A SURVEY ON CHALLENGES AND TECHNIQUES OF SENTIMENT ANALYSIS

Shiramshetty Gouthami[a], Dr. Nagaratna P Hegde[b]

[a]Research Scholar ,Osmania University, CSE, Hyderabad ,India, gouthami.shiramshetty@gmail.com

[b]Professor,Vasavi College of Engineering,CSE,Hyderabad, India,nagaratnaph@staff.vce.ac.in

Abstract: The billions of users share and exchange their opinions on web through different social platforms like twitter, Facebook, Amazon and other product review sites about different problems like products, events, persons or any organizations with the vast developing technology today.Thus the sentiments have been generated by the large number of users on different types of entities which are more useful for the organizations, businesses and even individual also. The sentiment analysis is thus required for this to extract Abstract:The useful information from the large number of resources by using the text analysis and natural language processing techniques. Sentiment analysis is the most significant aspect for the various business and government organization in order to achieve high and better accurate prediction on their future actions based on the opinions from users but the process of this sentiment analysis has been facing different challenges. These challenges become difficulty in analyzing the accurate meaning of sentiments and detecting the suitable sentiment polarity. In this paper a survey on challenges and techniques of sentiment analysis is presented. Various approaches and methodology used in Sentiment Analysis and challenges relevant to their approaches and techniques are covered in this paper. The focus is on Internet text like, Product review, tweets and other social media.
Keywords: Sentiment analysis, opinions, reviews.

1 . Introduction

Internet today has changed the way people interact with each other and know each other. Web, in general has presented us a platform in the form of Social Media, Forums, and Blogs etc. to raise our opinions publicly and to know their feedbacks about a certain topic from people almost anywhere in the world [1].It has changed the outlook ofpeople towards the internet from just being a "Read Only" platform to "Read-Write". The users hunger for and dependence on online advice and recommendations has drawn interest of researchers to research in this area. This need to analyze the thoughts of people and to gain wealth of information from it has led to the emergence of field of the Opinion Mining [2].Sentiment analysis is an interdisciplinary research field which depends on techniques from Natural Language Processing (NLP) [3], text mining [4], machine learning, statistics [5], and information retrieval [6], the main aim of sentiment analysis or opinion mining is study of people's opinions, behaviors, emotions, attitudes and beliefs about an entity such as product, event/topic, person or organization. The purpose of such analysis is to classify the polarity of user's sentiment and extract his opinion regarding an interested entity, which help in providing valuable information for decision making. Sentiment analysis has been classified into different levels, such as document level which classifies the whole document text into positive or negative polarity, sentence level which extract the polarity of each sentence of a document into positive or negative polarity, and aspect/feature level which classify the sentiment polarity of each entity's aspect or feature of a document [7]. There are many numbers of sentiment analysis and opinion mining applications and academic research studies that can perform.Sentiment Analysis is also popularly known as Opinion Mining. It can be defined as a sub discipline of Natural Language Processing and Computational Linguistics mainly concerned with the emotion, thought, or a mood expressed by a reader in any document. Here, the former term signifies evaluation of information when extracted and latter denotes drawing or outing of subjective information from text corpus or reviews [8]. Recently with the incremental growth of the users on social media sites where users daily share their content on different blogs, review sites, Twitter and Facebook. The huge availability of users' opinionated text online made sentiment analysis as one of interested topics either in academic researches or in applications domain, which helps in providing important decision making information for individuals and organizations in different domains. Many challenges are also there that required to efficiently highlight and handled even though sentiment analysis is a challenged task.

2. Challengesof Sentiment Analysis

Sections sentiment analysis mentioned early in the many previous challenges is nontrivial task still not

addressed and resolve efficiently. In this section, based on holistic perspective view of sentiment analysis challenges we highlight the most important challenges which are general for the sentiment analysis as critical field for researchers and industries. Bellow these challenges are discussed with some details.

Big Data-related Issues

The proliferation of web-enabled devices offers new mediums for people to create, communicate and share contents on social web platforms including blogs, social networks, forums, etc., at the same time enormous amount of heterogeneous data are generated by the users of these web communities, the generated data or as it called ‘‘big data’’ offers an unprecedented opportunity for individuals or organizations to mine and analytics big data content using advance technologies and analytics techniques, which enable in providing valuable information for decision makers. Sentiment analysis is one of the valuable text analytics techniques that extract the social web users’ opinions and classify sentiment polarity which is feasible and applicable in different domain. In general the analysis of big data is a challenging task due to volume, variety, velocity, variability and verity of data, which are the main characterize the big data. Sentiment analysis on big data are challenging by the common characteristics of big data [9]. Following are the common sentiment analysis challenges related to big data:

1. Data Collection:
2. Data Preprocessing
3. Data Storage and Analytics
4. Velocity of big data

Language-oriented Issues

Performing sentiment analysis on Non-English languages such as Hindi, Arabic, Chinese, etc., is one of the critical challenges in sentiment analysis due to the different characteristics of each language and the limited number of available researches in other languages comparing to English language which already have many number of corpus and dictionary lexicon available. Although performing sentiment analysis on non-English languages is essential due to the large percent of people around the world who are non native English speakers. Although some of researches tried to handling the language related issues using cross language sentiment classification in which non-English language are automatically translated into English language and the sentiment is performed based on English corpuses and dictionaries but the accuracy of automatic translation is still remarkable. Below are the common challenges for nonEnglish languages sentiment analysis.

1. Lack of Corpuses And Dictionaries.
2. Different Writing Style
3. Different Word Meaning

Domain-oriented Issues

Sentiment analysis is a highly domain-sensitive activity where the ranking of sentiment is highly dependent on the domain from which the training data comes, with the classifier trained on a domain's training dataset generally behaving bad when testing a dataset from another domain. The main challenge is that the opinion words and constructs used to describe an event in on domain often different from one domain to another. Also the orientation of opinion word may be recovered from one domain to another.

Spam and Fake Opinions on Social Sites

Social web communities are characterized by anonymity of their users, the anonymity of user’s identity may be used to fraud other users on web communities. Organizations may use opinion spammers to post fake positive opinions or reviews to promote their products, or fake negative opinions to discredit their competitors, this also true for individuals in political domain or any other domains where the posted opinions about targeted events can influence the evaluation of events from the reader. The challenge is that it is hard to differentiate the fake opinion from non spam opinions by reading it manually. The issue for sentiment analysis is to develop the appropriate techniques and advance algorithms for detecting and filtering out the faked opinions in the collected dataset. Supervised and unsupervised methods for spam opinions detections methods [10] have been discussed.

Opinionated Text Related Issues

The Comparative opinion, Subjective words that not expressed any opinion, Objective words that implicitly expressed opinion, Negation handling and Sarcasm and ironic detection are the common sentiment issues related to the opinionated text and should be addressed efficiently:

3. GeneraL Frameworkof Sentiment Analysis approach

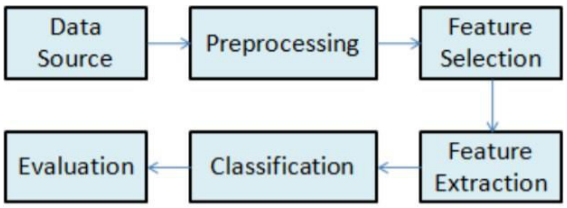

Figure 1: General Framework

Data Sources

The sources mentioned below are widely used by people for finding opinions and giving recommendations for a particular application.

1) Blogs: Blog pages contain expression of one's opinion related to any topic say an event, issues or a service written in a informal or conversational style. Blogs are regularly updated, effective and fast way to share your news and views with others. The popular blogging providers are WordPress.com, LiveJournal, and Blogger.

2) Review Sites: They are the direct sources to know the users opinion about businesses, products or services for decision making in a more clear and summarized way highly influencing consumers choices. Some of the well known sites are Angie's List, Yelp, Glassdoor, TripAdvisor etc.

3) Social Media: These are the highly interactive platforms through which individuals can share, co-create, discuss and modify user generated content. Some of the strongest media used by people over the world today are Facebook, Tumblr, YouTube containing tremendous amount of knowledge literally about any domain.

Preprocessing

Real world data is often incomplete and inconsistent and is likely to contain many errors. In this phase the raw data is processed [11].

1) Tokenization: It is the process of splitting a stream of text into words, phrases or some meaningful elements called tokens with the help of boundaries between words marked by special delimiting characters such as spaces, punctuations and symbols. This process is also known as word extraction, word segmentation or lexical analysis.

2) Stop words Removal: These are the very commonly occurring words in text having a very low discriminative value, serving only the syntactic meaning. It involves creating a list of stop words and then scanning the document so that word appearing in the stop list is removed.

3) Stemming: It is a process to reduce a word to its stem or root word. It is widely used in Information Retrieval to increase the recall rate.

4) Case Normalization: The text published contains both the uppercase and lowercase characters; this process converts the entire text in either Uppercase or Lowercase.

Feature Selection

Feature Selection is a method which reduces both the data and the computational complexity. The set of features that describe a particular object has a greater impact on the classification of objects and thus should be chosen smartly. Data along with useful features usually, may also contain redundant features i.e. the ones providing no extra information or irrelevant features providing no useful information. Thus it becomes a critical task to select a more appropriate subset of features. The most efficient subset would be the one which minimizes error rate making the task more effective and accurate [12].

Feature Extraction

Feature Extraction creates the new features from the functions of original features. When the input data is too large and contains so much of redundant data, this reduction technique is used to extract a reduced set of the most relevant features while still describing the data with sufficient accuracy. Feature Extraction can be combined with Dimensionality reduction using techniques like PCA (Principal component Analysis, Term Presence, Term Frequency, Standard Deviation by elimination of low level or unwanted linguistic features. Three main methods of feature extraction are Filter Techniques, Wrapper techniques and Embedded techniques. Filter methods select the best of features based on intrinsic criterion such as distance measures, Wrapper methods selection is based on generation and evaluation of different subsets in space of states and embedded method looks for an optimal subset of features via search in hypotheses and space of feature subset [13].

Classification

The machine learning techniques, lexicon based techniques and the hybrid techniques are the three types from which sentiment classification techniques arecategorized into. Machine learning techniques are the most primary one used for the sentiment classification analysiswhich involves making use of the linguistic features. Lexion based classification approach is the second one in which group of sentiment words are analyzed which were precompiled in the sentiment lexicon. The combined machine learning and lexicon based techniques are considered as the hybrid approach. The most widely used algorithms from the three of these approaches for the sentiment classification analysis are briefly discussed below.

Machine Learning Techniques

Various machine learning algorithms are used in this machine learning approach almost extensively and exclusively for conducting the sentiment analysis classification. The syntactic features are also used in these machine learning algorithms along withthe linguistic features.

1. Supervised Learning: In this supervised learning method datasets that are labeled clearly are shared with the various supervised learning models [14].

a) Decision Tree Classifiers: The hierarchical classification of training dataset can be accomplished with this classifier on the dataset which has been stored with the attribute values that has to be classified. According to the presence or absence of words in the dataset this method can be predicated and until registering the minimum number of records along with lead nodesthis method is recursively conducted which required for doing classification.

b) Linear Classification

Support vector machine (SVM):SVM is a form of most popular kind of linear classification which is focused on determining best linear separation and providing isolation among various classes.Finding the linear separation attributes in the search space that separate classes from various classes is the fundamental principle of SVM.

Neural Network (NN): As the name indicated that this neural network approach consists of the basic building block as neuron. Each neuron in this takes vector as their input which depicts the occurrences of words over a document across a line. The input functions included are enabled to compute in considering the group of weights by each neuron. Multi neural networks are implemented in the case of nonlinear boundaries. In addition to the multiple linear boundaries those multiple layers are used for approximation of enclosed regions involving particular class.The outputs generated from the previous neuron layers are given as input to the next layers.

Naïve Bayes Classifier (NB):This form of classification is almost widely used to classify text documents and run SA on such document modules.

2 Unsupervised Techniques: By going with the comparison sentiment classification can be carried out in this type of approach. Component comparisons are made with word dictionaries that have been assigned sentiment scores prior to use. Hierarchical and partial clustering techniques are the most popular types of this approach.

Lexicon-based Approach

The sentiment polarity has been determined in this type of approach by using opinion words from the sentiment dictionaries and comparing them with the data. Then the sentiment scores are labeled to the words in this approach to indicate them as positive, negative or objective type [15].Lexicon-based approaches are depending on lexicon sentiment, which includes a set of pre-compiled and well-known sentiment phrases, terms and idioms. There are two types of sub-classifications presentin this type of approach that are explained in the following sections.

1 Dictionary-based Approach:A manual approach that involves a series of instructions known in prior can enable the arrangements of words in this Dictionary-based Approach. By searching a specified corpus called WordNet, the result dataset has produced for the appropriate words and antonyms related to the sentiment analysis. This is aniterative process and can only stops whenever it detects no new words otherwise continues the subsequent iterations when the seed list is progressively added with the words. The evaluation and correction of errors can be done by performing the manual appraisal once the process stopped.
2 Corpus-Based: Dictionaries particularly for a given domain are involved as a key role in this approach. According to the root of opinion words that result from searching on related words using statistical or semantic methods.

35.3 Combination or Hybrid Method

Some of the technologies are there in the research studies for sentiment classification which includes combination of both the machine learning and lexicon based approaches instead of using individual machine learning and lexicon based models as discussed above. Those research studies involve the improved Naïve Bayes and Support vector machine algorithms. The bigrams and unigrams have been most widely used as a feature selection in order to fill the gap between positive and negative sentiments.From the majority of the studies it can be illustrated that the accuracy level of sentiment classificationcan significantly improvedby combined using the machine learning and dictionary-based approach.

3.6 Evaluation

Accuracy, precision, recall and F-measure are the most commonly used measures of evaluation. Accuracy gives a measure of how close a value is to the true (actual) value while the precision gives fraction of instances that are relevant from the retrieved instances and recall is the fraction of relevant instances that are retrieved. Then the harmonic mean of both precision and recall gives the F-measure.

4. Applications

Recently with the incremental growth of the users on social media sites where users daily share their content on different blogs, review sites, Twitter and Facebook. The huge availability of users' opinionated text online made sentiment analysis as one of interested topics either in academic researches or in applications domain, which helps in providing important decision-making information for individuals and organizations in different domains. Although, sentiment analysis is a challenged task and there are also various challenges required to be efficiently highlight and handle.

5. Conclusion

Many research studies and industries applications of sentiment analysis on social web users are available and incrementally receive attention due to its importance in providing valuable decision making information in different domains. Sentiment analysis task is involves many challenges need to be addressed to be performed accurately. This paper review and analysis the existing work related to the sentiment analysis challenges, many number of challenges need to be addressed, the most important challenges are highlighted and discussed. Big data analytics are the major challenges and advance technical and algorithms are required to handle the issues of

sentiment analysis on social web big data. More research works in non-English languages and corpuses-based other languages are needed. Domain transfer, fake and spam opinions detection, and issues related to opinionated text are needed to be handled efficiently. The highlighted challenges provide new directions in sentiment analysis both academic researchers and application industries.

References

[1] Guanghui Yu, Lei Jiang , Dong Zhou, Qian Zhou ,Qian Zhou,Xiansheng Yang ,Jujun Liu and Lei Jiang,"Predicting the Evolution of Hot Topics: *A Solution Based on the online Opinion Dynamics Model in Social Network"*, IEEE Transactions on Systems, Man, and Cybernetics:
Systems, Volume: 50, Issue: 10, 2020

[2] Deyu Zhou, Rui Wang, Jiasheng Si and Yang Yang, MingminJiang*, "A Survey on Opnion*

[3] Qiao Liu, Lan Wei, Yin Jia, Xiaolu Fei,
Pengyu Chen, Beier Zhao and HairongLv *"Automatically Structuring on Chinese Ultrasound Report of Cerebrovascular Diseases via Natural Language Processing*", IEEE Access, Volume: 7, 2019

[4] Gancheng Zhu, Shujie Zhang, Jun Wang,
Lihui Zhang, Gancheng Zhu, Xin Fang, Xiangping Zhan, Weixuan Meng and Peng Wang *"Assessment of Career Adaptability: Combining Text mining and Item Response Theory Method"*, IEEE Access, Volume: 7, 2019

[4] Desheng Dash Wu,Rui Ren and Tianxiang Liu, *"Forecasting Stock Market Movement Direction Using sentiment analysis and Support Vector Machine",*IEEE Systems Journal,
Volume: 13, Issue: 1, 2019

[5] Alexey D. Varlamov,Ruslan V. Sharapov, Ekaterina V. Sharapova, *"Method for Sentiment Text Analysis based on Statistical and Semantic Properties of Words*", 2019 International Russian Automation Conference (RusAutoCon), 2019

[6] Azreen Azman, ShyamalaDoraisamy, Eissa M. Alshari, Norwati Mustapha and Mostafa Alkeshr*" Effective Method for sentiment Lexical Dictionary Enrichment Based on Word2Vec for sentiment analysis"* ,2018 Fourth International Conference on information retrieval and Knowledge Management (CAMP), 2018

[7] TomoakiOhtsuki, Mondher Bouazizi, "*Multi-class sentiment analysis on twitter: Classification performance and challenges"*, Big Data Mining and Analytics, Volume: 2, Issue: 3

[8] Uliniansyah, Gunarso, Made Gunawan,Agung Santosa, Hammam Riza, Elvira Nurfadhilah *"Development of text and speech corpus for an Indonesian speech-to-speech translation system*"2017 20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA), 2017

[9] Penubaka Balaji and D. Haritha, O. Nagaraju, *"Levels of sentiment analysis and its challenges: A literature review"*, 2017 International Conference Big data Analytics and Computational Intelligence (ICBDAC), 2017

[10] Gabriella Pasi, Marco Viviani and Julien Fontanarava, *"Feature Analysis for Fake Review detection through supervised Classification",* 2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA), 2017

[11] Christos Troussas, AkriviKrouska, Maria Virvou"*The effect of preprocessing techniques on Twitter sentiment analysi"*, 2016 7th International Conference on Information, Intelligence,
Systems & Applications (IISA), 2016

[12] KonpusitKaewmak, Chakrit Pong-Inwong, *"Improved sentiment analysis for teaching evaluation using feature selection and voting ensemble learning integration"*, 2016 2nd IEEE International Conference on Computer and Communications (ICCC), 2016

[13] RajashreeShedge, Sneha Pasarate *"Comparative study of feature extraction techniques used in sentiment analysis "*, 2016 International Conference on Innovation and Challenges in Cyber Security (ICICCS-INBUSH), 2016

[14] , P C D. Kalaivaani, P. *Bharathi "Sentiment classification using weakly supervised learning techniques"* International Conference on Information Communication and Embedded Systems (ICICES2014), 2014

[15] Ali A. Ghorbani, Mostafa Karamibekr "*Verb Oriented sentiment classification"*, 2012 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology, Volume: 1, 2012