Sentiment analysis to social Media data for a historical event

Owais Muhammad, Yiding Li, Xintong Li.

Abstracts:

"What other people think" has always been an important piece of information for most of us during the decision-making process.

As social media is a generous and very auspicious part of Big Data, it is illustrated what different sources and types of social media data there are and how such data can easily be extracted. Social Media is becoming major and popular technological platform that allows users to dropping their opinion, discussing and sharing information and their opinion or views. Information is engendered and accomplished through whichever computer or mobile devices via one person and consumed by numerous other persons. Furthermost of these user generated content are textual information, as Social Networks (Facebook, LinkedIn), Microblogging (Twitter), blogs (BlogSpot, WordPress). Looking for valuable nuggets of knowledge, captures their ideas and thoughts from these massive amount of data could help users make informed decisions. In this article, we have study the algorithms of Sentiment Analysis of Social Media Data for Past Great Event.

This article also presents a comparison of different machine learning algorithms applied to the case of sentiment analysis in social media. Several machine learning algorithms were used during experimentation session like Support vector machine, Navies Bayes, decision tree classifier, Random forest classifier, K-nearest neighbor classifier, Gradient Boosting Classification, AdaBoost classifier. Excepting these, we have also used the clustering algorithm like K-Means and DBSCAN for checking their performance. The results show the method performed well for pre-processing the natural language and found those ones which impact on the building accurate classifiers. We have also used the neural network model to find out their accuracy and then compared with other models too.

The best performance was achieved by the Gaussian navies Bayes. We have compared the top three classifier like Gaussian navies Bayes, AdaBoost as well as Gradient Boosting Classifier, so the best accuracy of Gaussian navies Bayes. We have Chosen *the Gaussian Navies Bayes* is the best algorithm for our data sets.

This topic is associated to big event in history, by these types of data can discovers and learn new thing and idea which might be useful in future because from history can learn so much things.

Key Words: Social media, sentiment analysis, machine learning, nature language processing, scikit learn

Introduction:

In 21st century the Social Media has develop one of the best popular and prevalent platforms to tolerate users discussing, communicating, and sharing their interested topics as well as issues without having similar geo-location and same time, have free-hand to drop your opinions or discuss/debate for the topic or any discussion, even if a person agree or disagree about the issue also can take part in discussion too. Information can be generated and managed through either computer or mobile devices by one person and consumed by many other persons as in the form of different groups of peoples from anywhere. Different people could express different opinions on the same topic, as they wish as they want. Topics with wide variety and variation, ranging from current events and political debates to sports and entertainment are being keenly and intensely chatted and discussed on these social forums. The clout of social media as a marketing tool has been acknowledged, and is being assertively used by governments, major organization, schools and other groups to excellently and rapidly communicate with large numbers of people. Another imperative metric for business and commerce to measure their online reputation and standard is word of mouth persuasive. Word of mouth is the process of spreading information from person to person and is frequently done through social media networks. It also plays a key role in customer buying verdicts. Some emblematic examples are that Facebook users could comment campaigns airmailed by a company or like a company on Facebook, Twitter users could send tweets with a maximum length of 140 characters to instantly share and deliver their opinion and sentiments on politics, movies, sports, etc. Collecting and scrutinizing these data could assistance users or managers make conversant decisions and conclusion. Marketing leaders or product managers might collect, evaluate and analyze feedbacks, opinions and comments on campaigns launched by themselves from Facebook targeting to embrace efficient presenting scheme and progress product quality and worth. Furthermost of these user generated content are textual information. The swift growth in volume of web texts from major social network sites like Facebook and Twitter initiatives us to analyze and mine the data through computational methods. Ascertaining their sentiments has become a vital issue and fascinated many attentions and cares. Lately, there have been a number of studies struggling to model/predict real-world events using information from social media networks. Among these, Twitter has attracted further attention and devotion because of the enormous swell in its attractiveness. Jansen et al [1] worked on twitter's large-scale analysis of brand sentiment. Results concludes that 19% of tweets contain brand orientations, of which closely 20% cover sentiments approximately the brands. The key research efforts on sentiment analysis done beforehand can be classified into 2 branches. On one hand, they take state-of-the-art sentiment identification algorithms to solve problems in factual applications such as summarizing customer reviews [2], for product's ranking and classifying[3], finding product features that imply opinions [4], evaluates and analyzes tweet sentiments about movies and tries to attempts the expect box office revenue. The authors describe different metrics to measure the popularity/sentiment of a movie and then use a linear regression model to predict box-office revenue. Joshi ET all [5] use an analogous technique to envisage box-office revenue of movies using review text. On the other hand, researchers put their focus on discovering new sentiment algorithms. Bag-of-Words approach produces domain-specific lexicons. There is an enormous body of research which endeavors to integrate these interactions as features in a machine learning model [6]. Rule-based methods has been studied by many researchers. Propose compositional

semantics, which is centered on the assumption that the meaning of a composite expression is a function of the meaning of its parts and of the syntactic rules by which they are combined. In addition to these rules, we require a method of assessing and gaging the influence of these on the polarity of an expression. We also ripen our version of the Compose function for computing the polarity of an expression based on Compositional Semantic rules.

Sentiment analysis is one of the fastest growing research are as in computer science, making it challenging to keep track of all the deeds in the region, Natural Language Processing (NLP) is one of the most active research areas in Sentiment analysis or Opinion Mining, is used to extract the opinions that appear on the web to explicit the assertiveness and judgment of the opinion holder about a certain area either social or public topics else more. For the scrutinize reviews it is used comprehensively, social media, and blogs for sentiments and views expressed on products, services, individuals, and organizations. Polarity detection is the most common form of sentiment analysis, i.e., determining whether the sentiment expressed in a review is positive, negative or neutral. The social media, blogs, forums and e-commerce websites encourage people to share their opinions and feelings overtly. The People's assessments, capabilities and experiences are exquisite information in the decision-making process. Buying a new product or watching a new movie it is normal now-a-days to look at reviews first to read and examine the people's opinion about the product. However, reading a review fully is a time-consuming task and more than that most of them does not provide a final verdict. So it is desirable to have an automated sentiment analysis system that identifies the sentiment expressed in a review. In 2010, the Barbosa et al. [7] intended dual segment automatic sentiment analysis technique for classifying tweets. They classified tweets as objective or subjective and then in second phase, the subjective tweets were classified as positive or negative. The feature space used included punctuation, hashtags, retweets, link, and exclamation marks in conjunction with features like prior polarity of words and POS. The author [8] used bag-of-words method for sentiment analysis in which the relationships between words was not at all considered and a document is demonstrated as just a collection of words. To determine the sentiment for the whole document, sentiments of every word was determined and those values are united with some aggregation functions.

Related Work:

Fu Xianghua and Liu Guo have proposed unsupervised approach which called Multi-aspect Sentiment Analysis for Chinese Online Social Reviews (MSA-COSRs) to determine the multi-aspect sentiment of Chinese social reviews, which is called [9]

Clustering [10] is a most significant data mining technique for organizing data. In the clustering process, the data objects are grouped into number of groups and clusters for more resemblance of the objects, except some are unlike to objects in the another clusters. Clustering problem can be written as: Given

a) Dataset= A1, A2, A3……AN

b) Desired no. of clusters C

c) Function FN to find the clusters.

We need = B (1, 2……N) → (1, 2…..C)

The similarity measure is the key point to the clustering problem.

K-Means algorithm [11] is an algorithm for clustering algorithms. It takes input and then partitions it into k clusters. Partition is done in such a way that inter cluster distance between

similar clusters is very less. It is calculated using “center of gravity”.

Tuanqi Chen and Carlos Guestrin [12] have designed and build a highly scalable end to end tree boosting system, they proposed a theoretically justified weighted quantile sketch for efficient proposal calculation. As well as an effective cache-aware block structure for out-of-core tree learning.

Liang Li [13] used the supervised method, AdaBoost-Based Algorithm for the classification of alternatively spliced exons for show the accuracy and experiment result show that the accuracy is higher than current algorithms.

Text classification is the problem of automatically allocating predefined sorts or categories to permitted text documents, though additional textual information is available online, effective repossession is challenging devoid of good indexing and summarization is one solution to this problem. A growing amount of statistical sorting methods and machine learning techniques have been applied to text categorization in current centuries, containing multivariate regression models [14] nearest neighbor classification [15], decision trees [16].

Methodology:

Our approach consists of three steps namely data extraction, data processing and data classification. In addition to this, we have catered for colloquial words written differently by different users.

A. Data Extraction

The dataset was extracted using the Facebook, twitter, Youtube as well as quora. This data consists of revolutionary tenure, after it there was a war about 7-8 year between Iran and Iraq. All the extracted data are saved in an SQLite database and may also be exported to a csv file. Our dataset consists of comments extracted from different posts including videos, pictures and links. The trilingual dataset consists of 3175 comments each having only in English language.

B. Data Processing

Pre-processing’s aim to convert the original manuscript data in a data mining ready structure, in other words the new documents into the information retrieval system. The preprocessing is one of the important components in a characteristic text classification structure. This article aims to comprehensively observe the impression of preprocessing on text classification in terms of various aspects like text domain, classification accuracy, dimension reduction as well as text language. The following are the step for the Pre-processing the datasets.

After extracted all comment from the above source, we have worked on it manually to categorize into three polarities, which are Positive, Negative as well as Neutral. Peoples used the different kind of language, like English, Urdu, Persian or roman too, but we have only worked on English language comments and ignore other languages.

The preprocessing level of the study adapts the innovative textual data in a data-mining-ready structure, wherever the utmost important text-features that serve to distinguish amongst text-categories are recognized. Pre-processing is the process of combining a new document into an information retrieval system. An effective preprocessor represents the document efficiently in terms of both space (for storing the document) and time (for processing retrieval requests) requirements and maintain good retrieval performance (precision and recall). This point is the greatest critical and complex process that centrals to the depiction of each text or documents by a

choice set of index terms. The key impartial of preprocessing is to acquire the key features or key terms from online news text documents and to improve the relevancy amongst word and document and the relevancy among word and category. 'The enormous growing in the weighbridge of data has been pragmatic and observed in this eon actuality a key factor of the Big Data scenario. Big Data can be defined as high volume, velocity and variety of data that require a new high-performance processing. Addressing big data is a challenging and time-demanding task that requires a large computational infrastructure to ensure successful data processing and analysis. [17]. the data we have extracted and then categories into positive, negative and neutral.

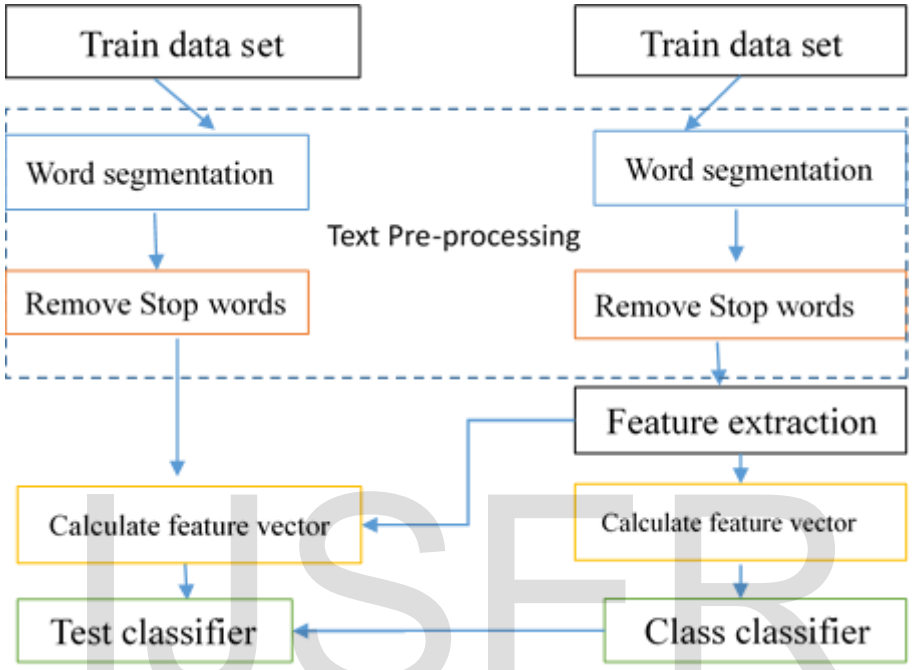


Figure1-1 Data Pre-Processing

The next step is to process and clean the data. The raw data is practically unreadable due to the presence of extra information such as punctuations and other symbols. Tokenization is first applied to break the sentences into distinct words. In the process, emoticons have also been captured. Our work comprised of finding a link and conversation between what people write and what their opinion as well as their smiley tend to depict, means emoji have been detected like smiley, angry etc.

2) Dictionaries

For the dictionary we use only one language, first I was thinking to take data set from different languages, like the Urdu, Persian etc. but I didn't success to get it. Then I only use the English comments which is sufficient for experiment. I make dictionaries

D. Machine Learning Algorithms

To understand the performance and reliability of our proposed algorithm, we also classified the dataset using two well-known machine leaning the supervised learning, two pre-annotated datasets are required, training set and test set. The training set is used to train our classifier while test set is used to evaluate the performance of the classifier. The first step is to collect the data for the training set and then classifier is trained accordingly with the help of the chosen techniques.

1. Navies Bayes
2. Support Vector Machine
3. Decision Tree Classifier
4. Random Forest Classifier
5. Neural Network
6. K-NN
7. Gradient Boosting
8. AdaBoost

For more satisfactions we have used the unsupervised learning to k-mean and DBSCAN.

1. K-Mean
2. DBSCAN

The supervised learning methods can be readily applied to sentiment classification, e.g., naïve Bayesian algorithm, and support vector machines (SVM) algorithm, etc. Pang et al. [18] took this method to categorize movie reviews into two classes, positive and negative. The multilayer perceptron has a large wide of classification and regression applications in many fields: pattern recognition, voice and classification problems. [19]

For the using of classification and regression the SVM constructs a hyper-plane or set of hyper-plan in high or infinite dimension space.

For the SCV, given the training vectors $x_i \in \square^p, i = 1, \ldots, n,$ in two classes, and a vector $y \in \{1, -1\}^n$, SVC solves the following primal problems.

$$\min_{w,b,\zeta} \frac{1}{2} w^T w + c \sum_{i=1}^{n} \zeta_i \qquad (1)$$

Subject to $y_i (w^T \phi(x_i) + b) \geq 1 - \zeta_i, \quad \zeta_i \geq 0, i = 1, \ldots, n$

Its dual is

$$\min_{\alpha} \frac{1}{2} \alpha^T Q \alpha - e^T \alpha \qquad (2)$$

$$\text{Subject to } y^T \alpha = 0$$

$$0 \leq \alpha_i \leq C, i = 1, \ldots n$$

Where e is the vector of all ones, $C > 0$ is the upper bound, Q is an n by n positive semidefinite matrix, $Q_{ij} \equiv y_i y_j K(x_i, x_j)$ where $K(x_i, x_j) = \varnothing(x_i)^T \varnothing(x_j)$ is the kernel, Here training vectors are implicitly mapped into a higher dimensional space by the function $\varnothing$.

The decision function is:

$$\operatorname{sgn}\left(\sum_{i=1}^{n} y_i \alpha_i K(x_i, x) + p\right) \quad (3)$$

Gradient Boost algorithm: One can arbitrarily specify both the loss function and the base-learner models on demand. In practice, given some specific loss function $\Psi(y, f)$ and/or a custom base-learner $h(x, \phi)$ the solution to the parameter estimates can be difficult to obtain. To deal with this, it was proposed to choose a new function $h(x, \phi_t)$ to be the most parallel to the negative gradient $\{g_t(x_i)\}_{i=1}^{N}$ along the observed data

$$g_t(x) = E_y\left[\frac{\partial \Psi(y, f(x))_{|x}}{\partial f(x)}\right] f(x) = f^{t-1(x)} \quad (4)$$

Instead of looking for the general solution for the boost increment in the function space, one can simply choose the new function increment to be the most correlated with $-gt(x)$. This permits the replacement of a potentially very hard optimization task with the classic least-squares minimization one:

$$(\rho_t, \theta_t) = \arg\min_{p,o} \sum_{i=1}^{N} \left[-g_t(x_i) + \rho h(x_i, \theta\right]^2 \quad (5)$$

Experiments and Result:

S.No	Model Name	Parameters	Accuracy(%)
1	Support Vector Machine	rbf	50.94
		linear	47.13
		poly	50.94
		sigmoid	50.94
2	*Navies Baysein*	*Gaussian Naives Bayes*	*54.31*
		Bernoumlli Naives bayes	*53.61*
		Multinomial Naives Bayes	*53.42*
3	Decision Tree	entropy	53.74
		gini	53.74
4	*Neural Network*	*relu*	*49.4*
		identify	*48.45*
		logistics	*48.20*
5	Random Forest	n_tree= 1	50.84
		n_tree= 50	53.99
		n_tree= 100	53.55
6	*K-Nearest Negibore*	*n-neighboor, n=1*	*43.36*
		n-neighboor, n=2	*43.67*

		n-neighboor, n=3	*39.58*
		n-neighboor, n=4	*43.72*
		n-neighboor, n=5	*41.78*
7	GDBT	Loss= 'deviance'	54.18
8	*ADABOOST*		*54.24*

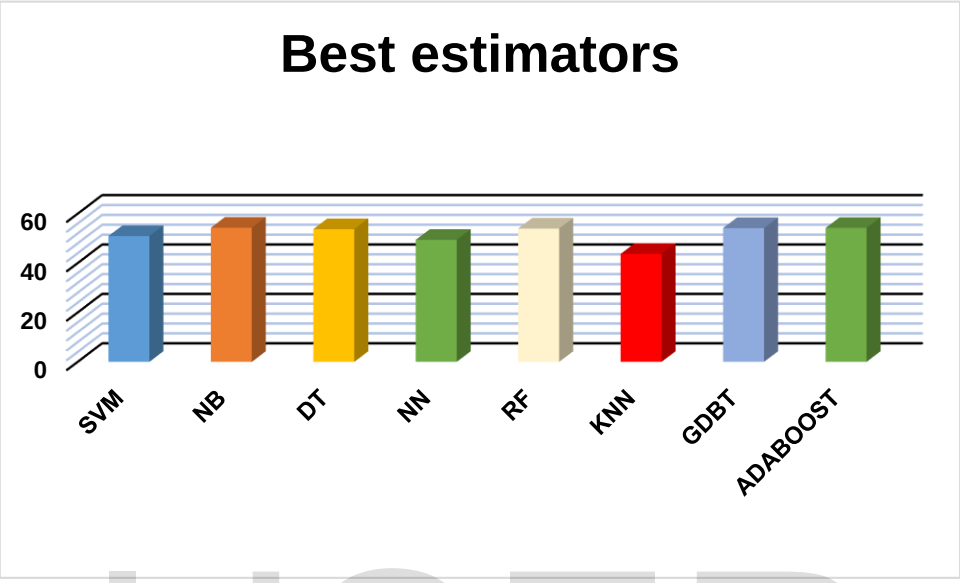
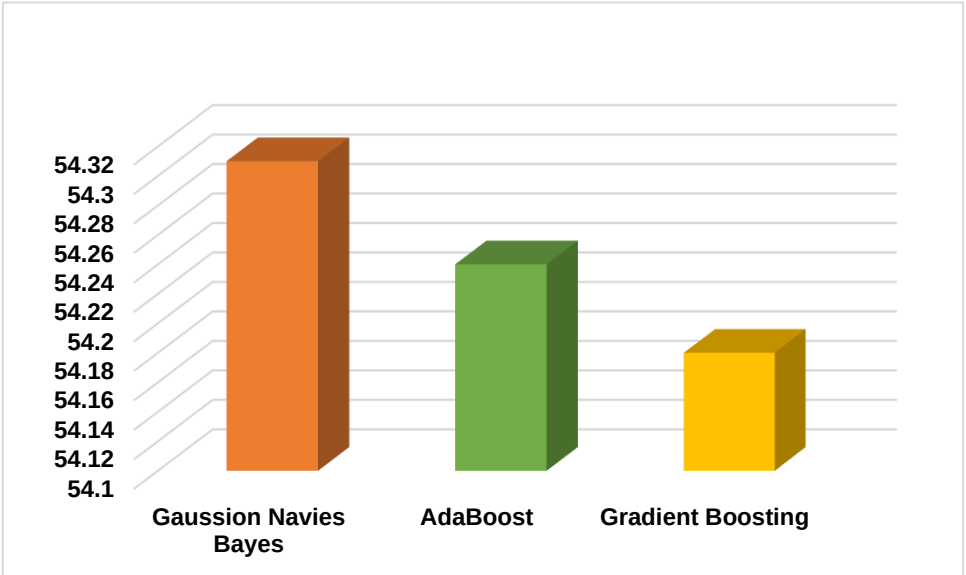


Figure2 Result of all estimator

The top three best results:

Table3-8 Top three models and their accuracy

S.No	Model	Accuracy (%)
1	Gaussian Navies Bayes	54.31
2	AdaBoost	54.24
3	Gradient Boosting	54.18



Top three algorithms and their best accuracy

Compare the Un-Supervised Method

K-Mean

In this method, we have changed number of clusters, as shown in figure below.

Table4-1 K-mean with different number of clusters

S.No	n_cluster	array
1	n_cluster=3	[0 0 2, . . . , 0 0 0]
2	n_cluster=5	[0 0 3, . . . , 0 0 0]
3	n_cluster=7	[1 1 6, . . . , 1 1 1]
4	n_cluster=6	[0 0 3, . . . , 0 0 0]

DBSCAN

Table4-2 DBSCAN and their different parameters, its result

S.No	Parameter	Array
1	eps=0.3, min_sample=10	[0 0 3, . . , 0 0 0]
2	Eps=0.5, min_sample=5	[0 0 3, . . , 0 0 0]

Conclusion

To analyses the opinion views of different peoples on a historical an event, analysis their tweets, facebook comments as well as some online quarries website and comment on different videos on Youtube about the event. We have used the two methods like supervised method and clustering method, in supervised method we used the algorithms like Super vector machine, Navies Bayes, decision tree classifier, Random forest classifier, neural network, K-nearest neighbor classifier, Gradient Boosting Classification, AdaBoost classifier. We have used the three models of Navies Bayes like Gaussian navies Bayes, multinomial navies' Bayes and Bernoulli Navies Bayes. The best result is Gaussian Navies Bayes 54.31. The second top accuracy result is the AdaBoost 54.24 and 3rd one is Gradient Boosting 54.18. So, for the best performance and good accuracy is Gaussian Navies Bayes 54.31, so we have selected the Gaussian Navies Bayes as the best algorithm for our data sets.

Future Work

The future work is to collect more data sets for find out the opinion, views and sentiments of peoples on the same topic, because if we have more data sets then can get more good results as well as the opinion of peoples are changing due to time for same topic. In future work it can be also include to test the data sets by clustering method. For better results can be test the data sets to develop the new algorithm like the deep learning. If do more struggle can be developed the software for preprocessing the data like past event.

References:

[1] B. Jansen, and M. Zhang, and K. Sobel, and A. Chowdury, Twitter power: Tweets as electronic word of mouth [J]. Am. Soc. Inf. Sci. Technol., pages 2169–2188, 2009.

[2] M. Hu, and B. Liu, Mining and summarizing customer reviews.[C] Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, pages 168–177, 2004.

[3] K. Zhang, and R. Narayanan, and W. Liao, and A. Choudhary, Voice of the Customers: Mining Online Customer Reviews for Product Feature-based Ranking. [J] 3rd Workshop on Online Social

Networks, 2010.

[4] A. Popescu, and O. Etzioni, Extracting product features and opinions from reviews. [C] Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing, pages 339–346, 2005.

[5] M. Joshi, and D. Das, and K. Gimpel, and N. A. Smith, Movie reviews and revenues: An experiment in text regression. [J] Proceedings of NAACL-HLT, 2010.

[6] B. Pang, and L. Lee and S. Vaithyanathan, Thumbs up? Sentiment Classification using Machine Learning Techniques[C]. Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 79–86, 2002.

[7] L. Barbosa, J. Feng. "Robust Sentiment Detection on Twitterfrom Biased and Noisy Data"[J]. COLING 2010: Poster Volume,pp. 36-44

[8] J. Kamps, M. Marx, R. J. Mokken, and M. De Rijke, "Using wordnet to measure semantic orientations of adjectives," [J] 2004

[9] Multi-aspect sentiment analysis for Chinese online social reviews based on topic modeling and How Net lexicon [J] Fu Xianghua, Liu Guo, Guo Yanyan, and Wang Zhiqiang

[10] Wenliang C, Xingzhi C, Huizhen W. Automatic word clustering for text categorization using global

Information. [J] ACM Digital Library. 2004; 3411:1–11.

[11] Anuradha A. Neural network approach for text classification using relevance factor as term weighing method. [J] International Journal of Computer Applications. 2013; 68(17):37 -41.

[12] Hassan Ramchoun, Multilayer Perceptron: Architecture Optimization and Training [J], International Journal of Interactive Multimedia and Artificial Intelligence, Vol. 4, N0:1

[13]. Tianqi Chen Aug 2017, XGBoost: A Scalable Tree Boosting System [J].

[14] N. Fuhr, S.Hartmanna, G. Lusting, M. Schwanter and K.Tzeras, "Rule based multistage indexing systems for large subject field", [J] in 606-623, editor, Proceedings of RIAO'91.

[15] R.H. Creecy, B.M. Masand, S.J. Smith and D.L. Waltz, "Trading mips and memory for knowledge Engineering", [J] classifying census returns on the connection machine comm. ACM, 35:48-63, 1992

[16] D.D. Lewis and M. Ringvette,"Comparison of two learning algorithm for text categorization", [J] In Proceeding Analysis and Information Retrieval (SDAIR'94) 1994.

[17] Big data preprocessing: methods and prospects Salvador García, Sergio Ramírez-Gallego, Julián Luengo, José Manuel Benítez and Francisco Herrera [C]

[18] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up? Sentiment classification using machine learning techniques,"[C] Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 79–86, 2002

[19] T.B Ludermir "Hybrid Optimization Algorithm for the Definition of MLP Neural Network Architectures and Weights" [C] Proceedings of the Fifth International Conference on Hybrid Intelligent Systems (HIS'05) 0-7695- 2457-5/05 20.00 2005 IEEE.