CENNLP at SemEval-2018 Task 2: Enhanced Distributed Representation of Text using Target Classes for Emoji Prediction Representation

Hariharan V, Naveen J R, Barathi Ganesh H. B., Anand Kumar M, Soman K P
Center for Computational Engineering and Networking (CEN)
Amrita School of Engineering, Coimbatore
Amrita Vishwa Vidyapeetham, India
cb.en.p2cen16007@cb.students.amrita.edu,
barathiganesh.hb@gmail.com , m_anandkumar@cb.amrita.edu

Abstract

Emoji is one of the "fastest growing language " in pop-culture, especially in social media and it is very unlikely for its usage to decrease. These are generally used to bring an extra level of meaning to the texts, posted on social media platforms. Providing such an added info, gives more insights to the plain text, arising to hidden interpretation within the text. This paper explains our analysis on Task 2, " Multilingual Emoji Prediction" sharedtask conducted by Semeval-2018. In the task, a predicted emoji based on a piece of Twitter text are labelled under 20 different classes (most commonly used emojis) where these classes are learnt and further predicted are made for unseen Twitter text. In this work, we have experimented and analysed emojis predicted based on Twitter text, as a classification problem where the entailing emoji is considered as a label for every individual text data. We have implemented this using distributed representation of text through fastText. Also, we have made an effort to demonstrate how fastText framework can be useful in case of emoji prediction. This task is divided into two subtasks, they are based on dataset presented in two different languages English and Spanish.

1 Introduction

The consumption of technology in industry delivers potential tools for communication. Messaging has turned into a critical method of communication all through the world and is expanding at a quick rate. Adding emoji in the text convey little more information about the person's emotion, which otherwise is absent in the normal text. Emotions contents of text is better expressed by usage of emojis. Emojis are fundamentally kind of image which are logically connected with the written text. Tweets and online social media platforms are investigated to assess the emotion depth of the several issues for sentiment analysis and Opinion mining in natural language platform. In recent times, the interest in theses area received is very much increased and made several classifications which are polarity based classification such as positive, negative and neutral. In case of emojis certain remarkable studies on emoji sematic and usage find out in papers (Aoki and Uchida, 2011) Relevant study into emoji (Barbieri et al., 2018) are limited in number. The common exploration about emoji has inspired (Barbieri et al., 2017) on descriptive analysis or used them as a indication the emotional affect (Rathan et al., 2017) on social media. That is too restricted in face emojis.

2 Corpus

The shared task (Barbieri et al., 2018)provided 20 most commonly used emojis in tweets English as well as Spanish. That are distinct in nature for English and Spanish corpus (Barbieri et al., 2016). For the simplicity the corresponding emojis are labelled from 0 to 19. The data given for the task is 500k tweet ids for Spanish and 1000k tweet ids for English. Using the tweet ids, tweets are crawled from twitter using 4 different accounts. The crawled data was in JSON format, the raw data from twitter is prepocessed and the labelled data is converted into format suitable for the learning algorithm. For tuning the hyperparameters of the model 20% from the training data set is made into validation set and only the rest 80% is used for the training the model.

3 Related Works

Any mathematical system or an algorithm need some form of numeric representation to work with. One of most naive way of representing word in vector form is one hot representation but it is very ineffective way for representing words in a

Proceedings of the 12th International Workshop on Semantic Evaluation (SemEval-2018), pages 486–490
New Orleans, Louisiana, June 5–6, 2018.

large corpus since the length of one hot vector grows as the vocabulary increases, so we need a better and more effective way which captures some semantic similarities (Ganesh et al., 2016) between nearby words, thus creating the representation for words bring beneficial info about the word and its actual meaning, the methods which encodes these information about the words are called word embedding models, they are categorized into count based and predictive word embedding models. Both embedding models at least some way share syntactic meaning (Soman et al., 2016). But count based word embedding models does not preserve the word order and learn about word semantics

Predictive models attempts to calculate the word vector which captures the both syntactic and semantic meaning (Ganesh H. B. et al., 2016) of the word. This is done by calculating the softmax of the word over the context window. The word embeddings provided by the predictive model not only gives a representation for words but it is also able to learn word similarities and interesting word analogies like "king"-"man"+"woman"="queen". The wor2vec and glove models are the popular predictive models used to learn the word embeddings in many NLP pipelines.The disadvantage of above predictive models are they does not form a sentence representation and morphology of the words is not considered. These shortcomings are overcome the FastText framework which is a recent development in predictive model. (Bojanowski et al., 2016) presented the fastText embeddings, which is devolopment on the word2vec model. Since FastText is considering the char n-gram of the words it learns a good representation for words when compared to word2vec embeddings.

4 Methodology

This work explores the FastText Framework for text classification. It is a fast and lightweight implementation written in C++ to learn word embedding. FastText is a unified framework for text representation and text classification.

Generally text classification is done to categories the class of sentences or documents for task such as sentiment analysis, spam detection etc. In these tasks vector representation is assigned either to the words, sentences or documents depending upon the task. There are various methods for getting the vector representation. In this work using distributed representation, the word embedding is assigned to the words in the vocabulary over the corpora. In FastText framework for doing text classification it extends the concept of Continuous Bag Of Words (CBOW) model introduced word2vec model. Since fastText is improvised version of word2vec model so to get a good understanding the full word2vec

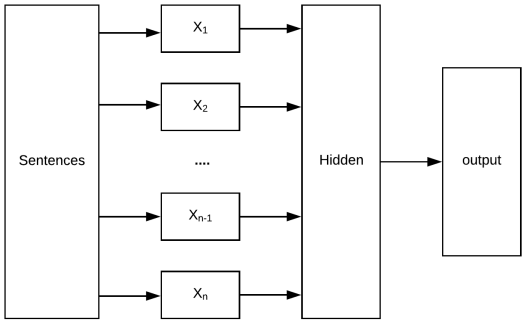

Figure 1: FastTex architecture for emoji and classification.

The word2vec model by proposed by (Mikolov et al., 2013) is shallow neural network architecture, where word embedding is learnt in an unsupervised fashion. It was proposed in the paper Distributed Representations of Words and Phrases and their Compositionality. This paper proposed two architectures for learning the word embeddings, they are continuous bag of words(CBOW) and skip-gram. In the continuous bag of words architecture given the surrounding words, the centre word is predicted, while in the skip gram model given the center word, the context word is predicted.

In a CBOW architecture, as in the figure 1. Let us consider this example sentence we built a sandcastle in the beach in this sentence every word is predicted by taking the surrounding words in the window as the input and softmax is applied over the output to predict the corresponding word (Poria et al., 2016). To reduce the computational complexity of the softmax with large vocabulary, two techniques namely negative sub sampling and hierarchical softmax are applied.

In FastText, for text classification CBOW architecture of Word2vec is slightly modified to form a representation of sentences or document. In Bag of words (BOW), instead of predicting the center word given the context words, the center

word is flipped with the label which it is associated to. Then the softmax is applied over the predefined class. For N set of classes the negative log-likelyhood is given by

$$\frac{-1}{N}\sum_{n=1}^{N} y_n log(\int (BAx_n))$$

where y are the labels and x is the normalized bag of words feature. B and A are the weight matrix from the hidden layer. The BOW does not take the word order into consideration, which otherwise may increase the computation complexity of the model. The word2vec does not predict the word which it has not seen in the training time, to overcome this char N-gram is done within the word and it is given (Brown et al., 1992) to model during training along with word to overcome the unknown words at the test time. The model is trained with a decaying learning rate using stochastic gradient decent. The hierarchical softmax can come in very handy if the no of classes is very large, it basically works like a binary search tree, all the classes are arranged using the Huffmans encoding. By selecting the element in the tree, the search space for the class get reduced into half at each node, this brings the computation complexity in the order of log.

5 Result

Hyperparameter	English corpus	Spanish corpus
learning rate	0.1	0.1
dimension	200	100
window size	3	4
word n-gram	1	1
loss	softmax	softmax
neg	10	10

Table 1: Hyper parameters for english and spanish.

In this work, a classifier based on fast text framework is applied on the Sem-Eval 2018 task 2 emoji detection data set, the classifier is trained to predict the emoji on the English emoji corpus and Spanish corpus. Our team got placed in the 21th position in the English corpus and 13th position on the Spanish corpus. To make our model perform better we evaluated the classifiers hyper-parameter with various values and found the following hyper-parameter values to best performing on the English and the Spanish emoji corpus.

Emo	P	R	F1	%
❤️	80.56	80.24	80.4	21.6
😍	24.61	48.55	32.67	9.66
😂	29.1	59.86	39.16	9.07
💕	23.57	24.34	23.95	5.21
🔥	45.29	43.97	44.62	7.43
😊	11.1	12.46	11.74	3.23
😎	19.88	13.18	15.85	3.99
✨	29.41	19.28	23.29	5.5
💙	23.71	8.59	12.61	3.1
😘	17.98	12.85	14.99	2.35
📷	31.19	48.67	38.01	2.86
🇺🇸	43.58	38.99	41.16	3.9
☀️	63.2	42.77	51.01	2.53
💜	35.48	0.99	1.92	2.23
😉	14.75	2.45	4.2	2.61
💯	22.6	11.17	14.95	2.49
😁	12.04	1.13	2.06	2.31
🎄	65.03	59.09	61.92	3.09
📸	39.31	8.9	14.51	4.83
😜	0.0	0.0	0.0	2.02

Table 2: Precision, Recall, F-measure and percentage of occurrences in the test set of each emoji for english.

Emo	P	R	F1	%
❤️	59.47	65.58	62.37	21.41
😍	24.32	49.08	32.53	14.08
😂	38.07	60.37	46.7	14.99
💕	10.66	5.97	7.65	3.52
😊	13.22	11.67	12.4	5.14
😘	25.48	13.35	17.52	3.97
💪	25.07	28.01	26.46	3.07
😉	10.91	1.32	2.36	4.53
👌	6.58	5.56	6.02	1.8
🇪🇸	26.39	41.51	32.26	4.24
😎	18.75	2.65	4.65	3.39
💙	0.0	0.0	0.0	4.13
💜	0.0	0.0	0.0	2.35
😜	0.0	0.0	0.0	2.74
💞	0.0	0.0	0.0	0.93
✨	35.56	7.69	12.65	4.16
🎶	15.69	15.09	15.38	2.12
💘	0.0	0.0	0.0	1.34
😁	0.0	0.0	0.0	2.09

Table 3: Precision, Recall, F-measure and percentage of occurrences in the test set of each emoji for spanish.

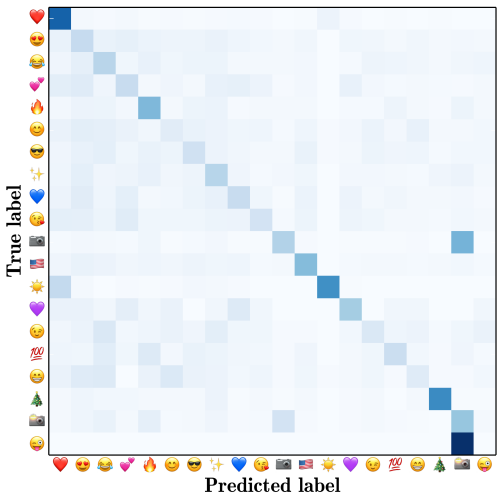

Figure 2: Confusion matrix set of each emoji for English.

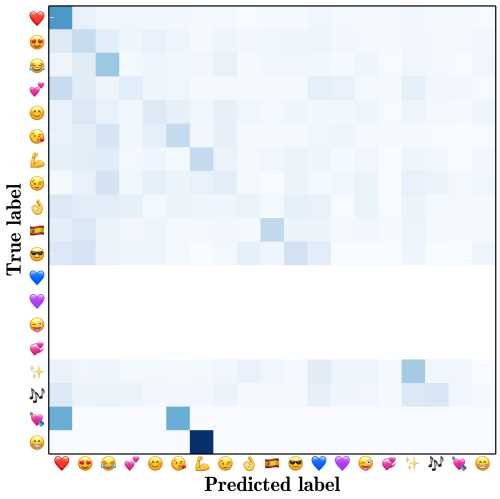

Figure 3: Confusion matrix set of each emoji for Spanish.

6 Conclusion and Future Scope

Every year we see a no of new words being added to the dictionary, most of these new words are the result of new culture trends, the same is applicable with the emojis, which expresses them visually and it is important for us to study its occurrence along with text in social media to better understand the sense of the message. A classifier trained on predicting the emoji from text is significant for understanding the interaction of emoji with the text. In this work a emoji is predicted for the sentences in English and Spanish using fastText framework which is known for being computationally efficient and the learned word representation is on par word representation learned from the standard models like Word2vec.

In this work, a single emoji is predicted for the sentence and since people usually people tend to use more the one emoji in tweets, comments and posts, so we can also extend this problem to a multi-label classification problem.

References

Sho Aoki and Osamu Uchida. 2011. A method for automatically generating the emotional vectors of emoticons using weblog articles. In *Proceedings of the 10th WSEAS International Conference on Applied Computer and Applied Computational Science*, ACACOS'11, pages 132–136, Stevens Point, Wisconsin, USA. World Scientific and Engineering Academy and Society (WSEAS).

Francesco Barbieri, Miguel Ballesteros, and Horacio Saggion. 2017. Are emojis predictable? In *Lapata M, Blunsom P, Koller A, editors. 15th Conference of the European Chapter of the Association for Computational Linguistics; 2017 Apr 3-7; Valencia, Spain. Stroudsburg (PA): ACL; 2017. p. 105-11.* ACL (Association for Computational Linguistics).

Francesco Barbieri, Jose Camacho-Collados, Francesco Ronzano, Luis Espinosa-Anke, Miguel Ballesteros, Valerio Basile, Viviana Patti, and Horacio Saggion. 2018. SemEval-2018 Task 2: Multilingual Emoji Prediction. In *Proceedings of the 12th International Workshop on Semantic Evaluation (SemEval-2018)*, New Orleans, LA, United States. Association for Computational Linguistics.

Francesco Barbieri, Luis Espinosa-Anke, and Horacio Saggion. 2016. Revealing patterns of twitter emoji usage in barcelona and madrid. *Frontiers in Artificial Intelligence and Applications. 2016;(Artificial Intelligence Research and Development) 288: 239-44.*

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606.*

Peter F Brown, Peter V Desouza, Robert L Mercer, Vincent J Della Pietra, and Jenifer C Lai. 1992. Class-based n-gram models of natural language. *Computational linguistics*, 18(4):467–479.

HB Barathi Ganesh, M Anand Kumar, and KP Soman. 2016. From vector space models to vector space models of semantics. In *Forum for Information Retrieval Evaluation*, pages 50–60. Springer.

Barathi Ganesh H. B., M. Anand Kumar, and K. P. Soman. 2016. Statistical semantics in context space : Amrita_cen@author profiling. In *CLEF*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Soujanya Poria, Erik Cambria, and Alexander Gelbukh. 2016. Aspect extraction for opinion mining with a deep convolutional neural network. *Knowledge-Based Systems*, 108:42–49.

M Rathan, Vishwanath R Hulipalled, KR Venugopal, and LM Patnaik. 2017. Consumer insight mining: aspect based twitter opinion mining of mobile phone reviews. *Applied Soft Computing*.

KP Soman et al. 2016. Amrita_cen at semeval-2016 task 1: Semantic relation from word embeddings in higher dimension. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 706–711.