Sentiment Analysis using CNN AND SVM

Dr. J.V.ANCHITA ALAGAMMAI[1], K. GURU PRIYA[2], T.R.DIVYA[3], P.K.PRIYANKA[4], K.S.RAGAVI[5]

Associate Professor, Dept. of CSE, Velammal College of Engineering and Technology, Madurai, Tamilnadu, India[1]

UG Student, Dept. of CSE, Velammal College of Engineering and Technology, Madurai, Tamilnadu, India[2,3,4,5]

ABSTRACT: Due to the increase usage of social media,a new stage of communication has been developed between human beings. People create and forward messages to show their interests in a particular information.The use of natural language processing, text analysis, computational linguistics and biometrics to identify, extract and to study states and subjective information is known as sentiment analysis.The algorithms for the detection of sarcasm on twitter is of great challenge. Here,a deep learning model namely Convolution Neural Network and machine learning algorithm namely Support Vector Machine has been used and their results have been compared.

KEYWORDS: Sarcasm, Sentiment Analysis, Twitter, natural language processing,SVM.

I.INTRODUCTION

The major protest in sentiment analysis function is to estimate the accuracy of the narrativein usual sense so as to coordinate text based on polarity(positive or negative).Sarcasm is a figure of speech which is very difficult to determine. Sarcasm can be defined as "a sharp, bitter, or cutting expression or remark".

The initial stage of web 2.0 and Online Social Networking(OSNs) sites have given advanced element to the global connection through everyone and has contributed large number of event for sharing common global idea to calculate and confirm patterns. Recently, Social network sites like Facebook, Instagram and Twitter has obtained large popularity and importance.

Understanding the proves about a given advent, permit the notice or perception of conflict between the objective polarity of the advent and its sarcastic exclusive by the author, as in "I love the pain of breakup".

"Sarcasm is a type of sentiment where human specific their glommy feelings using conclusive or enhanced conclusive words in the content"[1].It is a persuasive speaking or writing that presents a expression important discord between the certain position and the spoken text. For example, a post/tweet" , It is fabulous feeling to loss hours in traffic jams!" certainly express this disagreement between the certainposition of "being stuck in traffic jam" and the utterance text "fabulous".

The initial contribution of this paper involves:

Developing the act of sentiment analysis by constructing, practicing and checking a deep learning version, sAtt-BLSTM convNet ,a compound of soft attention based bidirectional long short-term memory(sAtt-BLSTM) and convolution neural network(convNet).

Aspect engineering for self-starting sarcasm detection using descriptive literary techniques such as ,emoticons,and punctuation marks.

Acting standard by analysis different kinds of versions and datasets for sarcastic tone classification in real time.

The deep learning version introduced here (sAtt-BLSTM convNet) has eight layers:

Input Layer:

The take in layer of a semantic network consists of artificial take in neurons and it carrys the beginning level input into the structure for another stages of transforming by subsequent layers of artificial neurons. This layer marks the beginning of the workflow for the artificial semantic network.

Embedding layer:

Every chats in tweet is summarized into a low aspect angle with the use of GloVe. This encodes the take in into original - valued vectors called embedding vectors.

BLSTM Layer:

The turn out of the previous chat-embedding layer is given as input into the BLSTM layer. The prime determination of this layer is to study long level characters from the earlier step.

Attention layer:

They demonstrate a stand –alone self attention layer which can be used to replace spatial convolutions and build a fully attentional model.

Convolution layer:

The convolution task is developed to gain a turn to character vector.It is the early layer to abstract character from an take in image. Convolution is a mathematical counting that takes two take in such as image array and a filter or kernel.

Activation layer:

This layer is an activation function that decides the final value of a neuron. For example, the cell value should be 1,however it has a value of 0.85,since the probability of 1 can never be achieved , activation function can be achieved.

Down-sampling Layer:

Down-sampling strategy is used by max-pooling operation in convNets.

Representation Layer:

A representation layer which comprises of a definite remodel layer and softmax activation operation to produce the turn out anticipate.

II.RELATED WORK

The net users expand to pursuit of public network has escalate analyzer attract to sharp mine the idea available , both significant and subjective. The renowned micro-blog, joins two person across the world and huge matched of people crisis in Twitter.

- The proposed version for feeling assay of football definite tweets using three classifiers random forest, SVM and multinomial Naives Bayes by Aloufi and EI Saddik
- The conclude of automobile auction by Sentimental Assay of Twitter input and store exchange values using least squares support vector regression by Pai and Liu.
- The equal-sided join of LSTM capable of catching complicated connotation depiction of natural language texts and assess on standard dataset, the Stanford Sentiment Treebank by Wu and Chi.
- The expanded idea of binary or ternary classification and suggested an approach to arrange text possessed from Twitter into seven sentiment classes by Bouazizi and Ohtsuki.

The field was done by(Tsur et al., 2010) in early stages based on 6600 dataset manually elucidate Amazon write-up using kNN-classifier over dagger based and pattern based features (i.e) the ordered arrangement of high frequency words.(Gonzalez-Ib ´ anez et al., 2011) felt manipulated support vector machine and logistic regression over a element of unigrams, reference-based literal features and logical features (e.g emoticons) and analyzed the performance of the classifiers with that of person.The admitting irony at a linguistic level is report a set of textual features, exclusively in short texts created via Twitter, and composed a new model that was judged along two dimension, representation and applicability by (Reyes et al., 2013).The present of positive sentiment in close contiguity of a opposing situation phrase as a feature of sarcasm detection used by(Riloff et al., 2013).The classification of Dutch tweets as sarcastic vs non-sarcastic used Balanced Window Algorithm by (Liebrecht et al., 2013); n-gram(uni,bi,tri) and intensifiers were used as feature of distribution.

The data generation is concentrated in large numberof studies on sarcasm detection. Gonzalez-Ibanez

et. al. generated corpus that includes only ironicutterances that have been clearly identified as such

by the message writer. The determination of sarcastictweets from tweets in difficulty level and reported

that are positive or negative sentiment. The machinelearning effectiveness consider impact of literal

and logical factors for identifying sarcastic utterances. The introduction of self-annotated—labeled by the author of the statement himself/herself by Khodak et.

AI using Reddit corpus for sarcasm research containing 1.3 million irony statements. They checked the corpus for efficiency, constructed benchmarks for irony detection and certainty baseline methods. For assess the corpora there are three benchmark: size, the proportion of sarcastic to non-sarcastic comments, and the rate of false positives and false negatives. Their work did not have any definite finding, rather the collection andassess of a large dataset was their main contribution

III.PROPOSED SYSTEM

In architecture can include of structure parts and the sub-structure envoled, that will effort cooperatively to implement the overall structure.There have been works to formalize languages to describe structure architecture, collectively these are called architecture description languages(ADLs).
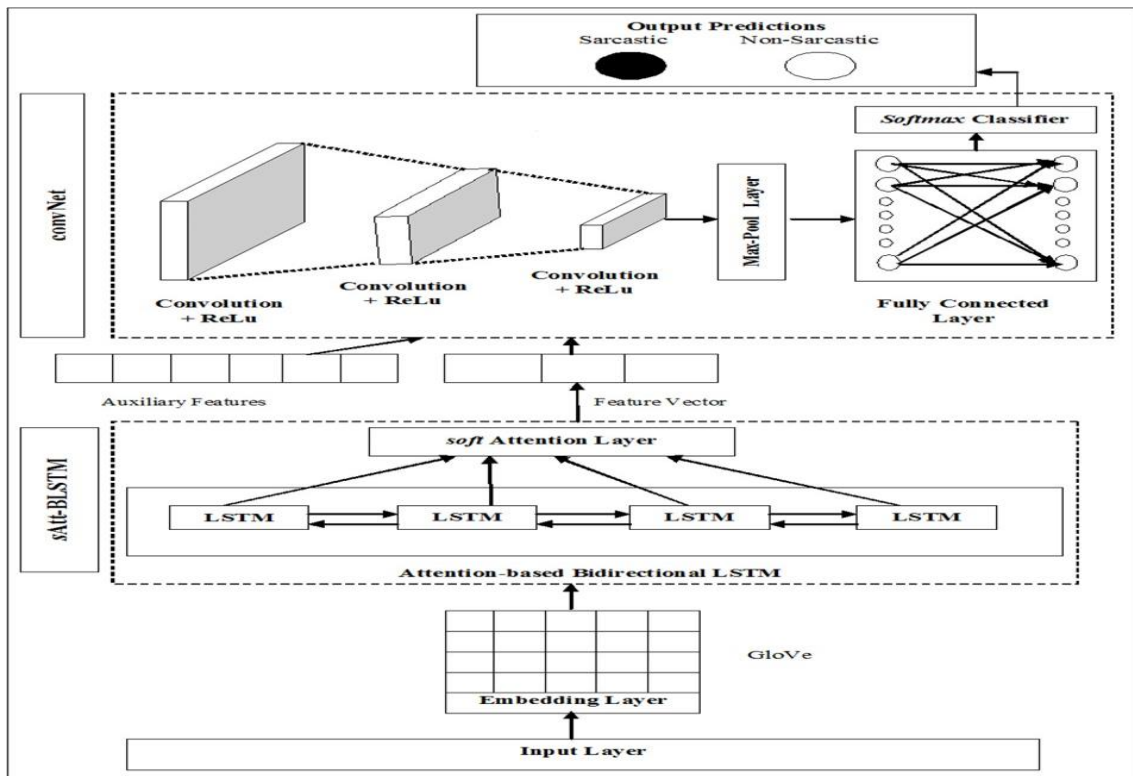
LAYER DESCRIPTION

A.INPUT LAYER

The tweet which has been collected from the dataset is given as a input to this layer which has been fed after the preprocessing. This layer is joined to next layer called embedding layer, which create word embeddings using Glove.

B.EMBEDDING LAYER

The fuction of this layer is to maps the input to real-valued vectors which has been taken from the look-up tables using encoding.Word embeddings is used to enhance learned word representations and it has a feature to detect sarcasm that have been recently reported.To construct word embeddings,a word vector table is used which is generated by GloVe.It is a calculate-based version of constitue words by feature vectors.The log-bilinear version research the communication of words by calculate repeatedly they co-occur. This version is to map the tokenized words in each tweet to their various phrase vector tables.To specify feature matrix,Proper padding is performed.It means that if the sum total of tweets is taken as Z and there is Y with N tokens,we can generate the word angle table with dimension d of the phrase vectors is completed using Glove. After the mapping has been done, each tweet is indicated as a angle of the word embeddings concatenation(E). Thus the feature vector array is attained as given in (1).

$F = U + B + T + P + E$, (1)

Where + is concatenation operator of the vector.

Each tweets has different portion, the tweets with the maximum length are used as threshold value to specify the vector matrix representation of tweets and to attach the portion of the tweet array. Zero padding was performed for every tweets that were lower than this threshold. This matrix is given as a input to the BLSTM Layer

C.BLSTM LAYER

This layer consists particular units in recurrent unseen layer which is called mind structure. In the memory cells,Self-connections exist.Each Memory cell contains an take in, take out, and forget gate. It contains unseen layer called LSTM cell.To plot long-term dependencies,each mind cell is defined with a set of gates 'd',where d is the mind dimension of unseen state. For each iteration, the LSTM cell has layer input x_t and the layer output $y_{t..}$ LSTM contains three gates take in gate, forget gate and take out gate and it can be calculated using equations (2)-(7).

$i_t = \sigma\ (w_ix\ x_t + w_ih\ h_{t-1} + b_i)$(2)

$f_t = \sigma\ (w_fx\ x_t + w_fh\ h_{t-1} + b_f\ )$(3)

$o_t = \sigma\ (w_o\ x_t + w_o\ \ h_{t-1} + b_o)$(4)

$s_t\ =\ tanh\ \ w_sx$ $\qquad\qquad x_t + w_sh\ h_{t-1} + b_s$ (5)

$c_t\ = f_t\ \circledS\ c_{t-1}$ $\qquad\qquad + i_t\ \circledS\ s_t$(6)

$h_t\ =\ tanh\ (c_t\ )$ $\qquad\qquad \circledS\ o_t$ (7)

In equations (2)–(7),

$\circledS$ is the element-wise product:

$w_i\ w_f\ w_o\ w_s$ are weighing factors used for mapping the

unseen layer take in to the three gates and the take in cell state;

$b_i\ b_f\ b_o\ b_s$ are bias vectors;

σ is the gate turn on role, which is normally the sigmoid function;

tanh is the hyperbolic tangent function.

The final take out of the LSTM layer is a angle of all the outputs, represented by (8).

$Y_t = [h_{t-n,,,,,,,,}\ h_{t-1}]$(8)

Two unseen layers is connected to a a single take out layer by bidirectional LSTM.For each iteration, the ahead-layer take out sequence is counted using take in a positive sequence from time t-n to time t-1, and the behind-layer take out series is counted using the alter take in from time t-n to t-1.Both the layers output are counted using the quality LSTM renovating equations.Finally this will create two copies of the unseen layer, one copy is fitted in the take in sequence and the other copy is used as altered copy of the take in sequence. These will be connected together by attention mechanism,The BLSTM layer produce an take out vector in which every factor is calculated using (9).

$y_t = \sigma\ (h{\rightarrow}t, h{\leftrightarrow}t)$(9) where σ is a soft attention function to combine the two output series. By this,our version can take salient words as possible from twain the directions.The final output can be represented by a vector(10).

$Y_t= [y_{t-n,,,,,,,,}\quad ,y_{t-1}]$(10)

D.ATTENTION LAYER

All the words and sentence does not convey exactly the same meaning.To observe this attention layer has been originated.Important and meaningful parts of the sentence can be identified by the user reading text. The attention system provides an further system to the internet,that can be studied as standard version via behind-propagtion.The soft attention like deterministic and differentiable attention mechanism is used,where the load on each position are usually given by a softmax, and the take out of the attention module is a weighted sum of representations at each location. The differentiable justification can describe the whole process, so that the training can merge with the rest of the network using back-propagation.

The phrases that have a closer semantic communication to the sentence meaning is focussed by word-level attention mechanism. The attention mechanism gives an attention score $e_{i,t}$ to every phrase t in the sentence i, as given by (11).

$e_{i,t} = g(Wh_t c),$ (11) where g is an activation function.

Auxilary aspect such as punctuation or again and again of phrases is used to designate descriptive content and symbolic hint within the content. The below example show that punctuation factors affect the concern which is given in the content.Best example is exclamation mark (!) increases the passionable content "*The flowers are so beautiful!!!!!!!!*" is more intense than "*The flowers are so beautiful*".Repeated letters or words (eg:happpppppppppppyyyyyyy) is used to express enthusiasm and happiness in communication. Capitalization using ALL CAPS indicate more stress compared to other relevant word in a document and it increases the escalate of sentiment conserving emantic adaptation. For eg: "simply LOVED the nature" is high intensify than "simply loved the nature".sarcasm detection can be easily done by detecting the patterns of tweet which contains question mark, exclamation mark etc.

E. CONVOLUTION LAYER

It has two initial structure, aspect studying(Conv, ReLU, and Pool) and classification (fully joined and softmax). The take out aspect of the sAtt BLSTM are merged with auxiliary aspect and provided as take in to the convent. For every refine,a transition matrix(T) will be generated.This filter consist of weight matrix. It is applied to take in array to acquire a one unit as take out.It is carry out for the bias.

$v_j = f\ T.F_{j:j+s-1} + b_a$ (13)

$F_{j:j+s-1}$ = The local angle from locate j to j+s-1 in F.

Each part will produces an output that will be merged in arrangement to acquire output called modified aspect map.This will be the output for the ensuing layer called ReLU Layer.

F. RELU LAYER

The turn on or ReLU Layer is used for pact with the nonlinearity in the convNet model. This layer developes a rectified aspect map, and this will fed into the pooling layer to decrease the dimensionality of the aspect map.

G. POOLING LAYER

To take the largest element from the rectified feature map, max pooling is used.

It helps to retain the most significant features and reduces the dimensionality of the feature map.To choose the top k aspect which obeys to various hidden layers,k-max pooling is performed, where p is expressed as (15)

$p = \max\ \{v_1, v_2,\ v_3,\ \ldots., v_{d-s+1}\}$ (15)

H. REPRESENTATION LAYER

The output layer consists of the softmax activation function and it is a fully connected layer.Fully connected softmax layer groups the tweet as sarcastic or non-sarcastic as an take out and calculates the probability of any take out phrase.The output angle of softmax layer is as (16)

$O(z) = T(z)O^{(z-1)} + b(z)$ (16)

where

$O^{(z-1)}$ denotes the output vector the pooling layer

$T^{(z)}$ is the transition matrix

$b^{(z)}$ represents the bias factor of softmax layer.

IV.PARAMETER SETTINGS

A.ACCURACY

Accuracy is the instrument to calculate the accurate value and it is a proportion of accurate state consideration to the total considerations.

Accuracy=TP+TN/TP+FP+FN+TN

B.PRECISION

Precision is the magnitude of accurate state positive
considerations to the total state positive consideration and it is a calculate of consistency and reproducibility.

Precision = TP/TP+FP

C.RECALL (SENSITIVITY)

Recall is the proportion of number of relevant instances to the total number of actual relevant instances.

Recall = TP/TP+FN

D.F1 SCORE

F1 Score is the mass mean of Precision and Recall. It gives a better calculate of the wrongly organized cases than the Accuracy metric. Accuracy is used when the accurate positives and accurate negatives are more important F1 Score is used when the inaccurate negatives and inaccurate positives are vital.

F1 Score = 2*(Recall * Precision) / (Recall + Precision).

- **TP-True Positives**

- **FP- False Positives**

- **FN-False Negatives**

- **TN-True Negatives**

V.COMPARISON WITH SVM ALGORITHM

It has been observed that CNN provides greater accuracy rate.Algorithms like Gradient Descent Ensemble performs good with an accuracy of 79% when compared with SVM which performs only 54%.Convolution Neural Network is non-linear classifier. Support Vector Machine is a linear classifier.CNN performs good with Visual image recognition where as SVM is used most probably for classification problems.Convolution Neural Network is a feed forward neural network which is generally used to analyse visual images by processing data with grid like topology.Support Vector Machine is an algorithm that analyse data used for classification and regression analysis.SVM is a supervised learning method that looks at data and sorts it into one of the two categories.
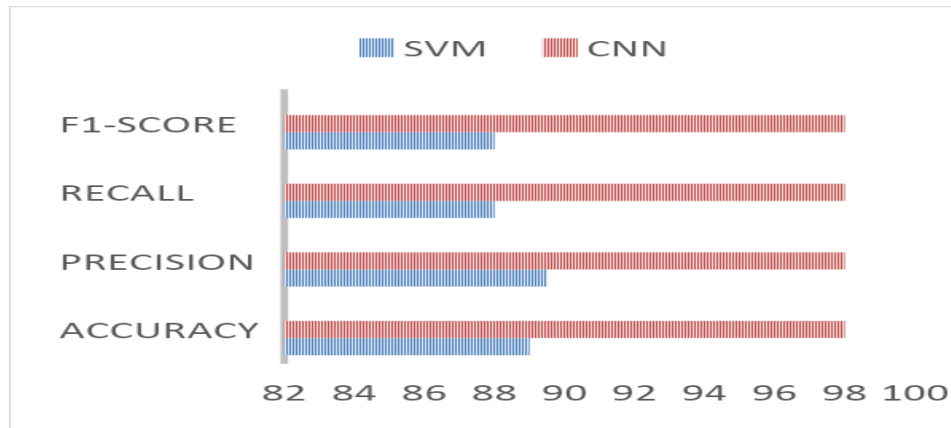
VI. CONCLUSION

The proposed approach uses convolution neural network for the classification of thestatements based on sentiments in the twitter dataset .The algorithm performs good withsingle layer convolution and it even out perform many techniques that exists.For the detection of sarcasm in an open statement ,there are many new dimensions to be analysed.Convolutionneural network performs good and has greater accuracy rate when compared with SupportVector Machine.

REFERENCES

[1]S. K. Bharti, B. Vachha, R. K. Pradhan, K. S. Babu, and   S.K. Jena, ''Sarcastic sentiment detection in tweets streamed in real time: A big data approach,'' *Digit. Commun. Netw.*, vol. 2, no. 3, pp. 108–121,2016.
[2]A. Kumar and A. Jaiswal, ''Empirical study of Twitter and Tumblr for sen- timent analysis using soft computing techniques,'' in *Proc. World Congr. Eng. Comput. Sci.*, 2017, pp.1–5.
[3]S. Aloufi and A. El Saddik, ''Sentiment identification in   football-specific tweets,'' IEEE Access, vol. 6, pp. 78609–78621,2018.
[4]P. F. Pai and C. H. Liu, ''Predicting vehicle sales by sentiment analysis of Twitter data and stock market values,'' IEEE Access, vol. 6, pp. 57655–57662,2018.
[5]D. Wu and M. Chi, ''Long short-term memory with quadratic connections in recursive neural networks for representing compositional semantics,'' IEEE Access, vol. 5, pp. 16077–16083,2017.
[6]D. Wu and M. Chi, ''Long short-term memory with quadratic connections in recursive neural networks for representing compositional semantics,'' IEEE Access, vol. 5, pp. 16077–16083,2017.
[7]M. Bouazizi and T. Ohtsuki, ''A pattern-based approach for multi-class sentiment analysis in Twitter,'' IEEE Access, vol. 5, pp. 20617–20639,2017.
[8]A. Joshi, P. Bhattacharyya, M. Carman, J. Saraswati, and R. Shukla, ''How do cultural differences impact the quality of sarcasm annotation?: A case study of indian annotators and american text,'' in Proc. 10th SIGHUM Workshop Lang. Technol. Cultural Heritage, Social Sci., Hum., 2016, pp.95–99.
[9]H. Salehinejad, J. Baarbe, S. Sankar, J. Barfett, E. Colak, and S. Valaee. (2017). ''Recent advances in recurrent neural networks.''[Online].Available:https://arxiv.org/abs/1801.01078
[10]Y. Kim. (2014). ''Convolutional neural networks for sentenceclassification.''[Online].Available:https://arxiv.org/abs/1408.5882

BIOGRAPHY

Dr. J.V.Anchita Algammai working as Associate Professor of Computer Science and Engineering Department, Velammal College of Engineering and Technology, Madurai-625009, Tamil Nadu, India.Undergraduate-B.E(CSE),Postgraduate-M.E(CSE),Phd-Information and Communication Engineering.

K.Guru Priya is an UG student in the Computer Science and Engineering Department, Velammal college of Engineering and Technology, Madurai-625009, Tamil Nadu, India.

T.R.Divya is an UG student in the Computer Science and Engineering Department, Velammal college of Engineering and Technology, Madurai-625009, Tamil Nadu, India.

P.K.Priyanka is an UG student in the Computer Science and Engineering Department, Velammal college of Engineering and Technology, Madurai-625009, Tamil Nadu, India.

K.S.Ragavi is an UG student in the Computer Science and Engineering Department,Velammal college of Engineering and Technology, Madurai-625009, Tamil Nadu, India.

.