SENTIMENTAL ANALYSIS IN MASHUP LANGUAGES

[1]Jagmeet Singh, [2]Karan Mahajan
[1]M.Tech Student, [2]Assistant Professor,
Information Technology, Chandigarh Engineering College, Landran, Mohali

ABSTRACT

In this study, we have solicited material related to the work done in area of fusion and mashup languages (languages which are mixture of two or more languages) for conducting opining mining. It has been found that limited work has been done in this context. We also found that no study related to brand product opinion/review mining has been done so far as per our knowledge. Neither much work is reported on the usage of machine learning algorithms like Multinomial Naive Bayes that use dictionaries, taxonomy, ontology based on mashup language like Hinglish to support sentimental systems.

I. INTRODUCTION

Sentimental Analysis is the study that analyses people's sentiments, evaluations, appraisals, attitudes, and emotions towards entities such as products, services, organizations, individuals, issues, events, the topics, and their attributes. Sentimental analysis helps us to find view of people towards any products, news, buildings; etc. There is also need to do sentimental analysis of mashup languages. Mashup languages are the mixture of two or more languages. Across the world thousands of languages are being spoken, but, when we talk about language used in computer English is most used. An example of it is that whenever a person is chatting on mobiles/computer they use the word of their mother tongue but write in English (HINGLISH MASHUP LANGUAGE - main aa reha hoon its Hindi written in English which in English means I am coming). Due to globalization or cultural invasion of world, English language is the most dominant language in digital world. English thrives people as it incorporates other language words in it like pyjama, chapatti are words of Hindi but now also used as English words. Most of these expressions occur on social media, which is two way form of communication as in this users are allowed to interact with others, the available information. It gives variety of contents as being a two way form of communication many individuals share/exchange their ideas or contents. Twitter, Facebook etc. are social media websites. So, we can say that for text mining, social media is starting step. It is at these places opinion is build to cultivate market good for products.

Social opinion is the public opinion about a product or issue. It is the collective response of many individuals towards any problem, social issues. Social opinions can be made in three forms i.e. Positive - when public is in favor, Negative – when public is against, Neutral – neither favor nor against. It is easy to build digital campaign as one requires a platform and professional experts, which are easily available and lot of people in routine, as a part of their life share lot of things. In this process express about products. One has to make their requirements clear to professionals as they will make everything with reference to the requirements. For digital campaign - website, its promotion and security is very important. With promotion of the websites, people come to know about thing an individual is campaigning for. Security is important as hackers can misuse the website for their own benefits by harming the contents in it. It can ruin the image of the company or organization. Sometimes due to social comments or opinions, people at large may be wrongly informed that may lead to problems. Hence, care must be taken to handle such situations and problems. This can be done by building digital awareness campaigns.

Internet is free for any individual to express their views on any topic/issue. There is no restriction to write anything. It brings two sides of the coin with it as people can comment good or bad / true or

false. It is mostly up to an individual's to choose their own path. Internet democracy also helps fighting against spam's or bad opinions as public can express their views against wrong events occurring in the city, state, country or even in the world. People who have important information to fight against such crimes need not to confront the world; they can message the right individuals just by sitting at their homes only. This way the wisdom of large number of people can be harvested.

Wisdom of crowd is response of crowd, aggregating in collective opinion by a group of people than to single individual/expert. Qualities of wise crowd shall be independent opinion (opinions shall not be inspired by people around them); Decentralization (process in which people are able to specialize and extract local knowledge); Diversity of opinions (each individual have private information even if it is conspicuous or unusual of known facts); Aggregation (Making private judgments into collective responses). These concepts are essential for understanding of opinion mining as a subject. However, more on this can be found in next section, describing contemporary work done in this area.

II. ORGANISATION OF PAPER

This paper discusses the issue of extracting & processing sensible Information/Data from the social networking websites written in mashup languages using machine learning algorithm. The main purpose of our work is to do sentimental analysis for product safety in mashup languages. The related work section includes various aspects of exploring large data at different levels, automatic generated reviews, spam detection. A proposed methodology is also given with which our algorithm shall provide us the better results. Conclusion and discussion is also provided before the future work.

III. RELATED WORK

Xindong Wu, Senior Member, IEEE, et. al.[3] According to these researchers to explore Big Data, one has analyze several challenges at the data, model, and system levels. To support, Big Data mining, high-performance computing platforms are required, which impose systematic designs to unleash the full power of the Big Data. At the data level, the autonomous information sources and the variety of the data collection environments, often result in data with complicated conditions, such as missing/uncertain values. In other situations, privacy concerns, noise, and errors can be introduced into the data, to produce altered data copies. Developing a safe and sound information sharing protocol is a major challenge. At the model level, the key challenge is to generate global models by combining locally discovered patterns to form a unifying view. This requires carefully designed algorithms to analyze model correlations between distributed sites, and fuse decisions from multiple sources to gain a best model out of the Big Data. At the system level, the challenge is that a Big Data mining framework needs to consider complex relationships between samples, models, and data sources, along with their evolving changes with time and other possible factors. A system needs to be carefully designed so that unstructured data can be linked through their complex relationships to form useful patterns, and the growth of data volumes and item relationships should help form legitimate patterns to predict the trend and future, all this is applicable for sentimental analysis also.

Leon French, et. al.[5] They created and applied a system for large-scale automatic extraction of connected knowledge. By analyzing over 20,000 abstracts, they found the neuroscience literature contains a wide diversity of terms, species, and brain region names. They also found that unfortunately, this diversity exceeds that of the existing formalized neuroanatomical lexicons. They found that it is difficult to create a clear set of annotation guidelines due to this diversity that extends to sentence structure and experiment design. While this diversity limits the automatic mining of neuroscience literature, they evaluated several methods that improve automatic extraction. They found great value in general-purpose and biomedical text mining tools. They also applied these tools with little or no tuning and report robust and extendable results. This allowed for more time for extensive manual evaluation and review. In addition to tested methods, their work provided a database of evaluated connectivity statements that can be used as a starting point for manual curation and to facilitate neuroscience text mining.

Leonard Barolli et. al.[9] With the continuous increase of data, scaling up to unprecedented amounts, generated by Internet-based systems,. The core of Big Data Science is the extraction of knowledge from data as a basis for intelligent services and decision making systems, however, it encompasses

many research topics and investigates a variety of techniques and theories from different fields, including data mining and machine learning, information retrieval, analytics, and indexing services, massive processing and high performance computing. Altogether the aim in this paper was the development of advanced data-aware knowledge based systems.

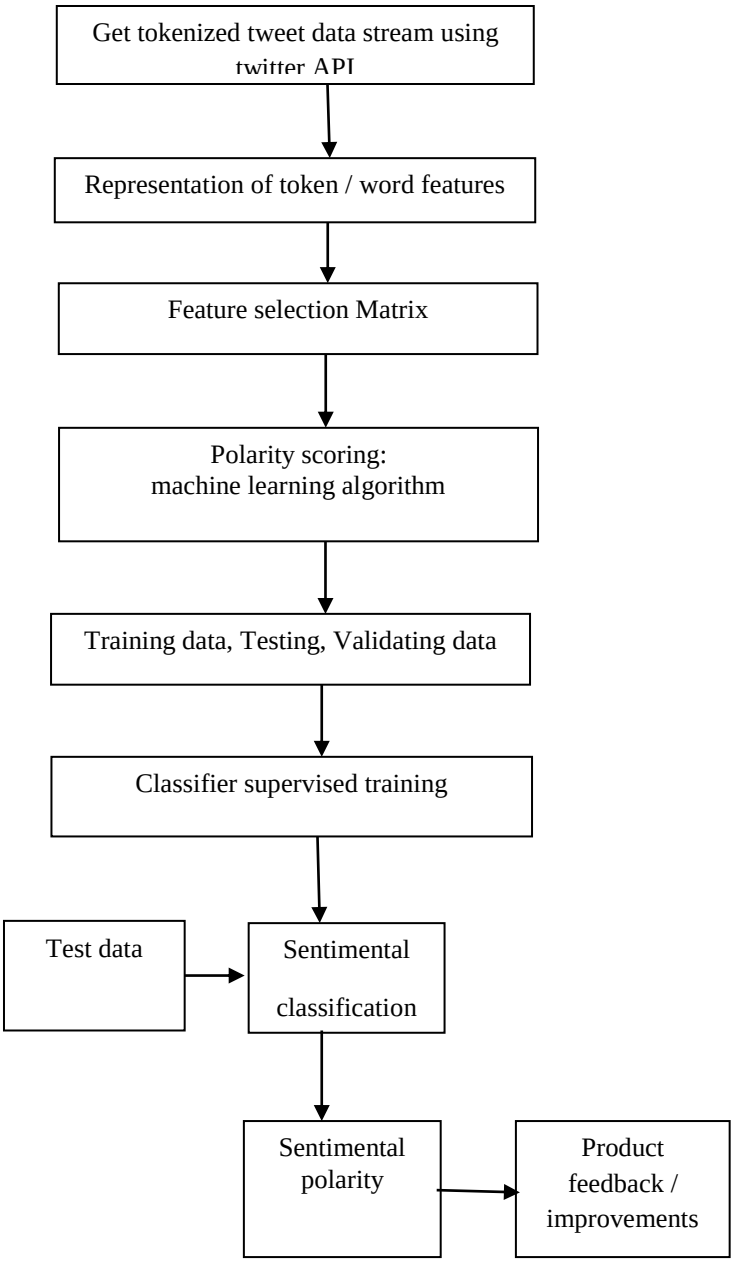
Vaishak Suresh, San Jose State University, et. al.[6] In this work, They proposed a system that automatically generated personalized reviews recommendations using two different approaches. Firstly, drawing inspiration from traditional collaborative filtering systems, the system generates user rating profiles and tailors the list of reviews to the preferences of each user. Secondly, they employed aspect-based opinion mining to identify the important features highlighted in each review. The amount of online reviews for products and services has grown to such extend that often makes it impossible to read all of them. In this work, they also proposed a system that personalizes the order in which the reviews were shown and provides an intuitive interface that allowed the users to see the important aspects of each review in a glimpse.

Kaipeng Liu, Yu Zhang, Harbin, China.[2] Their paper has demonstrated an approach to detect tag spam in social tagging systems. They uncovered the underlying tag scheme that is already presented in the way collaborative users are annotating resources with tags. This collaborative knowledge is then used to measure the quality of posts and tell spam posts from legitimate ones. An iterative spam detection algorithm is developed to identify spam posts by their information value. This method can be also extended to user level for detecting spammers. They have done many experiments on data set collected from real-world system, and the experimental results showed a convincing performance of their algorithm.

IV. RESEARCH GAP

Limited work has been reported in recent contemporary research studies related opinion mining specially, in context of languages like 'Hinglish' which are mashed up languages. The context written in Hinglish has not been used to find polarity classification of branded products as such. In this field of work limited efforts has been made especially when it comes to usage of machine learning algorithm for classification of products reviews/opinions. In this tweets written in Hinglish (Hindi written in English) can also be extracted to gain valuable information of branded products. With it, sentimental analysis of product can be done; which tells what does people think about it, what can be added or removed to a product? So that product/service providers may improve overall.

V. PROPOSED METHODOLOGY



Figure

Typical steps used in polarity analysis

The proposed model [Figure] may be implemented using the WEKA [10] machine learning tool and Twitter API. The product opinions will be collected from social media's application programming interface of Twitter/Facebook. Irrelevant words need to be removed which leads to tokenization and vectorization text. For sentimental analysis of the product safety ranking/classification algorithm for polarity scoring may be used to classify the sentiments in positive, negative and neutral form as often conducting systematic survey. The classification algorithm must be able to handle direct text sequence or vectorized sequence; along with probabilistic models like multinomial naïve bayes.

VI. CONCLUSION & DISCUSSION

In summary, we can say that the sentimental analysis is a field that also needs to consider the mashed up languages for its growth. The researchers in this field have limited their research to mainstream world languages; they have built dictionaries, taxonomies, ontologies etc, based on it. Infact, no work is reported where mashup colloquial is used for finding polarity of products and issues.

VII. FUTURE WORK

For future work, we suggest that machine learning algorithms may be used for text classification task, especially by the use of probability based algorithms like Multinomial Naive Bayes . Since, they are more suitable for Mashup Language text processing. The reason, why we are suggesting this is because the languages like Hinglish do have defined structure and grammar; therefore a probability based algorithm will be best for classification of polarity of comments /opinions and remarks.

REFERENCES

[1] Haruna Isah, Paul Trundle, Daniel Neagu, " *Social Media Analysis for Product Safety using Text Mining and Sentiment Analysis"* 2014 IEEE, Artificial Intelligence Research (AIRe).
[2] Kaipeng Liu, Binxing Fang, Yu Zhang, "*Detecting Tag Spam in Social Tagging Systems with Collaborative Knowledge***"** 2009 IEEE, Sixth International Conference on Fuzzy Systems and Knowledge Discovery.
[3] Xindong Wu, Xingquan Zhu, Gong-Qing Wu,and Wei Ding, "*Data Mining with Big Data*" 2014 IEEE, Transactions on Knowledge and Data Engineering.
[4] Joseph S. Kong, Behnam A. Rezaei, Nima Sarshar, and Vwani P. Roychowdhury, "*Collaborative Spam Filtering Using E-Mail Networks" 2006 IEEE* Computer Society.
[5] Leon French, Po Liu, Olivia Marais, Tianna Koreman, Lucia Tseng, Artemis Lai and Paul Pavlidis, May 2015, "*Text mining for neuroanatomy using WhiteText with an updated corpus and a new web application" University of Illinois, USC Information Sciences Institutes, USA.*
[6] Vaishak Suresh,Syeda Roohi, Magdalini Eirinaki, "*Aspect-Based Opinion Mining and Recommendation System for Restaurant Reviews***"** 2014, San Jose or vicinity, CA, USA.
[7] Anirban Mukhopadhyay, Ujjwal Maulik, Sanghamitra Bandyopadhyay and Carlos A. Coello. "*Survey of Multiobjective Evolutionary Algorithms for Data Mining*"2014 IEEE Transactions on Evolutionary Computation.
[8] Yuchun Tang Sven Krasser Paul Judge, Yan-Qing Zhang, "*Fast and Effective Spai Sender Detection with Granular SVM on Highly Imbalanced Mail Server Behavior Data* " 2006 IEEE.
[9] Fatos Xhafas Leonard Barolli, "*Semantics, intelligent processing and services for big data*"2014, Future Generation Computer Systems.
[10] Mark Hall Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer Peter Reutemann, Ian H. Witten, "*The WEKA Data Mining Software: An Update*",SIGKDD Explorations.

AUTHORS

Karan Mahajan is currently working as assistant professor at Global Institutes, Amritsar (Punjab). Mr. Karan Mahajan completed his post graduation Lovely Professional University, Phagwara in 2012 and he keeps interest in Data Mining and machine learning.

Jagmeet Singh completed graduation from D.I.ET, Kharar in 2012. And currently pursuing post Graduation at Chandigarh Engineering college, Landran, Punjab technical university, Punjab, (India) in 2013 and 2015.