WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Sentimental Analysis Using Emojis prediction

J.B.Jona[1], Ms.Chitirai Jothi[2], Sachin Dhana Paul[3], S.A.Gunasekaran[4], Sridhar[5]

[1]Dr. J.B.Jona, Associate Professor, Dept. Of Computer Applications, [2]Ms.Chitirai Jothi, Student, Dept. of Decision and Computing Sciences, [3]Mr. Sachin Dhana Paul, Student, Dept. of Decision and Computing Sciences, [4]Mr.S.A.Gunasekaran, Assistant Professor, Dept. Of Computer Applications, [5]Mr. Sridhar, Student, Dept. of Decision and Computing Sciences; all are attached with Coimbatore Institute of Technology

ABSTRACT

Emoticons is communicated to call attention to the profound and nostalgic sentiments as far as graphical images which might add weightage to the message explanation. Emoticons are generally utilized in web-based entertainment and visit discussion on the two networks furthermore as portable applications. With use of emoticons, the text based imparting explanations are more expressive and compelling. Accordingly, it's critical to mine the association between the text and emoticon images. By utilizing the high level AI strategies, related emoticons are frequently proposed naturally while grasping the significant setting of the assertion. during this paper, I present a brain network procedure to foresee fitting emoticon image for a given printed assertion and furthermore the AI model is built by utilizing word vector portrayals and profound learning systems. Model is shaped by utilizing Worldwide Vector (GloVe) implanting portrayal, mix of long memory (LSTM) organization and convolutional brain organization (CNN) and softmax layers. Calculation is prepared to sum up and relate words inside the test set finds the right emoticon image even. The model turns into a precise classifier planning from sentences to emoticon images, whether the words don't show up inside the preparation set.

Keywords: Emojis, LSTM, Deep Learning, Neural Networks, Sentimental Analysis

INTRODUCTION

Emoticons utilization is progressively famous on both web-based entertainment applications and company communicators like Twitter, Facebook, Whatsapp and Lync. Emoticons are normally utilized with the blend of printed data to convey the feelings and sentiments. It's a crucial undertaking to decide the relationship between's the emoticons and the literary substance. Normally, concentrates on use of emoticons basically target dissecting the semantics of the message based messages. Wijeratneet al. introduced the incorporating the emoticons along with the text portrayal to help the norm of text articulation. Nonetheless, these methodologies can't lay out the significant setting of full explanation; all things considered, the principal center is just around the singular words. In this manner, the presentation could likewise be sub-par. Peijun Zhao introduced on preparing of the Brain organization and depends on Twitter instant messages which are, the greater part of the days, of the blend numerous dialects, obscure words. Consequently, for the concealed words, this model can't connect to the emojis. In request to conquer the erroneous expectation issues, This paper proposes a profound learning put together methodology which is based with respect to GloVe data set, Bi-directional LSTM layers to perform on inconspicuous text words. This likewise speeds up model creation as the insignificant preparation informational collection is adequate to confirm the exactness of the expectation. The GloVe data set is a lot of helpful to finish the hyper boundaries to close the model design

PROPOSED APPROACH

The goal of this examination is to make a framework which will foresee the feeling of a client whether it is good or pessimistic utilizing RNN-LSTM. LSTM is a sort of RNN. Expectation is consecutively in RNN, and the concealed layer from one forecast is the secret layer of the following expectation which will dole out a memory to the organization. Results from past forecasts can work on future expectations. LSTM gives RNN an additional a perspective that gives it fine-grained command over memory. This angle controls how much the ongoing information matters in making the new memory, and how much the previous recollections matters in making the new memory, and what is significant in producing the result. Twitter is an incredible hotspot for assessments of different sorts of occasions and items. Distinguishing the opinion of these miniature web journals is a difficult undertaking which has drawn in expanded research interest as of late.

The clients post their tweets on twitter. These tweets are then separated progressively involving twitter Programming interface as crude information which are then saved in the data set. The crude dataset is changed over into target dataset through Information Pre-Handling and Component Extraction. The highlights of the words are chosen and afterward AI strategies are applied on removed elements to characterize them into feeling extremity is good or pessimistic.

Global Vector embedding Layer

Worldwide Vectors are utilized for word portrayal and in excess of 2000 papers are distributed. GloVe gives valuable data about the significance of individual words and is utilized to work out the connection between's the various words. Semantic vector organized portrayal of a language assesses each word with a genuine esteemed vector. These vectors are of explicit weighted least squares model which trains on an informational index with word co-event and recurrence counts and thus takes advantage of measurements. The model creates a word vector space with various layered base like 50, 100 and 150 aspects. For instance,

cosine_similarity(father - mother) = 0.890903844289 and cosine_similarity(ball, crocodile) = 0.274392462614. Words father-mother are having higher comparability though ball-crocodile are having lower similitude. Having this co-social vector portrayal of all of the English word reference words, the proposed model can perform well on inconspicuous or new words. The input statement will be converted into GloVe formatted vector

input proclamation will be changed over into GloVe designed vector portrayal and the normal of the installing layer as a solitary encoded vector will be taken care of to the LSTM layer. fig 2.1 addresses data set design of proposed framework.

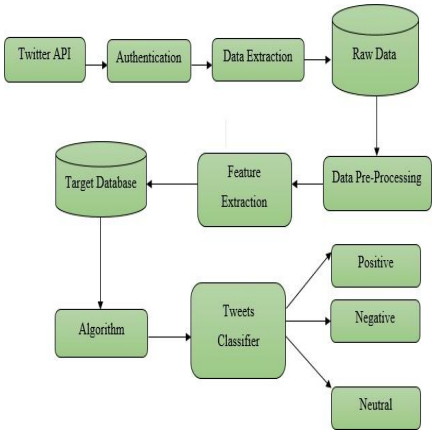

Fig 2.1: Database structure of proposed system

Long-short term memory layer (LSTM)

The LSTM layer has the incredible component of recalling the example of the assertion for long lengths of time to assess the setting of text based data. Significant benefit of LSTM is that this organization comprises of different memory blocks called cells. The phones are answerable for recalling things. Cell state and secret state are the two states which get moved to the following cell. There are three doors through which controls to the memory blocks are being finished. Disregard entryway is utilized to eliminate the data from the cell state. Input entryway is utilized to add the data to the cell state. Yield is utilized to choose the helpful data from the ongoing cell and give it as an out to the following cell. Every one of the three entryways cooperate consecutively to deal with the data and recognize the genuine setting of the information proclamation. In light of the assessed setting or significant word, the proposed model predicts the legitimate emoticon image.

MODEL CREATION

The contribution of the model is a string relating to a sentence (for example "I love you") and gets changed over into lower case for working on the model creation. The initial step is to change over the information sentence into the word vector portrayal, which then gets found the middle value of together. The model has thought of.pre-prepared 100-layered GloVe embeddings. The model makes word_to_index, index_to_word and word_to_vec_maps for all of the jargon accessible in the English word reference. The GloVe dataset consolidated vectors for 400,001 words ordering from 0 to 400,000. Hence, each expression of the info message gets addressed as a 100 esteemed vector and this large number of vectors get found the middle value of. Since I considered 20 groupings of emoticon images, the model makes an OneHotEncoding portrayal for yield names.Word implanting layer is utilized as contribution to LSTM layer and the preparation of the information is being finished as little groups. Since LSTM has an imperative that the length of all test and train dataset ought to be reliable, taking into account the length of the assertion as 10 which is a respectable size to work out the model execution. In the event that the length of the information proclamation is under 10, cushioning is utilized to make it a proper size. As a feature of the underlying stage, convert all preparing sentences into a rundown of records and afterward zero-cushion in the event that the length is under 10. The Implanting layer takes a number grid of size (cluster size, max input length) as info and these sentences are changed over into arrangements of whole numbers organized files as displayed in the figure 3.1

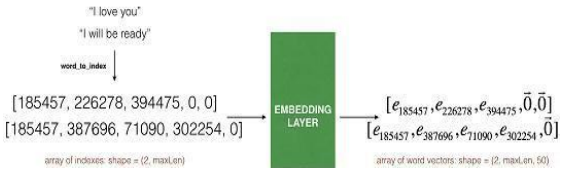

Fig 3.1: Embedding layer using Glove

The textual information is represented as a list of words which is denoted as W = [w1, w2, ..w10] and these words get translated into GloVe formatted numerical data as E = [exx, eyy,..].

The LSTM - 1 layer is developed with a 128-layered secret state returning as a bunch of groupings. Result of LSTM-1 is given to the Dropout layer with the likelihood of 0.25. Another LSTM - 2 is incorporated with a 128-layered secret state with returning result as a solitary secret state. Presently the result of LSTM-2 is a contribution to the Dropout layer with the likelihood of 0.25. The result is then taken care of to the Thick layer with Softmax enactment capacity and size of 20 which is the last result name which are emoticon arrangements. At lastSoftmax layer is given the contribution with size of 20 which are probabilities. Based on higher probability, the softmax layer finalizes the predicted emoji index which is then mapped to the corresponding emoji symbol.

Number of hyper parameters involved in model creation:

Vocabulary size: 400,001

non-trainable parameters: 400,001*100 = 40,000,100

Trainable params: 223,877

Total params: 40,223,977

In the wake of making the model, it should be arranged with different determinations like misfortune, streamlining agent and measurements (precision). Model is prepared with bunch size of 32, 50 ages and with mix empowered. The softmax layer fills in as a standardized remarkable capacity which is a speculation of the calculated capacity.

Sentiment and Emotion

Opinion and feeling are the two terms connected with human subjectivity, so they are some of the time utilized conversely in research without adequate separation, which might prompt unfortunate misgiving and disarray. As this study includes both feeling and feeling, It needs to separate them obviously. Meaning of Feeling and Feeling. Opinion alludes to a mentality, thought, or judgment provoked by an inclination. Generally, in the NLP people group, feeling is considered to have three polarities, i.e., good, pessimistic, and unbiased . Feeling alludes to a cognizant mental response emotionally experienced as unmistakable inclinations. Not quite the same as feeling, feeling is more refined. Up until this point, there has not been one standard hypothesis on ordering feelings. Feeling Models. In the brain research space, various observational and logical speculations about feelings have been proposed. A famous model is a tree-organized order model of feelings, which is portrayed in the Shaver structure. The main level of the tree comprises of six essential feelings, i.e., love, bitterness, outrage, happiness, shock, and dread. Such a tree-organized order model has been generally embraced in SE studies. Another normal inclination model is VAD model, which projects feelings into a bi-layered space, where the even aspect shows the profound polarities and the upward aspect demonstrates the degrees of sensitivity (i.e., excitement. As indicated by this model, feelings can. In this study, sentiment detection is considered as determining whether a text expresses a positive, negative, or neutral sentiment, while emotion detection is to identify the presence of specific emotion states from texts. Compared to sentiment detection, emotion detection is more challenging, because texts expressing a specific emotion are much scarcer than those conveying a kind of sentiment.

Word Embedding

Sentiment and emotion detection are both typical NLP tasks. In NLP, to eliminate the discrete nature of words, word embedding techniques, such as skip-gram algorithm and GloVe , are proposed to encode every single word into a continuous vector space as a high dimensional vector. Through these techniques, words that commonly occur in a similar context are represented as similar vectors, which can capture the semantic relationship among words. In practice, these techniques are usually performed based on a large amount of natural language texts by utilizing co-occurrence statistics of words in the corpus. For example, the skip-gram algorithm scans each sample in the corpus and uses each word that it has scanned as an input to predict words within a certain range before and after this word; GloVe is learned based on a global word-word co occurrence matrix, which tabulates how frequently words co-occur with one another in a given corpus. Compared to GloVe, the skip-gram algorithm is demonstrated to be more robust and utilize less system resources. In practice, the skip-gram algorithm is widely adopted in SE tasks, such as creating SE-specific word embeddings, enhancing software traceability, and localizing bugs.

Methodology

Since sentiment and emotion detection are both relatively new tasks in SE and labeled data for them are not so sufficient, employ transfer learning to tackle the two problems. Specifically, Emojis as indications of sentiment/emotion and employ emoji prediction as the source task. On one hand, emojis are able to express various emotions. The rich emotional information contained in emoji usage makes emoji prediction a suitable source task of sentiment and emotion detection. However, emojis are widely used in social media and developers' communication and thus can be easily collected. The large-scale emoji usage data can complement the scarce manually labeled data for the two target tasks. SEntiMoji, an emoji-powered transfer learning approach for sentiment and emotion detection in SE. First, It learns sentiment- and emotion-aware representations of texts by using emoji prediction as an instrument. More specifically, It uses emojis as indications of sentiment/emotion and learns vector representations of texts by predicting which emojis are used in them.

Texts that tend to surround the same emoji are represented as similar vectors. Then these informative representations are used as features to predict the true sentiment/emotion labels. Through these representations, sentiment knowledge contained in emoji usage data is transferred from the emoji prediction task into the sentiment/emotion classifiers. Since Felbo et al. have released such a representation model that is pre trained based on 56.6 billion Tweets, directly build SEntiMoji upon the off-the-shelf DeepMoji in a transfer learning way. Specifically, our approach takes two stages: fine-tune DeepMoji using GitHub posts that contain emojis to incorporate technical knowledge. The fine-tuned model is still a representation model based on emoji prediction, and is called DeepMoji-SE. Use DeepMoji-SE to obtain vector representations of labeled texts, and then use these representations as features to train the sentiment/emotion classifier. Call the final sentiment/emotion classifier SEntiMoji. Next, describe the existing DeepMoji model and the two-stage learning process in detail.

DeepMoji Model Felbo et al. learned DeepMoji through predicting emojis used in Tweets. To this end, they collected 56.6 billion Tweets (denoted asT), selected the top 64 emojis in this corpus, and excluded the Tweets that do not contain any of these emojis. For each remaining Tweet, they created separate samples for each unique emoji in it. Finally, they balanced the created 1.2 billion samples (denoted as ET) using up sampling and then performed the emoji prediction task. The model architecture is illustrated in Figure 2. First, for a given sample, words in it are inputted into the word embedding layer that is pre-trained onT. In this step, each word can be represented as a unique vector. Then these word vectors are processed by two bi-directional LSTM layers and one attention layer. Through these steps, the sample can be represented as one vector instead of several word vectors. Finally, the softmax layer treats the vector as the input and outputs the probabilities that this sample may contain each of the 64 emojis.The word embedding layer of 256 dimensions is pre-trained based on T with the skip-gram algorithm. A hyperbolic tangent activation function is used to enforce a constraint of each embedding dimension being within [−1, 1]. Through this layer, each sample in ET can be denoted as (x, e), where x = [d1,d2,...,dL] denotes the word vector sequences of the plain text removed emoji (di as the vector representation of the 'ith' word) and e denotes the emoji contained in the sample. To take context information (i.e., both past and future words) of the current word at each time step into consideration, DeepMoji employs

bi-directional LSTM with 1,024 hidden units (512 in each direction) instead of the traditional LSTM. Each bi-directional LSTM network contains two sub-networks (i.e., a forward network and a backward network) to encode the sequential contexts of each word in the two directions respectively.

Given the input x = [d1,d2,...,dL], it computes an encoded vector hi of each word vector di by concatenating the latent vectors from both directions:
hi = → hi ‖ ← hi

Emojis in Sentiment and Emotion Detection

Traditional sentiment and emotion detection in NLP is mainly performed in unsupervised or supervised ways. Unsupervised tools (e.g., SentiStrength) simply make use of lists of words annotated with sentiment polarity to determine the overall sentiment/emotion of a given text. However, fixed word lists cannot cope with the dynamic nature of the natural language. Then researchers started to use labeled texts to train sentiment/emotion classifiers in a supervised way, where deep learning techniques are widely adopted. However, it is time-consuming to manually annotate texts on a large scale, thus resulting in a scarcity of labeled data. To tackle this problem, many researchers attempted to perform sentiment or emotion detection in a distantly supervised way. For example, they used emoticons and specific hashtags as a proxy for the emotional contents of texts.

Recent studies extended the distant supervision to emojis, a more diverse set of indications, and demonstrated the superiority of emojis compared to emoticons. As emojis are becoming increasingly popular and have the ability to express emotions, they are considered benign indications of sentiments and emotions. The emotional information contained in the emoji usage data can supplement the limited manually labeled data. In this work, also employ emojis as indications of sentiment/emotion and tackle sentiment and emotion detection tasks using deep learning in a distantly supervised way. However, different from previous work, applying such a method on SE-related texts and proposing an SE-customized approach. Recently, to address the challenge of sentiment and emotion detection in SE, researchers also started to analyze emoticons and emojis in software development platforms so as to find some potential solutions. Claes et al. investigated the use of emoticons in open source software development. Lu et al. analyzed the emoji usage on GitHub and found that emojis are often used to express sentiment on this platform. Furthermore, Imtiaz et al. directly used emojis as the indicators of developers' sentiment on GitHub. Calefato et al. and Ding et al. took emoticons into account in their proposed SE-customized sentiment detection techniques. All of them demonstrated the feasibility of leveraging these emotional cues to benefit sentiment and emotion detection in SE. Following this line of research, this study leverages large-scale emoji usage from both technical and open domains to address sentiment and emotion detection in SE.

PERFORMANCE EVALUATION

In this section, The model compares the performance of my proposed approach with the available neural network techniques.

- SVM, using K independent word data set, model to predict for each emoji.
- CNN, using Convolution Neural Network as the word encoder and predict the emojis.
- LSTM with GloVe, proposed model.

According to my experiments done, observations as below:

SVM model performs well as it takes independent words and starts mapping words to emoji. Even though, SVM model is giving good accuracy, this cannot consider the real context of the full statement & performance is poor in case of negative words like 'I did not win the match', 'I am not happy'.

Similar to SVM performance, CNN model works on only individual words and cannot remember the earlier important words to derive the real meaning of the full statement.

The proposed approach gives decent accuracy with both on unseen or new words as well as by considering the full context of the textual statement to derive the real meaning of the full statement. table 3.1 represents the F1 score comparison

loss='categorical_crossentropy', optimizer='adam', metrics=['accuracy']

Method	Precision	Recall	F1-score
SVM	0.325	0.326	0.235
CNN	0.354	0.379	0.304
LSTM – GloVe	**0.380**	**0.402**	**0.379**

Table 3.1: F1 score comparison

Long Short Term Memory neural network works – usually just called "LSTMs" – are a special kind of RNN, capable of learning long-term dependencies. LSTM models are a variety of RNN. In RNN the prediction in sequence, where the hidden layer from one prediction is the hidden layer of the next prediction, will assign a memory to the neural network work, therefore, results from earlier estimation could lead to improve future predictions. LSTM gives RNN more features to an extreme control over memory; these aspects control how much the present input matters for forming the new memory, also how much the past memories matters in creating the new memory, and what parts of the memory are essential in producing the output.

CONCLUSION

In this paper, the proposed model introducesneural network model with LSTM with a combination of GloVe embeddings to predict the emoji symbol while understanding the real context of the full statement & to create the model in a simple and faster way. With the use of GloVe embeddings, tweaking the hyper parameters has become drastically faster to evaluate the model performance as well as on fixing the hyper parameters. The machine learning models that were implemented were Bernoulli Bayes, Multinomial Bayes, Regression and SVM. The models were trained only with the text message removing the Emoticons and Emojis and then tested and their performance was evaluated. To study the effect of Emoticons the models were

also trained with text and Emoticons and their performance were analyzed.In this model, It also proposes SEntiMoji, an emoji-powered transfer learning approach for sentiment and emotion detection in SE. It is developed based on an existing representation model called DeepMoji. DeepMoji is pre-trained on Tweets and can represent texts with sentiment- and emotion-aware vectors. As technical knowledge is highlighted in current SE-customized sentiment and emotion detection, we also use GitHub data to incorporate more technical knowledge into DeepMoji. Then the fine-tuned representation model, as well as manually labeled data, is used to train the final sentiment/emotion classifier.

References

1. C. Bell, Essays on the Anatomy and Philosophy of Expression. J. Murray, 1824.
2. C. Darwin and P. Prodger, The expression of the emotions in man and animals. Oxford University Press, USA, 1998.
3. F. H. Rachman, R. Sarno, and C. Fatichah, "Cbe: Corpus-based of emotion for emotion detection in text document," in 2016 3rd International Conference on Information Technology, Computer, and Electrical Engineering (ICITACEE). IEEE, 2016, pp. 331–335.
4. D. Tanna, M. Dudhane, A. Sardar, K. Deshpande, and N. Deshmukh, "Sentiment analysis on social media for emotion classification," in 2020 4th International Conference on IntelligentComputing and Control Systems (ICICCS). IEEE, 2020, pp. 911–915.
5. F. M. Shah, A. S. Reyadh, A. I. Shaafi, S. Ahmed, and F. T. Sithil, "Emotion detection from tweets using ait-2018 dataset," in 2019 5th International Conference on Advances in Electrical Engineering (ICAEE). IEEE, 2019, pp. 575–580.
6. M. Deshpande and V. Rao, "Depression detection using emotion artificial intelligence," in 2017 International Conference on Intelligent Sustainable Systems (ICISS). IEEE, 2017, pp. 858–862.
7. H. Tuli, M. Singh, and N. Singh, "Facial emotion recognition system using machine learning."
8. A. Bag, "Real time facial expression recognition using convolution neural network algorithm."
9. D. Y. Liliana and T. Basaruddin, "Review of automatic emotion recognition through facial expression analysis," in 2018 International Conference on Electrical Engineering and Computer Science (ICECOS). IEEE, 2018, pp. 231–236.