Chapter 9

Classification of Sentiment of Reviews using Supervised Machine Learning Techniques

Abinash Tripathy

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

Santanu Kumar Rath

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

ABSTRACT

Sentiment analysis helps to determine hidden intention of the concerned author of any topic and provides an evaluation report on the polarity of any document. The polarity may be positive, negative or neutral. It is observed that very often the data associated with the sentiment analysis consist of the feedback given by various specialists on any topic or product. Thus, the review may be categorized properly into any sort of class based on the polarity, in order to have a good knowledge about the product. This article proposes an approach to classify the review dataset made on basis of sentiment analysis into different polarity groups. Four machine learning algorithms viz., Naive Bayes (NB), Support Vector Machine (SVM), Random Forest, and Linear Discriminant Analysis (LDA) have been considered in this paper for classification process. The obtained result on values of accuracy of the algorithms are critically examined by using different performance parameters, applied on two different datasets.

1. INTRODUCTION

Sentiment analysis, also known as opinion mining, analyzes people's opinion or emotions towards entities such as products, organization and their attributes (Liu, 2012). With the presence of large number of social media on the web, the information about any product can be conveniently obtained from reviews, blogs and comments. These comments not only help the analyst to study the responses on any item or product, but also probably the customer gets an opportunity to gain knowledge about its demand in any

locality. In order to carry out analysis in an optimal way from people's sentiments, various machine learning techniques are often applied by researchers and practitioners.

There are mainly two categories of machine learning techniques which are mostly used in sentiment analysis, i.e., supervised and unsupervised learning. In supervised learning, the dataset is labeled during analysis, and is trained to obtain a reasonable output that helps in decision making (Gautam & Yadav, 2014). Unlike supervised learning, unsupervised learning does not provide any label data which is quite difficult for the process. In order to solve this problem, various clustering algorithm are often considered. This study highlights the impact of supervised learning method on labeled data.

In general, sentiment analysis has been investigated mainly in three levels such as document level, sentence level and aspect level (Feldman, 2013). *Document level* classifies the whole opinion document expresses a positive or negative sentiment. *Sentence level* classification helps to find out as to whether the sentence expresses any negative, positive or neutral opinion. *Aspect level* classification process focuses on all sentiment expressions, present within given document and the aspect to which it refers. During the course of this study, concentration has been made on document level sentiment.

In this study, an attempt has been further made to transform the labeled data, which is often available in text format, to numeric values arranged in the form of matrix. The derived matrix is considered as input for machine learning algorithms, which helps to classify the files into either positive or negative polarity. Using different performance parameters, the accuracy of the approaches are studied and compared with the result obtained by various researchers, available in literature

This paper is structured in the following manner: Section 2 presents literature survey. Section 3, indicates the methodology about the classification algorithm and its details. In section 4, the proposed approach is highlighted. In Section 5, performance evaluation of the proposed approach is carried out. Finally, section 6 concludes the paper and presents the scope for future work.

2. LITERATURE SURVEY

Pang *et al.*, have considered sentiment classification as a special case of topic based on categorization aspect with positive and negative sentiments (Pang, Lee, & Vaithyanathan, 2002). They have undertaken the experiment with three standard algorithms i.e., Naive Bayes classification, Maximum Entropy classification and Support Vector machine being applied over the n-gram technique.

Pang and Lee have labeled sentences in the document as subjective or objective (Pang & Lee, 2004). They have applied machine learning classifier to the subjective group which prevents polarity classification from considering any misleading data. They have explored extraction of methods on the basis of minimum-cut formulation, which provides an effective way for integration of inter-sentence level information with bag of words.

Matsumoto *et al.*, have considered the syntactic relationship among words as a basis of document level sentiment analysis (Matsumoto, Takamura, & Okumura, 2005). In this paper, frequent word sub-sequence and dependency sub-trees are extracted from sentences and they act as feature for SVM algorithm. They have used unigram method, bigram method and combination of both the methods for classification.

Read have proposed a source of training data based on language used in conjunction with emotions, could function independent of topic, domain and time (Read, 2005). Naive Bayes (NB) and Support Vector Machine (SVM) algorithms are applied to classify the polarity of the dataset.

Related Content

Data Mining for Multicriteria Single Facility Location Problems
Seda Tolunand Halit Alper Tayal (2020). *Cognitive Analytics: Concepts, Methodologies, Tools, and Applications (pp. 1248-1271).*
www.irma-international.org/chapter/data-mining-for-multicriteria-single-facility-location-problems/252080

Sentiment Analysis with Text Mining in Contexts of Big Data
Carina Sofia Andradeand Maribel Yasmina Santos (2020). *Cognitive Analytics: Concepts, Methodologies, Tools, and Applications (pp. 922-942).*
www.irma-international.org/chapter/sentiment-analysis-with-text-mining-in-contexts-of-big-data/252063

Investigation of Software Reliability Prediction Using Statistical and Machine Learning Methods
Pradeep Kumarand Abdul Wahid (2020). *Cognitive Analytics: Concepts, Methodologies, Tools, and Applications (pp. 1640-1660).*
www.irma-international.org/chapter/investigation-of-software-reliability-prediction-using-statistical-and-machine-learning-methods/252104

Classification of Sentiment of Reviews using Supervised Machine Learning Techniques
Abinash Tripathyand Santanu Kumar Rath (2020). *Cognitive Analytics: Concepts, Methodologies, Tools, and Applications (pp. 143-163).*
www.irma-international.org/chapter/classification-of-sentiment-of-reviews-using-supervised-machine-learning-techniques/252024

A Novel Anti-Obfuscation Model for Detecting Malicious Code
Yuehan Wang, Tong Li, Yongquan Cai, Zhenhu Ning, Fei Xueand Di Jiao (2020). *Cognitive Analytics: Concepts, Methodologies, Tools, and Applications (pp. 1556-1576).*
www.irma-international.org/chapter/a-novel-anti-obfuscation-model-for-detecting-malicious-code/252098