Expert Systems With Applications 57 (2016) 117–126

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Classification of sentiment reviews using n-gram machine learning approach

Abinash Tripathy[*], Ankit Agrawal, Santanu Kumar Rath

Department of Computer Science and Engineering, National Institute of Technology Rourkela, India

ARTICLE INFO

Article history:
Received 3 March 2015
Revised 15 March 2016
Accepted 16 March 2016
Available online 24 March 2016

Keywords:
Sentiment analysis
Naive Bayes (NB)
Maximum Entropy (ME)
Stochastic Gradient Descent (SGD)
Support Vector Machine (SVM)
N-gram
IMDb dataset

ABSTRACT

With the ever increasing social networking and online marketing sites, the reviews and blogs obtained from those, act as an important source for further analysis and improved decision making. These reviews are mostly unstructured by nature and thus, need processing like classification or clustering to provide a meaningful information for future uses. These reviews and blogs may be classified into different polarity groups such as positive, negative, and neutral in order to extract information from the input dataset. Supervised machine learning methods help to classify these reviews. In this paper, four different machine learning algorithms such as Naive Bayes (NB), Maximum Entropy (ME), Stochastic Gradient Descent (SGD), and Support Vector Machine (SVM) have been considered for classification of human sentiments. The accuracy of different methods are critically examined in order to access their performance on the basis of parameters such as precision, recall, f-measure, and accuracy.

1. Introduction

Sentiment analysis, also known as opinion mining, analyzes people's opinion as well as emotions towards entities such as products, organizations, and their associated attributes. In the present day scenario, social media play a pertinent role in providing information about any product from different reviews, blogs, and comments. In order to derive meaningful information from people's sentiments, different machine learning techniques are applied by scholars and practitioners Liu (2012).

Sentiment analysis is observed to be carried out in three different levels such as document level, sentence level, and aspect level Feldman (2013). *Document level* classifies whether the document's opinion is positive, negative or neutral. *Sentence level* determines whether the sentence expresses any negative, positive or neutral opinion. *Aspect level* focuses on all expressions of sentiments present within given document and the aspect to which it refers. In this study, document level sentiment analysis has been taken into consideration.

There are mainly two types of machine learning techniques, which are very often used in sentiment analysis, i.e., the technique based on supervised and unsupervised learning. In supervised learning technique, the dataset is labeled and thus, trained to obtain a reasonable output which help in proper decision making Gautam and Yadav (2014). Unlike supervised learning, unsupervised learning process do not need any label data; hence they can not be processed at ease. In order to solve the problem of processing of unlabeled data, clustering algorithms are used Hastie, Tibshirani, and Friedman (2009). This study presents the impact of supervised learning method on labeled data.

The movie reviews are mostly in the text format and unstructured in nature. Thus, the stop words and other unwanted information are removed from the reviews for further analysis. These reviews goes through a process of vectorization in which, the text data are converted into matrix of numbers. These matrices are then given input to different machine learning techniques for classification of the reviews. Different parameters are then used to evaluate the performance of the machine learning algorithms.

The main contribution of the paper can be stated as follows:

i. Different machine learning algorithms are proposed for the classification of movie reviews of IMDb dataset IMDb (2011) using n-gram techniques viz., Unigram, Bigram, Trigram, combination of unigram and bigram, bigram and trigram, and unigram and bigram and trigram.
ii. Four different machine learning techniques such as Naive Bayes (NB), Maximum Entropy (ME), Support Vector Machine (SVM), and Stochastic Gradient Descent (SGD) are used for classification purpose using the n-gram approach.
iii. The performance of the machine leaning techniques are evaluated using parameters like precision, recall, f-measure, and

[*] Corresponding author. Tel.: +91 9437124235
E-mail addresses: abi.tripathy@gmail.com (A. Tripathy), agrawala96@gmail.com (A. Agrawal), skrath@nitrkl.ac.in (S.K. Rath).

http://dx.doi.org/10.1016/j.eswa.2016.03.028
0957-4174/© 2016 Elsevier Ltd. All rights reserved.

accuracy. The results obtained in this paper indicate, the higher values of accuracy when compared with studies made by other authors.

The structure of the paper is defined as follows: Section 2 presents literature survey. Section 3, indicates the methodology about the classification algorithm and its details. In Section 4, the proposed approach is explained. Section 5, indicates the implementation of the proposed approach. In Section 6, performance evaluation of the proposed approach is carried out. The last section i.e., Section 7 concludes the paper and presents the scope for future work.

2. Literature survey

The literature on sentiment analysis indicates that a good amount of study has been carried out by various authors based on document level sentiment classification.

2.1. Document level sentiment classification

Pang *et.al.*, have considered the aspect of sentiment classification based on categorization study, with positive and negative sentiments Pang, Lee, and Vaithyanathan (2002). They have undertaken the experiment with three different machine learning algorithms, such as, NB, SVM, and ME. The classification process is undertaken using the n-gram technique like unigram, bigram, and combination of both unigram and bigram. They have used bag-of-word features framework to implement the machine learning algorithms. As per their analysis, NB algorithm shows poor result among the three algorithms and SVM algorithm yields the result in a more convincing manner.

Salvetti *et.al.*, have discussed on Overall Opinion Polarity (OvOp) concept using machine learning algorithms such as NB and Markov model for classification Salvetti, Lewis, and Reichenbach (2004). In this paper, the hypernym provided by wordnet and Part Of Speech (POS) tag acts as lexical filter for classification. Their experiment shows that the result obtained by wordnet filter is less accurate in comparison with that of POS filter. In the field of OvOp, accuracy is given more importance in comparison with that of recall. In their paper, the authors presented a system where they rank reviews based on function of probability. According to them, their approach shows better result in case of web data.

Beineke *et.al.*, have used NB model for sentiment classification. They have extracted pair of derived features which are linearly combinable to predict the sentiment Beineke, Hastie, and Vaithyanathan (2004). In order to improve the accuracy result, they have added additional derived features to the model and used labeled data to estimate relative influence. They have followed the approach of Turney which effectively generates a new corpus of label document from the existing document Turney (2002). This idea allows the system to act as a probability model which is linear in logistics scale. The authors have chosen five positive and negative words as anchor words which produce 25 possible pairs and they used them for the coefficient estimation.

Mullen and Collier have applied SVM algorithm for sentiment analysis where values are assigned to few selected words and then combined to form a model for classification Mullen and Collier (2004). Along with this, different classes of features having closeness to the topic are assigned with the favorable values which help in classification. The authors have presented a comparison of their proposed approach with data, having topic annotation and hand annotation. The proposed approach has shown better result in comparison with that of topic annotation where as the results need further improvement, while comparing with hand annotated data.

Dave *et. al.* have used a tool for synthesizing reviews, then shifted them and finally sorted them using aggregation sites Dave, Lawrence, and Pennock (2003). These structured reviews are used for testing and training. From these reviews features are identified and finally scoring methods are used to determine whether the reviews are positive or negative. They have used a classifier to classify the sentences obtained from web-search through search query using product name as search condition.

Matsumoto *et.al.*, have used the syntactic relationship among words as a basis of document level sentiment analysis Matsumoto, Takamura, and Okumura (2005). In their paper, frequent word subsequence and dependency sub-trees are extracted from sentences, which act as features for SVM algorithm. They extract unigram, bigram, word subsequence and dependency subtree from each sentences in the dataset. They used two different datasets for conducting the classification i.e., IMDb dataset IMDb (2011) and Polarity dataset Pang and Lee (2004). In case of IMDb dataset, the training and testing data are provided separately but in Polarity dataset 10-fold cross validation technique is considered for classification as there is no separate data designated for testing or training.

Zhang *et.al.* have proposed the classification of Chinese comments based on word2vec and SVM^{perf} Zhang, Xu, Su, and Xu (2015). Their approach is based on two parts. In first part, they have used word2vec tool to cluster similar features in order to capture the semantic features in selected domain. Then in second part, the lexicon based and POS based feature selection approach is adopted to generate the training data. Word2vec tool adopts Continuous Bag-of-Words (CBOW) model and continuous skip-gram model to learn the vector representation of words Mikolov, Chen, Corrado, and Dean (2013). SVM^{perf} is an implementation of SVM for multi-variate performance measures, which follows an alternative structural formulation of SVM optimization problem for binary classification Joachims (2006).

Liu and Chen have proposed different multi-label classification on sentiment classification Liu and Chen (2015). They have used eleven multilevel classification methods compared on two microblog dataset and also eight different evaluation matrices for analysis. Apart from that, they have also used three different sentiment dictionary for multi-level classification. According to the authors, the multi-label classification process perform the task mainly in two phases i.e., problem transformation and algorithm adaptation Zhang and Zhou (2007). In problem transformation phase, the problem is transformed into multiple single-label problems. During training phase, the system learns from these transformed single label data, and in the testing phase, the learned classifier makes prediction at a single label and then translates it to multiple labels. In algorithm adaption, the data is transformed as per the requirement of the algorithm.

Luo *et.al.*, have proposed an approach to convert the text data into low dimension emotional space (ESM) Luo, Zeng, and Duan (2016). They have annotated small size words, which have definite and clear meaning. They have also used Ekman Paul's research to classify the words into six basic categories such as anger, fear, disgust, sadness, happiness and surprise Ekman and Friesen (1971). They again have considered two different approaches for assigning weight to words by emotional tags. The total weight of all emotional tags are calculated and based on these values, the messages are classified into different groups. Although their approach yields reasonably a good result for stock message board, the authors claim that it can be applied in any dataset or domain.

Niu *et.al.*, have proposed a Multi-View Sentiment Analysis (MVSA) dataset, including a set of image-text pair with manual annotation collected from Twitter Niu, Zhu, Pang, and El Saddik (2016). Their approach of sentiment analysis can be categorized into two parts, i.e., lexicon based and statistic learning. In case of lexicon based analysis, a set of opinion words or phrases are