International Journal of Engineering Applied Sciences and Technology, 2020
Vol. 5, Issue 2, ISSN No. 2455-2143, Pages 202-209
Published Online June 2020 in IJEAST (http://www.ijeast.com)

SENTIMENT ANALYSIS AND PREDICATION MODEL

Vaishnavi Vats, Parth Garg, Vishwajeet Singh, Shikhar Yadav
IMSEC, AKTU University,
Ghaziabad, India

ABSTRACT: **Sentiment analysis on social media is a crucial a part of today's need for operation. Different machine learning techniques are utilized in recent years, and usage of an emoticon analytical to automatically annotate training sets has been a well-liked advent. As emojis are getting more popular to use in text-based elucidation this presumption explores the viability of an emoji training analytical for multi-class sentiment analysis employing a Multinomial Naive Bayes Classifier. Training sets consisting of 4000 to 400 000 tweets were wont to train the classifier using various configurations of N-grams. The results show that an emoji analytical accomplishes well compared to emoticon- or hashtag- based analytics. However, classifier befuddled is very hooked in to class selection and emoji depictions when multi-class sentiment analysis is accomplished.**

I. INTRODUCTION

Analysis of the emoticon and sentiments is additionally referred to as "data mining of the sentiment" or "emojis Artificial Intelligence" and affiliated to the use lingual processing (NLP), text mining, linguistics, and bio measurements to deliberately perceive, liberate, assess, and inspect sentimental states and subjective statistics. Sentiment analysis is usually concerned with the review in viewers aspects; for instance, surveys and reviews on the online and internet-based social media. In an aspect, sentiment analysis attempts to work out the inclination of a speaker, essayist, or other subjects in terms of theme via extreme sentimental or passionate responses to an archive, eludication, or occasion. The inclination could be a perception or appraisal, filled with emotion (in other words, the passionate condition of the creator or speaker) or an expectation of enthusiastic responses (in other words, the impact intended by the creator or buyer). Vast numbers of client surveys or recommendations on all topics are available on the online lately and audits may contain surveys on items like on clients or fault-findings of films, and so on. Surveys are expanding rapidly, on the idea that individuals wish to provide their views on the online. Large quantities of surveys are accessible for solitary items which make it problematic for clients as they need to peruse all so as to form a choice. Subsequently, mining this statistic, distinguishing client valuation and organizing them may be a vital undertaking. Sentiment mining may be a task that takes advantage of NLP and knowledge extraction advents to research an in-depth number of archives so as to collect the emotions of comments posed by different authors [1, 2]. This process incorporates various strategies, including computational and knowledge retrieval [2]. the essential idea of sentiment investigation is to detect the polarity of text documents or short sentences and classify them on this premise. Sentiment polarity is categorized as "optimistic", "pessimistic" or "impartial" (neutral). it's important to spotlight the very fact that sentiment mining is often accomplished on three levels as follows [3]:

- Document-level sentiment classification: At this level, a document is often classified entirely as "optimistic", "pessimistic", or "neutral".
- Sentence-level sentiment classification: At this level, each sentence is assessed as "optimistic", "pessimistic" or unbiased.
- Aspect and have level sentiment classification: At this level, sentences/documents are often categorized as "optimistic", "pessimistic" or "non-partisan" in light of certain aspects of sentences/archives and commonly referred as "perspective-level assessment grouping".

The main purpose of this paper is to review the prevailing sentiment analysis methods of Twitter data and supply theoretical comparisons of the state-of-art advents. The paper is organized as follows: the primary two subsequent sections discuss the

definitions, motivations, and classification techniques utilized in sentiment analysis. variety of document level sentiment analysis advents and sentence- level sentiment analysis advents also are expressed. Various sentiment-analysis advents used for Twitter are depicted including supervised, unsupervised, lexicon, and hybrid approach. Finally, discussions and comparisons of the latter are highlighted.

A study by Twitter in 2015 shows that 15% of tweets during TV clock time contain a minimum of one emoji which the foremost popular emojis aren't and but rather and. Multi-class sentiment analysis aims to utilize these statistics by using quite two classes of sentiment. Whereas traditional sentiment analysis determines whether a text is optimistic or pessimistic (polarity), multi-class sentiment analysis uses categories or clusters like excited, happy, bored and angry to raised understand the emotions expressed within the text.

Problem statement:

This report investigates whether multi-class sentiment classification of tweets can be achieved by automatically annotating training sets using an analytical based on emojis. The results will be evaluated against a testing set annotated by hand. The result will be used to answer the question of whether multi-class sentiment of tweets can be determined by using emojis as a training analytical? The contribution to the field of sentiment analysis will be a proof of concept with proposals for future research.

Definition and Motivation:

This analysis is often used to artifice for evaluation of estimation of people or clusters; for instance, a portion of a brand's followers or an individual customer in correspondence with a customer supports representative. With regard to a scoring mechanism, sentiment analysis monitors discussions and assesses dialogue and voice affectations to evaluate moods and feelings, especially those associated with a business, product or service, or theme. Sentiment analysis is a means of assessing written or spoken languages to decide whether articulation is optimistic, pessimistic or neutral and to what degree. The current analysis tools in the market are able to deal with tremendous volumes of customer criticism reliably and precisely. In conjunction with contents investigation, sentiment analysis discovers customers' opinions on various topics, including the purchase of items, provision of services, or presentation of promotions.

Furthermore, this study will only be concerned with tweets in English even though the analytical can be considered language agnostic. As the testing set used to evaluate accomplishance will be annotated by hand, knowledge of the language is required. All people involved in this study is proficient in English and therefore English was selected. This study only evaluates the generated training sets using a Naive Bayes classifier. Naive Bayes classifiers have been proved to work well in previous studies [12].
Furthermore, only unigram and bigram models, as well as a combination of both will be tested and evaluated.

II. RELATED WORK

In this work, the area of focus is on how a machine learning algorithm uses supervised learning techniques to accomplish sentiment analysis on data that is either collected and stored in a comma separated values file or retrieved online. A broad view of this work can be understood by the System Block Diagram as shown in figure 1.

International Journal of Engineering Applied Sciences and Technology, 2020
Vol. 5, Issue 2, ISSN No. 2455-2143, Pages 202-209
Published Online June 2020 in IJEAST (http://www.ijeast.com)

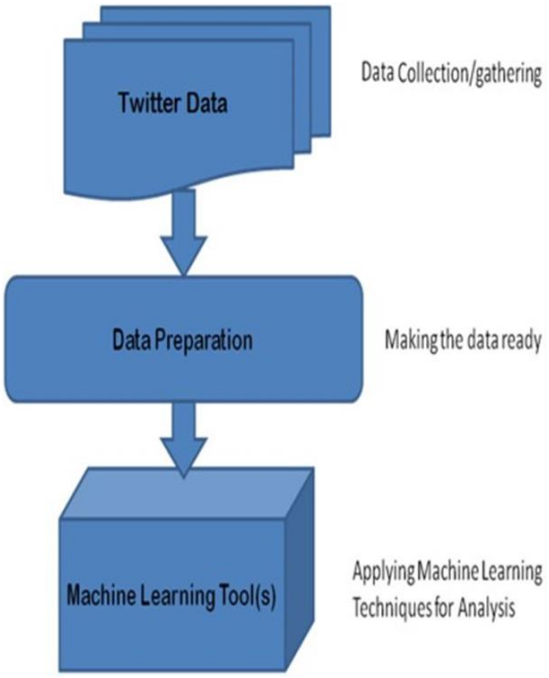


Figure 1. System Block Diagram

The "Supervised learning algorithms" are trained with the assistance labeled instances, like an input that the output expected is understood. Consider an instance, a bit of kit may have data points labeled "0" as "inactive" or "1" as "active". Set of inputs with the respective correct outputs are supplied to the training algorithm, and therefore the algorithm learns by comparing its correct output with the particular outputs to seek out errors, then it modifies the model accordingly. Using methods like classification, regression, prediction, and gradient enhancement. The supervised learning uses models to predict tag values for extra unlabeled data. Assisted learning is usually utilized in applications where earlier data or historical data predicts events likely to occur in future. for instance, to classify feelings as optimistic, pessimistic, neutral or unsure.

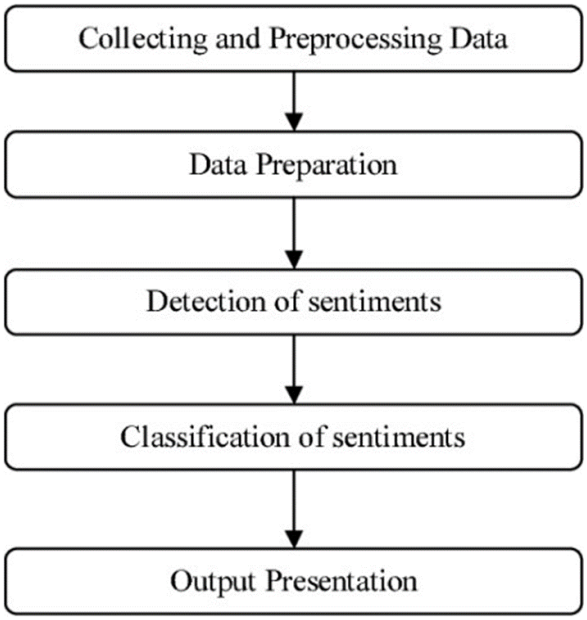


Figure 2. Module Design

III. MODULE DESIGN

Module design or modularity in design is a designing advent that splits a larger system in smaller parts known as modules. In this work important modules included are Training Data & Testing Data, Pre- processing, Feature Extractor, Features, Machine Learning Algorithm, Classifier Model and Label. Every module plays a significant role in overall implementation.

1. Training Data: Contains set of attributes and its instances which are given as input to classification algorithm to train model.
2. Test Data: Test Data is the dataset used to test the trained model which is in the similar form as that of training data.
3. Preprocessing: Initializing the data by accomplishing specific tasks and make the data ready in order to process further.
4. Feature Extractor: Uses technique(s) to draw the important features possessed by the tweet.
5. Features: Features are the important numerical facts & figures drawn by the extractor
6. Classifier Model: Classifies the features extracted using classification algorithm.

International Journal of Engineering Applied Sciences and Technology, 2020
Vol. 5, Issue 2, ISSN No. 2455-2143, Pages 202-209
Published Online June 2020 in IJEAST (http://www.ijeast.com)

Preprocessing Steps:
The aim of the following preprocessing is to create a Bag-of-words data depiction. The steps will execute as follows:

1. Cleansing
    a. Removing URLs
    b. Removing usernames (mentions)
    c. Removing tweets with Not Available text
    d. Removing special characters
    e. Removing numbers (digits)
2. Text processing
    a. Tokenization
    b. Transformation to lowercase
    c. Stem
3. Build word list for Bag-of-words

IV. DATA FLOW DIAGRAM:

This diagram shows what quite statistics is going to be input to and output from the system, where the info will come from and attend, and where the statistics is going to be stored. It doesn't show statistics about the timing of process or statistics about whether processes will operate in series or in parallel.

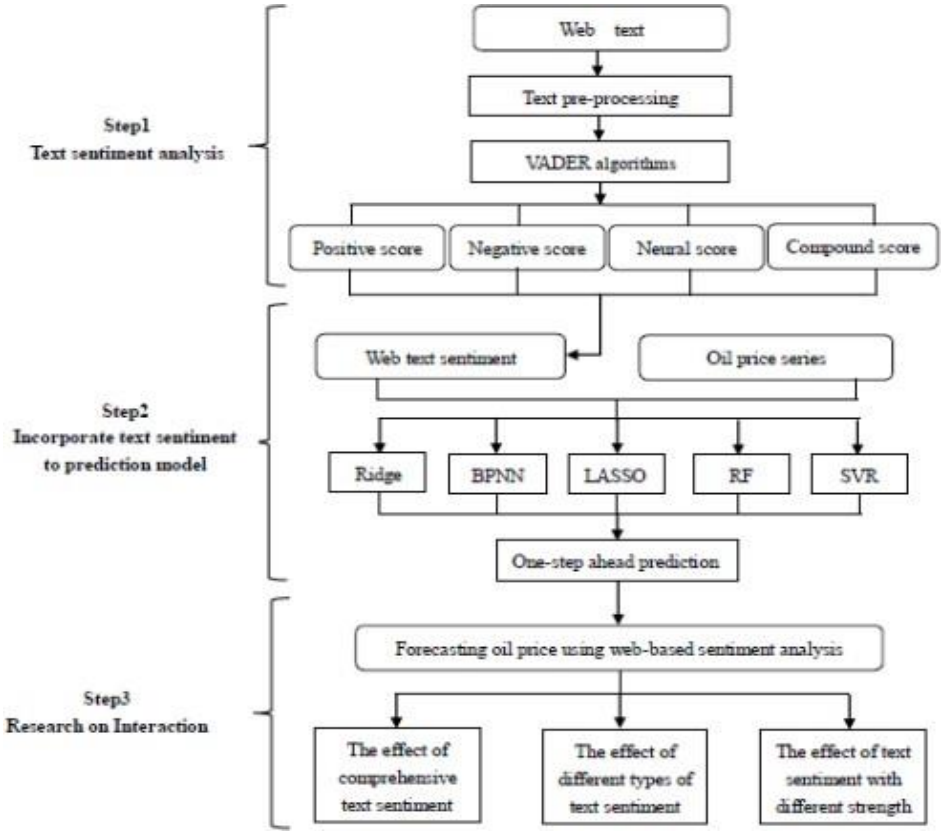


Figure3. Data Flow Diagram

V. METHODOLOGIES:

Tools and Techniques:

PYTHON:

In today's era vast number of mastery generation is unstructured which are often developed for perception. Blogs, social media's post, news articles, search history on internet are some instances related to unstructured facts and figures. NLP is the lingua processing for the strategy for analysing the lingua. It is used to classify the texts and information into pre-defined emotions, which is a frequent function. To research textual facts and figures we are often used NLPK lingua processing kit in python as library.

International Journal of Engineering Applied Sciences and Technology, 2020
Vol. 5, Issue 2, ISSN No. 2455-2143, Pages 202-209
Published Online June 2020 in IJEAST (http://www.ijeast.com)

GLOVE.TXT:

For retrieving vector depictions, we can use an unsupervised algorithm is known as Glove.txt.Corpus is used to retrieve average co-occurrence of texts in information through training and as outcome depictions provides the linear substructure of word vector space.

KERAS:

Neural network blocks which are activation function, layers, optimizers, numbers of tools and techniques which are used for implementation of the form of data, text, image, etc. which comprised in keras. It is used to accomplish the code easy and fast with deep learning.

NUMPY:

For highly accomplished multidimensional arrays object and tools which help in these arrays we can use a main library of python known as NumPy. It provides important mathematical functions.

PANDAS:

Wes McKinney developed a high-level information manipulation tool known as Pandas. It works on data frame which developed on NumPy package in python. It is key data structure which store and change the facts and figures in row observations and column variables which analysis facts and figures of specific cluster.

TENSORFLOW:

Google provide us a python library for easy and fast numerical calculations is known as TensorFlow. We can use wrapping libraries and also create deep learning models which simplify the process of it.

Recurrent Neural Network:

- According to the neural network outcome of the previous input is work as hidden layer in the network which further used to attain the next output, this series takes place in each layer. This composition of the inputs attains the happening time t.
- RNN are important for the values of the hidden layers which cannot be postier which store facts about previous input.

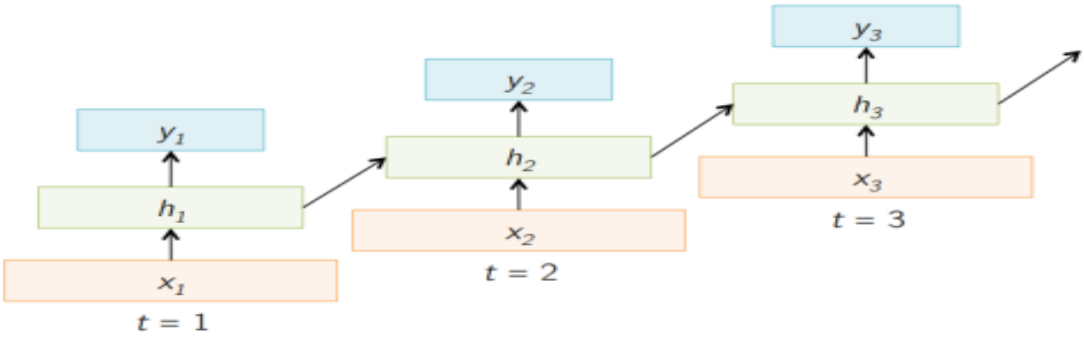
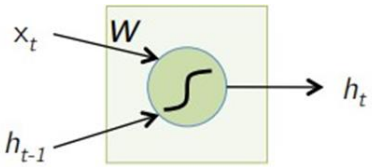


Figure 4 Sample RNN network

RNN Cell:

RNN cell are used as duplicates with different time set of the inputs made by unfolding and unrolling of the layers.



$$h_t = \tanh W \begin{pmatrix} x_t \\ h_{t-1} \end{pmatrix}$$

International Journal of Engineering Applied Sciences and Technology, 2020
Vol. 5, Issue 2, ISSN No. 2455-2143, Pages 202-209
Published Online June 2020 in IJEAST (http://www.ijeast.com)

Sentiment Classification:

- Categorize a tweet or comment as optimistic or pessimistic.
- Inputs: one or more than one sentences.
- Outcomes: Optimistic / Pessimistic category.
- "Indian soldiers are real heroes."
- "The hotdog crossed the road because it was uncooked."

Input Output Scenarios:

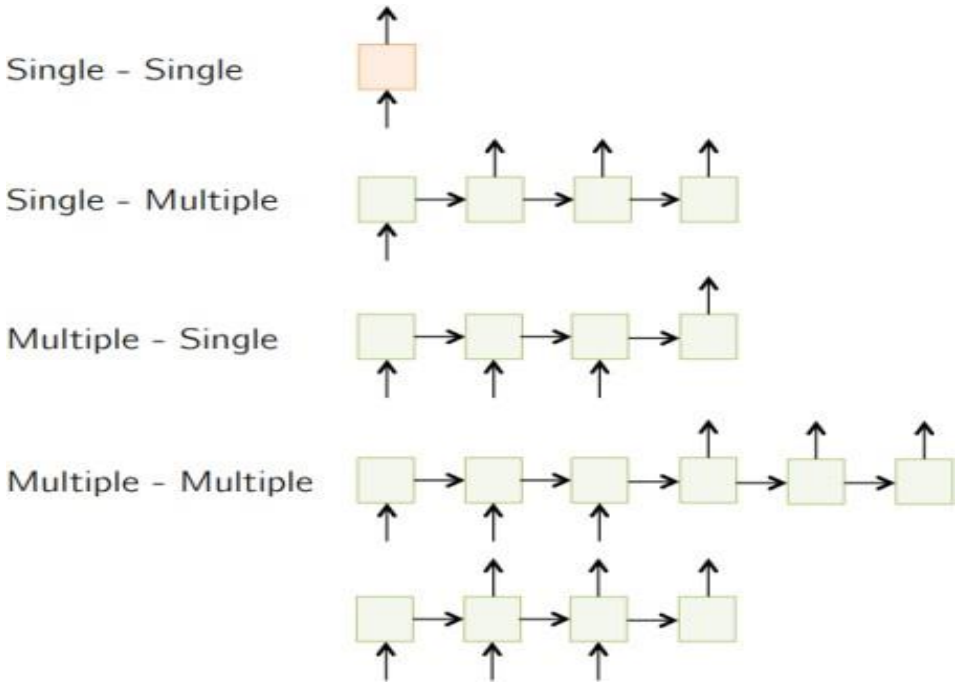


Figure 5 Input and Output

VI. RESULTS

Analysis of the emotions are the most necessary aspects of the today's era of the texts and information of social media. What the viewers view on their incoming texts or the blogs on the social media are often analyse whether it is optimistic, pessimistic or neutral. By taking any sentence as a demo to play the emotion of that corresponding sentence.

Objective Analysis

Objective of this analysis is to describe the intention of the texts or information receive by viewer by step by step of the message. This information is often an opinion, marketing, social media, suggestions related to any statement.

University: A deep analysis

Education is the important part in the sharing economy. University is the providence of the education in all over world. Universities get lot of the suggestions, comments, feedback, complains by the students, parents, faculty and people in all over world. Hereby social media is the vast platform for reviewing the such issues. These vast numbers of incoming facts and figures provide analysing, classifying and generating insights challenging undertaking.

We analyse the few online comments and conversation happen on social media as like facility, fee structure, environment, affiliations, services, placements and education system.

For vast coverage of facts and figures of the sources, we are often defining the university official Facebook official page on YouTube, tweets mentioning the university and latest news articles around university.

Analysing the comments and the conversations on the social media can help in overall brand perceptions. Further to dig deeper contextual sentiment search will help in classification of facts and figures.

FACEBOOK

Facebook is the vast platform where comments are surrounded in common conversation, news articles, blogs, marketing and promotional contents and spam/junk/unrelated content. For dig deeper we have to analyse the intentions of the comments on the post of the universities. Therefore, comments are optimistic and pessimistic according to the people which are studied in university or faculty or having some experience incident person. Comments are on fee structure, placements report etc.

TWITTER

Tweets on twitter describe the sentiments related to university which is mention in the related tweet. For the safety category optimistic tweets are classify in the related keyword which describes the perception regarding the university.

Fee structure, placements offers, facilities inside the campus, further curricular activities are vast area where people peek a view in the tweets and comments on twitter. This describe that people talk about the placement offers and the fee structure of the university.

University rely the fee structure and placement offer as optimistic comments or tweets of low percent and pessimistic of high percent. So, university analyse them to improve the placement offers and fees structure.

NEWS

Articles of the news are most described features for optimistic in overall and individually in any topics. It is safe about to most talked topic of the university.

News articles are defined on the sharing score which are scored by the sharing numbers or the posting of the people regarding that news article. Now we can classify the news on this basis which describe the optimistic, pessimistic and neutral of the university.

For instance

Ten students of the university topped in the state.

Football Team of university scored first position in national level competition.

Placement percentage of the university is 100% this year.

VII. CONCLUSION AND ACKNOWLEDGEMENT

Thereafter it is analysis that emoticons are used by the twitter, facebook, youtube and Instagram user. The execution that has been carried out illustrates the figures and information that the twitter, facebook, news article data fetched offline and online is put under accomplishing pre-processing tasks such as removal of stop words etc. and only the text along with emoticons essential to draw the sentiment is examined further for feature extraction. This includes accomplishing operations like stemming, using Porter-Stemmer, Lemmatizing, and removal of Punctuations. Several analyses were accomplished to analyze the sentiments of the comments and tweets by using tools and techniques of machine learning. Eventually, Accomplishance analysis is carried out, that computes overall accomplishance of the testing data, from the datasets available without internet source, to assess their "sentiments", "befuddled matrix" and "accuracy" based on 'Optimistic', 'Pessimistic', 'Neutral' or 'Not Sure' values, which concludes that outcomes obtained in analysis “with emoji” are vast accurate.

When contrasted to analysis “without emoji”, as produced in graphical depiction, i.e. bar graph, pie chart etc. for the contrasted and better aiming of the accomplished analysis and for better perceptible appearance. Hence, in today’s social media era, the emojis play significant part in expressing the sentiment of a comment, conversation, tweet. For future work, we plan to annotate updating the Python-Code with more optimized and efficient code. For more accuracy in results with less error rate, and fast processing of the inputs, tools and techniques of the machine learning can be replaced by some advance algorithms and techniques to compute and analyse and bring out outcomes in little bit of time.

VIII. REFERENCES

[1] Hao Wang, Jorge A. Castanon Silicon Valley Lab, IBM, USA. “Sentiment Expression via Emoticons on Social Media”.
[2] Ms. Payal Yadav, Prof. Dhatri Pandya “SentiReview: Sentiment Analysis based on Text and Emoticons” International Conference on Innovative Mechanisms for Industry Applications. (ICIMIA 2017)
[3] Alexander Hogenboom, Daniella Bal, Flavius Frasincar, “Exploiting Emoticons in

Sentiment Analysis".

[4] Waghode Poonam B, Prof. Mayura Kinikar, "Twitter Sentiment Analysis with Emoticons" International Journal of Engineering And Computer Science ISSN: 2319-7242 Volume 4 Issue 4 April 2015, Page No. 11315-11321

[5] Varsha Sahayak, Vijaya Shete, Apashabi Pathan, "Sentiment Analysis on Twitter Data" International Journal of Innovative Research in Advanced Engineering (IJIRAE). Issue 1, Volume 2 (January 2015)

[6] A. Beryl Joylin, Aswathi T, Nancy Victor, "Sentiment Analysis based on Word-Emoticon cluster" International Journal of Pharmacy and Technology (IJPTFI). ISSN: 0975-766X.

[7] Molly Redmond, Sadegh Salesi and Georgina Cosma, "A Novel Advent Based on an Extended Cuckoo Search Algorithm for the Classification of Tweets which contain Emoticon and Emoji", International Conference on Knowledge Engineering and Applications (ICKEA).

[8] Fred Morstatter, Kai Shu, Suhang Wang, Huan Liu, "Cross-Platform Emoji Interpretation - Analysis, a solution and Applications".

[9] Katarzyna Wegrzyn-Wolska, Lamine Bougueroua, Haichao Yu, Jing Zhong, "Explore the Effects of Emoticons on Twitter Sentiment Analysis".

[10] Harsh Thakkar, Dhiren Patel, "Adventes for Sentiment Analysis on Twitter: A State-of-Art study", Department of Computer Engineering, National Institute of Technology, Surat, India.

[11] Wiesław Wolny, "Sentiment Analysis of Twitter data using emoticons and emoji ideograms", University of Economics w Katowice.

[12] AnalyticsVidhya. (2016). A Complete Tutorial to Learn Data Science with Python from Scratch. Available:

[13] Lena Kallin Westin, "Receiver operating characteristic (ROC) analysis Evaluating discriminance effects among decision support systems", Department of Computing Science, Umeå University, SE-90187 Umeå, Sweden.

[14] R. Pascanu, T. Mikolov, and Y. Bengio, On the difficulty of training recurrent neuralnetworks, ICML 2013

[15] S. Hochreiter, and J. Schmidhuber, Long short-term memory, Neural computation, 1997 9(8), pp.1735-1780

[16] F.A. Gers, and J. Schmidhuber, Recurrent nets that time and count, IJCNN 2000

[17] K. Greff , R.K. Srivastava, J. Koutník, B.R. Steunebrink, and J. Schmidhuber, LSTM: A search space odyssey, IEEE transactions on neural networks and learning systems, 2016

[18] K. Cho, B. Van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, Learning phrase depictions using RNN encoder-decoder for statistical machine translation, ACL 2014

[19] R. Jozefowicz, W. Zaremba, and I. Sutskever, An empirical exploration of recurrent network architectures, JMLR 2015