ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 5, May 2017

Sentiment Analysis using Machine Learning Algorithms: A Survey

Kaushik Hande[1], Prof. A. G. Phakatkar[2]

M.E Student, Dept. of Computer Engineering, Pune Institute of Computer Technology, Pune, Maharashtra, India[1]

Assistant Professor, Dept. of Computer Engineering, Pune Institute of Computer Technology, Pune, Maharashtra, India[2]

ABSTRACT: With the advent of Web 2.0, people became more eager to express and share their opinions on web regarding day-to-day activities and global issues as well. Evolution of social media has also contributed immensely to these activities, thereby providing us a transparent platform to share views across the world. Sentiment analysis (SA) is a computational study of opinions, sentiments, emotions, and attitude expressed in texts towards an entity. Sentiment Analysis identifies the polarity of extracted public opinions. This survey paper tackles a overview in this field and presents a survey which covers Opining Mining, Sentiment Analysis, techniques, tools and classification.

KEYWORDS: Sentiment Analysis, Opinion Mining, Machine learning

I. INTRODUCTION

Sentiment is an attitude, thought, or judgment prompted by feeling. Sentiment analysis is also known as opinion mining, it involves studying of peoples sentiments towards certain entities. Internet is a resourceful place with respect to sentiment information. From a perspective of a user, people are able to express their views through various social media, such as forums, micro-blogs, or on-line social networking sites.

With the advent of Web 2.0 techniques, users started preferring to share their opinions on the Web. These user-generated and sentiment-rich data are valuable to many applications like credibility analysis of news sites on the Web, recommendation system, business and government intelligence etc. At the same time, it brings urgent need for detecting overall sentiment inclinations of documents generated by users, which can be treated as a classification problem. Sentiment analysis includes several subtasks which have seen a great deal of attention in recent years:

1) To detect whether a given document is subjective or objective.
2) To identify whether given subjective document express a positive opinion or a negative opinion.
3) To determine the sentiment strength of a document, such as strongly negative, weakly negative, neutral, weakly positive and strongly positive.

Besides individuals on social media marketers also need to monitor all media for information related to their brands whether it is for public relations activities, fraud violations, or competitive intelligence. Thus, aside from individuals, sentiment analysis is also the need of companies which are anxious to understand how their products and services are perceived by the public.

The rest of the paper is organized as follows. Section 2 presents the work of different researches on detecting sentiment from text. Section 3 describes feature selection in sentiment classification. Section 4 describes the different machine learning algorithms used for sentiment analysis. Section 5 presents tools of sentiment analysis classification. Finally, in section 6 we conclude the paper.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 5, May 2017

II. RELATED WORK

Pang and Lee pioneered in applying machine learning viz. NB, Maximum Entropy (ME), and SVM for binary sentiment classification of movie reviews. Pang et.al. considered the aspect of sentiment classification based on categorization study, with positive and negative sentiments . They have studied sentiment analysis with three different machine learning algorithms, such as, Naive Bayes, Support Vector Machine, and Maximum Entropy. The classification process is undertaken using the n-gram technique like unigram, bigram, and combination of both unigram and bigram. The bag-of-word features framework is used to implement the machine learning algorithms. Naive Bayes algorithm shows poor result among the three algorithms and SVM algorithm yields the best results of the three machine learning algorithms [1].

Peter Turney presents a simple unsupervised learning algorithm for classifying reviews as recommended (thumbs up) or not recommended (thumbs down).The first step includes a part-of-speech tagger to identify phrases in the input text that contain adjectives or adverbs. The second step includes estimate the semantic orientation of each extracted phrase. The classification of a review is predicted by the average semantic orientation of the phrases in the review that contain adjectives or adverbs [3].

In Xia et. al. a review text is represented by a pair of bags-of-words with opposite views (i.e., the original and antonymous views). By making use of two views in pairs, a dual-view co-training algorithm is proposed for semi-supervised sentiment classification. The dual-view representation is in a good accordance with the two co-training requirements [6].

Xia et. al. proposed intensive study of the effectiveness of ensemble techniques for sentiment classification tasks. Rather than an ensemble of different data re-sampling methods (e.g. bagging and boosting), focus was on ensemble of feature sets and classification algorithms. Two schemes of feature sets that are particular to sentiment analysis are part-of-speech (POS) based and the word-relation (WR) based. For each scheme, different machine learning algorithm such as Naive Bayes, Maximum Entropy, and Support Vector Machine as the base classifiers were used to predict classification scores. Ensemble stage consists of three types of ensemble method (fixed combination, weighted combination, and meta-classifier combination) with three ensemble strategies (ensemble of feature sets, ensemble of classification algorithms, and ensemble of both feature sets and classification algorithms). However, the problem of polarity shift still persists and increase computation requirements were the drawbacks of this approach [5].

III. FEATURE SELECTION IN SENTIMENT CLASSIFICATION

Sentiment Analysis task is considered a sentiment classification problem. The first step in the SC problem is to extract and select text features. Some of the current features are:

1) **Terms presence and frequency:** These features are individual words or word n-grams and their frequency counts. It either gives the words binary weighting (zero if the word appears or one if otherwise) or uses term frequency weights to indicate the relative importance of features.

2) **Parts of speech (POS):** Finding adjectives, as they are important indicators of opinions.

3) **Opinion words and phrases:** These are words commonly used to express opinions including good or bad, like or hate. On the other hand, some phrases express opinions without using opinion words. For example: cost me an arm and a leg.

4) **Negations:** the appearance of negative words may change the opinion orientation like not good is equivalent to bad [4].

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 5, May 2017

IV. MACHINE LEARNING ALGORITHMS USED FOR SENTIMENT ANALYSIS

Naive Bayes: In machine learning, naive Bayes classifiers are simple probabilistic classifiers based on applying Bayes' theorem with independence assumptions. Because they consider independence between features, they are called as naive. Naive Bayes is a simple technique for constructing classifiers: It is not a single algorithm for training such classifiers, but a family of algorithms using 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. For example, a car may be considered to be anMaruti if it is red, round and small. A naive Bayes classifier considers each of these features to contribute independently to the probability.

SVM: A Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyper plane. In other words, given labelled training data (supervised learning), the algorithm outputs an optimal hyper plane which categorizes new examples. Text data are ideally suited for SVM classification because of the sparse nature of text, in which few features are irrelevant, but they tend to be correlated with one another and generally organized into linearly separable categories. SVM can construct a nonlinear decision surface in the original feature space by mapping the data instances non-linearly to an inner product space where the classes can be separated linearly with a hyper plane.

Random Forest: Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of over fitting to their training set.

V. TOOLS OF SENTIMENT ANALYSIS CLASSIFICATION

There are so many open-source text-analytics tools used for natural language processing for sentiment classification. The following are tools used for Sentiment Classification:

NLTK: NLTK is a leading platform for building Python programs to work with human language data. It provides easy-to-use interfaces to over 50 corpora and lexical resources such as WordNet, along with a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning, wrappers for industrial-strength NLP libraries, and an active discussion forum. NLTK has been called "a wonderful tool for teaching, and working in, computational linguistics using Python," and "an amazing library to play with natural language."

OpenNLP: OpenNLP supports the most common NLP tasks, such as tokenization, sentence segmentation, part-of-speech tagging, named entity extraction, chunking, parsing, and co reference resolution.

Scikit Learn: Scikit-learn (formerly scikits.learn) is a free software machine learning library for the Python programming language. It features various classification, regression and clustering algorithms including support vector machines, random forests, gradient boosting, *k*-means and DBSCAN, and is designed to interoperate with the Python numerical and scientific libraries NumPy and SciPy.

RTextTools: RTextTools is a machine learning package for automatic text classification that makes it simple for novice users to get started with machine learning, while allowing experienced users to easily experiment with different settings and algorithm combinations. The package includes nine algorithms for ensemble classification (svm, slda, boosting, bagging, random forests, glmnet, decision trees, neural networks and maximum entropy), comprehensive analytics, and thorough documentation.

VI. CONCLUSION

This survey paper presented an overview on sentiment analysis using machine learning algorithms. Naïve Bayes and Support Vector Machines are the most frequently used machine learning algorithms for solving sentiment classification

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 5, May 2017

problem. It is important to consider the context of the text and the user preferences. That is why we need to make more research on context-based sentiment analysis. Using Natural Language Processing tools to reinforce the sentiment analysis process has attracted researchers recently and still needs some enhancements. Overall, sentiment analysis has found various promising applications like market prediction, political sentiment determination, equity value prediction, box office prediction etc. But, a lot work still remains to be done.

REFERENCES

1. Pang, Bo, Lillian Lee, and ShivakumarVaithyanathan. "Thumbs up?: sentiment classification using machine learning techniques." *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*. Association for Computational Linguistics, 2002.
2. Pang, Bo, and Lillian Lee. "Opinion mining and sentiment analysis." *Foundations and Trends® in Information Retrieval* 2.1–2, pp.1-135, 2008.
3. Turney, Peter D. "Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews." *Proceedings of the 40th annual meeting on association for computational linguistics*. Association for Computational Linguistics, 2002.
4. Das, Sanjiv R., and Mike Y. Chen. "Yahoo! for Amazon: Sentiment extraction from small talk on the web." *Management science*, Vol.53, Issue no.9, pp.1375-1388, 2007.
5. Xia, Rui, ChengqingZong, and Shoushan Li. "Ensemble of feature sets and classification algorithms for sentiment classification." *Information Sciences*,Vol.181, Issue no.6, pp.1138-1152, 2011.
6. Xia, Rui, FengXu, ChengqingZong, QianmuLi, Yong Qi, and Tao Li. "Dual sentiment analysis: Considering two sides of one review." *IEEE transactions on knowledge and data engineering*, Vol.27, Issue no. 8, pp. 2120-2133, 2015.
7. Medhat, Walaa, Ahmed Hassan, and HodaKorashy. "Sentiment analysis algorithms and applications: A survey." *Ain Shams Engineering Journal*, Vol.5, Issue no.4, pp.1093-1113, 2014.
8. Pradhan, Vidisha M., Jay Vala, and PremBalani. "A Survey on Sentiment Analysis Algorithms for Opinion Mining." *International Journal of Computer Applications*,Vol.133, Issue no. 9, 2016.
9. SuadAlhojely , "Sentiment Analysis and Opinion Mining: A Survey" in International Journal of Computer Applications (0975 – 8887) Volume 150 – No.6, September 2016.

BIOGRAPHY

Kaushik Hande is a student pursuing M.E. in the Computer Engineering Department, Pune Institute of Computer Technology, Pune. His research interests are Data Mining, Data Analysis and Machine Learning.

Prof. A. G. Phakatkar is a Assistant Professor in the Computer Engineering Department, Pune Institute of Computer Technology, Pune. Her research interests are Data Mining and Data Warehouse, Information retrieval and image processing.