e-ISSN:2582-7219

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH

IN SCIENCE, ENGINEERING AND TECHNOLOGY

Volume 6, Issue 5, May 2023

Impact Factor: 7.54

6381 907 438 6381 907 438 ijmrset@gmail.com @ www.ijmrset.com

Social Media Content Processing Issues

Dr. Parul Verma

Assistant Professor, Amity Institute of Information Technology, Amity University, Lucknow, India

ABSTRACT*:* Social media these days is overflowing with thoughts. People take social media as a platform to express their personal and social views. Nowadays with such a busy lifestyle sometimes we have nobody to interact with and in those situations everyone find it easy to share their views on social media. SocialMedia content is nowadays being used for many applications like Opinion Mining, Sentiment Analysis, Observing mental state of a person and many more. However it is not an easy task to performprocessing of Social Media content. It involves various steps and each and every step involves some complications. The paper will discuss steps involved in processing of social media content. The paper will also focus on various challenges involved at each step of processing.

KEYWORDS: Natural Language Processing, Social Media Content, Feature Selection, Feature Extraction, Sentiment Analysis, Opinion Mining

I. INTRODUCTION

Nowadays people express themselves on social media. They do not have time to socialize and meet friends. So the easiest available way to share their views, ideas and feelings is the social platform. Due to such lifestyle where people don’t have time to interact with their friends and relatives, because of many reasons lot of negative energy and thoughts are generated which leads to depression, anxiety and many more mental sicknesses.

Such kind of mental sickness is a curse to our human being. During a survey in 2010 by WHO (World Health Organization), it has been mentioned that 350 million of people are suffering from depression worldwide. Depression leads to the suicidal attempts and most of the victims harm themselves.

Identifying mental state of person by processing its posts on social media can help lot many people who are suffering from mental sickness. It can help people by alarming their friend and relative that their dear ones are in need of help or counseling.

Keeping in mind such alarming situation researchers have been working to find out mental state of people by observing their posts on social media. The kind of vocabulary they use, the thoughts they share need to be processed to identify their mental state. It is observed that people suffering from any such kind of negative mental state; do not share their views and problems openly. So the need is to automate the process of identification of such persons who are undergoing or facing these situations and society, friends and their relatives can help them.

It is a cumbersome task to identify such victims by manually going through thousands of posts. Sometimes person is in a very bad mental state and failing to identify his or her mental state at the right time will lead to irrecoverable disaster. Hence it is a challenging task to identify such victims and help them promptly. The processing of social media is not an easy task. It has lot of complications in it.

Besides it there are many other applications that rely too much on social media content. Be it a sentiment analysis, opinion mining or Analysis of various unstructured data all these applications nowadays are dependent on social media content. The social media content is unstructured in nature hence lot of complications arise while handling that data. Rest of the paper will discuss about literature survey, Steps for processing of Social Media Content and Challenges for processing of Social Media Content

II. LITERATURE SURVEY

Various researchers have worked in the field of processing social media contents posted in natural language and put their efforts in drawing various inferences from it. This section will brief the work of such researchers. Many researchers have worked in this filed and have got success in processing persons thoughts on social posts by using NLP techniques and drawing inferences from it.

Overflowing social media content has opened up new opportunities for the analysis of the content and drawing some patterns or inferences from it. Social media data can be analyzed for various purpose like - gain insights into issues, trends, influential actors and other kinds of information. Twitter data has been analyzed by Golder and Macy [1] to study how people's mood changes with time of day, weekday and season. Social media data can also be used by Information Systems to study various questions like the influence of network position on information diffusion [2]

Social media nowadays is a platform where everyone can express their views to anyone. Not only business experts or critics can give their views any person can share his/her views regarding anything. [3]. Opinion Mining is a new stream of research where enterprises and business organizations are more and more dependent on the surveys and opinion polls [4] Sentiment analysis is not considered as a new research stream lot of work has been already done in the earlier year[5, 6,7, 8, 9]

With the popularity of internet and smart mobile devices now huge population is on web and they are ready to share and express their views on social media which is being used for sentiment analysis. In the former years also various researchers have showcased their work in this field which is related to the interpretation of metaphors, sentiment adjectives, subjectivity, viewpoints, affects and related areas. [10, 11, 12,13, 14, 15, 16, 17, 18]. Due to several factors like rise in machine learning in NLP processing, training datasets access for machine learning techniques and huge applications in versatile industries has promoted rapid growth of sentiment analysis

III. NATURAL LANGUAGE PROCESSING AND SOCIAL MEDIA CONTENT

Natural Language Processing is one of techniques of Artificial Intelligence. It is quite obvious that people express themselves in natural language. For humans it is easier to understand it but when those views are posted on some social forum processing it is a very difficult task. For the processing of such views we need to exploit Natural Language Processing Techniques. These techniques require various steps for processing. The processing of the posts of users on these social sites include following steps –

a) Data collection and accessing from social media
b) Preprocessing of textual data
c) Presentation of Data
d) Knowledge Extraction

A. Data Identification from social media

First step of this processing is to collect data from social networking sites. One can use various API's of respective social networking sites to collect data from these sites which are posted by their users. Following API's are available to collect data-

Twitter API

Twitter API is being introduced in three variations: Standard, Premium and Enterprise. Standard API is available free of cost and is public in nature. It can be used to access only last 7 days tweets. Premium is a mix of free and paid where if one demands to access last 30 days post then it is free otherwise if one wants to access posts since inception of Twitter than it is not free of cost. Enterprise one gives access to complete functionality of Twitter API by providing direct account management support. It serves JSON data, supporting POST requests with JSON data bodies.

Facebook API

Facebook API also provides access to the public posts of the users. “Public Feed API” of Facebook gives you the facility to access the status updates that are posted with the privacy setting as public. The feed isn’t available via an HTTP API endpoint, instead updates are sent to your server over a dedicated HTTPS connection. The feed only includes basic data about the given post.

B.Pre-processing of textual data

The contents posted by the end-users in natural language needs to be processed first. There are some traditional steps for the pre-processing of text which are as follows-

Stop Words Removal

In NLP there are some words like a, an, the, is etc. are considered as stop words. Any application based on NLP techniques removes the stop words from the contents and then it proceeds to the actual processing. Usually any application like Text Classification, Disambiguation, Clustering, Searching of Text ignore these stop words while main processing task. This is the reason they need to be removed. Number of stop words varies from language to language. There are some utilities like IBM stop word which gives you the facility to add or modify the list of stop words as per language and then update stop word dictionaries by using the stop word tool. Cleanse stop words is also used to filter out stop words before or after of the text processing. Besides this on can write its own script for Stop Words Removal using nltk of Python, Java language or in any language that supports Unicode text processing to support Natural Language Processing.

Tokenization

Breaking the contents in a form of tokens is called tokenization. The boundary of the words needs to be identified. Token are segregated by some characters like space in Hindi, English, French. However there are some language which does not have boundary characters like Chinese. In such languages tokenization is quite difficult. Languages which are morphologically rich also face some challenges while tokenizing.

Stemming

The process of reducing words to its root form is called Stemming. Words used in different inflectional forms in many documents. For example – organizing, organized, organization all are variants of root word “organize”. The necessity to convert words to its root form is to relate semantically or search in dictionary for many NLP tasks. Words usually exist in their root form in dictionaries or lexical database.

C. Presentation of Data

The pre-processed data in NLP applications are presented using various models. The data posted on the sites need to be presented in a certain fix pattern. These patterns are called models. The most generalized form of presenting data is in the form of numeric vectors on which one can perform linear algebra operations. Following are the general models that are used for presenting documents-

- Bag of Words
- Vector Space Model
- Neural Networks

D. Knowledge Extraction

NLP allows us to extract useful information from text. NLP has a great potential. It is just not being used only for the purpose of Machine Learning or Translation but it’s usage scope is quiet wide and can be utilized to judge the mental state of a person. Be it a speech or posts on social networking sites. People convey their emotions in many ways. Most of the time we are not able to understand it directly, however we need to do some processing to identify the sentiments of a person. Besides that content posted on social media for different domain can be analyzed in order to draw useful inferences like popularity of product based on knowledge extraction from review of the products, review of movies or some political issue as well.

IV. CHALLENGES FOR PROCESSING OF SOCIAL MEDIA CONTENTS

A. Language

There are different languages spoken by people in different countries, states or regions. In India itself there are 22 different major languages that are being used by people in different parts of the country. Every language has its unique morphological structure. The morphology of the language makes it complicated to be processed by machines. There are many languages which are morphologically very complicated for example- Hindi, Sanskrit and Chinese to name a few. Nowadays people are in practice to post their views on social media in their native language. To automate the processing of such contents understanding the morphological aspect of these languages is the biggest hindrance. Automated system to process social media content needs to be trained for different languages spoken by the people all over the world. Hence complex language structure is one of the barriers in implementation of automated systems.

B. Data Collection

The processing of social media content is in fact a challenging job because of the amount of content generated by the massive number of users by passing every minute. Twitter boasts 316 million monthly active users with 500 million tweets per day. (Source: https://wasimahmed.org/page/5/)

The biggest challenge is at the ethical end. For reproducing and processing of the contents the researchers need to acquire consent of the users.

At legal end Twitter's API Terms of Service prohibits researchers to share their datasets. However tweet identification number can be further utilized by other researchers to obtain Twitter datasets.

Retrieval of dataset is limited due to usage of certain keywords which may not retrieve all the data related to a topic.

Data Retrieval from Twitter is pretty costly as using the free API ecosystem one can only access 7 days back data. Hence lot of time is required to collect data and do further processing.Accessing the data from authentic user is also a challenge. How one can identify that the data accessed from Twitter account is real or posted by some fake person?

C. Feature Extraction

Feature Extraction is an important step for processing of social media data. The contents of social media need to be categorized on the basis of various features. These features can be demographic, lexical, behavioral and social as well.

The mental health of any person cannot be diagnosed by taking some simple parameters and keywords. Besides that there are numerous features which are required to extract from the content to give extremely correct diagnosis about the mental state of a person.

Feature extraction is in itself a challenging task like identification and connecting geophysical position of a person with the type of contents posted by it. Other than that behavioral aspect like time of post, time taken in replying a post also helps in feature classification which leads to the correct diagnosis.

Researchers like (Schwartz et al.,2013) [19], (Sadilek et al. (2013)) [20], (Strapparava and Mihalcea, 2014; Pennebaker, 2011) [21,22] and many more have worked on various categories of feature extraction which supports feature classification and supports diagnostics.

D. Feature Selection

Feature selection is the process of selecting features which is a subset of the training corpus terms which can be further used for the classification purpose. The basic purpose of feature selection is to make data with multiple dimensions that can be utilized well for data mining purpose. A feature selection is a challenging job for the social media content. The reason behind is that the data produced from the social media is voluminous and versatile (in form of tweets, comment, image or text) in nature hence selection of the feature become a tedious job. The nature of social media also determines that its data is massive, noisy, and incomplete, which exacerbates the already challenging problem of feature selection.

Feature selection from social media content faces two major issues- (1) relation extraction which means to identify unique relations from the linked data of social media. (2) Representation of such data in some mathematical formula for the selection of feature.

Feature selection can be broadly classified into wrapper model, filter model and embedded model [23] . The wrapper model use a mining algorithm which is already defined and its performance is been used as one of the evaluation criteria of selection. However filter model does not use any mining algorithm besides that it uses some general characteristics of dataset for the selection of feature subsets. There are some popularly used filter model based feature selection methods like multi-class extension, Information Gain, t-test and ReliefF. In embedded model the process of feature selection is entirely different from the rest two. In this model feature selection is embedded in training process itself.[24,25,26]

Social media is entirely different from regular data. It is not structured in nature and quite versatile too this makes application of data mining techniques quite difficult. This is the reason it puts challenges for feature selection too.

V. CONCLUSION

The overflow of social media content and its use in our daily life has forced researchers to analyze the content posted on social media. The analysis of such data can be used to draw major inferences for many domain be it entertainment industry, product industry, politics, sports to name a few. The paper discussed steps involved in the processing of social media content. The paper also outlined the various challenges that researchers face while processing of data on social media and has categorized those challenges into four categories –Language, Data Collection, Feature Classification and Selection.

REFERENCES

[1] Golder A. S. and Macy W. M., Diurnal and seasonal mood vary with work, sleep and daylength across diverse cultures Science 333, 1878 (2011); DOI: 10.1126/science.1202775. 2011

[2] Susarla Oh and Tan , Social Networks and the Diffusion of User-Generated Content: Evidence from YouTube, Information Systems Research Information Systems Research Vol. 23 No. 1, pp-23-41, DOI: 10.2307/23207870, 2012

[3]Pang B and Lee L. Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval, 2(1-2):1–135, 2008.

[4] Liu B. Sentiment Analysis and Opinion Mining. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers, 2012

[5] Sanjiv Das and Mike Chen. Yahoo! for amazon: Extracting market sentiment from stock message boards. In Asia Pacific Finance Association Annual Conf. (APFA), 2001.

[6] Morinaga S, Yamanishi K., Tateishi K., and Fukushima T. Mining product reputations on the web. In Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), pages 341–349, 2002. Industry track.

[7] Pang B., Lee L, and Vaithyanathan S., Thumbs up?: sentiment classification using machine learning techniques. In Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10, pages 79–86. Association for Computational Linguistics, 2002.

[8]Tong M. Richard. An operational system for detecting and tracking opinions in on-line discussion. In Proceedings of the Workshop on Operational Text Classification (OTC), 2001.

[9] Turney D. P. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. CoRR, cs.LG/0212032, 2002.

[10]Kouloumpis E., Wilson T., and Moore J., Twitter sentiment analysis: The good the bad and the omg! ICWSM, 11,538–541, 2011.

[11] Qiu G, Liu B. , Bu J., and Chen C., Expanding domain sentiment lexicon through double propagation. In CraigBoutilier, editor, IJCAI, 1199–1204, 2009.

[12] Thelwall M, Buckley K, and Paltoglou G. Sentiment in twitter events. Journal of the American Society for Information Science and Technology, 62(2),406–418, 2011.

[13] Tan C, Lee L., Tang J., Jiang L., Zhou M., and Li P., User-level sentiment analysis incorporating social networks. In Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 1397–1405. ACM, 2011.

[14] Hu X, Tang L., Tang J, and Liu H., Exploiting social relations for sentiment analysis in microblogging. In Proceedings of the sixth ACM international conference on Web search and data mining, 537–546. ACM, 2013.

[15] Zhao J, Dong L, Wu J, and Xu K.,Moodlens: an emoticon-based sentiment analysis system for chinese tweets. In Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 1528–1531. ACM, 2012.

[16] Hu X, Tang J, Gao H., and Liu H., Unsupervised sentiment analysis with emotional signals. In Proceedings of the 22nd international conference on World Wide Web, pages 607–618. International World Wide Web Conferences Steering Committee, 2013.

[17] Tang D, Wei F, Yang N., Zhou M, Liu T, and Qin B., Learning sentimentspecific word embedding for twitter sentiment classification. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 1: Long Papers, pages 1555–1565, 2014.

[18] Saif H., He Y, and Alani H., Semantic sentiment analysis of twitter. In The Semantic Web–ISWC 2012, pages 508–524. Springer, 2012.

[19] Schwartz, H. A., Eichstaedt, J. C., Margaret L. Kern, L., Dziurzynski, M. A., Park, G. J., Lakshmikanth, S. K., Jha, S., Seligman, M. E. P. and Ungar, L. 2013. Characterizing geographic variation in well-being using tweets. In Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media, pp. 583–591.

[20] Sadilek, A., Homan, C., Lasecki, W., Silenzio, V. and Kautz, H. 2013. Modeling FineGrained Dynamics of Mood at Scale. In WSDM, pp. 3–6.

[21] Strapparava, C. and Mihalcea, R. 2014. Affect Detection in Texts. In R. A. Calvo, S. D'Mello, J. Gratch, and A. Kappas (Eds.), The Oxford Handbook of Affective Computing, Chapter 13, pp. 184–203. New York: Oxford University Press.

[22] Pennebaker, J. W. 2011. The secret life of pronouns: How our words reflect who we are. New York, NY: Bloomsbury Press

[23] .H. Liu and L. Yu. Toward integrating feature selection algorithms for classification and clustering. IEEE Transactions on Knowledge and Data Engineering, 17(4):491, 2005.

[24] H. Peng, F. Long, and C. Ding. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. IEEE Transactions on pattern analysis and machine intelligence, pages 1226–1238, 2005.

[25] M.Robnik-Sikonja and I. Kononenko. Theoretical and ˇ empirical analysis of ReliefF and RReliefF. Machine learning, 53(1):23–69, 2003.

[26] X. He, D. Cai, and P. Niyogi. Laplacian score for feature selection. NIPS, 18:507, 2006.

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor
7.54

INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA

निस्केयर
NISCAIR

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH IN SCIENCE, ENGINEERING AND TECHNOLOGY

| Mobile No: +91-6381907438 | Whatsapp: +91-6381907438 | ijmrset@gmail.com |

www.ijmrset.com