Journal of Computing & Biomedical Informatics
ISSN: 2710 - 1606

Volume 05 Issue 01
2023

Research Article
https://doi.org/10.56979/501/2023

Identification of Age and Gender on Twitter Using DenseNet and LSTM

Khalid Masood[1,*], Aneela Mehmood[1] and Mariyum Nizami[1]

[1]Lahore Garrison University, Lahore, Pakistan.
[*]Corresponding Author: Khalid Masood. Email: khalid.masood@gu.edu.pk.

Received: November 07, 2022 **Accepted:** May 24, 2023 **Published:** June 05, 2023.

Abstract: This study presents a novel and automated analysis on social media platforms to distinguish between the gender and identify age group of users. The personality of a person can be determined through his or her social behavior. The proposed design employs dense deep learning framework to automatically identify the gender whenever users express their thoughts and comments on public forums. The system aims to provide gender and age group information to organizations to plan future needs of people and may focus on a particular gender or age group. By learning from the proposed sentiment model, it is observed that deep features are strongly associated with gender discrimination. Sentiment analysis is a technique that identifies emotions in a text and computes scores for the post to lsbel it as positive, negative, or neutral. Users' messages or posts are identified into gender, age, or location, and a detailed assessment is provided to segment increasing population into respective groups and categories. In this work, the proposed gender detection model using sentiment analysis and LSTM (long short term memory) algorithm achieves 91% classification accuracy whereas age group assessment model performs at 85% accuracy level.

Keywords: Social Media Analysis; LSTM; CNN; Gender.

1. Introduction

Social media is a platform for the people to express their personal thoughts and discuss daily life issues. The ideas of people can be termed as positive, negative or neutral [1, 2]. Considering Web 2.0, social media has become the most powerful medium for information dissemination and the spread of news. Facebook is the most popular social media platform along with Twitter and Instagram. Recently TikTok has emerged as the most growing platform in terms of users' activation and has surpassed WhatsApp group. The data on social media is either unstructured or semi structured that needs to be processed to extract useful information for business and government applications [3, 4]. Apart from positive use, social media can be used for negative activities such as hate speeches and dark web. The criminals promote anti-social activities and spread or plan terrorist attacks. The law enforcement agencies use social media to identify anti-social activities. However, in this research we employ social media to promote business and assist government agencies to benefit masses [5, 6]. The Business process management (BPM) increases the output of an organization and it is helpful in reducing the operational cost. BPM lifespan comprises designing, enacting, managing, and analyzing business processes [7, 8]. Business Process has become mandatory of an organization because they are able to improve the operations and fundamental competency. A business process model is defined as an organized way, to check all set of events planned to create a definite output for a specific client or market. Text summarization methods are either abstractive or extractive methods [9, 10]. In extractive summarization, important sentences and paragraphs are selected and short documents are designed whereas in extractive summarization, a detailed summary is presented that contains keywords and important phrases from the original document [11]. The extractive summarization contains popular sentences of the original document [12].

Our study is based on a real-life application of twitter data. In the proposed research, we have collected data from twitter using the Scrapy API developed for Python [13, 14]. The hypothesis states that not only there is a considerable variation in the expression of emotions for men and women but choice of words is also different between the two categories. A model is proposed that extracts emotions, words and phrases from the social media messages and identifies the gender and age group of a user. Using a learning framework for the gender and age discrimination, the hidden gender and age information of a user is determined based on his and her text messages [15]. Our aim is to distinguish such posts by the sentiment analysis on the written social media posts. The objective is to identify relevant contents concerning personal preferences and subsequently compare them to the disclosed information in user tags [16]. The overall flow diagram of the proposed algorithm is shown in Figure 1. The posts from the social media are analyzed and processed to achieve the objective of the proposed study.

The rest of the paper contains the following sections. The related work is presented in the next section. In section 3, pre-processing techniques are introduced. The methodology containing the deep learning framework along with LSTM (long short term memory) analysis is presented in section 4. The experimental setup is described in Section 5. The work is concluded in section 6 and future work is also suggested.

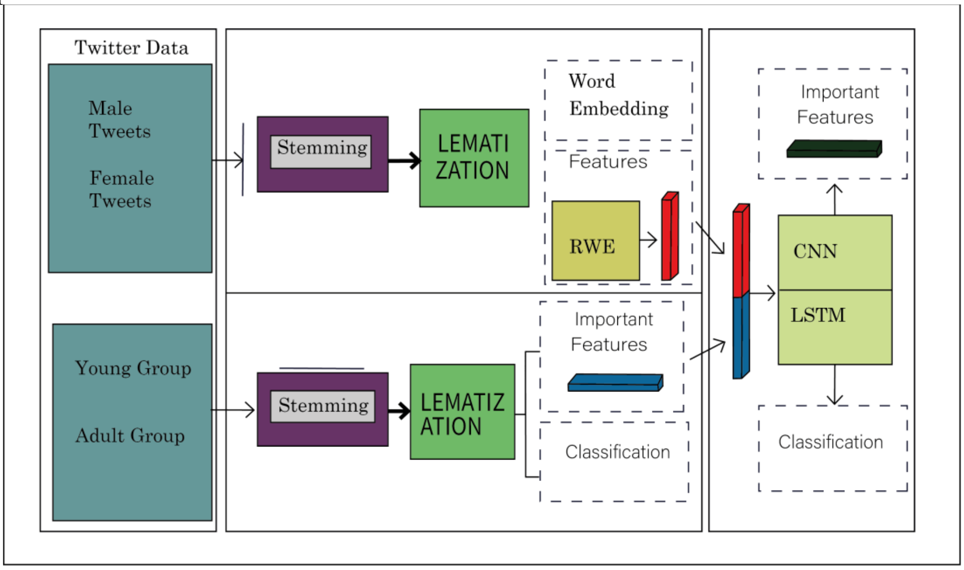

Figure 1. Block Diagram for the proposed model

2. Literature Review

There are many domains such as computer vision, speech recognition, natural language processing and health informatics in which deep learning methods have overwhelming success [17, 18]. In the past few years, deep learning has revolutionized the healthcare and cancer treatment has significantly reduced the cancer related diseases [19]. Deep features from deep learning method are used in neuropathology. In comparison to traditional methods, deep learning techniques provide end to end learning. These methods do not require image pre-processing by manual selection and are able to automatically classify raw images and perform the classification task [20].

There are number of studies that explore semantic analysis to explore the gender of a person. In [21], ensemble classification is used to detect gender on social media posts. In the study, multimodal deep learning architectures are used to extract features from users' posts and employ them in deep convolutional networks to classify the gender of the user. In [22], deep learning techniques are applied to extract the features from the ear of a person to identify his or her age. The aim of the research is to control the contents of the social media by detecting the age of the user and thus allowing or denying the access of a particular content. The convolution neural network (CNN) with multiple layers is used in the classification stage to identify the age of a user. The work in [23] presents a gender detection model using machine learning and deep learning techniques. Sentiment analysis is performed on the posts of patients who are members of a health web forum. Most of the users have sexually transmitted diseases, hence they tend to hide their identities. Analyzing the texts of social media chats, gender of user is detected that can be used for establishing health policies and detecting problems and needs according to the particular gender. In [24], the gender of twitter users is identified using machine independent classifiers. The blog profile pages of twitter users are used on a large dataset and 92% classification accuracy is achieved. The research presented in [24] presents classification of gender in web forums. The writing styles and topics of interest are analyzed using features of texts in online web forums. The feature set contains both context free and content specific features and it is reported that content free features perform better than content specific features set.

3. Methodology

We used twitter to extract the data from a medical blog forum and identified the gender of users. The emptions and their frequency are analyzed. The emotions complexity is also computed that represents varying emotions in a post. The pre-processing stage plays an essential part in text mining strategies and applications. Generally, it is the first stage in the text mining techniques. In our proposed algorithm, there are three key strides of preprocessing specifically, stop words removal, stemming, and lemmatization based on TF-IDF algorithms [25]. Following are the key stages of our proposed pre-processing algorithm.

1. Stop words removal
2. Stemming
3. Lemmatization
4. Stop words and stemming
5. Stop words and lemmatization
6. Stemming and lemmatization

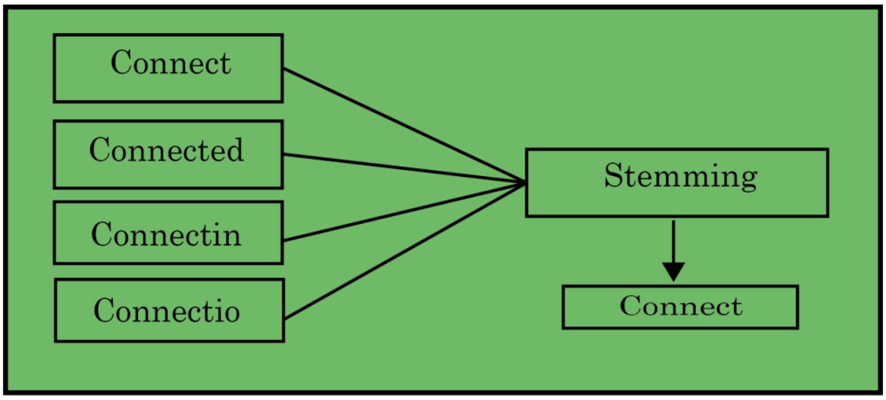

Figure 2. A Complete Stemming Process

Stop words removal is a first step for natural language processing. The purpose of removing these words is to make the text less heavy and make it more readable for analyst. Eliminating stop words decreases the dimensionality of term space [26]. Articles are the most common stop-words that a text document can contain, along with prepositions, and pro-nouns, etc. Such words do not give meaning to a document. These words are stop words. Instances for stop words: are the, in, an, a, with, and so forth. Text mining applications do not consider stop words as keywords. An exemplary technique is used in our algorithm that depends on expelling stop words acquired from pre-ordered words.

The morphological forms of a word are translated to its root the process is carried on for words that are semantically related in stemming method. While using a stemmer, the following two observations should be considered.

- Similar words should be kept separate from other words
- Various morphological forms of a word are in the same category and are mapped to its root word

The above principles are great and adequate in text mining or language processing applications. Stemming is generally considered as a recall-enhancing device. For the languages using moderate and basic morphology, the stemming is not as efficient as for those who have a more complicated morphology. The greater part of the stemming is used in English as compared to other west European languages.

Stemming algorithms are generally classified into three groups including truncating methods, statistical methods, and mixed methods. Each group has its own certain way of finding the stem variants as shown in Figure 3. In Truncating methods, removing suffixes and prefixes (affixes) of a word is called affix removal [27]. Truncate stemmer is the fundamental stemmer that truncates the word at the nth symbol. Smaller words have the higher probability of over stemming. S-stemmer is another approach which is an algorithm that combines the singular and plurals forms of nouns. S-stemmer algorithm was proposed by Donna Harman [28]. The algorithm removes suffixes in plurals so as to convert them to the singular forms. The other method of stemming is based on statistical analysis and techniques. Affixes are removed using a few statistical techniques.

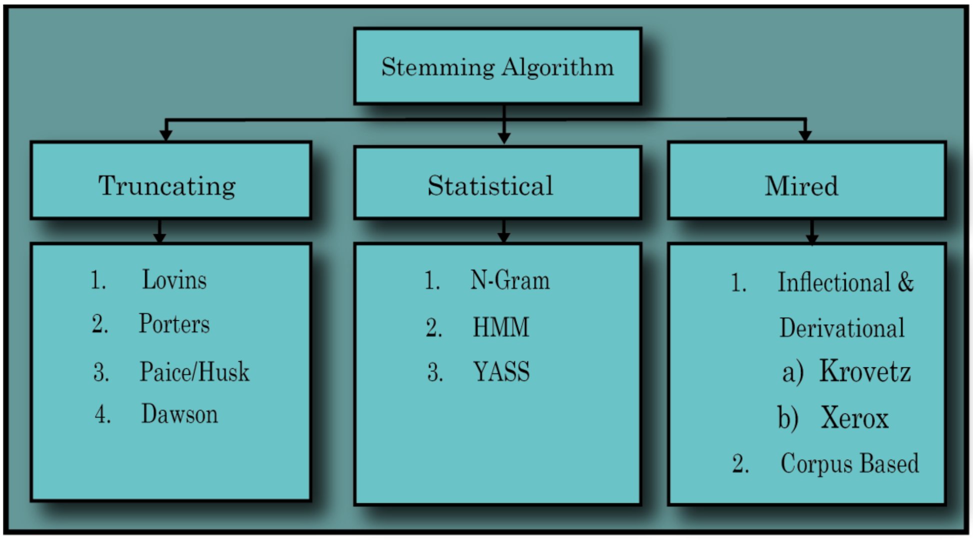

Figure 3. Flowchart for the Stemming Algorithm

Lemmatization is also closely related to stemming. The process of clubbing different forms of a single word so that it can be treated as one is Lemmatization. The method is popular in data mining for linguistics. As an example, "computers" is the plural form of "computer", and using the same principal, "dogs" being an inflected form of "dog". In simple words, lemmatization is the linking of a word to its root form. Lemmatization is not only applicable to nouns instead it works on adjectives and action verbs also. Following are the few more examples of lemmatization.

- For constructing, after applying lemmatization, construct is formed.
- For extracts, the extract is formed.
- For singing, lemmatization forms sing.

Initially it seems very simple but confusion arises when some complicated words are used for lemmatization. For example, the word, worker, is not considered as an infected form of work, neither the word speaker is considered as an infected form for the word speak. The reason is that the speaker talks about a topic or delivers a speech whereas speak (verb) is the process for talking.

The frequency of terms in a document is denoted by term frequency [29]. The behavior of TF varies from long to short documents i.e. a term would appear much less times in short documents as compared to long ones. That's why the TF is often determined by using the document length that is the total terms in the document. In this way the TF is normalized as follows.

$$TF(w_f, d_z) = \frac{Count(w_f, d_z)}{size(d_z)} \quad (1)$$

The Inverse Term Frequency (IDF) represents the importance of a term. When we compute Term Frequency, we consider all the terms are of same importance. But in reality, there are many terms that are of little importance but appear a lot of times in the document, such as "is", "of" and "that". The following formula is thus used to measure the weight of frequently appeared terms, and scale up the weight of rarely used terms.

$$IDF(w_f, D_w) = \log \frac{N_t}{d \times w} \quad (2)$$

TF-IDF composite weight for each term in each document can be computed by taking the product of Term Frequency and Inverse Document Frequency. The value is maximum when it happens to be critical to scoring and ranking. Firstly, the overlap score measure is introduced, which is defined as the sum (of all query terms) of the number of occurrences of each query term. This can be enhanced by adding TF-IDF weight for each term. The value becomes lesser when the occurrence of term is less in a document, or the term occurs in many documents (therefore presenting a less pronounced relevance signal). The document is considered a vector which has a component that corresponds to each term in the dictionary and a weight for each component [30]. The weight is zero for the dictionary terms that are not in a document. It is lowest for the terms which appeared in all the documents. The average results are produced using word-based scoring techniques, which assign score to those words that are fundamentals. Initially preprocessing steps such as stop words removals, stemming and lemmatization are performed.

4. Experiments

Deep learning is quiet popular in image analysis and computer vision. It has vast applications in social media and natural language processing. In our research, we have used social media data to identify age and gender of users. To extract features from the data, we employed 2D DenseNet as our initial framework. The model enhances features from lower layers and retains discriminant features while discards redundant features [31]. In our proposed 2D Densenet model, 2D convolutions are performed to extract features from pooling layers as shown in the Figure 4. The output for every neuron is connected to the fully connected layer and dropout layers only transmits selected neurons and discards a suuficiient number of neurons. The last layer is the classification layer and contains the number of classes and in our case, there are 4 classes that contains male, female, young and adult. The probability of belonging to each class can be computed by the softmax function [32].

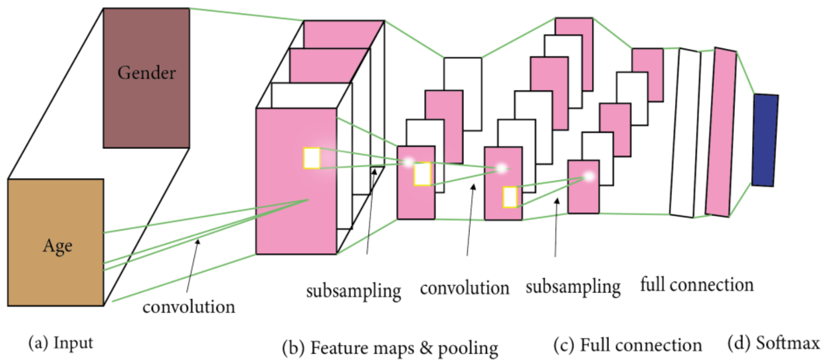

Figure 4. The Proposed Deep Learning Framework

LSTM (Long Short term Memory) belongs to RNN (recurrent neural network) that learns the backforound informatioon or weights of the previous layers. It consists of a series of logical gates that either store the information or discard it. The complexity of the model is increased in LSTM but the important information or weights are retained. The information is used in successive iterations and during backward propagation, redundant information is deleted. In order to produce global characteristics described through a set of global input layers, convolutional layers are used. Both the local and global characteristics are then fed into a regular affine network to identify specified objects.

The deep CNN model is used to extract features using many layers of convolutions and pooling. The training process involves tuning of hyper parameters and updating the weights in each layer for forward propagation and minimizing the error in the backward propagation, optimum accuracy is achieved. Pooling layers filter the initial features and discriminant features are selected whereas redundant features are discarded. Finally, the activation function connects the input neurons to the output layers and probabilities are computed for input signal to be labeled according to the majority weight. The data (3450 posts and 18000 comments) are collected using twitter posts. There are 1050 male users and 850 female users whereas 1300 users did not disclose their gender. For feature extraction, word2vec and random word embedding. The filter size of CNN structures are varied from 5 to 8 digits that means that means that 5 words or 8 words are aggregated. It is found that random word embedding achieves more accuracy than the word2Vec technique. The random word embedding performed at 91% accuracy levels for gender classification and 85% accuracy is achieved in age group population, as shown in Table 1, Table 2 and Figure 5. The sentiment analysis can be improved by using the transformer based networks wher auto encoders are used for extraction of features. The feature set is also varied from word features into senti-word features. In the senti-word features, there is small number of features as compared to word features set. The performance of senti-word features is almost equal to the word features set but its size is only 20% of the words features set. Compared to machine learning techniques, deep learning and transformer based learning techniques perform superior.

5. Results

In sentiment analysis, each word is assigned a score, ranging from -1 to +1. The score +1 shows that the word is positive, -1 score shows that it is negative whereas 0 represents neutral words. The quality metrics accuracy, precision and recall are used to test the model's performance. There are ywo classification tasks in our experiments, identify the gender to be either male or female and classify the young or adult age group.

Let Yi = 1,..,n be the set of gender type labels in the dataset for the i-th record, where n is the total number of subject records. Let Zi be the set of predicted results for the i-th twitter record. The accuracy, precision and recall are defined as follows,

$$Accuracy = \frac{|Y_i \sqcap Z_i|}{Y_i} \tag{3}$$

$$\text{Precision} = \frac{1}{n}\sum_{i=1}^{n} \frac{|Y_i \sqcap Z_i|}{Z_i} \tag{4}$$

$$\text{Recall} = \frac{1}{n}\sum_{i=1}^{n} \frac{|Y_i \sqcap Z_i|}{Y_i} \tag{5}$$

Considering Table 1 and Table 2, an accuracy of 87% is achieved for CNN in identifying male users whereas LSTM achieves 90% accuracy. In female gender, CNN and LSTM perform at 89% and 91%, respectively. In age group detection, CNN achieves an accuracy of 83% and LSTM performs at 85% accuracy levels for Young age groups. To identify adult age group, CNN achieves 82% accuracy and LSTM performs at 84% accuracy level.

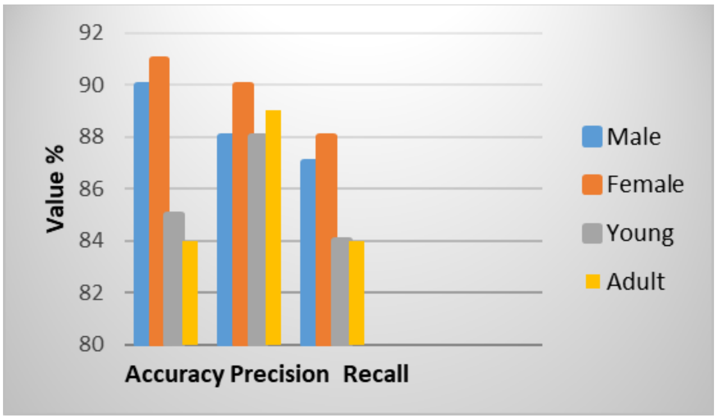

Figure 5. Accuracy Metrics for Gender and Age Groups

As shown in Table 1 and Table 2, our automated analysis represents a more accurate gender and age group determination based on the users' comments on social media platforms.

Table 1. Gender Classification Scores

Male			Female	
Quality Metrics	CNN (%)	LSTM (%)	CNN (%)	LSTM (%)
Accuracy	87	90	89	91
Precision	86	88	88	90
Recall	85	87	87	88
F1 Score	83	86	84	87

Table 2. Age Group Classification Scores

Young	Adult

Quality Metrics	CNN (%)	LSTM (%)	CNN (%)	LSTM (%)
Accuracy	83	85	82	84
Precision	86	85	80	88
Recall	82	84	81	87
F1 Score	81	86	82	86

6. Conclusions

The deep features are extracted from users' text on twitter and WhatsApp data. The proposed 2D DenseNet is used for modeling of five stacks of CNN whereas back propagation is used in bilateral LSTM to enhance the accuracy. It is evident that LSTM technique using TF-IDF features achieves superior classification accuracy. The deep extreme features are obtained using DenseNet that carries weights from previous layers and attenuates the phenomenon of vanishing gradient. Considering the performance of the proposed sentiment model, it is observed that deep features are strongly associated with gender discrimination. We further investigate the role of features for the age groups of young and adult classes. The quality metrics showing precision and recall are better in female group in comparison to male sub class. One possible explanation for this phenomenon is that for the female group, data is dense and compact as women use more compliment phrases than men.

In future, a more detailed analysis will be performed on a more extensive data set that spans at least one year. Also, it is suggested that in addition to adding new trends of social media in data collection and data generation, reinforcement learning techniques will be used for efficient utilization of the provided datasets.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Alsaeedi, A. and Khan, M. A study on sentiment analysis techniques of Twitter data. International Journal of Advanced Computer Science and Applications, 2019, 10(2), 361-374.
2. Zheng, D., Chen, H., Lusch, R. and Li, S. Social media analytics and intelligence. IEEE Intelligent Systems, 2010, 25(6), 13-16.
3. Zainuddin, N., Selamat,A. and Ibrahim, R. Hybrid sentiment classification on Twitter aspect-based sentiment analysis. Applied Intelligence, 2012, 48(2), 21-30.
4. Basiri, N., Ghasem-Aghaee, N. and Naghsh-Nilchi, A. Exploiting reviewer's comment histories for sentiment analysis. Journal of Information Science, 2014, 40(3), 313-320.
5. Asghar, M., Khan, A., Ahmad, S. and Qasim, M. Lexicon-enhanced sentiment analysis framework using rule-based classification scheme. PLOS ONE, 2017, 1(12), 1-22.
6. Cochran, A., Skao, L., Gusani, N., Suliburk, J. and Nwomah, B. Use of Twitter to document the 2013 academic surgical congress. Journal of Surgical Research, 2014, 190(1), 36-40.
7. Hawkins, C., Duszak, R. and Rawson, J. Social media in radiology: Early trends in Twitter microblogging at radiology's largest international meeting. Journal of the American College of Radiology, 2014, 11(4), 387-390.
8. Matta, R., Doiran, C. and Leveridge, M. The dramatic increase in social media in urology. The Journal of Urology, 192(2), 494-498.
9. Xiang, Z. and Gretzel, U. Role of social media in online travel information search. Tourism Management, 2010, 31, 179-188.
10. He, U., Zha, S. and Li, L. (2013) Social media competitive analysis and text mining: A case study in the pizza industry. International Journal of Information Management, 2013, 33, 467-462.
11. Denecke, K. and Nejdl, W. How valuable is medical social media data? Content analysis of the medical web. Information Sciences, 2009, 179(12), 1870-1878.
12. Tang, L. and Liu, H. Community detection and mining in social media. Synthesis Lectures on Data Mining and Knowledge Discovery, 2010, 2, 1-137.
13. Nahn, Y. and Mooney, R. Text mining with information extraction. In AAAI Springer Symposium on Mining Answers from Texts and Knowledge Bosses, New York, USA, 2002.
14. Pang, B. and Lee, I. Thumbs up: Sentiment classification using machine learning techniques. In Proceedings of the AC102 Conference on Empirical Methods in Natural Language Processing, London, UK.
15. Smith, T. The social media revolution. International Journal of Market Research, 2009, 51(4), 559-561.
16. Barbier, G. and Liu, H. Data mining in social media. Social Network Data Analytics, Springer, 2011, 2, 327-352.
17. Zarrinkalam, F. and Bagheri, E. Event identification in social networks. Encyclopedia with Semantic Computing and Robotic Intelligence, 2017, 1, 163-170.
18. Pang, B. and Lee, L. Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval, 2008, 2, 1-135.
19. Tang, J. and Liu, H. Unsupervised feature selection for linked social media data. IEEE Transactions on Knowledge and Data Engineering, 2014, 26(12), 1914-1927.
20. Gundecha, P., Barbier, G., Tang, J. and Liu, H. User vulnerability and its reduction on a social networking site. Journal of Transactions on Knowledge Discovery Forum Data, 2014, 2(1), 150-160.
21. Bollen, J., Huina, M. and Xiaojun, Z. Twitter mood predicts the stock market. Journal of Computational Science,011, 2, 1-8.
22. Hirudkar, A. and Sherekar, S. (2013) Comparative analysis of data mining tools and techniques for evaluating performance of database system. International Journal of Computing Science and Applications, 6(2), 232-237.
23. Masood, K. and Al Ghamdi, M. Modeling mental stress using a deep learning framework. IEEE Access, 2019, 7(1), 68446 – 68454.
24. Khan, A. Baharudin, B., Lee, L. and Khan, K. A review of machine learning algorithms for text-documents classification. Journal of Advances in Information Technology, 2010, 1, 4-20.
25. Kowsari, K., Heidarysafa, M., Odukoya, T., Potter, P., Barnes, L. and Brown D. Gender detection on social networks using ensemble deep learning. Proceedings of the Future Technologies Conference (FTC), 2020, 1, 346-358.
26. Alghieth, M., Alhuthail, J., Aldhubiay, K. and Alshowaye, R. Smart age detection for social media using deep learning techniques via ear shape. International Journal of Advanced Computer Science and Applications, 2020, 10(11), 272-280.
27. Park, S. and Woo, J. Gender classification using sentiment analysis and deep learning in a health web forum. Applied Sciences, MDPI, 2019, 9(6), 1249-1260.
28. Burger, J., Henderson, J., Kim, G. and Zarrella, G. Discriminating gender on twitter. Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2011, 1, 1301-1309.
29. Zhang, Y., Dang, Y. and Chen, H. Gender classification for web forums. IEEE Trans. Syst. Man Cybern. Part A Syst. Hum., 2011, 41, 668–677.
30. Gavilanes, M., Martinez, J., Mendez, S., Montenegro, E. and Gonzalez, F. Creating Emoji lexica from unsupervised sentiment analysis of their descriptions. Expert Systems with Applications, 2018, 10(3), 74-91.
31. Esuli A. and Sebastiani, F. SENTIWORDNET: A publicly available lexical resource for opinion mining. In Proceedings of the 5th Conference on Language Resources and Evaluation (LREC'06), 2009, 417–422, Genova, IT.
32. Masood, K. Adaptive Wavelet packets for minimization of Error in computation of Mental Stress. International Journal of Wavelets, Multiresolution and Information Processing (IJWMIP), 2015, 13(1), 82-90.