International Journal of Psychosocial Rehabilitation, Vol. 24, Issue 05, 2020
ISSN: 1475-7192

Social Media Sentiment Analysis for Opinion Mining

M.K. Sudha and Dr.R. Priya

Abstract--- *The sentiment analysis is a digital epidemiology can support faster response and deeper understanding of public health threats than traditional methods is a rapidly growing area. There are numerous social media sentiment analysis over twitter data and other similar microblogs faces several new challenges due to the typical short length and irregular structure of such content sites available on internet. Which provides options to users to give feedback about names of diseases and their symptoms. This paper discusses an approach the health expertise trends and sentiments of users using Twitter their emotional content as positive, negative and irrelevant an attempt to observe the public's opinions and identify their issues. In existing top-down approaches, necessary but unknown information, such as disease names and symptoms, is mostly unidentified in social media data until national public health institutes have formalized that disease. In this paper we present a methodology for early detection and analysis of epidemics based on mining Twitter messages. In order to reliably trace messages of patients that actually complain of a disease, we adopt a symptom-driven, rather than disease-driven, keyword analysis. In this paper various algorithms for sentiment analysis are studied and challenges and applied machine learning techniques appear in this field are discussed.*

Keywords--- *Sentiment Analysis, Opinion Mining, Twitter Mining, Sentiment Classification, Machine Learning.*

I. INTRODUCTION

The research topic of sentiment analysis has seen an increase of interest with the fast growth in health technology trends in the recent years. Among several ways of exploring these trends, social media analysis has surfaced as a useful methodology. The user information to consumers, health consumers are aggressively using social media to share their experiences and collect opinions from others of available text data containing opinions, critics and recommendations on the web (movie reviews, forum debates, tweets and other entries in social networks) [1]. There are large amounts of unstructured, free-text information about quality of health care available on the Internet in blogs, social networks, and on physician rating websites that are not captured in a systematic way. New analytical techniques, such as sentiment analysis, may allow us to understand and use this information more effectively to improve the quality of health care [2]. The variety of the data and of the industrial applications using sentiment analysis increases various scientific issues that have yet to be fully addressed by the existing systems. A stimulating area is the development of opinion detection methods relying on these new sources. Opinion detection systems using sentiment analysis have been developed to target patient and evaluate the success of medical treatment, to know the patient expression with certain symptoms or their feeling or to predict diesis. Another growing research field is the development of embodied conversational agents (ECAs), virtual characters able to interact with humans [3].

M.K. Sudha, Research Scholar, Department of Computer Applications, VISTAS, Chennai, India. E-mail: mksudha2k@yahoo.com
Dr.R. Priya, Professor, Department of Computer Applications, VISTAS, Chennai, India. E-mail: priyaa.research@gmail.com

DOI: 10.37200/IJPR/V24I5/PR202075
Received: 03 Mar 2020 | Revised: 25 Mar 2020 | Accepted: 02 Apr 2020

One of the research challenges of this field is to integrate the affective component of the interaction. First, the ECA has to take into account the human emotional behaviors and social attitudes. Sentiment analysis of the user's verbal content is crucial for ECAs in order to determine the user's emotions and attitudes and to adapt its behavior accordingly. Second, they must be able to convey them appropriately [4] [5].

The aim of this research paper is to provide a state of the art on sentiment analysis from the perspective of the opinion mining and conversational agent's communities, identifying the most relevant advances of both communities and discussing the open research questions and prospects. We have contextualized the discussion in the development of a sentiment analysis module and its integration in an ECA platform dealing with multimodal socio-emotional interactions. The final goal is to determine which reaction an ECA should have according to the user's detected socio-emotional behaviors and sentiments**.**

The ability to analyze various sentiments of social media users, provides candid insight to users' feelings and opinions. Social-media based digital epidemiology can support faster response and deeper understanding of public health threats than traditional methods. Knowing the sentiment toward diabetes is fundamental to understanding the impact that such information could have on people affected with this health condition and their family members. The objective of this study is to analyze the sentiment expressed in messages on diabetes posted on Twitter. The publicly available data on Twitter has created a new intersection of public health, health informatics, and data science. Automatic systems can probabilistically infer what is happening around the world by using the data of what people are thinking and doing. With roughly 3281 million active users, Twitter is seen as a reliable data source to provide real-time feedback and opportunity to understand users' concerns.

II. SUBMISSION LITERATURE REVIEW

2.1 Review Stage

Twitter draws researcher's attention on different problems and has been used for a variety of purposes, such as marketing, communication, business, and education. In this review, we will discover some of the related research work on twitter data extraction, its meaningful processing, and twitter based developed applications. Different analysis tools are presented to collect twitter data. Archivist is a service that uses the Twitter Search API to find and archive tweets having specific keyword [2] [3]. The extracted data was stored in ontologies like: SIOC, FOAF, and OPO. Garin Kilpatrick introduced list of all twitter tools to collect and analyze Twitter data. He divided all Twitter tools into 53 categories. These tools provides facility in backup tweets, trend analysis, tweets translation, voice tweet, and Twitter statistics.

Jeonghee presented a model to extract sentiments about particular subject rather than extracting sentiment of whole document collectively [5]. This system proceeded by extracting topics, then sentiments, and then mixture model to detect relation of topics with sentiments. Whereas, Namrata [6]. The introduction of a sentiment analysis system for news and blog entities. This system determined the public sentiment on each of the entities in posts and measured how this sentiment varies with time. They used synonyms and antonyms to find path between positive and negative polarity to increase the seed list. The same way, the extracted tweet information is used [7]. The personalized news recommendation. They collected tweets of individual user and based on user interest

DOI: 10.37200/IJPR/V24I5/PR202075
Received: 03 Mar 2020 | Revised: 25 Mar 2020 | Accepted: 02 Apr 2020

recommended them news article. performed analysis of Twitter as electronic word of mouth in the product marketing domain. They analyzed filtered tweets for frequency, range, timing, content, and customer sentiments.

The author Bharath Sriram proposed an approach to classify tweets into news, opinions, deals, events and private messages with better accuracy [9]. They used eight basic features from tweets. They did not apply noise reduction techniques which may degrade introduced TweetStand to classify tweets as news and non-news. Naive Bayes classier was trained on a training corpus of tweets that had already been marked as either news or junk. After filtering news tweets they clustered tweets into different topics. They also extracted geographic content from each tweet, to determine the clusters overall geographic focus.

The analysis of Tweets can help allied health professionals identify users' negative sentiments as it relates to the characteristics of DDEO. This research identifies the prominent topics of users as it relates to the negative sentiments (feelings). Statements themselves may be negative and understanding prominent topics among negative sentiments (corpus of negative statements) is an alternative approach to identifying the topics related to DDEO based solely on negative expressions (Nasukawa & Yi, 2003).

III. METHODOLOGY

A very basic step of opinion mining and sentiment analysis is feature extraction. Figure 1 shows the process of opinion mining and sentiment analysis.

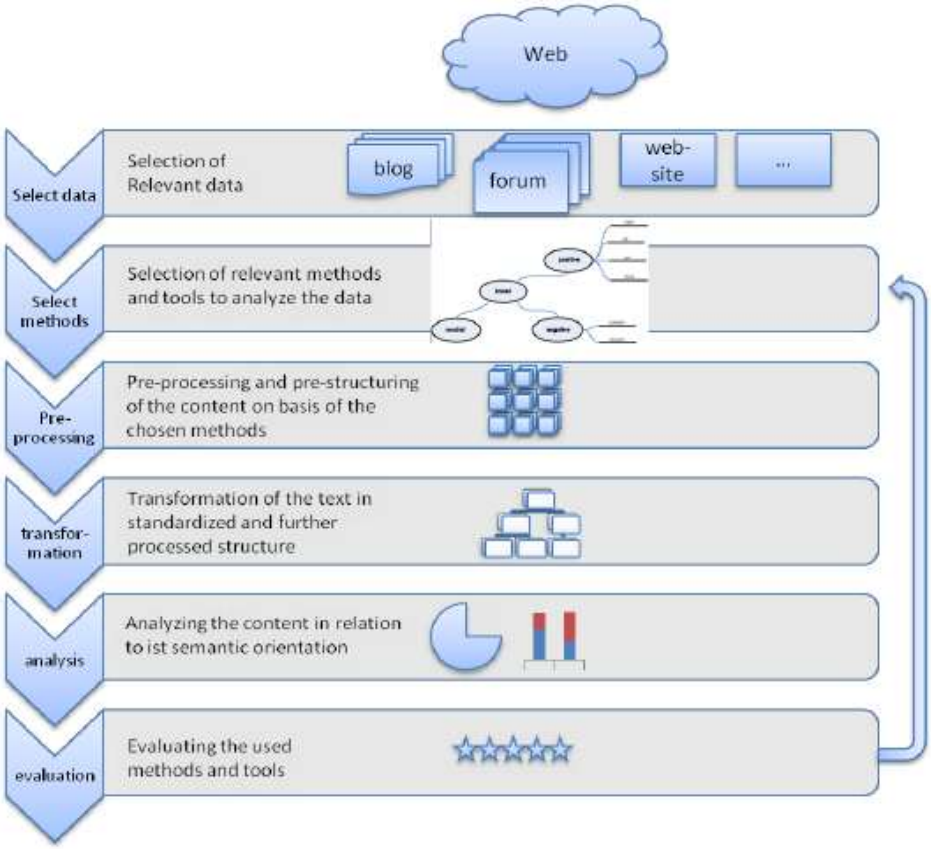

Figure 1: Process of Opinion Mining and Sentiment Analysis

DOI: 10.37200/IJPR/V24I5/PR202075
Received: 03 Mar 2020 | Revised: 25 Mar 2020 | Accepted: 02 Apr 2020

In conjunction with data collection, the proposed framework utilizes sentiment analysis and topic modeling Figure 2.

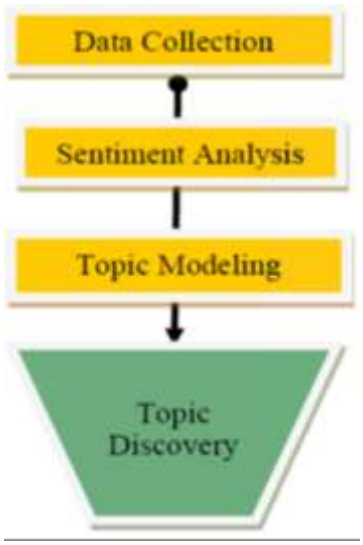

Figure 2: The Classification Process

3.1 Data Collection

This study collected the Twitter data using a real-time data collection method. By using this method, we are able to collect 10% of the publicly available English tweets.

However, the data collected for this study will focus on the four areas key terms Diet, Diabetes, Exercise, and Obesity. From these key terms, the search terms for the Twitter API were created. Eight queries were used to collect nearly 6 million tweets with 61 million tokens between Jan 1, 2018 and Jan 30, 2019 (Table 1).

Table 1: Case Definitions for the Five Common Syndromes

DDEO TOPICS TWITTER API QUERIES	***DDEO TOPICS TWITTER API QUERIES***	***SUB-TOPICS***
DIABETES #diabetes OR diabetes	DIABETES #diabetes OR diabetes	Hypertension; Kidney; Cancer; Food;
DIET #diet OR diet	DIET #diet OR diet	Food; Fastfood; Medications; Wellness; Alternative Diets; Religious Diets; Sweets
EXERCISE #exercise OR exercise	EXERCISE #exercise OR exercise	Lifestyle; Weightless; Body Image; Diet; Mental Health
OBESITY #obesity OR obesity	OBESITY #obesity OR obesity	Diet; Diabetes;
Non-Health Topics	People;	Emotions; Celebrity; Government; Events

As part of the data collection process, post processing of the data collected requires cleaning by removing stop words such as and, of, the (based on a standard list of stop words). This will allow the topic modeling toolkit used for the discovery of topics, to correctly identify the topics for analytics purposes.

3.2 Sentiment Analysis

Sentiment analysis shows subjectivity and polarity in text data with two main approaches: learning-base and lexicon base. Learning-based approach applies machine learning techniques to build classifiers from data. Lexicon-

DOI: 10.37200/IJPR/V24I5/PR202075
Received: 03 Mar 2020 | Revised: 25 Mar 2020 | Accepted: 02 Apr 2020

based approach uses a pre-defined dictionary of positive and Negative words to find the frequency of positive and negative words[12].

3.3 Discovery of Topics

Topic discovery among the negative tweets to identify themes is done using topic modeling. As a commonly used method, the Latent Dirichlet Allocation model (LDA) was appropriate to use for this type of experiment.

IV. CLASSIFICATION TECHNIQUES

There are various methods used for opinion mining and sentiment analysis the important one is Naïve Bays Classifier. In this research work, categorization of work done for feature extraction and classification in opinion mining and sentiment analysis is done [3]. In addition to this, performance analysis, advantages and disadvantages of different techniques are appraised.

4.1 Naïve Bayes Classifier

The Naive Bayes classifier is a classical demonstration of how generative assumptions and parameter estimations simplify the learning process. Consider the problem of predicting a feature $y \in \{0, 1\}$ on the basis of a vector of features $x = (x1, \ldots, xd)$, where we assume that each xi is in {0, 1}.

The probability function $P[Y = y|X = x]$ is estimated that corresponds to $P[Y = 1|X = x]$ for a certain value of $x \in \{0, 1\}$ d. This implies that the instances grows exponentially with the number of features [9]. According to this theorem, if there are two events say, e1 and e2 then the conditional probability of occurrence of event e1 when e2 has already occurred is given by the following mathematical formula.

$$P(e1 \mid e2) = (p(e1|e1)p(e1))/e2 \qquad (1)$$

This algorithm is implemented to calculate the probability of a data to be positive or negative. So, conditional probability of a sentiment is given as:

$$P(\text{Sentiment} \mid \text{Sentence}) = (P(\text{Sentiment})\, P(\text{Sentence}| \text{Sentiment}))/P(\text{Sentence}) \quad (2)$$

And conditional probability of a word is given as:

P(Word| Sentimen)t=(Numberof wordoccurencein class+ 1)/(NumberofwordsbelongingtO aclass+TotalnosofWord)

Algorithm:
S1: Initialize P(positive) ←num -popozitii (positive)/ num_total_propozitii
S2: Initialize P(negative) ←num -popozitii (negative) / num_total_propozitii
S3: Convert sentences into words
for each class of {positive, negative}:
for each word in {phrase}
P(word | class) < num_apartii (word | class) 1 | num_cuv (class) +
num_total_cuvinte P (class) ←P (class) * P (word | class) Returns max {P(pos), P(neg)
}

4.2 Random Forests

A RF is a classifier containing a decision trees collection in which each tree is built by applying the algorithm with a training set and an additional random vector which is a sampled i.i.d. from some distribution. The prediction of the random forest is obtained by a majority vote over the predictions of the individual trees.

DOI: 10.37200/IJPR/V24I5/PR202075
Received: 03 Mar 2020 | Revised: 25 Mar 2020 | Accepted: 02 Apr 2020

Algorithm
Step 1: procedure (DT, F, N).
Step 2: A = ∅
Step 3: for i = 1 to N do
Step 4: D(i) T is bootstrap instance from DT
Step 5: ai is Randomized Tree Learn (D (i) T ,F)
Step 6: A ∈ [ai]
Step 7: If not return A
Step 8: end
Step 9: Begin function Random forest (DTF)
Step 10: Each node f, belong to F subset
Step 11: Segment best feature in F
Step 12: Return learned tree

4.3 Proposed HCNN-LSTM Algorithm

Recurrent neural networks with Long Short-Term Memory have developed as a scalable and effective model for learning several problems associated to sequential data. The RNN are a sub-class of neural networks that were constructed to generate the long-range inherent correlation among data samples. Though the normal NN do not detail the temporal input data order, the RNN eludes this issue by having the time built notion into it. Related to other NN architectures, the RNNs have a hidden layer and update its hidden layer after each time-step process in the input. This confirms that the input sequence temporal structure is valued. The existing RNN may subject to local minimal solutions during the iterations through the layers.

Algorithm for Proposed HCNN back propagation
Step 1: procedure()
Step 2: A constant A is employed at the output unit and teh backwards propogation of network starts
Step 3: Arriving nodal data is added and the outcome is multiplied with the stored value in the left side of unit
Step 4: The outcome is diffused to the left side unit
Step 5: The outcome composed at the input unit is the network function derivative corresponding to x
Step 6: Set RNN parameter, F and shuffle the dataset DT
Step 7: Set i = 0
Step 8: For each feature belong to A, update F

V. PERFORMANCE EVALUATION

To evaluate the algorithm following measures are used:

- Accuracy
- Precision
- Recall
- Relevance

Following contingency table is used to calculate the various measures

	Relevant	***Irrelevant***
Detected Opinions	True Positive (tp)	False Positive (fp)
Undetected Opinions	False Negative (fn)	True Negative (tn)

Now,

$$\text{Precision} = tp/(tp + fp) \quad (3)$$

$$\text{Accuracy} = (tp + tn)/(tp + tn + fp + fn) \quad (4)$$

DOI: 10.37200/IJPR/V24I5/PR202075
Received: 03 Mar 2020 | Revised: 25 Mar 2020 | Accepted: 02 Apr 2020

$$F=(2*Pr\ ecision\ *Re\ call)/(Pr\ ecision\ Re\ call) \quad (5)$$

$$Recall =tp/(tp + fn) \quad (6)$$

Accuracy

The accuracy of the prediction model was the prime parameter as it establishes the correctness of the ML techniques in predicting the occurrence of diabetes.

S. No	*Prediction models*	*Accuracy*	*Precision*	*Recall*	*F1 score*	*AUC in ROC*
1	Navies Bayes [22]	0.763	0.759	0.763	0.76	0.819
2	Random Forest [23]	0.747	0.5075	0.694	0.5875	0.806
3	HCNN- LSTM	1	1	1	1	1

VI. RESULT AND DISCUSSION

In overall analysis, it was evident that the proposed HCNN-LSTM algorithm was very efficient in predicting the diabetes. Even though the accuracy may not reflect the effectiveness of algorithm in many cases, the other performance metrics also exhibited the appropriateness of the proposed algorithm to be employed in predicting both blood pressure and diabetes among the female patient.

VII. CONCLUSION

Examining social media users' responses to new health technology can be useful to understand the trends in rapidly evolving fields. Recent research has mainly focused on the 'diabetes keyword itself, which is rather broad. The novel HCNN-LSTM prediction model was generated by integrating CNN and LSTM was proposed to predict the diabetes over the Pima Indian diabetes. Twitter negative sentiments dataset. The HCNN-LSTM Classifier methods discussed in the paper are actually applicable in different areas like clustering is applied biological reviews & analysis. From a convergent point of view HCNN-LSTM is best suitable for textual classification, clustering for reading. In overall analysis, it was evident that the proposed HCNN-LSTM algorithm was very efficient in predicting the diabetes. Even though the accuracy may not reflect the effectiveness of algorithm in many cases, the other performance metrics also exhibited the appropriateness of the proposed algorithm to be employed in predicting diabetes based on twitter dataset.

REFERENCES

[1] Clavel, Chloe, and Zoraida Callejas. "Sentiment analysis: from opinion mining to human-agent interaction." *IEEE Transactions on affective computing* 7.1 (2015): 74-93.

[2] Ji, Xiang, et al. "Twitter sentiment classification for measuring public health concerns." *Social Network Analysis and Mining* 5.1 (2015): 13

[3] Saif, Hassan, Yulan He, and Harith Alani. "Semantic sentiment analysis of twitter." *International semantic web conference*. Springer, Berlin, Heidelberg, 2012.

[4] Singh, Pravesh Kumar, and Mohd Shahid Husain. "*Methodological study of opinion mining and sentiment analysis techniques*." *International Journal on Soft Computing* 5.1 (2014): 11

[5] Lee, Jisan, et al. "Health Information Technology Trends in Social Media: Using Twitter Data." *Healthcare informatics research* 25.2 (2019): 99-105

[6] Gabarron, Elia, et al. "Diabetes on Twitter: a sentiment analysis." *Journal of diabetes science and technology* 13.3 (2019): 439-444.

[7] Pradhan, Vidisha M., Jay Vala, and Prem Balani. "A survey on Sentiment Analysis Algorithms for opinion mining." *International Journal of Computer Applications* 133.9 (2016): 7-11.

DOI: 10.37200/IJPR/V24I5/PR202075
Received: 03 Mar 2020 | Revised: 25 Mar 2020 | Accepted: 02 Apr 2020

[8] Kaur, Amandeep, and Vishal Gupta. "A survey on sentiment analysis and opinion mining techniques." *Journal of Emerging Technologies in Web Intelligence* 5.4 (2013): 367-371.

[9] Shaw Jr, George, and Amir Karami. "Computational content analysis of negative tweets for obesity, diet, diabetes, and exercise." *Proceedings of the Association for Information Science and Technology* 54.1 (2017): 357-365

[10] Pak, Alexander, and Patrick Paroubek. "Twitter as a corpus for sentiment analysis and opinion mining." LREc. Vol. 10. No. 2010. 2010.

[11] Batool, Rabia, et al. "Precise tweet classification and sentiment analysis." 2013 IEEE/ACIS 12th International Conference on Computer and Information Science (ICIS). IEEE, 2013

[12] Dr.K. Abirami Dr.K. Dharmarajan, Farhanah Abuthaheer, "Sentiment Analysis on Social Media," *Journal of Emerging Technologies and Innovative Research (JETIR)*, pp. 210-217 Vol.6 Issue 3.

DOI: 10.37200/IJPR/V24I5/PR202075
Received: 03 Mar 2020 | Revised: 25 Mar 2020 | Accepted: 02 Apr 2020