A New Framework for Social Media Content Mining and Knowledge Discovery

Prashant Bhat
Asst Professor, School of Computational Science & IT, Garden City University, Bengaluru

Pradnya Malaganve
Research Scholar, School of Computational Science & IT, Garden City University, Bengaluru

Prajna Hegde
Research Scholar, School of Computational Science & IT, Garden City University, Bengaluru

ABSTRACT

Social media has come up with many popular websites such as Facebook, Twitter, Instagram, LinkedIn etc for the use of the generation to share each other's views. Social Media Content Mining is the process of extracting useful information i.e. Text, Video, Audio, Images from the Web by applying Data Mining techniques such as classification, clustering, regression, Outlier Detection and association rules etc can be applied to discover knowledge from web data. This paper presents some existing social media content mining techniques and proposed a new approach for efficient Data Mining frame work to extract useful knowledge from the web data.

Keywords

Data Mining, Multimedia, Data Classification, Data Clustering, Outlier Detection

1. INTRODUCTION

Social Media Web has enormous amount of information and it will proceed with the increased size and complexity. It is a biggest task to search relevant information from such large amount of data. In recent years, Social Media such as Facebook, Twitter, LinkedIn and Instagram etc have become popular platforms to share the Multimedia data (Text, Image, Audio and Video) among others. To mine web data we use some methods such as Classification, Clustering, Outlier Detection and Association rules etc. In recent years many works have been implemented to discover knowledge from web Multimedia using Data Mining techniques and tools[15] such as Unsupervised Classification(Sentiment Lexicon, Opinion definition and Summarization, Sentiment Orientation, Opinion Extraction and Basic Clustering Techniques), Semi supervised Classification and Supervised Classification(Support Vector Machine, Naïve Bayers, Neural Network, K-nearest Neighbours, Decision Tree, chi-square automatic interaction selection, Text Mining) As the count of 2018, Facebook had 2.23 billion monthly active users make it the first social network ever. Active users are those who have logged into Facebook during the last 30 days. And many useless or waste information have been stored on social media. Hence using social media mining techniques such as Data Classification, Data Clustering, Outlier Detection and Association rule we can extract useful information. A generalized Framework for Social Media Content Mining is shown in the figure1.

2. PRIOR WORK

The omnipresence of the internet is with media of different types such - 'Sports', 'Entertainment', 'Politics and news' etc. Nowadays web Multimedia-video classification has got raised attractiveness. It is the combination of different components of Multimedia such as image, text, audio, video etc.The metadata can be classified based on video bit rate kbps, maximum bit rate kbps, width pixels, height pixels etc. Authors Siddu P. Algur, Basavaraj A. Goudannavar, Prashant Bhat[1] have introduced Decision Tree and Support Vector Machine (SVM) approach for classification process. The metadata has been extracted from the Multimedia components and the same will be stored in the database for experiment. Based on the number of components present in the domain, the web Multimedia data are labelled and that will be classified as KDD process.

In recent years, social network marketing has gained extremely large popularity. As consideration, Facebook and Twitter etc are attracting the web-based companies in such a way that all these companies are focusing their marketing strategies on such social network platforms. Even though social network marketing has got a huge success, little or nothing is known about how well such social network-based marketing campaigns perform. To contribute in this field of research, Authors Christoph Trattner and Frank Kappe[2] have presented the results of an ad-driven social network-based marketing campaign centred on Facebook. And demonstrated which kind of ads generated by the Facebook tools and applications can make more counts of visits and ROI of a web-based platform i.e. VirWoX. After finding these things, Authors[2] have presented an analysis of simple real-time measures to find out the most 'valuable' users on Facebook.

Because of connecting the people all over the world, the Social Media has become more popular from years and it has given the flexible platform to share the important information, but it has also become a way for people who misuse it to humiliate others. To overcome from such issues, Authors Niloofar Safi Samghabadi, Suraj Maharjan, Alan Sprague, Raquel Diaz-Sprague| and Thamar Solorio[3] have presented NLP approach to find out reproachful posts as an initial step to eventually discover and dissuade cyber humiliation. First step is to collect the data containing abusive language and then finding whether or not the collected data is containing dissuade. Then comments or metadata added to this data are improved by in-lab annotations and crowd sourcing. Authors[3] have followed some NLP approaches containing typical and new techniques to classify the use of swear words, which people have used in an abusing way. Authors[3] confirms that, the introduced model not only works for their chosen data set, but also can be applied to different data sets.

Named Entity determining for social media data is challenging because of its inherent noisiness. With that to improper grammatical structures, it includes spelling inconsistencies and many informal abbreviations. Authors Gustavo Aguilar, Suraj Maharjan, A. Pastor L´opez-Monroy and Thamar Solorio[4] have proposed a novel multi-task approach by employing a more general secondary task of Named Entity (NE) segmentation together with the primary work of fine-grained Named Entity categorization. The multi-task neural network architecture learns higher order feature representations from word and character sequences along with basic Part-of-Speech tags and gazetteer information. This neural network acts as a feature extractor to feed a Conditional Random Fields classifier.

Using Social media websites people can communicate and share their thoughts with each other through different tools such as comments , chats, discussion forums etc. The nature of these information on social media websites can be categorized as unstructured and fuzzy. In regular day-to-day discussions, spellings, grammar and sentence structure are usually neglected. This may cause various types of ambiguities, for example, lexical, syntactic, and semantic, which makes it difficult to analyse and extract data patterns from such datasets. Authors Said A. Salloum , Mostafa Al-Emran and Khaled Shaalan 's[5] study aims at analyzing textl data from Facebook and attempts to find interesting knowledge from such data and represent it in different forms.

In these years many people have affected by natural disasters such as floods, earth quake, hurricanes, blizzards, tsunami etc. By keeping these things in mind, the Authors Zahra Ashktorab, Christopher Brown, Manojit Nandi and Aron Culotta Culotta[6] have proposed Tweedr- "Twitter for Disaster Response". The aim of this tool is to extract information which is related to first responders from tweets generated during and after the disaster has occured. The Tweedr consists of three main parts: classification, clustering and extraction. The classification phase has a variety of classification methods (sLDA, logistic regression and SVM) to classify the tweets as disaster damage or casualty information. The clustering phase is used for filtering and to merge tweets which are similar to each other. And extraction phase is used for extracting tokens and phrases to report specific information about different classes of infrastructure damage, damage types, and casualties.

Applying Data Mining (DM) in the field of education is an interdisciplinary research also known as educational Data Mining (EDM). It is concerned with developing methods for exploring the unique types of data that come from educational environments. Authors Cristobal Romero and Sebastian Ventura's[7] aim is to understand students way of learning and find out the settings in which they learn to improvise educational outcomes and to pursue insights into and describe educational phenomena. Educational information systems can store huge amount of informational data from many sources which can be in different formats and at different levels. Each and every educational problem has a particular objective with special characteristics that require a different treatment of mining problem. It means that traditional techniques of Data Mining cannot be applied directly to these types of data and problems. The process of Knowledge discovery has to be adapted and some specific Data Mining techniques are needed.

Big data and Data Mining methods have attracted people's attention widely in information industry in recent years, due to the availability of large amounts of data and the quick need for turning such information into useful knowledge. It affects different company dynamics in many sectors, mainly social media services have become much important for the marketing and Crime departments of companies. In such way, communication has been established with the customers and the use of Big Data in these fields is seen as the important step of the companies to become a big brand. Methodology used by the Authors Umman Tugba Gursoy, Diren Bulut and Cemil Yigit[8] is Social Media Mining and Sentiment Analysis.

Social networks are the vast used communication media nowadays. As of the second quarter of 2018, the micro-blogging service averaged at 335 million monthly active users. Twitter is a social networking and microblogging service, enabling registered users to read and post their short messages, which are called tweets. These tweets contain the users thoughts and can get reply from other users of tweeter. Hence its necessary to analyze these tweets, identify and classify trends of different users. . Authors Mashael Saeed Alqhtani, M. Rizwan Jameel Qureshi's[9] goal is to classify social network to anomaly groups such as: Terrorist and dissident; by analyzing tweets data on the Twitter and identify an anonymous user's affiliation to these groups. Author[9] has used different Data Mining techniques to extract the data such as: Text mining, sentiment analysis, and opinion mining. Extracted features and characterized groups will be then stored in the database. . The objective of data extraction is to measure the similarity of selected user tweets with respect to extracted features. It will enable to determine high percentage of similarity between the user tweets and group characteristics to expose his/her affiliation to this group.

In all sectors of society, Internet has found its place. Most of the companies wants/needs to do marketing on social media network to introduce their products and services. In several cases, people find it boring to watch advertisements on social media so they may ignore it. Such advertisements might be considered as spam. Taking this into consideration, Authors Saman Forouzandeh, Heirsh Soltanpanah and Amir Sheikhahmadi[10] have made a research where user's interests, attitudes, and behavior on Facebook are specified through data mining techniques, based on which content marketing is conducted. Author's[10] study is conducted on the social network of Facebook, where content marketing, a new form of marketing, is utilized and instead of introducing the goods, the contents of different goods are presented. The data utilized in the study are actual and related to Facebook users.

3. PROPOSED SOCIAL MEDIA CONTENT MINING FRAMEWORK

The proposed Framework in the figure1 indicates the Data Mining process for Multimedia data available in different Social Media such as Facebook, Twitter, LinkedIn, Instagram etc. The data will be collected from different datasets Which contain the information of Multimedia data. Then the data will be extracted using several data extraction methods such as Data Classification(Supervised, Semi supervised and Unsupervised), Data Clustering, Outlier Detection, Regression and Association rules etc. Mining result will be evaluated and knowledge discovery will be made.

3.1 Web

Billions of people are connected with a sing source i.e. Social Media. The storage of Multimedia is getting increased day by day and it continues the same. Large datasets will be generated by social media resources such as Facebook, Twitter, LinkedIn and Instagram etc.

3.2 Resource (dataset) Discovery

Extracting Information from datasets is the task of automatically extracting structured information from unstructured or/and semi structured documents

3.3 Data Repository

Different types of information extracted from Social Media and stored in Information Repository with respect to particular text, video, image, audio.

3.4 Pre Processing the Data

Most of the times, real-world data will be incomplete, inconsistent and may contain errors hence Data preprocessing involves some techniques such as, Data Selection, Data Cleaning, Data Handling etc. These Data Mining techniques transform raw data into an understandable format.

3.5 Data Mining Process

By using different methods such as Classification, Clustering, Regression, Outlier detection and Association rules, the data is mined.

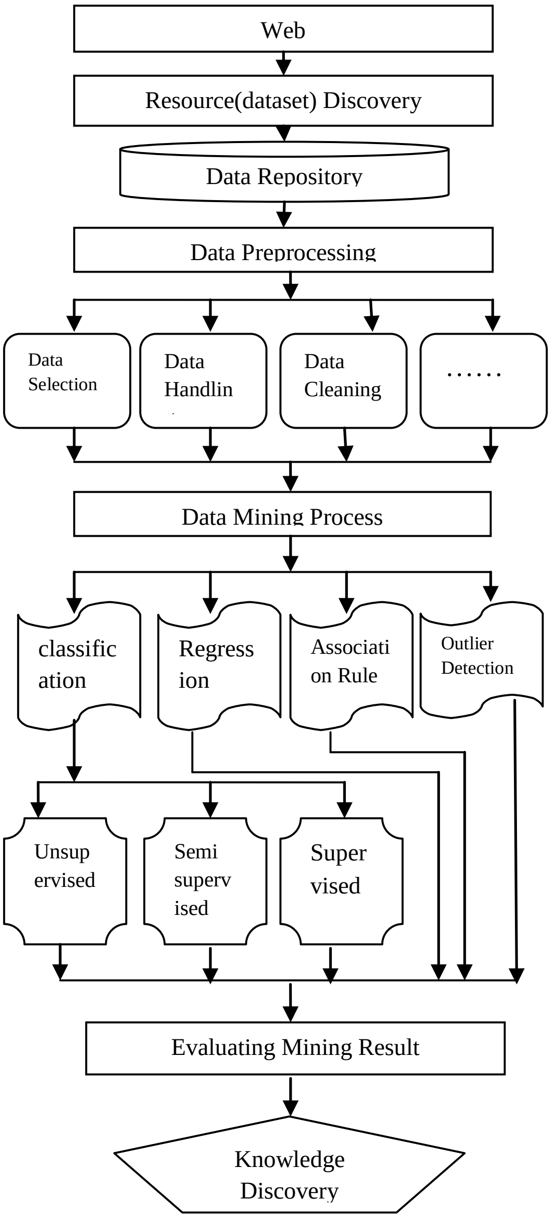

Fig1: Proposed Framework for Social Media Content Mining

3.6 Classification/Clustering

Classification is one of the data mining technique used for prediction analysis. Classification analyses the data into discrete categories. The data uploaded on to Social Media belongs to different categories such as sports, entertainment, politics, education, Food, medical etc. Sports category contains the information about several sports such as cricket, hockey, football, Kabbaddi etc. Entertainment category contains all the movies, videos, audios etc. Politics category contains the recent and previous updates regarding all politics parties. Education category contains information about several kind of courses and degrees etc. Food category contains different types of dishes and its recipe and also other related information. Medical category contains about psychology, physiology, anatomy etc. These all things are differentiated using Classification method .Sever classification methods are:

Unsupervised Classification(Sentiment Lexicon, Opinion definition and Summarization, Sentiment Orientation, Opinion Extraction and Basic Clustering Techniques), Semi supervised Classification and Supervised Classification(Support Vector Machine, Naïve Bayers, Neural Network, K-nearest Neighbours, Decision Tree, chi-square automatic interaction selection, Text Mining)

3.7 Regression

To predict a range of numerical or continuous values from a particular dataset, Regression method is used. Many industries, companies, marketing sectors, business and finance sectors are using Regression for predicting numeric values.

3.8 Outlier Detection

Something that is situated away from or classed differently from a main or related body. A statistical observation that is marked different in value from the others of the sample.

3.9 Association Rule

Is a data mining technique which is used for finding frequent patterns, structure, correlation and association structure from the dataset i.e. to determine which kind of data is mostly or most of the time related or got together with which another data. Example Shampoo and Conditioner are brought together by most of the customers.

3.10 Evaluating Mining Results and Knowledge Discovery

Using data mining methods such as Classification, Clustering, Outlier detection, Regression and Association rule, mining results are evaluated.

4. CONCLUSION AND FUTURE WORK

In this paper we reviewed some of existing techniques for Multimedia Mining and observed that there is no proper Framework for Multimedia Mining. This paper explains the efficient Framework for knowledge discovery from Multimedia.

5. REFERENCES

[1] Siddu P. Algur, Basavaraj A. Goudannavar, Prashant Bhat "Web Multimedia Classification Using Dt and Svm Models". Web Site: www.ijettcs.org Email: editor@ijettcs.org Volume 5, Issue 4, July - August 2016

[2] Christoph Trattner, Frank Kappe" Social stream marketing on Facebook: a case study". International Journal of Social and Humanistic Computing · March 2013

[3] Niloofar Safi Samghabadi, Suraj Maharjan, Alan Sprague, Raquel Diaz-Spragu and Thamar Solorio ” Detecting Nastiness in Social Media”. Proceedings of the First Workshop on Abusive Language Online, pages 63–72, Vancouver, Canada, July 30 - August 4, 2017. Association for Computational Linguistics

[4] Gustavo Aguilar, Suraj Maharjan, A. Pastor L´opez-Monroy and Thamar Solorio “A Multi-task Approach for Named Entity Recognition in Social Media Data”. Proceedings of the 3rd Workshop on Noisy User-generated Text, pages 148–153 Copenhagen, Denmark, September 7, 2017. Association for Computational Linguistics

[5] Said A. Salloum , Mostafa Al-Emran and Khaled Shaalan “Mining Social Media Text: Extracting Knowledge from Facebook”. IJCDS Journal · March 2017

[6] Zahra Ashktorab, Christopher Brown, Manojit Nandi and Aron Culotta “Tweedr: Mining Twitter to Inform Disaster Response”. Proceedings of the 11th International ISCRAM Conference – University Park, Pennsylvania, USA, May 2014 S.R. Hiltz, M.S. Pfaff, L. Plotnick, and P.C. Shih, eds.

[7] Cristobal Romero and Sebastian Ventura “Data mining in education” Volume 3, Januar y / Februar y 2013

[8] Umman Tugba Gursoy, Diren Bulut and Cemil Yigit “Social Media Mining and Sentiment Analysis for Brand Management”. Global Journal of Emerging Trends in e-Business, Marketing and Consumer Psychology (GJETeMCP) An Online International Research Journal (ISSN: 2311-3170) 2017 Vol: 3 Issue: 1

[9] Mashael Saeed Alqhtani, M. Rizwan Jameel Qureshi “Data Mining Approach For Classifying Twitter’s Users”. International Journal of Computer Engineering & Technology (IJCET) Volume 8, Issue 5, Sep-Oct 2017, pp. 42–53, Article ID: IJCET_08_05_006

[10] Saman Forouzandeh, Heirsh Soltanpanah and Amir Sheikhahmadi, “Content marketing through data mining on Facebook social network”. Volume 11, June 2014

[11] Siddu P. Algur, Prashant Bhat*, Suraj Jain, “The Role of Metadata in Web Video Mining: Issues and Perspectives.© International Journal of Engineering Sciences & Research Technology, 2015

[12] https://hackernoon.com/what-steps-should-one-take-while-doing-data-preprocessing-502c993e1caa

[13] https://www.lifewire.com/regression-1019655

[14] https://ieeexplore.ieee.org/document/4739542

[15] Thabit Zatari, “ Data Mining in Social Media”. International Journal of Scientific & Engineering Research, Volume 6, Issue 7, July-2015 152 ISSN 2229-5518