]Formal concept analysis. In Wikipedia. Retrieved November 4, 2020, from https://en.wikipedia.org/wiki/Formal_concept_analysis

[4] Wille R. Restructuring lattice theory: an approach based on hierarchies of concepts. In: I. Rival, Reidel, Dordrecht-Boston; 1982, p. 445–70.

[5] Hu, M., & Liu, B. (2004, August). Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 168-177).

[6] Zhang, W., Xu, H., & Wan, W. (2012). Weakness Finder: Find product weakness from Chinese reviews by using aspects based sentiment analysis. *Expert Systems with Applications*, 39(11), 10283-10291.