Journal of Ubiquitous Computing and Communication Technologies (UCCT) (2020)
Vol.02/ No.03
Pages: 178-186
https://www.irojournals.com/jucct/
DOI: https://doi.org/10.36548/jucct.2020.3.006

Emotional Analysis of Bogus Statistics in Social Media

Dr. Wang Haoxiang
Director and lead executive faculty member,
GoPerception Laboratory,
NY, USA.
Email: hw496@goperception.com

Abstract:- Fake info or bogus statistics is a new term and it is now considered as a greatest threat to democracy. Since the world is full of surprises and humans have developed their delicate nature to detect unexcepted information. Social media plays a vital role in information spreading, since the impact towards fake information has gained more attention due to the social media platforms. Trending the hot topic without analyzing the information will introduce great impact over millions of people. So, it is essential to analyze the message and its truthfulness. Emotional analysis is an important factor in bogus statistics as the information gets reshared among other based on individual emotions. Considering these facts in social media information analysis, an efficient emotional analysis for bogus statistics in social media is proposed in this research work using recurrent neural network. In an emotional perspective, fake messages are compared with actual message and false messages are identified experimentally using recurrent neural network.

Keywords:- Bogus statistics, social media, Recurrent neural network, data analysis,

1. Introduction

Social media provides a great platform for sharing and exchanging user views, opinions, and information through various social media sites like Facebook, Twitter, and Instagram, etc. Social media is purely based on mobile and web-based applications and many of the researchers are tending to analyze the social media using some tools for analyzing the emotional and sentimental analysis of the users based on their activities and the availability of API that is provided on the sites like Twitter and Facebook (Brandt *et.al.* [1]). This leads to the evolution of tools for social media analysis and scrapping. Business is also carried out using social media and thus helps to develop their product and services by communicating and understanding their requirements.

Figure 1 represents the social media analytics cycle. Social media is mainly composed of raw and unprocessed data such as audio, video, images, and text in a huge amount uploaded by the user. These data are converted into valuable information using sentimental analysis and it can be done using machine learning and artificial intelligence. The process of collecting, combining, and analyzing the data is a challenging task for the researchers. For implementing the social media methodology, it is divided into analytics, data, and facilities (Brooker. *et.al.* [2]). The data can be either real-time or historic data, news data, and public data. Social media data analytics is carried out using java, MATLAB, or python language. Social media techniques consist of computational science techniques and Sentimental analysis.

ISSN: 2582-337X
Submitted: 18.08.2020
Accepted: 12.09.2020
Published: 18.09.2020

Journal of Ubiquitous Computing and Communication Technologies (UCCT) (2020)
Vol.02/ No.03
Pages: 178-186
https://www.irojournals.com/jucct/
DOI: https://doi.org/10.36548/jucct.2020.3.006

Fig.1 Social Media Analytics Cycle

In Sentimental analysis, it takes a large amount of news content and user-generated texts online. Most of the data generated from the four main social media platforms are Twitter, Facebook, Snapchat, and Instagram. It is stated that most youngsters are using frequently and expressing their joyful and opinions. For buying a product, people are nowadays checking the product reviews and then they opt for buying based on sentimental analysis. It consists of both positive and negative comments given by the user. Figure 2 shows the analysis of Social media. This analysis is not only based on the words and concepts but also sentence syntactical tree.

The main components for sentimental analysis are semantic engine, search engine, crawler, machine translation engine, classification engine, and geo-referentiation engine. The social media data resources are classified as freely available data, accessing data through tools such as Google Trends and then data accessing through API. One of the open-source social media that provides free content to users is Wikipedia and World Bank Databank (Wang *et.al*.[3]). These are providing freely available data for researchers. In social networking media, not all the sites need to provide API access some sites do have API access like LinkedIn, Bing, and Skype and it does not permit data scraping. In Twitter, most of the user default account setting is public and it is their choice to change the account settings. It is proved that only 10% of user accounts are private on Twitter and it is available in JSON format. It consists of two API such as Search API and Streaming API. The Search API is used for processing of the past data and a part of Twitter API v1.1. This requires an authorized application before processing any results. The Streaming API is used for processing the real-time data such as keyword entered by user and location, user ID, etc., the data can be accessed by the filtering process.

ISSN: 2582-337X
Submitted: 18.08.2020
Accepted: 12.09.2020
Published: 18.09.2020

Journal of Ubiquitous Computing and Communication Technologies (UCCT) (2020)
Vol.02/ No.03
Pages: 178-186
https://www.irojournals.com/jucct/
DOI: https://doi.org/10.36548/jucct.2020.3.006

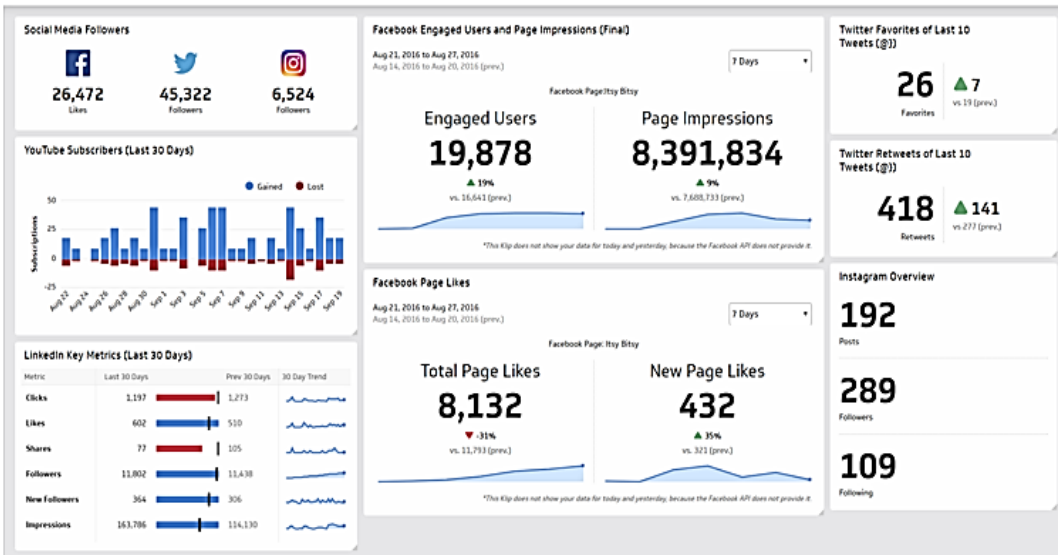

Fig. 2 Analysis of Twitter, Facebook, and Instagram

On Facebook, it has a lot of privacy issues than Twitter but on Facebook, it stores everything as objects and consists of a series of API. If the user wants to access an object, the user must know the unique ID to make an API call. It consists of a graph and a public API. The graph API needs an access token and must be attached to the request. An app access token is required for searching the pages and places if the user is searching for other types it requires a user access token. Sentiment analysis is based on the context and the level of sentiment then sentiment subjectivity, orientation/polarity, and strength. The social media analytics platform comprises data analytics tools, data feeds, and data mining. It is further classified as social network media platforms and news platforms. Social network media platforms are Brandwatch, Sysomos MAP, Attensity these are some examples and that are used to measure the sentiments and demographics that include analysis on sentiments, text analytics, and provide a user-friendly interface.

In order to access the real-time and historical data from the popular social networks DataSift is used. It makes the client filter and aggregate, and discovering trends from the millions of public conservations on social media. News platforms are used to provide new feeds and the analytics associated with the targeted companies for monitoring the market sentiment in the news. The two business news providers are Bloomberg and Thomson Reuters. This uses Natural Language Processing (NLP) to seek news items over the companies. It consists of relevance, uniqueness, volume analysis, author sentiment, and headline analysis.

2. Related Works

A detailed survey has been conducted to obtain the issues in existing emotion recognition and fake detection mechanism on social media. Zulfadzli *et.al.* [4] conducted a study on Sentiments Analysis on Social media. The study analysis comprised above 1000 Facebook posts on newscasts. Comparison of sentiment analysis between Rai-the Italian public broadcasting service and the emerging dynamic private company La7. Sentiment analysis on social media not only exhibiting the positive and negative oppositions of words and concepts, but it also analyzed the syntactical tree of the sentence. The assessment shows between Rai-the Italian public broadcasting service and La7 company the performance is denoted normally above 87% and 97%. By this analysis, the results reveal the reality as explore by Osservatorio di Pavia and Auditel about Facebook as a platform for online marketing in Lee *et.al.* [5] research model. This study was performed by using knowledge mining through the security sector-related government institution in Italy to minimize the overload in OSINT and web mining.

Luis *et.al.* [6] proposed a study on Sentiment analysis of Twitter data. This study focuses on tweets in English about the thoughts and impressions of people about which is better in reviews on McDonald's or KFC. The main aim to analyse the opinion on these products and identify the valid popularity to maintain the business. Many people tweet in different time during a day but the lowest tweet mostly in the time of morning and the time and tweets changes by day to day among the population. The tweets are directly extracted from Twitter API. In this

ISSN: 2582-337X
Submitted: 18.08.2020
Accepted: 12.09.2020
Published: 18.09.2020

Journal of Ubiquitous Computing and Communication Technologies (UCCT) (2020)
Vol.02/ No.03
Pages: 178-186
https://www.irojournals.com/jucct/
DOI: https://doi.org/10.36548/jucct.2020.3.006

project R language is used that is a programming language for statistical computing and machine learning algorithms. The number of tweets about 7000 both in McDonald's and KFC based on positive, negative, and neutral opinions, and each response described separately by analysis programming. In McDonald's tweets about positive is 2184, negative is 1589, and the neutral 3227. While in KFC positive shows 2076, negative is about 1311, and neutral 3613. After this analysis, the reviews of positive and negative showed by McDonald's and KFC had slight variations in reviews, most of the positive and negative review tweets obtained from McDonald's when comparing KFC. Many people tweet about likes and dislike about McDonald's but KFC had more Neutral tweets. Using of R programming language to obtain reviews about popularity and negative or positive opinion helps to future use in business marketing.

The story of four major platforms of social media is Twitter, Facebook, Snapchat, and Instagram among college students. It is explained in terms of how they use it and the amount 7of time daily they spent on each site, motivation, and its uses in Alhabash *et.al.* [7] research work. A survey reported that 396 college students participated in the survey and they measured the intensity of the usage. It is showed that they spent most of the time on all these social media sites and all are under the age of 29. They also found the difference in all the media by uses and gratifications (U&G) for understanding the uniqueness of each site. U&G and Social networking sites (SNS) are considered as a theoretical framework that is often used. The main critique in U&G such as lack of measuring the use of social media, limited power, assumption of problem-related to the users, concentrating on needs, and motivations. SNS includes both positive and negative results. It helps to maintain and build relationships with the people socially.

Moe *et.al.*[8] performed social media analysis using college students, 33 students were failed in the test and it is reported that 97.2 percent of people having accounts on Facebook and 87.1 percent on Instagram, 79.1 percent on Twitter, and 84.3 percent on Snapchat. To find the difference among the various platforms, the reports were submitted on analysis of variance (ANNOVA) for measuring the usage. They have measured the time daily spent and mean the difference in intensity and motivations of using Instagram, Facebook, Twitter, and Snapchat by considering attributes like social interaction, self-expression, convenience, entertainment, time passing, and self-documentation. It will differ according to various platforms and the ranking is given for each social media site. The common motivation in all the social networking sites is social interaction and it's in the 7th position in the four major platforms. The implications of U&G and SNS are more reflected in different ways among college students. The major time spent on each platform is 506 minutes per day. This is the standardized approach of bringing the analysis, motivations, and uses among different social media.

The Sentimental analysis in social media and applications is analyzed by Xiang *et.al.* [9] through various methods, usage of social media platforms. All the social media content is available in the raw format that is used by the user. Then the required data is converted into meaningful information for further process of sentimental analysis. It uses the opinion-lexicon method for analyzing the text in social media. Nowadays, peoples are more frequently using social media and uploading their information through images, text, audio, or video. These all are considered raw and unprocessed data. The sentimental analysis uses Natural Language Processing (NLP) to extract the user opinion from the given text and then it is sent for analyzing whether it is positive or negative comments.

Srishti Vashishtha *et.al.* [10] research model used the given comments and review the user, analyzing the customer sentiment is quite easier and the decision is taken immediately based on the comments, and the product and services are also improved. It is composed of two methods for analyzing sentiments such as the lexicon-based approach and machine learning. Using the machine learning algorithms, it used to detect or extract sentiment from the data and lexicon works by identifying the positive and negative approach in Drus *et.al.* [11] research work. Social media information is classified into four types based on the usage and services providers such as Blogs and Micro-Blogs, Content Communities, and Social networking. Twitter is considered as the most valuable sites because most of the people expressing their views and user interaction is more when compared to other sites. Twitter has nearly 500 million tweets every day and provides easy public access through API. Facebook is not so popular for analyzing the sentiments and it is not arranged in structure and contains a spelling error. The applications of sentiment analysis are based on public action on health, politics, business, and marketing. In business, it is based on customer popularity and brands.

ISSN: 2582-337X
Submitted: 18.08.2020
Accepted: 12.09.2020
Published: 18.09.2020

Journal of Ubiquitous Computing and Communication Technologies (UCCT) (2020)
Vol.02/ No.03
Pages: 178-186
https://www.irojournals.com/jucct/
DOI: https://doi.org/10.36548/jucct.2020.3.006

Stieglitz et.al.[12] discussed about Lexicon method which uses Senti Wordnet and for machine learning, Naïve Bayes and SVM are used. It is important to choose the method wisely for analyzing and data is also important. To increase the accuracy and quality of the data, it is better to combine the methods. Every social media is different, each people has different views on different information. This method helps to understand the different customer views and provides better decision making. From the survey it is observed that emotional analysis in social media is important for identifying the fake information. Conventional models are lags in performance while classifying the mixed data. Considering these issues, proposed news classification model is developed in this research work using recurrent neural network.

3. Proposed Work

This section provides the architectural information of proposed emotional analysis model using recurrent neural network (RNN). It is a deep learning model which run multiple times and the output of each run is fed into input for next run. RNN is suitable for evaluating the sequences and the hidden layers are used to learn the data from its previous layers of the network as a sequence. Rather than perform classification over single perceptron, recurrent neural network uses entire sequence to train and predict the results. It has an advantage of processing dynamic input with varying lengths and model doesn't change the size based on the input. This provides better analysis performance over historical information. The weight function in the recurrent neural network depends on time function which helps to improve the classification performance with minimum computation time. The training process of RNN requires input data in an encoded format so that each word in the information is assigned with an unique integer. Embedding layer is the initial layer which is used to initialize the random weights and learn all the words in the information. Dense layer is used to classify the connected neural network layer and it is used to connect the input to output node. Similarly, dropout layer is used to activate random nodes in the network to prevent data overfitting.

Sequential analysis of given information is the vital idea for utilizing recurrent neural network in the proposed research work. Network attaches a time step for the given sequence and the maximum input length is proportional to this time step function. From this it is clear that recurrent neural network delineates the identified tasks within the group and outputs are calculated to define the memory cell. The information might be negative or positive, so that a features vectors are assigned and then it is proceeded into input layers of recurrent neural network and it is given as

$$hs_t = f(xI_t + w_n hs_{t-1} + b) \quad (1)$$

$$o_t = \text{Max}(y, hs_t) \quad (2)$$

where o_t is the output for time stamp t, I_t is the input for time stamp t and f is the non-linearity function and it is considered as Rectifier liner unit function (ReLU). In case emotional analysis recognize the word, truthfulness is essential for a specific sentence. Conventional recurrent models focus over the step size and the proposed recurrent neural network doesn't require these step function and using its hidden state the sequential information is updated to the network model. Further the network is refined using long short-term memory network (LSTM) which deals the dependency issues in the network. It addresses the gradient issues at training instances and reduces the issues using backpropagation. The hidden states and the outputs are formulated and obtained as follows

$$i_t = \rho\{c_i(hs_{t-1}, I_t + p_i)\} \quad (3)$$

$$o_t = \rho\{c_o(hs_{t-1}, I_t + p_o)\} \quad (4)$$

$$f_t = \rho\{c_f(hs_{t-1}, I_t + p_f)\} \quad (5)$$

The adaption rate of the network model is derived as

$$r_t = \tanh\{c_r(hs_{t-1}, I_t + p_r)\} \quad (6)$$

The state update based on the adaption rate and vector function is obtained as

$$R_t = f_t \times \{r_{t-1} + I_t \times r_t\} \quad (7)$$

$$hs_t = o_t \times \tanh(r_t) \quad (8)$$

where o_t is the output gate vector, I_t is the input gate vector and f are the non-linearity function gate vector, the cell parameters are given as c and p, input vectors are given as I_t, and ρ is the sigmoidal function. Figure 3 depicts the proposed recurrent neural network architecture

ISSN: 2582-337X
Submitted: 18.08.2020
Accepted: 12.09.2020
Published: 18.09.2020

Journal of Ubiquitous Computing and Communication Technologies (UCCT) (2020)
Vol.02/ No.03
Pages: 178-186
https://www.irojournals.com/jucct/
DOI: https://doi.org/10.36548/jucct.2020.3.006

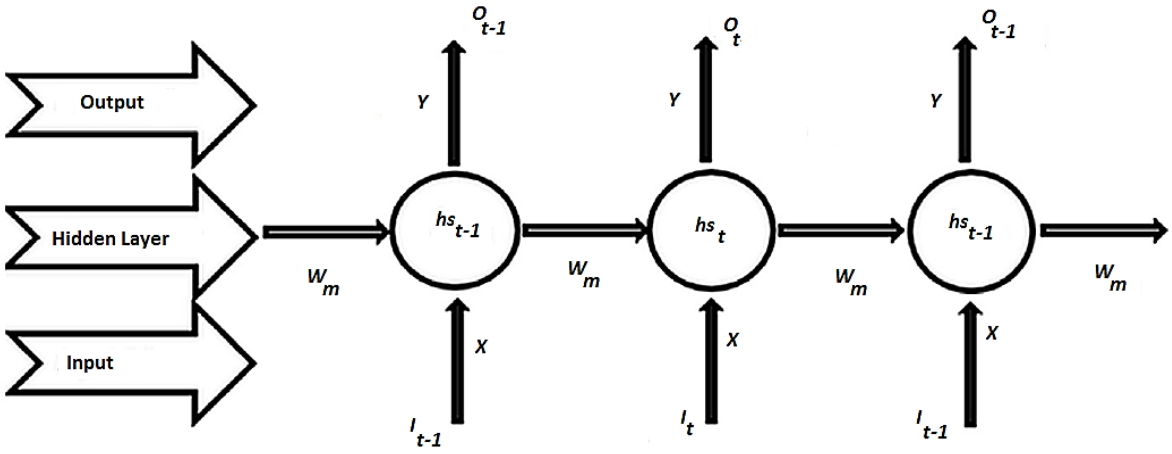

Fig. 3 Proposed Recurrent Neural Network

The proposed recurrent neural network is trained using backpropagation algorithm in which the error function is measured using repetitive connections. The time step in the design is neglected, instead chain rule is used to define the error which has advantages over hidden layer for data recall. So that the proposed neural network is computationally controllable based with respect to gradients. The long-term dependencies in the sentence attempt to process the full information using hidden state vectors. These hidden state vectors are fixed and the outputs of neural network is filtered using hidden states. The sentiment features are extracted using global average pooling layer which provides positive and negative outputs. The steps followed in preprocessing the information is summarized as follows

---

Step 1. Handling the uniodes and non-English words is the initial step in data pre-processing which removes the unnecessary terms from the sentence.
Step 2. Removal of URLs and user mentioned sites need to be removed
Step 3. Replace the abbreviations with exact terms for words minimization and avoid word similarity
Step 4. Common communal words replacement
Step 5. Numeric and numberings used in the information are need to be removed
Step 6. Convert the Emoji into suitable text
Step 7. Replace the Contraction
Step 8. Repeated contents are need to be identified
Step 9. Replace the negations and antonyms into correct word.
Step 10. Replace the stop words
Step 11. Similarity check in letter cases
Step 12. Replace the orthographic words
Step 13. Replace the incorrect words with correct spelling
Step 14. Tag the necessary parts of speech
Step 15. Perform lemmatization and stemming to avoid lengthy sentences

---

4. Result and discussion

The proposed model is experimentally verified and the performance metrics are validated through comparative analysis. ANN and DNN models are compared with proposed RNN model through the parameters such as accuracy, F-score. Since f-score is need to be obtained from precision and recall, those two parameters are also included for comparative analysis. The batch size is fixed into 3520 and number of epochs is 5. Three datasets are used in the experimentation such as keras, tensorflow and SS-TDS. All the three dataset has positive and negative information. Table 1 depicts the detailed information about dataset used in the proposed research model.

ISSN: 2582-337X
Submitted: 18.08.2020
Accepted: 12.09.2020
Published: 18.09.2020

Journal of Ubiquitous Computing and Communication Technologies (UCCT) (2020)
Vol.02/ No.03
Pages: 178-186
https://www.irojournals.com/jucct/
DOI: https://doi.org/10.36548/jucct.2020.3.006

Table 1 Dataset Description

Dataset	Total Information	Positive information	Negative information
Keras	11,228	6542	4686
Tensorflow	50000	25000	25000
SS-TDS	4242	1252	1037

From the analysis the results are categorized as fake, real and discernment based on the emotions such as excited, upset, guilty, scared, inspired, distressed and interested. Table 2-4 depicts the observed results for all the three data sets respectively.

Table 2 Analyses for each emotion for Keras dataset

	Excited	Upset	Guilty	Inspired	Strong	Distressed	Interested
Real	0.003	0.01	-0.03	-0.0002	-0.01	0.002	0.04
Fake	0.14	0.12	0.10	0.16	0.10	0.12	0.05
Discernment	-0.12	-0.12	-0.10	-0.16	-0.11	-0.12	-0.02

Table 3 Analyses for each emotion for Tensorflow dataset

	Excited	Upset	Guilty	Inspired	Strong	Distressed	Interested
Real	0.002	0.02	-0.02	-0.002	-0.03	0.001	0.02
Fake	0.12	0.11	0.09	0.14	0.12	0.10	0.04
Discernment	-0.11	-0.11	-0.09	-0.14	-0.12	-0.10	-0.04

Table 4 analyses for each emotion for SS-TDS dataset

	Excited	Upset	Guilty	Inspired	Strong	Distressed	Interested
Real	0.001	0.03	-0.01	-0.0001	-0.002	0.021	0.04
Fake	0.12	0.11	0.12	0.15	0.13	0.10	0.02
Discernment	-0.11	-0.11	-0.12	-0.15	-0.13	-0.10	-0.02

Comparative analysis of precision and recall is depicted in figure 4 and 5. The values are observed for dataset percentage values from 0 to 100% and observed values are compared with ANN and DNN models. It is observed that proposed RNN model attains better performance in precision and recall compared to other models due to its hidden layer properties and LSTM features.

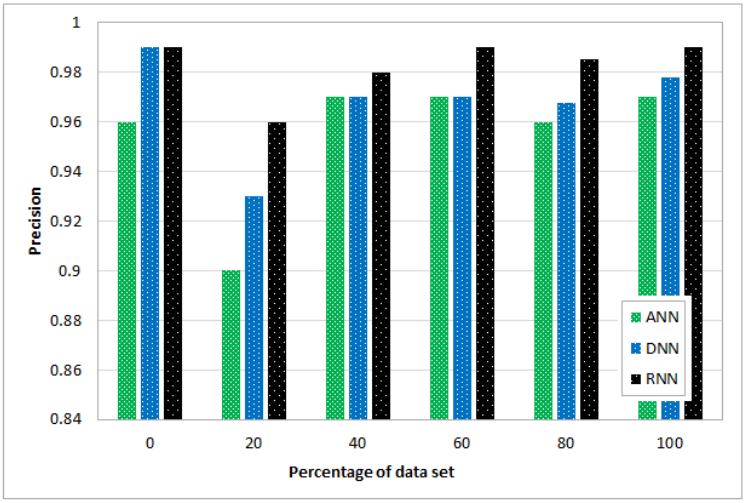

Fig. 4 Comparative analysis of precision

ISSN: 2582-337X
Submitted: 18.08.2020
Accepted: 12.09.2020
Published: 18.09.2020

Journal of Ubiquitous Computing and Communication Technologies (UCCT) (2020)
Vol.02/ No.03
Pages: 178-186
https://www.irojournals.com/jucct/
DOI: https://doi.org/10.36548/jucct.2020.3.006

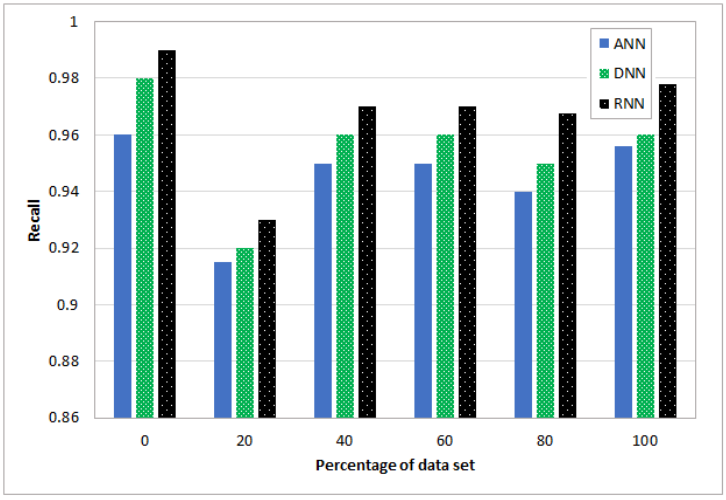

Fig. 5 Comparative analysis of Recall

Accuracy comparison of proposed model and other models performance against the dataset is depicted in table 5. Average accuracy is obtained to validate the system performance and it is observed that proposed RNN model attains better accuracy average of 94.97% which is 2% greater than DNN and 5% greater than ANN model.

Table 5 Proposed model accuracy comparison

Dataset	ANN	DNN	Proposed RNN
Keras	89.94	92.55	95.22
Tensorflow	88.62	92.42	94.86
SS-TDS	89.66	91.28	94.83
Average Accuracy (%)	89.41	92.08	94.97

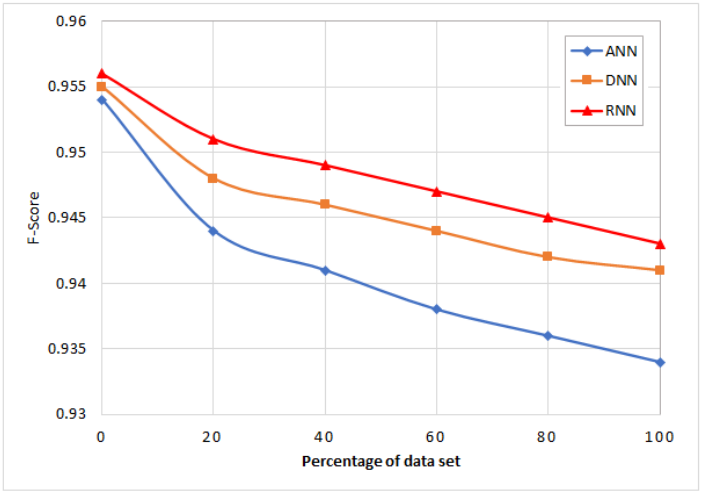

Fig. 6 Comparative analysis of F-score

Finally, the proposed model performance is compared against other neural network models in terms of f-score and it is depicted in figure 6. From the figure it is clearly visible that proposed RNN has better f-score compared to other network models which describes the proposed model attain maximum efficiency in emotional recognition and data classification performance.

ISSN: 2582-337X
Submitted: 18.08.2020
Accepted: 12.09.2020
Published: 18.09.2020

Journal of Ubiquitous Computing and Communication Technologies (UCCT) (2020)
Vol.02/ No.03
Pages: 178-186
https://www.irojournals.com/jucct/
DOI: https://doi.org/10.36548/jucct.2020.3.006

5. Conclusion

Emotional analysis of bogus statistics in social media is proposed in this research work to analyze the fake information in social media platforms. Fake information misguides the public by influencing them to believe bogus messages. In order to identify the truthfulness of shared information, the positive and negative portion of the information are need to be identified. Based on the information and its effects over human emotions are analyzed in this research work through recurrent neural network with long short-term memory. Three different social media datasets are used to experiment the proposed model and it is observed that proposed RNN model attain better detection and classification accuracy of 95% which is much better than conventional artificial neural network and deep neural network models. Further this research work could be improved using nature inspired optimization models to enhance the system accuracy.

References

1. Brandt, T., Bendler, J., & Neumann, D. (2017). Social media analytics and value creation in urban smart tourism ecosystems. *Information & Management*, *54*(6), 703-713.
2. Brooker, P., Barnett, J., & Cribbin, T. (2016). Doing social media analytics. *Big Data & Society*, *3*(2), 2053951716658060.
3. Wang, Z., & Ye, X. (2018). Social media analytics for natural disaster management. *International Journal of Geographical Information Science*, *32*(1), 49-72.
4. Zulfadzli Drus, Haliyana Khalid (2019). Sentiment Analysis in Social Media and Its Application: Systematic Literature Review. Procedia Computer Science. 161:707-714.
5. Lee, I. (2018). Social media analytics for enterprises: Typology, methods, and processes. *Business Horizons*, *61*(2), 199-210.
6. Luis Terán, José Mancera (2019). Dynamic profiles using sentiment analysis and twitter data for voting advice applications. *Government Information Quarterly*. 36(3):520-535
7. Alhabash, S., & Ma, M. (2017). A tale of four platforms: Motivations and uses of Facebook, Twitter, Instagram, and Snapchat among college students?. *Social Media+ Society*, *3*(1), 2056305117691544.
8. Moe, W. W., & Schweidel, D. A. (2017). Opportunities for innovation in social media analytics. *Journal of Product Innovation Management*, *34*(5), 697-702.
9. Xiang, Z., Du, Q., Ma, Y., & Fan, W. (2017). A comparative analysis of major online review platforms: Implications for social media analytics in hospitality and tourism. *Tourism Management*, *58*, 51-65.
10. Srishti Vashishtha, Seba Susan (2019). Fuzzy rule based unsupervised sentiment analysis from social media posts. *Expert Systems with Applications*. 138:1-15.
11. Drus, Z., & Khalid, H. (2019). Sentiment Analysis in Social Media and Its Application: Systematic Literature Review. *Procedia Computer Science*, *161*, 707-714.
12. Stieglitz, S., Mirbabaie, M., Ross, B., & Neuberger, C. (2018). Social media analytics–Challenges in topic discovery, data collection, and data preparation. *International journal of information management*, *39*, 156-168.

ISSN: 2582-337X
Submitted: 18.08.2020
Accepted: 12.09.2020
Published: 18.09.2020