Twitter Sentiment Analysis using Natural Language Processing

Suhashini Chaurasia
Computer Science
S.S. Maniar College of Computer and Management
Nagpur, India
ssuhashinic@gmail.com

Dr. Swati Sherekar
Department of Computer Science
SGBAU
Amravati, India
ss_sherekar@rediffmail.com

Dr. Vilas Thakare
Department of Computer Science
SGBAU
vilthakare@yahoo.co.in

2021 International Conference on Computational Intelligence and Computing Applications (ICCICA) | 978-1-6654-2040-2/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICCICA52458.2021.9697136

***Abstract*— Social media is the richest source of text generated by the user. So there is a necessity to automate the system to help organizing and classifying the opinions posted on social media sites. Proposed methodology framework using Artificial Recurrent Neural Network (ARNN) with bi-directional long short term memory (LSTM) has been used for the classification of sentiments. Structure for RNN with bidirectional LSTM is depicted. US airline Twitter sentiment dataset has been analysed using bidirectional LSTM model. Text with varying length is taken for the experiment. Graphical representation of the analysis has been depicted in this paper. Confusion matrix shows the result. At the end it is concluded that the sentiments are analysed and classified as positive, negative or neutral.**

Keywords—Sentiments, sentiment analysis, deep learning, Natural Language Processing, ARNN, bidirectional LSTM

I. INTRODUCTION

Researcher interests in the increase in social media generated text messages. These text messages are collected and processed by Natural Language Processing (NLP). With the rapid growth of various social media web sites like Twitter which provides rich text of data in large scales for various research opportunities [1]. With increase in development of social media platforms and various devices which provides means of sharing views, provides users a platforms where they can share their views is one of the source in contributing to big data. Twitter social media creates millions of tweets by millions of users daily [2]. So there is a necessity to organize and categories the data. This paper is one approach towards collecting Twitter data, analyzing and classifying. Here we address Twitter sentiment analysis through machine learning. Twitter is a social media where user can read and write short text messages called tweets. Systematically and scientifically study of semantic contents of these tweets is called sentiment analysis. It is a method for categorizing the polarity of an uploaded text on social media. The goal is to determine about a particular message of text about its polarity as positive, negative or neutral according to the standard of categorization [3].

II. LITERATURE REVIEW

There are various research paper which discuss classification algorithm. Some of the classification algorithms are Support Vector Machines, Aspect Based Sentiment analysis, Multinomial Naïve Bayes, ANN, CNN, LSTM. Here we have discussed ANN-CNN with LSTM for the classification of sentiments.

III. SOCIAL MEDIA SENTIMENTS

Social media is a virtual platform for posting text by the users. Text posted on social media like Twitter, Facebook, WhatsApp, etc. is called as social media sentiments. Sentiments are also called as opinion. These sentiments play a vital role in predicting people's mind. This will help in organizing and facing challenges.

With the increase popularity of social media posts in current era, finding people's attitude for a specific topic, text interaction or events has drove research interest in NLP and a new area of research has been introduced called sentiment analysis.

The model has been designed and tested using ARNN with LSTN on Twitter US airline dataset. Results and conclusion are drawn which classify those sentiments as positive, negative or neutral.

IV. SENTIMENT ANALYSIS

Sentiment analysis is the popular and growing area of research. Sentiment analysis is also called as opinion mining. Sentiment analysis means finding and classifying opinionated parts of text. These subjective parts needs to be recognized by various NLP and machine learning techniques and should be detached from part of the text. This type of technique is typically applied in the search for words which expresses opinion of a person or an individual.

There are two ways of classifying sentiment analysis technique - symbolic and non-symbolic. Non-symbolic sentiments are in the form of static text whereas symbolic sentiments are in the form of emoji. Sentiment analysis is a field of research in data analysis which determines about other people thinking towards an individual, topic, events issues or any other relevant text. It identifies the polarity of text as positive, negative or neutral [5]. Sentiment analysis deals with analyzing feeling, emotion and attitude of a speaker or a writer from a given text. Sentiment analysis or opinion mining is the application of text processing, computational linguistics and NLP to find and extract the information from the posted user generated social media big data. It involves apprehending of user's action, individual's likes and dislike from the sentiments posted on social media. Sentiments reflect thoughts, views and attitude of an individual focusing on emotions rather than reasons. Sentiments are considered as manifestation of the emotions and feeling. This paper deals with analyzing and determining the hidden information which is stored in the user generated text messages on social media. The hidden text gives the information regarding the user's intentions, tastes, likes and dislikes. Another way of classifying text is subjective or objective . Subjective text shows the text have opinion contents and objective shows that the text does not have specific objective to post. These techniques which are

978-1-6654-2040-2/21/$31.00 ©2021 IEEE

used to determine the opinion mining are deep learning and NLP.

V. DEEP LEARNING

With the rapid development in the field of deep learning, text processing has been increasing its attention [6]. Deep learning is a part of machine learning stimulated by ANN.

In previous research works machine learning has been employed directly in branch of Neural Network which provide a deep learning answers to various sentiment glitches. The learning environment is comprehended in three instants as simple RNN, LSTM and multilayer perceptron (MLP), which can learn extended dependencies. With influence of social media, the model employs online learning from the errors which are encountered and recover itself in the process [7].

Deep learning offers multiple means of learning data representations. Supervised and unsupervised learning are two different ways of deep learning. Deep learning helps in different layers which allow multiple processing. Deep learning in sentiment analysis have been initiative by feature learning where they can learn themselves and determine input representations from data dependencies. Their learning has been inspired by more number of training data with multiple class and word embedding [8].

VI. PROPOSED MEHTODLOGY

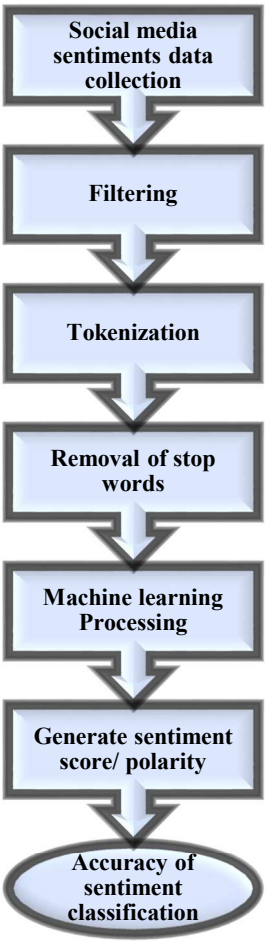


Fig. 1 Methodology for Sentiment Analysis on Social Media

Natural Language Processing

NLP is a technique which requires pre-processing of text or document or web content. Tokenization is a first step of sentiment analysis technique for most NLP tasks. Sentences are extracted from the text. Then a text is split into tokens called a phrase or a word. While splitting words by spaces some precautions should be taken like that of opinion phrases, name identities or like that. Likewise, some stop words does not provide any useful information, these must be removed [9]. Recent work in applying Convolutional Neural Network (CNN) to sentiment learning is discovered where low level features were shown. It has been noticed that increase in number of layers will increase the performance of CNN. In the previous research work, the very first layer makes an array of vector from the words of the sentence. Data which is taken as input is represented in the form of a matrix where each row contains words of a sentence [12].

In this research paper CNN with increase in number of layers has been proposed in a new methodology framework. The framework has been depicted in Fig. 1. Following are the steps involved –

A. Social media sentiments dataset classification

Datasets containing social media sentiments are already available on various web sites. In this research the dataset are used to analyze those sentiments and proceed for research.

B. Filtering

Filtering is the second step in the methodology. It means cleaning of raw data. Noise data which is not relevant in the context will be filtered out. Segregating the data according to the topic framed.

C. Tokenization

Tokenization means segmentation of sentences. Segmentation or tokenization of text is done by breaking text by punctuation or space to form container of words. Tokenization refers to the breaking of text into small tokens before forming an array of vectors. The unnecessary tokens are also filtered out using tokenization. So a text is converted into paragraphs. A paragraphs into sentences and sentences into words.

D. Removal of stop words

Stop words are useful syntactically and grammatically but they don't tell anything about the document type. These are neutral to the topic. These don't have enough descriptive power to distinguish between relevant and not relevant documents. These words need to be removed in the fourth step.

E. Machine learning processing

Machine learning is a branch of AI which offers the system an ability to lean automatically and expand from the learning without being unambiguously programmed. Machine learning provides accessing of data and use it to learn from experience. Various sentiment analysis methods are available but RNN with LSTM has been used in this

paper. This helps in analyzing the sentiments on social media network.

F. Generate sentiment score/ polarity

The main goal of sentiment analysis is to analyze user generated text on social media data. Categorizing these opinions into positive, negative or neutral value is called polarity. The results generated by it can be used to decide the polarity of the sentiments.

Sentiments will be classified according to the vocabulary used. This vocabulary is called as lexicon. This lexicon database is already available on various we sites which are used in this paper.

Collection of words or phrases labeled sentiment is called as Lexicon database. Additionally, a sentiment annotated database which contains from social media networking sites like Tweets, texts or messages labeled with concept or phrase from a general purpose lexicon. Words or phrases have two kinds of sentiments – relative and absolute. Depending on the polarity of the sentiments a database containing sentiment will be generated and can be provided for further research.

G. Accuracry of Sentiment classification

The final result showed the accurate and desired output of sentiment analysis on social media text. The accuracy will be measured quantitatively or qualitatively. Quantitative results are depicted in the form of graphs, charts and table.

VII. Structure of RNN and LSTM

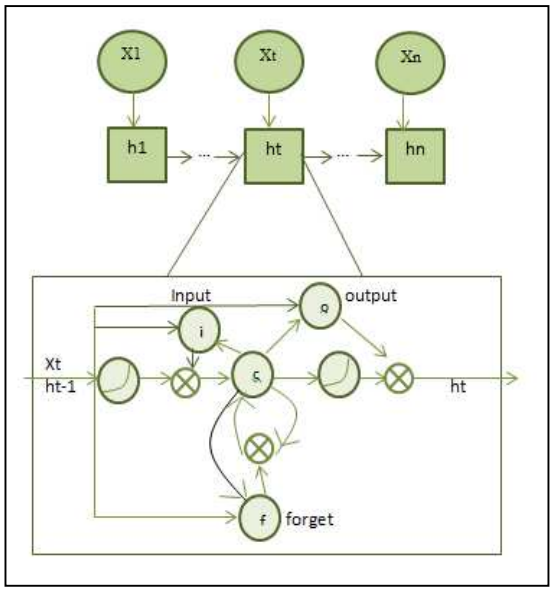


Fig. 2. Structure of RNN with LSTM[14]

A. Artificial Recurrent Neural Network (ARNN)

Recurrent neural network (RNN) is a branch of ANN. Most popularly accepted variant of RNN is LSTM which operates memory cell to tackle the aforementioned problems of basic RNN. RNN is a type of cell containing memory power to store information which can be calculated. These gated cells have the feature to capture and stores more information as compared to RNNs [13].

In this paper, ARNN with LSTM model has been used for analyzing sentiments.

B. Bidirectional LSTM

LSTM is a type of ARNN. It is plays vital role in learning long term dependencies. In Fig 1, i represents input gate, o represents output gate and a represents forget gate. These gates represent the flow of information into and out of the cell. Following mathematical formulas trains the neurons.

$$\begin{bmatrix} c \\ o \\ i \\ f \end{bmatrix} = \begin{pmatrix} \tanh \\ \sigma \\ \sigma \\ \sigma \end{pmatrix} \left(w \begin{bmatrix} xt \\ ht\text{-}1 \end{bmatrix} + b \right)$$

Fig. 2 depicts the structure of RNN with LSTM. Following abbreviations are used

c- class
o- output gate
i-input gate
f-forget gate
w- weight
b-bias
x-text sequence
h-update of LSTM unit
σ-logistic sigmoid function

VIII. Experimental Results

A. Data collection

Twitter US airline sentiment data has been collected and analyzed. Dataset contains 182329 sentiments of varying length. The datasets has been pre-processed to train a bidirectional LSTM model which is in turn used to predict the sentiments behind the tweets fetched in real time using tweepy. These are further categorized as positive, negative or neutral sentiments. It is programmed in Python.

B. Following steps are involved for data processing

1. Install and import dependency- tweepy has been installed
2. Cleaning and prepping dataset
3. Data visualization
4. Data pre-processing
5. Train and test split
6. Bag of words(BOW) feature extraction
7. Tokenizing and padding
8. Saving tokenized data
9. Train and test split
10. Bidirectional LSTM using Neural network
11. Model accuracy and loss
12. Develop confusion matrix

C. Sentiment analysis

Sentiment analysis functions at the juncture of information retrieval, ANN and NLP [15].

Fig. 3 depicts the graphical representation of US Twitter airline negative sentiments text length whereas Fig. 4 shows the graphical representation of the Twitter US airline negative sentiments text length.

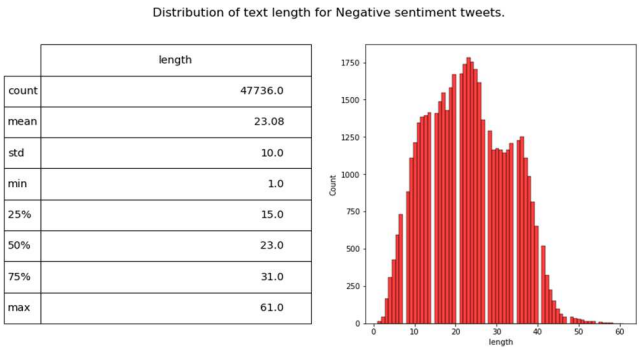


Fig. 3. Negative sentiments text length

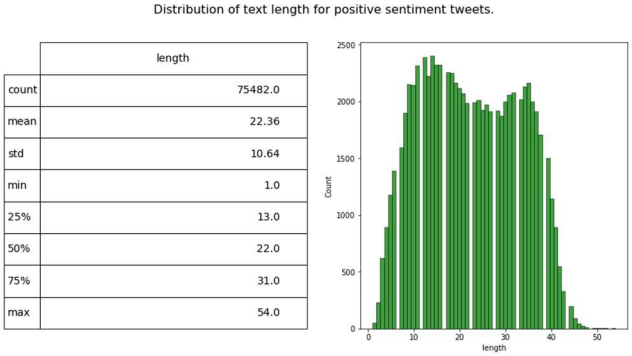


Fig. 4. Positive sentiments text length

A sample tweet:

Original tweet: when modi promised “minimum government maximum governance” expected him begin the difficult job reforming the state why does take years get justice state should and not business and should exit psus and temples

Processed tweet: ‘modi’, ’promis’, ’minimum’, ’govern’, ’maximum’, govern’, ‘expect’, ‘begin’, ‘difficult’, ‘obj’, ‘reform’, ‘state’, ‘take’, ‘get’, ‘justice’, ‘state’, ‘busi’, ‘exit’, ‘psu’, ‘ templ’

D. After tokenization and padding

[41 1 349 73 1911 1180 44 2465 2 1259 219 2 236 32
165 102 53 55 1184 236 50 3 6 533 3 50 3833 3
0 0 0 0 0 0 0 0 0 0 0 0 0 0]

E. Train and test split

Train set: [109397,50] [109397,3]
Validation set: [36466,50] [36466,3]
Test set: [36466,50] [36466,3]

F. Bidirectional LSTM using Neural Network

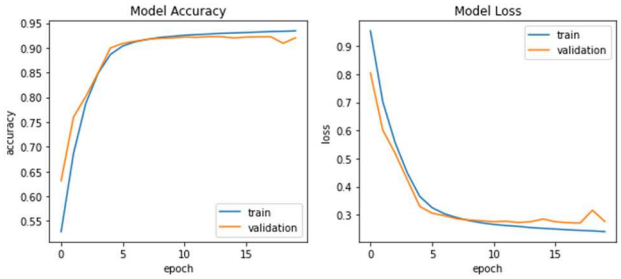
TABLE 1
ARNN with bidirectional LSTM

Layer(type)	**Output shape**	**Parameter**
Input layer	[none,50]	
Embedding	[none,50,32]	160000
Convolutional 1D	[none,50,32]	3104
MaxPolling1D	[none,25,32]	0
Bidirectional LSTM	[none,64]	0
Dropout	[none,64]	0
Dense	[none,3]	195

Total parameters: 179,939
Trainable parameters:179,939
Non-trainable parameters:0

G. Model accuracy and loss

Accuracy: 0.9152
Precision:0.9175
Recall:0.9127
F1 score:0.9151



H. Confusion matrix

The above result expresses sentiment analysis of Twitter data categorization as positive, negative or neutral. Traditional and most commonly used evaluation metrics is confusion matrix. It shows precision, accuracy, F-score and recall of the US airline sentiments which are analyzed [16].

Confusion matrix is a table which depicts the true values of classification model on a dataset of test data for which the performance are as shown below. The confusion matrix is very simple matrix and has been developed to predict the sentiment classification. Sentiments are then analyzed and classified as positive, negative or neutral. Fig. 5 shows the confusion matrix of US airline Twitter sentiments.

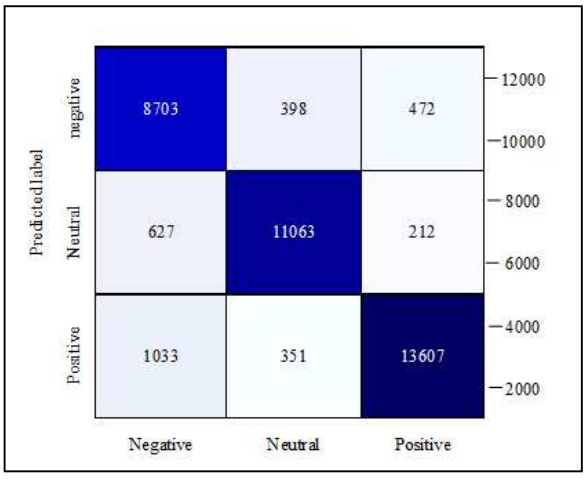


Fig. 5. Confusion Matrix

Input: I hate when I have to call and wake people up
Output: negative

Input: The food was meh
Output: neutral

Input: He is a best minister India ever had seen
Output: positive

VIII. CONCLUSION

Skinner et al. states that positive, negative or neutral sentiments are not only good or bad rather knowledgeable for emerging or challenging event [17]. In this paper , we have conducted experiments on the US airline Twitter dataset. Dataset are cleaned and then processed for analysis. We showed that the dataset is trained by ARNN with bidirectional LSTM using neural network. After passing neuron through embedding input layer, Convolution 1 D, max polling, bidirectional LSTM, dropout and dense output shape results were amazing. There is no dropout. Total parameter and trainable parameters are same. ARNN with bidirectional LSTM using neural network is presented in Table 1. Result shows that when ARNN with bidirectional LSTM model is used the accuracy is 0.91. Model accuracy is depicted in the graph. From 1 to 15 epoch there is relative increase accuracy. After 15 epoch there the graph shows linear behaviour. Model loss after 5 epoch. There is linear behaviour which shows that our model results give more accuracy with minimum loss. Confusion matrix table shows predicted sentiments are classified as positive, negative or neutral.

REFERENCES

[1] Yuxiao Chen, Jianbo Yuan ,"Twitter Sentiment Analysis via Bi-sense Emoji Embedding and Attention-based LSTM", International conference on Multimedia., pp -117-125,ACM, October 2018

[2] Shahid Shayaa, Noor Ismawati Jaafar, Shamshul Bahri, Ainin Sulaiman,Phoong Seuk Wai, Yeong Wai Chung, Arsalan Zahid Piprani And Mohammed Ali Al-Garadi, "Sentiment Analysis of Big Data Methods, Applications, and Open Challenges", Open Access journal, pp - 37807-37827, vol 6, IEEE Access , June 2018

[3] Gonzalo A. Ruz, Pablo A. Henriquez, Aldo Mascareno, "Sentiment analysis of Twitter data during critical events through Bayesian networks classifiers", Furture Generation Computer Systems, vol 106, pp - 92-104, Elsevier, May 2020

[4] Gonzalo A. Ruz, Pablo A. Henriquez, Aldo Mascareno, "Sentiment analysis of Twitter data during critical events through Bayesian networks classifiers", Furture Generation Computer Systems, vol 106, pp - 92-104, Elsevier, May 2020

[5] Mauro Dragoni, Soujanya Poria, Erik Cambria, "OntoSenticNet: A Commonsense Ontology for Sentiment Analysis – onto SenticNet", pp -77-85, IEEE intelligent system , vol 33, issue 3, June 2018

[6] Malik Khizar Hayat, Ali Daud, Abdulrahman A. Alshdadi, Ameen Banjar , Rabeeh Ayaz Abbasi, Yukun Bao, Hussain Dawood, "Towards Deep Learning Prospects Insights for Social Media Analytics", Open Access, pp – 36958-36979, vol – 7, IEEE Access, 6 March 2019

[7] Shubhankar Mohapatra , Nauman Ahmed , Paulo Alencar , "KryptoOracle: A Real-Time Cryptocurrency Price Prediction Platform Using Twitter Sentiments", International Conference on Big Data, IEEE, 24 February 2020

[8] S. Olivier Habimana, Yuhua LI, Ruixuan LI, Xiwu GU & Ge YU, "Sentiment analysis using deep learning approach an overview", Science China Information Sciences, vol 63, Springer, January 2020

[9] Shiliang Sun, Chen Luo, Junyu Chen, "A review of natural language processing techniques for opinion mining systems", Information Fusion, vol 36, pp - 10-25, Elsevier, July 2017

[10] Victor Camposa, Brendan Joub, Xavier Giro-i-Nieto, "From Pixels to Sentiment: Fine-tuning CNNs for Visual Sentiment Prediction", Image and Vision Computing, vol 65, pp 15-22, Elsevier, September 2017

[11] Seungwan Seo, Czangyeob Kim, Haedong Kim,Kyounghyun Mo, Pilsung Kang, "Comparative study of deep learning based sentiment classification", Open Access, pp - 6861-6875, vol - 8, IEEE Access, 1 Jan 2020

[12] Khuong Vo, Tri Nguyen, Dang Pham, Mao Nguyen, Minh Truong, Trung Mai, Tho Quan, "Combination of Domain Knowledge and Deep Learning for Sentiment Analysis", International Workshop on Multi-disciplinary Trends in Artificial Intelligence , pp - 162-173, Springer, 19 October 2017

[13] Sahar Sohangir, Dingding Wang, Anna Pomeranets, Taghi M. Khoshgoftaar , "Big data deep learning for financial sentiment analysis", pp -1-25, Springer, 2018

[14] Jinlong Ji, Changqing Luo, Xuhui Chen, Lixing Yu, andPan Li , "Cross-Domain Sentiment Classification via a Bifurcated-LSTM", Journal of Big Data, IEEE, 24 February 2020

[15] Kim Schouten, Flavius Frasincar, "Survey on Aspect-Level Sentiment Analysis", IEEE Transactions on Knowledge and Data Engineering, vol 28 , Issue 3, pp 813 – 830, 1 March 2016

[16] Anastasia Giachanou, Fabio Crestani , "Like It or Not A Survey of Twitter Sentiment Analysis Methods", ACM transaction on Computing Surveys, vol – 49, issue 2, Article No. 28, June 2016

[17] Rui Gaspar, Cláudia Pedro, Panos Panagiotopoulos, Beate Seibt, "Beyond positive or negative: Qualitative sentiment analysis of social media reactions to unexpected stressful events", Computers in Human Behavior, vol 56, pp -179-191, vol 56, Elsevier, March 2016