IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified

Vol. 6, Issue 12, December 2017

Machine Learning Impact on Sentiment Analysis of Tweets: A Review

Punita Bhardwaj[1], Aman Kumar[2], Astha Gautam[3]

MTech Scholar, Dept., of Computer Science and Engg LRIET Solan, H.P[1]

Assistant professor, Dept., of Computer Science and Engg LRIET Solan, H.P[2,3]

Abstract: Tweet sentiment analysis is an effective and valuable technique in the sentiment analysis domain. It is the most extensively used approach for tweet sentiment analysis. Machine learning algorithms and Sentiment analysis of tweets are an application of mining Twitter and it is growing in popularity as a means of determining public opinion. Machine learning algorithms are used to perform sentiment analysis; however, data quality issues such as high dimensionality, class imbalance or noise may negatively impact on classifier performance. Machine learning techniques are for targeting these problems but it has not been applied to this domain or studied in detail.

Keywords: Machine learning, sentiment,optimization,tweets

I.INTRODUCTION

The most common definition describes characteristics of big data as volume, velocity and variety. Volume refers to the massive size of big datasets. Velocity refers to the rate at which data are generated and must be acted upon, such as filtered, reduced, transferred and analyzed, as opposed to stored for future processing [7]. Variety refers to the diverse data forms in big data, including structured (tabular such as in a spreadsheet or relational database), unstructured (such as text, imaging, video, and audio), and semi-structured (such as XML documents)[5]. Today, the textual data on the internet is growing rapidly. Several kinds of industries are trying to use this massive textual data for extracting the people's views towards their products. Social media is a crucial source of information in this case. It is not possible to manually investigate the heavy amount of data. This is where the requirement of automatic classification becomes clear. x The popularity of micro blogging stems from its distinctive communication services such as portability, immediacy, and ease of use, which allow users to instantly respond and spread information with limited or no restrictions on content. Twitter is currently the most popular and fastest-growing microblogging service, with more than 140 million users producing over 400 million tweets per day—mostly mobile— as of June 2012.Twitter enables users to post status updates, or tweets, no longer than 140 characters to a network of followers using various communication services. Tweets have reported everything from daily life stories to latest local and worldwide events. Twitter content reflects real-time events in our life and contains rich social information and temporal attributes. Monitoring and analyzing this rich and continuous flow of user-generated content can yield unprecedentedly valuable information.
Online social media sites (FaceBook, Twitter, YouTube, etc.) have revolutionized the way we communicate with individuals, groups, and communities and altered everyday practices (Boyd and Ellison 2007). Several recent workshops, such as Semantic Analysis in Social Media (Farzindar and Inkpen 2012), are increasingly focusing on the impact of social media on our daily lives. For instance, Twitter has changed the way people and businesses perform, seek advice, and create "ambient awareness" and reinforced the weak and strong tie of friendship.

Unlike other media sources, Twitter messages provide timely and fine-grained information about any kind of event, reflecting, for instance, personal perspectives, social information, conversational aspects, emotional reactions, and controversial opinions.

A major challenge facing event detection from Twitter streams is therefore to separate the mundane and polluted information from interesting real-world events. In practice, highly scalable and efficient approaches are required for handling and processing the increasingly large amount of Twitter data (especially for real-time event detection). Other challenges are inherent to Twitter's design and usage. These are mainly due to the short length of tweet messages, the frequent use of (dynamically evolving) informal, irregular, and abbreviated words, the large number of spelling and grammatical errors, and the use of improper sentence structure and mixed languages.

Machine learning is a method of data analysis that automates analytical model building. Using algorithms that iteratively learn from data, machine learning allows computers to find hidden insights without being explicitly programmed where to look.

DOI 10.17148/IJARCCE.2017.61236

II.LITERATURE REVIEW

Tripathy et al. [1] The author proposed machine learning algorithms Naïve bayes (NB), Maximum entropy(ME), Stochastic Gradient Descent (SGD), and Support Vector Machine (SVM) for the sentiment classification. These algorithms are further applied using n-gram approach on IMDb dataset. It is observed that as the value of 'n' in n-gram increases the classifi-cation accuracy decreases i.e., for unigram and bigram, the result obtained using the algorithm is remarkably better; but when tri-gram, four-gram, five-gram classification are carried out, the value of accuracy decreases.

Luo et al. [2] In this paper author proposed emotion space model for sentiment classification. This method classifiy the investors emotions in public. Lexical semantic extension and correlation analysis methods to extend the scale of emotion words, which can capture more words with strong emotions for ad hoc domain, like network emotion symbols. ESM is applied on the messages of stock message board TheLion. ESM model is also compared with information gain and mutual information. It provides effective results in accuracy.

Niu, Teng, et al. [3] The author proposed a Multiview Sentiment Analysis Dataset (MVSA) that includes a set of images with text pairs collected from the Tweeter. Data set can be used as single view and multi view sentiment analysis. With this dataset, many State-of-art approaches are evaluated. Effectiveness of the correlation between different views is also evaluated. Results shows that the performance is also boost by using textual and images.

Gautam et al. [4] Sentiment analysis of customers review is done in this paper by using machine learning algorithms Naïve Bayes, SVM and Maximum Entropy. The customer data collected from the Tweeter which is unstructured and not easy to understand that the review is positive or negative regarding the product. In this firstly the preprocessing of the tweets dataset then extract adjective from the dataset that have some meaning which is called feature vector, then selected the feature vector list and thereafter applied machine learning based classification algorithms.

Balahur et al.[5] In this paper author proposed Machine translation method for the classification of the sentiments. In this three different languages was used for the analysis. These languages are English, German, Spanish and French. Three machine translation method Google, Bing and Moses Supervised learning algorithms for various types of features. Meta classifiers are used to remove the noise introduced by translation.

Neethu et al. [6] Knowledge base approach and Machine learning approach are the two strategies used for analyzing sentiments from the text in this paper. Author analyzed the reviews related to electronics devices like cell phones, laptops etc. Author proposed a new feature vector for classifying the tweets as positive, negative and extract peoples' opinion about products. Stacked Denoising Auto-Encoders with sparse rectifier units can perform an unsupervised feature extraction which is highly beneficial for the domain adaptation of sentiment classifiers.

Glorot et al. [7] Deep Learning approach is proposed in this paper which used to extract the meaningful representation of each review. Sentiment classifiers trained with this high-level feature representation clearly outperform state-of-the-art methods on a benchmark composed of reviews of 4 types of Amazon products.

Glorot et al. [8] In this paper, the author examined the data mining on the Spanish Tweeter data . Author discussed the effect of various settings on the precision. The experiments done on the Naïve Bayes, Decision Tree and Support Vector Machine. Also presented a novel resources for analysis of emotions in text.

Li, Nan et al. [9] In this paper author proposed the sentiment analysis of the online forums hotspot detection and forecast. Firstly author analyzed the emotional polarity of the text and obtained value for each piece of text. Secondly, combined algorithm K-mean clustering and Support vector Machine (SVM) for text mining. Experimental results demonstrate that SVM forecasting achieves highly consistent results with K-means clustering.

Yao et al. [10] In this paper author proposed a kernel based sentiment analysis technique to classify the user's view regarding to any topic. In this paper author analysis the Chinese sentences for the review polarization. In this Paper features was extracted from lexical and semantic levels. The result of this paper shows that approach is effective and outperforms the very competitive n-gram method.

Arantxa Barrachina Arantxa Duque et.al. [11]: Technical Support call centres frequently receive several thousand customer queries on a daily basis. Traditionally, such organisations discard data related to customer enquiries within a relatively short period of time due to limited storage capacity. This paper proposes a Proof of Concept (PoC) end to end solution that utilises the Hadoop programming model, extended ecosystem and the Mahout Big Data Analytics library

DOI 10.17148/IJARCCE.2017.61236

for categorising similar support calls for large technical support data sets. The proposed solution is evaluated on a VMware technical support dataset.

Chen Min, et.al. [12]: They review the background and state-of-the-art of big data. They first introduce the general background of big data and review related technologies, such as could computing, Internet of Things, data centers, and Hadoop. Then focus on the four phases of the value chain of big data, i.e., data generation, data acquisition, data storage, and data analysis. For each phase, they introduce the general background, discuss the technical challenges, and review the latest advances. Finally examine the several representative applications of big data, including enterprise management, Internet of Things, online social networks, medial applications, collective intelligence, and smart grid..

Hashem Ibrahim Abaker Targio et al[13]: Massive growth in the scale of data or big data generated through cloud computing has been observed. Addressing big data is a challenging and time-demanding task that requires a large computational infrastructure to ensure successful data processing and analysis. The rise of big data in cloud computing is reviewed in this study. The definition, characteristics, and classification of big data along with some discussions on cloud computing are introduced.

Ioannis Partalas et al[14]: This paper provides an overview of the workshop Web-Scale Classification: Web Classification in the Big Data Era which was held in New York City, on February 28th as a workshop of the seventh International Conference on Web Search and Data Mining. The goal of the workshop was to discuss and assess recent research focusing on classification and mining in Web-scale category systems.

Jonathan Stuart Ward and Adam Barker [15]: The term big data has become ubiquitous. Owing to a shared origin between academia, industry and the media there is no single unified definition, and various stakeholders provide diverse and often contradictory definitions. The lack of a consistent definition introduces ambiguity and hampers discourse relating to big data. This short paper attempts to collate the various definitions which have gained some degree of traction and to furnish a clear and concise definition of an otherwise ambiguous term.

Literature Review Findings

Reference No.	Author Name	Year	Algorithm Used	Description
[1]	Tripathy et al.	2016	Naïve Bayes (NB), Maximum entropy(ME), Stochastic Gradient Descent (SGD), and Support Vector Machine (SVM)	Proposed machine learning algorithms Naïve Bayes (NB), Maximum entropy(ME), Stochastic Gradient Descent (SGD), and Support Vector Machine (SVM) for the sentiment classification. These algorithms are further applied using n-gram approach on IMDb dataset.
[2]	Luo et al. [2]	2016	Low dimensional emotion space model (ESM)	Author proposed emotion space model for sentiment classification. This method classify the investors emotions in public.
[3]	Niu, Teng, et al.	2016	Multiview Sentiment Analysis Dataset	Proposed a Multiview Sentiment Analysis Dataset (MVSA) that includes a set of images with text pairs collected from the Tweeter.
[4]	Gautam et al.	2014	Naive Bayes, Maximum entropy and SVM	Sentiment analysis of the customers review.
[5]	Balahur et al.	2014	Machine Translation and Supervised methods	Sentiment analysis of different languages using machine translation system.
[6]	Neethu et al.	2013	Support Vector Machine	Identifying and classifying opinions or sentiments expressed in source text
[7]	Sidorov, Grigori, et al.	2012	Naïve Bayes, Decision tree, Support Vector Machine	Examines the working of classifiers on Spanish data.
[8]	Glorot et al.	2011	Deep learning approach	Propose a deep learning approach which learns to extract a meaningful Representation for each review in an unsupervised fashion.
[9]	Li, Nan et al.	2010	K-means clustering and support vector machine (SVM)	Proposed text mining approach to group the forums into various clusters, with the center of each representing a hotspot forum within the current time span.
[10]	Yao et al.	2009	kernel-based sentiment classification	Kernel based classification for the Chinese Sentences.

DOI 10.17148/IJARCCE.2017.61236

CONCLUSION

In this paper review the sentiment on Twitter using Machine Leaning Techniques. Another consideration is that we applied Bigram, Unigram, Object-oriented features as an effective feature set for sentiment analysis. Used a good memory for resolving features better. However, we chose an effective feature set to enhance the effectiveness and the accuracy of the classifiers shows the comparative analysis of accuracy and precision between four algorithms showing the effect of features optimization

REFERNCES

[1] Tripathy, Abinash, Ankit Agrawal, and Santanu Kumar Rath. "Classification of sentiment reviews using n-gram machine learning approach." *Expert Systems with Applications* 57 (2016): 117-126.

[2] Luo, Banghui, Jianping Zeng, and Jiangjiao Duan. "Emotion space model for classifying opinions in stock message board." *Expert Systems with Applications* 44 (2016): 138-146.

[3] Niu, Teng, et al. "Sentiment analysis on multi-view social data." *International Conference on Multimedia Modeling*. Springer, Cham, 2016.

[4] Gautam, Geetika, and Divakar Yadav. "Sentiment analysis of twitter data using machine learning approaches and semantic analysis." *Contemporary computing (IC3), 2014 seventh international conference on*. IEEE, 2014.

[5] Balahur, Alexandra, and Marco Turchi. "Comparative experiments using supervised learning and machine translation for multilingual sentiment analysis." *Computer Speech & Language* 28.1 (2014): 56-75.

[6] Neethu, M. S., and R. Rajasree. "Sentiment analysis in twitter using machine learning techniques." *Computing, Communications and Networking Technologies (ICCCNT), 2013 Fourth International Conference on*. IEEE, 2013.

[7] Sidorov, Grigori, et al. "Empirical study of machine learning based approach for opinion mining in tweets." *Mexican international conference on Artificial intelligence*. Springer, Berlin, Heidelberg, 2012.

[8] Glorot, Xavier, Antoine Bordes, and Yoshua Bengio. "Domain adaptation for large-scale sentiment classification: A deep learning approach." *Proceedings of the 28th international conference on machine learning (ICML-11)*. 2011.

[9] Li, Nan, and Desheng Dash Wu. "Using text mining and sentiment analysis for online forums hotspot detection and forecast." *Decision support systems* 48.2 (2010): 354-368.

[10] [10]Yao, Tianfang, and Linlin Li. "A kernel-based sentiment classification approach for chinese sentences." *Computer Science and Information Engineering, 2009 WRI World Congress on*. Vol. 5. IEEE, 2009.

[11] Arantxa Duque Barrachina, Aisling O'Driscoll. A big data methodology for categorising technical support requests using Hadoop and Mahout .Journal of data 2014: doi: 10.1186/2196-1115-1

[12] Chen, Min, Shiwen Mao, and Yunhao Liu. "Big data: a survey." *Mobile Networks and Applications* 19.2 (2014): 171-209.

[13] Hashem, Ibrahim Abaker Targio, et al. "The rise of "big data" on cloud computing: Review and open research issues." *Information Systems* 47 (2015): 98-115.

[14] Ioannis Partalas,, et al. "Web-scale classification: web classification in the big data era." *Proceedings of the 7th ACM international conference on Web search and data mining*. ACM, 2014.

[15] Jonathan Stuart, and Adam Barker. "Undefined by data: a survey of big data definitions." *arXiv preprint arXiv:1309.5821* (2013).

DOI 10.17148/IJARCCE.2017.61236