International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-8 Issue-11, September 2019

Deep LSTM for Emoji Twitter Sentiment Exploration via Distributed Embeddings

Anupama Angadi, Satish Muppidi, Satya Keerthi Gorripati

Abstract*: *Social media's sentimental data is the most vital digital marketing platform that can help us to reveal the real world events including qualitative insights to understand people's visibility about brands, politics, emotional status, and so on. With today's interrelated world, a public relations disaster can be initiated with one post or a tweet. Conventional sentimental analysis is the process of defining whether the shared post on social media is neutral, positive or negative and has been focused by the Dealers, Administrations to understand public feelings of their products and corporation. However, extensive usage of emoji in social media has attracted an increasing interest. In this proposed framework, we suggest a novel scheme for Twitter sentiment method on emojis by considering pre-trained word and emoji embeddings. We first train our model to learn word, emoji embeddings under positive and negative tweets; later a classifier passes them through a neural network combining LSTM to achieve better performance. Our tests show that the proposed model operational for extracting sentiment-aware emojis and outperforms the state-of-the-art simulations.

Keywords*: *emoji, distributed embeddings, LSTM, classifier, opinion mining.

I. INTRODUCTION

In latest years, Social Network Method (SNS) such as Instagram, Facebook, and Twitter has become well-known. With more than millions of active users, communicating individual messages on posts or tweets. Social media has become one of the best platforms for information, news, and collaboration around the world. Twitter is an online social network allows the users to create short message called tweets with a limit of 280 characters. This distinctive feature makes tweets good contestants for natural language processing tasks, machine learning task of sentiment method. The objective of sentiment method is to outline automatic tools able to assess subjective information, such as sentiments and opinions from posts, tweets and images. When monitoring the contents of social media, like Instagram, Facebook and Twitter, we need to control new types of formats and flow models that need to be demonstrated obviously beginning from the language.

The features that distinguish attractive blogging (e.g. blogs) from microblogs text are to narrate user's attitude and emotion on instant messaging. However, method of tweets or posts on social media leads towards substantial communicators like voice inflection, facial expressions and body stance. These informal linguistics and non-standard spelling sociophonetics has urged to move forward by more complex scenarios. Natural language processing, Machine learning models on microblog for predicting polarity rather than textual is the recent attention. Emojis, are ideograms, smileys and are considered to be trustworthy indicators of opinions. Because of this reason, emoji prediction is moderately new model [7].

The rest of the paper is organized as follows: section 2 designates method of sentimental tweets with emoji, section 3 describes the format of the data and steps for pre-processing, section 4 provides models methodology and finally conclusion in section 5.

II. RELATED WORK

Sentimental method objective is to determine independent information from dissimilar formats such as tweets, posts, audios and videos. Conventional sentiment method classifies statement's polarity completely based on words. Words are treated as measures of the statements opinion. In the case of sentiment method negation words are very significant, for example, the difference between "great job" and "not a great job". The classifier of the conventional model classifies statements polarity as negative when it detects a negative word in the sentence deprived of considering the context. Dictionary of all negative words are typically not negative in statements context. For example, an instance of a negative polarity statement "she pushed the anxious uncertainties out of her mind", where the word "uncertainties" negating the opinion and marks the whole statement as negative although it carries a positive context. Better completion of sentiment classification is often obtained through new technical perspectives such as NLP, machine learning tasks [11, 12].

In recent years, researchers and analysts claim to have observed the growing usage of emoji in al-most every social media because the emojis carry sentimental and semantical data and expressively useful in knowing the embedded emotional gesture in texts. Facial emoji have become a regular for condense feelings of objects and emotions. Being fixed in Unicode, they have no language obstructions and are spread on the Web quickly. The advent of deep learning techniques, many researchers have endeavored to use progressive neural network models for sentiment method [3, 11]. Recent research shows a long short-term memory (LSTM) based model for predicting sentiments than simple words polarity aggregation with the pre-trained embedding features. Embedding vectors of NLP represents words in a continuous vector space where all semantic words are mapped to nearby data points [6, 9].

Manuscript published on 30 September 2019.
*Correspondence Author(s)
Anupama Angadi*, Department of CSE, Raghu Engineering College (Autonomous), Dakamarri, Visakhapatnam, 531162, A.P, India.
Satish Muppidi, Department of IT, GMRIT, Rajam, India.
Satya Keerthi Gorripati, Department of CSE, Gayatri Vidya Parishad College of Engineering (Autonomous), Visakhapatnam, A.P, India.

Researchers have been using computationally efficient word2vec predictive model in various applications like word, sentence and document classification [4, 5, 7]. However, previous literatures lack in concerns of both words, emoji embedding on LSTM to handle sentence semantics and assortment of emoji [9].

A. Method of sentimental tweets with emoji

Information enclosed in Social media, became vital for the communication, and makes them a striking foundation of data for opinion method. Linguistic sociophonetics elements are used not only in tweets but also in posts, to elicit a textual message, can be used to enhance sentiment method.

The accuracy of identifying emotions can improve and increase with the exploration of emojis. They deliver a crucial data and this is vital for politicians, marketers, organizations in order to realize their people and customer's feelings towards their schemes and products. Thanks to the constantly changing technology for communication which replaces paragraphs to text then to images. Due to emojis visual representation they convey sentiment more precise than a tweet and this makes researchers divert towards this topic.

Emoticons are popularized as animated, non-verbal communication components in tweets and posts, matching human reactions based on facial expressions or speech. Their part is mainly practical: emoticons provide a negative or positive sense to tweets or posts by a graphic expression. Hence, there exists sentiment coordination between the message and emoji. This sentiment interpretation is articulated as a positive, neutral or negative category. Instances of positive emojis are "ready for thanks giving party", "picture is awesome", while examples negative ones are "it's a boring day". These emojis certainly are an significant basis of information for opinion based polarity classification. In fact, on social networks positive and negative messages have a high percentage of emojis, a sample of basic emojis are shown in Table 1.

Table 1: Basic emojis

.CHAR	IMAGE	UNICODE	UNICODE NAME
		0x1f633	Flushed face
		0x1f60e	Smiling face with sunglasses
		0x1f60d	heart-shaped eyes
		0x263a	White smiling face
		0x1f609	Winking face
		0x1f602	Face with tears of joy
		0x1f62d	Lousy face
		0x1f61c	Face with stuck-out tongue and winking eye
		0x1f621	Pouting face

III. DATASET

In recent days, Twitter social media is undergoing higher engagement rates associated with emoji usage. According to the latest method A/B test on social network tweets one with emojis and one without. As an outcome, tweets with emoji received more percentage of attention than pure textual tweet. Moreover, the emojis in a tweet is more expressive and had a lower cost per arrangement. The proposed model proceeds by mining the tweets with emoji which should express pure emotions , facial feelings or objects ; we discover consistency of 60% to 80% exists between text and emoji. Nevertheless it is an occurrence that the sentiments of text and emoji are weakly correlated.

To extract the valuable information presented on Twitter and how to custom it to appeal significant perception. To attain this, at initial stage of the model a precise analyser for tweets was build. Twitter permits software developers to extract data via a third party APIs such as Twitter REST API and The Streaming API. Twitter has abundant guidelines and frequency limits forced on its APIs, and for this purpose it obliges that all users must register an account and provide confirmation details when they query the API. One of the preeminent choices of connecting to Twitter Streaming API and downloading the data is by using R, Python using the libraries TwitteR, Tweepy [2, 8]. These APIs exports data in JSON format, the sample data of the dataset is shown below.

"http://farm6.staticflickr.com/5260/5536366901_d00048ac6c_z.jpg","id": 562121},{"license": 1,"file_name": "000000360661.jpg","coco_url":
"http://images.cocodataset.org/val2017/000000360661.jpg","height": 480,"width": 640,"date_captured": "2013-11-18 21:33:43","flickr_url":
"http://farm4.staticflickr.com/3670/9709793032_f9ee4f0aa2_z.jpg","id": 360661},{"license": 3,"file_name": "000000016228.jpg","coco_url":
"http://images.cocodataset.org/val2017/000000016228.jpg","height": 440,"width": 640,"date_captured": "2013-11-19 00:09:53","flickr_url":
"http://farm4.staticflickr.com/3737/10031812195_372ae7538f_z.jpg","id": 16228},{"license": 1,"file_name": "000000382088.jpg","coco_url":
"http://images.cocodataset.org/val2017/000000382088.jpg","height": 426,"width": 640,"date_captured": "2013-11-19 02:48:18","flickr_url":
"http://farm9.staticflickr.com/8253/8637472246_0783196c6f_z.jpg","id": 382088},{"license": 3,"file_name": "000000266409.jpg","coco_url":

Figure 1: Sample JSON dataset

Pre-trained emoji embeddings are accessible for download and practice. Embeddings of size 100 and 300 dimensions are available for public access in the github repository, but up to the specific dimension also can be trained manually [6]. For this model, we used a distributed vector space model of same size for both emojis and words [8]. This distributed vector space model idea developed by Einster and it provided us an enormously appropriate and semantically accurate tool for displaying the linear association between English words to emoji characters.

The whole emoji2vec weights, embeddings and visualizations (suitable for dissimilar dimensions and accomplished on both architectures) are accessible in gensim python's library [1]. Gensim seems to be a standard NLP package, and has some good documentation, comprising for word2vec. Word2vec is a very influential tool to examine the unseen semantic of emojis and to discover the most related sentence or a word for a given input sentence or word [4]. This is how we can generate a vocabulary of emojis by relating their corresponding "word synonyms".

Glove, Google's distributed vectors are two well-known pre-trained models of size 1.5 GB comprises word vectors for a vocabulary of 3 million phrases and words that they accomplished training on approximately 100 billion words from a Google News dataset [5]. This dataset has a vector length of size 300 features (or dimensions).

Our proposed method maps words, emoji symbols into the same distributed space as the 50-dimensional Google News word2vec embeddings. Thus, the subsequent emoji2vec embeddings can be used in addition to 50-dimensional word2vec embeddings in any solicitation [1, 4]. To this conclusion we crawl relevant emoji, their name and their keyword phrases for a given input tweet.

A. Pre-processing

To organize the data for method we initiated cleaning tweets by eliminating stop words, hash tags, user mentions, hyperlink, trailing three dots and urls. Then we used NLTKs libraries stop-words, pos_tag to apply lemmatization and part-of-speech tagging (PoS). Part-of-speech tags were limited to only three (e.g. noun (NN), verb (VB), adjective (JJ) and removed certain kinds of PoS that don't carry meaningful information.

B. Computed Features

For obtaining the sentiment annotations, we used VaderSentiment library in python which is a rule-based algorithm for sentiment method to produce weak polarity labels of the textual tweets. This algorithm produces sentiment ratings ranging from 1 (positive) to -1 (negative) with neutral in the central. We deliberate this examination as a binary classification task (with positive and negative polarity), we removed all the examples with fragile prediction scores between −0.70 to 0.70 and keep the textual tweets with high sentiment scores. Emoji existences are computed distinctly for positive and negative tweets.

SentiStrength [13] algorithm can also be used for Polarity prediction was also supplemented by using positive and negative scores. The benefit being that the model has polarity intensity, so it could be beneficial even if all categories of emojis are neutral or positive.

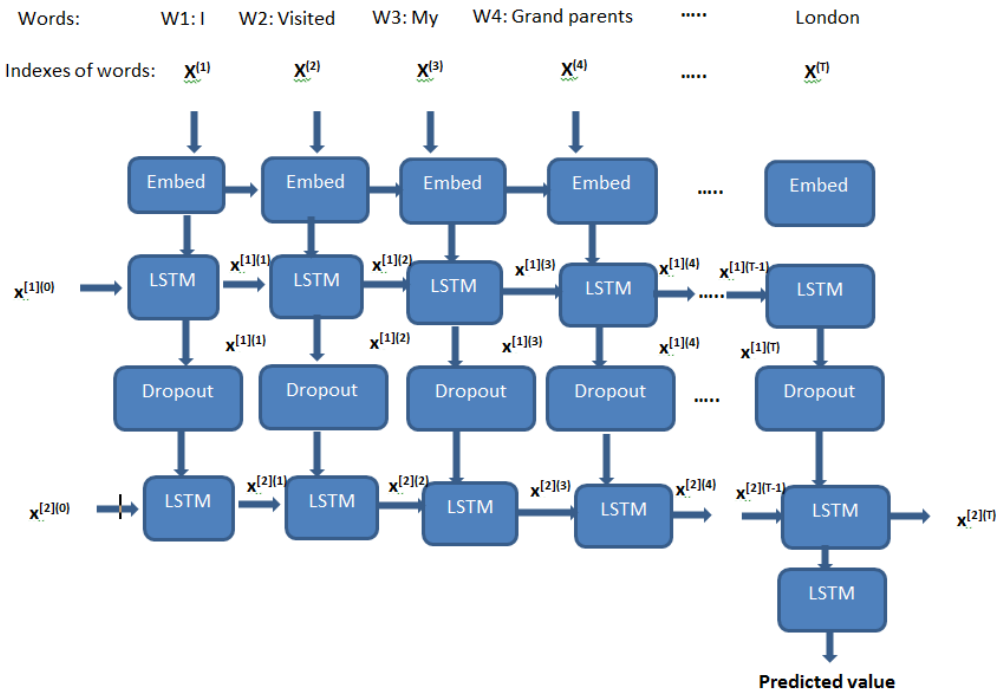
IV. Models' Methodology

Our model receives a textual sentence or a tweet (such as "Let's go for birthday celebration!") as an input and discovers the most relevant emoji to be used with this sentence (🎂). In various social network or mobile interfaces, you need to think of that (✌) is the "well done and victory" symbol rather than the "well done" symbol. With the help of distributed word and emoji embeddings on a same vector space, training set of tweets relates to a very limited number of specific emojis, our proposed model simplify words in the test set to map to the similar emoji as in training set [2,7,8]. This allows building a precise classifier mapping from tweets or posts to emojis.

Each tweets changes into one-hot representation and then into the distributed word vector representation [8]. We use pre-trained 50-dimensional GoogleNews-SLIM word embeddings to represent words, and will feed them into an LSTM, whose job is to calculate our model on an essential emoji-description classification and twitter sentiment method [12].

LSTM components have been widely used to encode textual substances. The elementary encoder model comprises of a textual embedding layer [9], LSTMs layer, and fully-connected layers for supplementary chores such as text classifications based on the encoded feature. The following the architecture and operations of LSTM component for the initial time step and is formulated as given below in Figure 2:

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)



$$in_1 = \sigma(W_{in} x_1 + U_{in} h_0 + b_{in})$$
$$fo_1 = \sigma(W_{fo} x_1 + U_{fo} h_0 + b_{fo})$$
$$out_1 = \sigma(W_{out} x_1 + U_{out} h_0 + b_{out})$$
$$g_1 = \tanh(W_{cell} x_1 + U_{cell} h_0 + b_{cell})$$
$$cell_1 = fo \odot c_o + in_1 \odot g_1$$
$$h_1 = out_1 \odot \tanh(cell_1)$$

Figure 2: Architecture of LSTM components

Where h_0, h_1 signify the current and previous hidden states, in_1, fo_1 and out_1 denote input, forget and an output gates of the initial component, x_1 denotes the recent LSTM and x_t is the word embedding of the current word w_t, W and U denote the weight matrices of the input, recurrent connections and σ is the sigmoid function.

The proposed model comprises a three layer LSTM possessing with 128-dimensional hidden states. This method is based on effective existing literature implementations of neural network-based methods to the task of text processing [3]. As for software, we chose to use Tensorflow to build LSTM. Typically, the default model run till 50 epochs and then save the weights from the emoji embedding layer of the model checkpoint with the maximum categorical accuracy. For loss calculation, we simply add the measure of the cross-entropy loss for each of the image. For optimization, we used the Adam optimizer. After several trails of various learning rates, a value of 0.001 was found to work best. Moreover, the 2nd and 3rd LSTM layer uses dropout of 0.5.

The embedding vector is denoted as a "layer", and maps real integers into compact vectors of fixed size as shown in Figure 3. It can be initialized or trained with a pre-trained embedding [6]. This layer is initialized it with the GoogleNews-SLIM 50-dimensional vectors loaded earlier in the model and requires the description of the vocabulary size, length of input document. Vocabulary size is equivalent to the largest words index in the corpus.

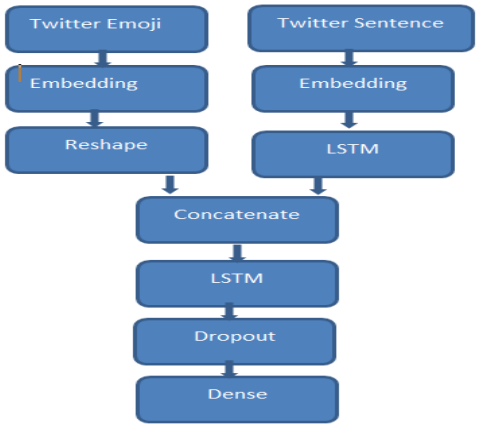


Figure 3: Network design for predicting emoji

Training takes roughly 20 minutes, using Tensorflow as a backend for Keras on an Amazon Web Services EC2 p2-xlarge CPU compute instance.

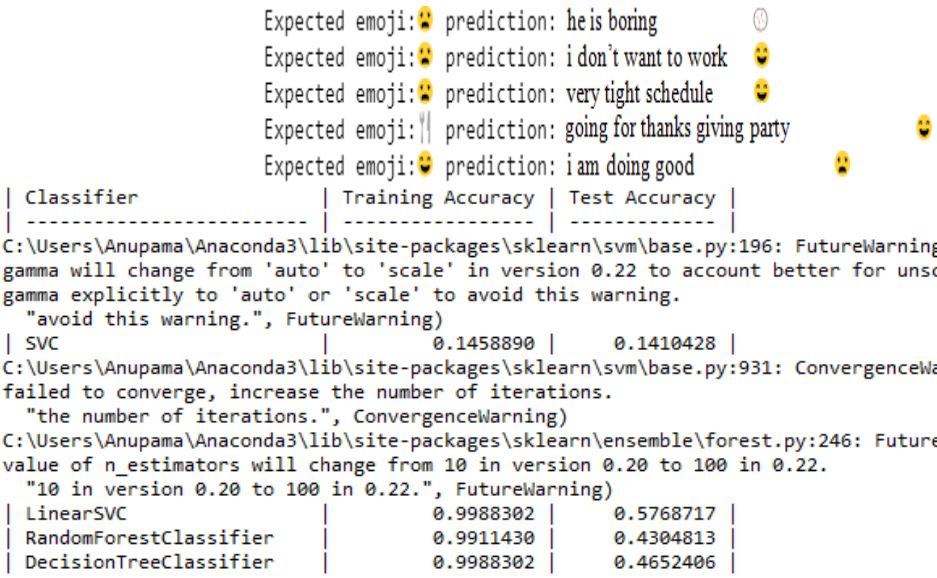


Figure 4: Sample outcome and accuracy of classifiers on emoji prediction

V. CONCLUSION

In this proposed framework, we present a novel method for sentiment method is a field within NLP and attain state-of-the-art performance. Dissimilar from the preceding literature, our proposed method represents both word, emoji embeddings on a same distributed vector space to map the relevant emoji for the given tweet or post, with LSTM deep learning network handle sentence sentiment semantics and assortment of emoji. The predicted classification is also done by linear svc, random forest and decision tree algorithms for GoogleNews-SLIM and emoji2vec embeddings.

REFERENCES

1. Eisner, B., Rocktäschel, T., Augenstein, I., Bošnjak, M., & Riedel, S. (2016). emoji2vec: Learning emoji representations from their description. arXiv preprint arXiv:1609.08359.
2. Guibon, Gaël, Magalie Ochs, and Patrice Bellot. "LIS at SemEval-2018 Task 2: Mixing Word Embeddings and Bag of Features for Multilingual Emoji Prediction." 2018.
3. Lei Zhang, Shuai Wang, and Bing Liu. 2018. Deep learning for sentiment method: a survey. Wiley Interdiscip. Rev. Data Min. Knowl. Discov. 8, 4 (2018).
4. Rong, Xin. "word2vec parameter learning explained." arXiv preprint arXiv:1411.2738 (2014).
5. Pennington, Jeffrey, Richard Socher, and Christopher Manning. "Glove: Global vectors for word representation." Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). 2014.
6. Liu, Pengfei, Shafiq Joty, and Helen Meng. "Fine-grained opinion mining with recurrent neural networks and word embeddings." Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. 2015.
7. Francesco Barbieri, Miguel Ballesteros, and Horacio Saggion. 2017. Are emojis predictable? In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, pages 105–111. Association for Computational Linguistics.
8. Francesco Barbieri, German Kruszewski, Francesco Ronzano, and Horacio Saggion. 2016. How cosmopolitan are emojis?: Exploring emojis usage and meaning over different languages with distributional semantics. In Proceedings of the 2016 ACM on Multimedia Conference, pages 531–535. ACM.
9. Sepp Hochreiter and Jurgen Schmidhuber. 1997. Long ¨ short-term memory. Neural Comput., 9(8):1735– 1780.
10. Fernández-Gavilanes, Milagros, et al. "New methodologies to evaluate the consistency of emoji sentiment lexica and alternatives to generate them in a fully automatic unsupervised way." (2018).
11. Finn ˚Arup Nielsen. A new anew: Evaluation of a word list for sentiment method in microblogs. arXiv preprint arXiv:1103.2903, 2011.
12. Petra Kralj Novak, Jasmina Smailovic, Borut Sluban, and ´ Igor Mozetic. 2015. Sentiment of Emojis. ˇ PLoS ONE, 10(12):1–22, 12.
13. Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., & Kappas, A. (2010). Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology*, *61*(12), 2544-2558.

AUTHORS PROFILE

Dr.A.Anupama received Ph.D. in Computer Science and Engineering, from AdikaviNannaya University, A.P, India. And M.Tech in Computer Science Technology from Andhra University. She is currently working as Associate Professor in the department of CSE at Raghu Engineering College (Autononomous). Her research interests include trust evaluation models, game theory, recommender systems and algorithms in online social networks. Most of her publications and guidance to the students relate to community detection algorithms and recommender systems from social networks and Recommender Systems

Satish Muppidi M.Tech., is working as a Senior Assistant Professor in the Department of Information Technology at GMR Institute of Technology, Rajam, Andhra Pradesh. His areas of interest are Cloud Computing, Soft Computing, and Big Data Analytics. He has 11 years of teaching experience. He presented 15 papers in International / National Conference, and published 12 papers in International journal and attended several workshops and FDPs. He is a life member of CSI, ISTE, and IE.

Dr.Satya Keerthi Gorripati received Ph.D in Computer Science and Engineering from Andhra University, India and M.Tech in Computer Science and Technology from GITAM University. He is currently working as Assistant Professor in the department of CSE at Gayatri Vidya Parishad College of Engineering (Autonomous). His Research Interests include Data Analytics, Social Network Analysis and Recommender Systems