THE CHALLENGES OF SENTIMENT ANALYSIS ON SOCIAL WEB COMMUNITIES

Osamah A.M Ghaleb[1],Anna Saro Vijendran[2]

[1]Ph.D Research Scholar, Department of Computer Science, Sri Ramakrishna College of Arts and Science,(India)

[2]Dean, School of Computing, Sri Ramakrishna College of Arts and Science, (India)

ABSTRACT

With the explosive growth of social platforms on web including blogs, products review sites, forums, Twitter and Facebook, millions of users daily share and exchange their opinions about different issues like products, events, persons or organizations on these sites. Sentiment analysis on social users' data considered as a valuable analysis for automatically extract people opinions regarding some interested topic issues which enables to provide important information for informed decision making in different domains. With the noticed importance of sentiment analysis on social sites many applications and techniques are available. Although, performing such analysis efficiently is not travail task which can be done easily. There are a number of challenges related to sentiment analysis which needs to address and resolve. In this paper the most important challenges of sentiment analysis on social sites were highlighted and discussed with the aim to provide new directions for the interested researchers and industries by handling theses challenges and performing sentiment analysis efficiently.

Key Words: Natural Language Processing (NLP), Sentiment Analysis, Social Media, Opinion Mining

I.INTRODUCTION

Sentiment analysis is an interdisciplinary research field which depends on techniques from Natural Language Processing (NLP), text mining, machine learning, statistics, and information retrieval, 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. There are many numbers of sentiment analysis and opinion mining applications and academic research studies that can perform different related tasks such as polarity classification which classify the user's sentiment or opinion into positive, negative or natural polarity; subjectivity classification which classify the document as objective document that describe real facts and not includes opinion words, or subjective document in which opinion or sentiment words are shown in the document sentence(s); another task are called features extraction which are essential task in sentiment analysis, features including Parts of Speech (POS), Opinion words, unigram, bigram, n-gram, negations, etc. such task are basically considered as NLP task which helps in extracting the important features of text and then classifying the sentiments in text.

Sentiment classification can be done using either machine learning approach (supervised vs. unsupervised techniques), or lexicon/knowledge-based approach in the need for domain knowledge for building and annotating corpus and dictionary are required which consume more time efforts comparing with machine learning approach. Large number of research studies is performed by the two approaches [12].

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 many challenges need to be highlights and handled efficiently. The reset of this paper are structured as: section 2, survey the existing research studies on social users' sentiment analysis and opinion mining challenges, then analyze it in coherent way (Table 1.) , in section 3, the most important challenges are highlighted with discussion. Finally, section 4 is the conclusion of our work

II.LITERATURE REVIEW

According to the importance of sentiment analysis in providing valuable decision making information in different domains, sentiment analysis/opinion mining is an interested research field in text mining and analytics domain, many numbers of sentiment analysis applications and academic research studies are available today and continues in growth, among of those researches some researchers have analyzed the sentiment analysis challenges of the existing researches [1] [6] [9], while others have tries to identify and resolve the unaddressed issues that related to the sentiment analysis task. In [1], author has survey a forty seven research articles, and based on two comparisons, first comparison was addressed the relationship between review structure and sentiment challenges. Second comparison was examining the importance of resolve the addressed challenges in order to improve the accuracy of sentiment analysis. Based on those two comparisons the most important sentiment challenges were highlights.

The challenges of sentiment analysis on dynamic event have been discussed by [2], using multi-class classifier they have conducting sentiment analysis on real time tweets for predicting election results, the developed model achieve high level of accuracy in predicting the results by using deep learning-based model. Other researches were addressed the challenges of multi language issues (non-English languages) [3] [4] [5] [10] [11], among of those [3] have survey on opinion mining in Hindi language and mentioned a number of challenges related to language issues when performing sentiment analysis. Arabic language is the native language for hundreds of millions people in Middle East countries and hundreds of, sentiment analysis of Arabic text also involves many challenges related to the language. In [4], authors have addressed many challenges of sentiment analysis in Arabic language social media, then they have conducting experimental study on Egyptian Arabic microblogs, they achieve reasonable accuracy level of Arabic sentiment analysis taking into consideration handling of the language based challenges. Using lexicon based model [5] have conducting a sentiment analysis on topical Chinese microblogs posts, a Webo-lexicon with representative topic words and Out-of-vacuolar (OOV) words have been constructed, and they have addressed the challenges related to post text in Chinese language with better performance accuracy. Many other researchers [6] [7] [8] [9], have been discussed the common challenges of sentiment analysis and opinion mining in general. Table 1; summarize the survey of sentiment analysis challenges in previous studies and listing the addressed challenges in each of them.

Table 1. Summary and Analysis of the Previous Studies on Sentiment Analysis & Opinion Mining

Ref. NO.	Language-related	Domain-related	Used Technique(s)	Addressed challenges
[1]	N	N	Non (Empirical study)	Huge lexicon, bi-polar, Extracting features, NLP Overheads, World knowledge, Negation, Domain dependence, and Spam and fake opinion
[2]	N	Y; Politcal	Support	Fast-paced change in dataset, Candidate-dependence, Content-related challenges (hashtags), The importance of identifying the user’s political preference, Content-related challenges (links),

		Domain	vector machine	Content-related challenges (sarcasm), Interpretation-related challenges (Sentiment Analysis versus Emotion Analysis), Interpretation-related challenges (Vote counting versus engagement counting), Location importance, and Trustworthiness-related challenges (Bots)
[3]	Y; Hindi language	N	Non	Word order, Morphological variations, Handling spelling variations, lack of resources, and co-reference resolution
[4]	Y; Arabic language	N	Sum polarity & Double polarity	Unavailability of colloquial Arabic parsers, Unavailability of Sentiment Lexicons, The need for person name recognition, and Handling compound phrases and idioms
[5]	Y; Chinese language	N	Weibo Lexicon with OOV & Propagation algorithm	Length of content in Chinese character-based language (same number of characters contain more information than English language), and Chinese word Segmentation
[6]	N	N	Non	Detection of spam and fake reviews, Limitation of classification filtering, Asymmetry in availability of opinion mining software, Incorporation of opinion with implicit and behavior data, Domain-independence, and

				Natural language processing overheads
[7]	N	N	Non	Key word selection, Sentiment is domain Specific, Multiple opinions in a sentence, Negation handling, Sarcasm detection, Implicit Opinion, Comparative Sentences, and Opinion spam
[8]	Y; German language	N	Support Vector Machine & Rule-based approach	Relevance, Target identification, Negation, Contextual information, Volatility over time, and Opinion aggregation and summarization
[9]	N	N	Non	Object identification, Features extraction, grouping synonyms, Writing style, Opinions change with time, Sarcastic and ironic statements, and Spam opinions
[10]	Y; Arabic language	N	Quantitative, Qualitative analysis & Smoothness analysis	Limited number of research in Arabic language, Morphological complexities, and dialectal varieties
11	Y; Arabic language	N	Naïve Bayes algorithm.	Different meaning for same word, Variations in lexical category, Morphological characteristics, and Vowelization or diacritization

III.SENTIMENT ANALYSIS CHALLENGES

As we mentioned early in the previous sections sentiment analysis is nontrivial task, many challenges 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.

1.1. 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 feasible and applicable in different domain. In general the analysis of big data is a challenging task due to volume, variety, velocity, variability and veracity of data, which are the main characterize the big data.

Sentiment analysis on big data are challenging by the common characteristics of big data. Following are the common sentiment analysis challenges related to big data:

1.1.1. Data Collection

Data collection is a preliminary step for any sentiment analysis task but is one of the main challenges for researchers. Benchmark data set are not available free for the interested researchers in sentiment analysis field; most of the available social user’s data are commercial. Some of social networks sites including Twitter and Facebook provide APIs for enabling data collection from their sites. Although, due to the volume, variety, velocity of big data the collection of data set through using APIs is still challenging task, since the APIs like Twitter API enables user to retrieve only 100 tweets each time, comparing to the volume of data available online regarding the selected user’s keyword/target the retrieval of relevant data from a very huge volume data using APIs is difficult task and the relevancy of the collected data set is a major issues for researches in sentiment analysis.

1.1.2. Data Preprocessing

Preprocessing is another essential task for sentiment analysis and one of major challenges in big data world. Data volume restricts the filtering of relevant data from non relevant data which may compromise the sentiment analysis results. Big data variety and velocity limiting the feature extractions which are one of critical task in preprocessing of sentiment analysis data set. Extraction of opinion words and sentences, POS tagging challenge when the volume of dataset is so huge and the data are diverse with

1.1.3. Data Storage and Analytics

Another of sentiment analysis issues in big data is the memory size required to the preprocessed dataset for analytic. With the abundant size of the data with different format storage is one of technical issues that addressed by some or advanced storage techniques. Another challenges is velocity of big data since sentiment analysis on dynamic and real time events in big data world is challenging task need to be addressed efficiently taking into consideration the people opinions are changes over time

1.2. 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, for example hundreds of people in Middle East countries are Arabic language native and sentiment analysis on Arabic social sites is critical for political and economic events. Although some of researches try 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 non-English languages sentiment analysis.

1.2.1. Lack Of Corpuses And Dictionaries Lexicon

Due to the different characteristics of non-English languages the number of other languages corpuses and dictionaries lexicons is limited comparing with English language building language-oriented corpuses and dictionaries is difficult task based on the difficulty of each language morphologies, characters but still required. More numbers of researches in other languages are needed.

1.2.2. Different Writing Style

Writing style is another issue of non-English languages when performing sentiment analysis, in some of these languages like Arabic language writing style is from right-to-left and the same word is written in different styles or format, this issue also applicable in other languages and need to be addressed efficiently

1.2.3. Different Word Meaning

This is the case when the same word has different meaning in different contexts, this also another important issue in sentiment analysis since it extends the efforts when building language-oriented lexicons and dictionaries, and it may comprise the accuracy of translation when sentiment analysis is performing by translating other languages into English language.

1.3. Domain-oriented Issues

Sentiment analysis is highly domain sensitive task in which the sentiment classification is highly depending on the domain the training data has been extracted from, where the classifier trained using training dataset from one domain is usually performs poorly when testing on test dataset from another domain. The 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 revered from one domain to another. Existing researches are trying to overcome domain dependence challenge using domain transfer [13] where small amount of training data are labeled from the new domain which is called the target domain where it used for testing the original/source domain training dataset

1.4. 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 in 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 [13] have been discussed.

1.5. Opinionated Text Related Issues

Following are the common sentiment issues related to the opinionated text and should be addressed efficiently:

- Comparative opinion
- Subjective words not expressed any opinion
- Objective words implicitly expressed opinion
- Negation handling
- Sarcasm and ironic detection

IV.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 is 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] D. Hussein, A Survey on Sentiment Analysis Challenges, *Journal of King Saud University - Engineering Sciences*, 2016. DOI: http://dx.doi.org/10.1016/j.jksues.2016.04.002

[2] M. Ebrahimi, A. Yazdavar, and A. Sheth, On the Challenges of Sentiment Analysis for Dynamic Events, *IEEE Intelligent Systems*, 32(5), 2017

[3] R. Sharma, S. Nigma, and R. Jian, Opinion Mining In Hindi Language: A Survey, *International Journal in Foundations of Computer Science & Technology (IJFCST)*, 4(2), 2014

[4] S. El-Beltagy, and A. Ali, Open Issues in the Sentiment Analysis of Arabic Social Media: A Case Study, *9th International Conference on Innovations in Information Technology (IIT)*, 2013

[5] C. Anqi, Z. Haochen, L. Yiqun, Z. Min, and MA. Shaoping, Lexicon-based Sentiment Analysis on Topical Chinese Microblog Messages, *Semantic Web and Web Science, Springer Proceedings in Complexity, pp. 333–344, Springer*, New York, NY, USA, 2013

[6] H. Rahamath, Opinion Mining and Sentiment Analysis -Challenges and Applications, *International Journal of Application or Innovation in Engineering & Management (IJAIEM), 3(5),* 2014

[7] A. Kumar, and T. Sebastein, Sentiment Analysis: A Perspective on its Past, Present and Future, *I.J. Intelligent Systems and Applications, 10(1),* 2012

[8] D. Maynard, K. Bontcheva, and D. Rout, Challenges in developing opinion mining tools for social media, *In Proceedings of the 24th ACM Conference on Hypertext and Social Media. ACM*, 2013

[9] B. Seerat, and F. Azam, Opinion Mining: Issues and Challenges (A survey), *International Journal of Computer Applications, 49(9), 2012*

[10] A. Assiri, A. Emam, and H. Aldossari, Arabic Sentiment Analysis: A Survey, *International Journal of Advanced Computer Science and Applications, 6(12),* 2015

[11] S. AlOtaibi, and M. Khan, Sentiment Analysis Challenges of Informal Arabic Language, *International Journal of Advanced Computer Science and Applications, 8(2),* 2017

[12] W. Medhat, A. Hassan, and H. Korashy, Sentiment analysis algorithms and applications: A survey, *Ain Shams Engineering Journal 5*, 2014, 1093–1113.

[13] B. Lue, Sentiment Analysis and Opinion Mining (Morgan & Claypool Publishers, 2012).