Available online at www.sciencedirect.com
ScienceDirect
Procedia Computer Science 132 (2018) 167–173
Procedia Computer Science
www.elsevier.com/locate/procedia

International Conference on Computational Intelligence and Data Science (ICCIDS 2018)

Opinion Mining on Emojis using Deep Learning Techniques

Valmeekam Karthik[a], Dheeraj Nair[a], Anuradha J[a]

[a] *School of Computing Science and Engineering , VIT University, Vellore, 632014, India*

Abstract

Opinion mining is a trending research topic right now. These techniques are used to know the status of the performance of a business product or the way in which information is spread (positive or negative). It is also being used in the domain of journalism to know the pulse of the people in a famous public action. Even with such huge applications, opinion mining has been restricted to text mining. This can lead to wrong results whenever these texts are used with other tools like emojis. On platforms such as Twitter and Facebook, there is a significant impact of emojis on the opinion that the posts or tweets convey along with the text. This paper aims to provide insight into mining opinions through emojis on tweets using different techniques such as Machine Learning Classifiers, Artificial Neural networks, and Convolutional Neural networks. Additionally, an analysis is made as to which is the best technique to get the proper polarities for the emojis. Through this analysis, the emojis and the text of a tweet are mined separately and then aggregated to generate the polarity of the opinion conveyed in the tweet. This procedure works on a datastore which contains the unicode data of the emojis and the associated weights/polarities. The tweets contents are searched for these emojis and the emoji polarities are given. A relation between the text polarity and emoji polarity is found to detect cases of sarcasm in some tweets.

Keywords: Opinion Mining; Twitter; Artificial Neural Networks; Convolutional Neural Networks

1. Introduction

An important part of a person's life is the process of decision making. This process occurs at each and every second with the frequency of major decisions increasing over time. In this decision- making process, one very important aspect is the idea of "What other people think". Before the Internet era, people used to ask their friends and well-wishers in order to take a decision in their life. But the development of the Internet and Web have helped to get opinions about a product from a vast group of people. In such a vast pool of audience for a product, opinions matter. The web has made it possible to find out about the opinions and experiences of the people who are in no way related to the buyer or are in no way a professional critic. In a similar manner, the reverse is also possible. People are making their opinions

* Corresponding author. Valmeekam Karthik. Tel.: +91-900-374-3098.
E-mail address: valmeekam.karthik2015@vit.ac.in

1877-0509
10.1016/j.procs.2018.05.200

available to a vast number of strangers via the Internet. In such a scenario, on platforms like Twitter and Facebook, the use of emojis has risen. Emojis also have a huge impact on the opinions that people want to express on a service, product or a public action. Unfortunately, due to the lack of attributes in an emoji (technically), very less research has been done in this particular field. In this paper, the aim is to use various Machine Learning and Deep Learning techniques which are at our disposal to find the best way to mine the emojis in order to get the polarity of the emojis in a given tweet.

1.1. Sentiment Analysis

Sentiment analysis, or opinion mining, is the computational study of peoples opinions, sentiments, emotions, and attitudes. It is one of the most active research areas in natural-language processing and is also extensively studied in data mining, web mining, and text mining. The importance of sentiment analysis is growing with the growth of social media networks like Facebook and Twitter.

1.2. Emojis

Facial expressions are one of the most powerful expression techniques. During the time where computers and technology were just developing, there was a trial to imbibe such expressions into our text. Thereby, emoticons were introduced. They are shorthands for facial expressions in the text format. It allows a person to show his feelings and moods regarding a particular topic, object or service. Taking these emoticons a bit further and making them a bit more expressive, an emoji was introduced. Emojis are graphical representations rather than textual representations which make them more powerful. Emojis have become extremely popular on social networking and messaging sites. An immense amount of emojis has been used on all the popular platforms like Twitter, Facebook etc. However, to the best of our knowledge, no large-scale sentiment analysis using emojis has been conducted so far.

1.3. Motivation

Emojis, which are powerful tools of expression, have a huge impact on the opinions that people express. Such powerful tools have not been explored and utilized for mining opinions. The difficulty in mining emojis is relatively much easier than that of mining texts. Even then, the research done in mining opinions through emojis is very less. This was the main driving force behind the paper. This paper aims to provide an insight in mining opinions through emojis using different machine learning and deep learning techniques.

2. Literature Survey

With the ever growing amount of data on the internet, it has become very important to use this data for understanding the audience on a particular application. What people think about a particular change is important for a company to better understand its customers and develop products customised according to their needs. One of the pioneers about this concept was explored in [15] where the author works on a learning algorithm which could classify a review as either a thumbs-up or a thumbs-down according to the sentiment of the review text. The usage of sentiment analysis in a social media presence such as Twitter can be seen in [11]. In this, various avenues are discussed where the method of sentiment analysis could be done. These previous techniques used to work on just words rather than understanding the phrase as a whole. The concept of phrase-level analysis was done in [17] where an entire expression is classified as a whole rather than just a combination of scores for the individual words. The paper [14] talked about the introduction of lexicon-based analysis to sentiment analysis. This was the method in which every word was given a value according to its positivity/negativity and this concept improved the generic sentiment detection compared to the other algorithms at that time. In the paper [7], the authors work on a technique for detection of the tweets into three categories: positive, negative and neutral which was run on a stream of data from Twitter. Most of the work on sentimentanalysis focuses on the English language. There is however a lexical resource which has been used extensively, which is knows for its explicit divisiveness to support sentiment analysis and opinion mining, SentiWordNet 3.0 [2]. SentiWordNet is an extension to its predecessor known as WordNet where each synset is associated with three scores describing the positive, negative and objective nature of the terms in the synset. The paper [5] talks about the possibility of using emotions by emoticons in the tweet to allow sentiment analysis over multi-lingual tweets. Rather than just using the tweet contents, the paper [16] discussed about a method which used the hashtags associated with the tweet to determine sentiment analysis for a particular topic. The paper [8] uses a hybrid approach in their algorithm to classify the twitter feed. The framework of the paper is based on 3 different classifiers, an emoticon classifier, an improved polarity classifier and a SentiWordNet classifier. The paper [3] presses on the issue of the inconsistency of sentiment analysis which can be improved by using different kinds of meta-data of the tweet which can include data like location, time and more. The effects of emoji on sentiment analysis has been talked about in the paper [13]. They had concluded that the usage of emoji can help improve the overall score of the sentiment values for a tweet. The authors of [1] have manually mapped emoticons from Unicode 8.0 to nine emotional categories and then have had a sentiment analysis performed using the emoticons and bag-of-words as features. Another state-of-the-art performer in the field of emojis is DeepMoji [6]. DeepMoji is a model trained on 1.2 billion tweets with emojis to understand how text can be expressed as emojis.

3. Procedure

The procedure involves 6 major steps in order to classify the emojis, polarize the emojis, polarize the text and finally aggregate the polarities. These 6 steps are applied to the tweets that are extracted from the Twitter platform.

3.1. Extracting Tweets and their emojis

The tweets in Twitter was accessed by using the TwitterAPI of Python. A search was made for tweets of a particular hashtag(Example: #iphone") through the Tweepy Cursor of the tweepy library and then the tweets were obtained. Different components of the tweet, such as text, date, number of retweets etc., were fed into different columns of a single pandas dataframe. From the text component of the tweet, the emojis were then extracted by their UNICODE and were put in a LIST. Tweets without any emojis were discarded. Due to the restriction by the TwitterAPI regarding the number tweets that can be retrieved at a time, the tweets were continuously been added to a csv file.

3.1.1. Additional procedure for machine learning classifiers and Artificial Neural Networks

All the emojis in the tweets are then converted into a special code as mentioned above. This is referred to as the emoji code. The procedure for the formation of the emoji code for an emoji is illustrated with an example as follows:

Example: UTF code of a happy emoji \xf0 \x9f \x98 \x81

i) Slashes were removed - xf0x9fx98x81

ii) The xs were removed - f 0 9 f 9 8 8 1

iii) Each character of the string was converted into ASCII form - 102 48 57 102 57 56 56 49

iv) The code was formed - 1024857102577565649

A new column is added into the dataframe consisting of these codes. The reason for forming such codes is because the training of the classifiers has been done on a limited number of emojis. If an emoji, retrieved from a tweet, is not present in the training set, the classifier needs to have a standard representation of the emoji to apply its learning on. This equation to find a special code for each emoji depends on the standard UNICODE representation. This will aid the classifiers in applying their learning in classifying emojis.

3.2. Training

In this step, various machine learning classifiers and neural networks were trained. The neural networks that were used were Artificial Neural Networks and Convolutional Neural Networks.

3.2.1. Training the machine learning classifiers

A csv file consisting of 109 hand-classified emojis, their UNICODES and their polarities, is put into a dataframe. Using the above procedure, the emoji codes for these emojis are obtained and are put into a separate column in the same dataframe. This dataframe with the emoji codes as the input and the polarities as the output was used as the training set. The 10-fold cross validation technique was used to evaluate different machine learning models on this training set with a validation size of 0.5. Once the cross validation was done, the model with the highest accuracy, was chosen to predict the polarities of the emojis.

3.2.2. Training the Artificial Neural Network model

Using the above dataframe where the contents of the csv file are stored, the emoji codes are stored as input and the polarities are stored as the output. The polarities are binarized into 8 labels which correspond to the range of the polaritiy. By doing this, the output layer will have 8 nodes and depending on the polarity the corresponding node is highlighted by the ANN(Artificial Neural Network)[12]. This process aids the ANN in classifying the emojis into 8 different categories. The ANN has 2 hidden layers and an output layer[10]. These layers use the rectifier linear unit function as the activation function. The adam optimizer,[9] an algorithm for first-order gradient-based optimization of stochastic objective functions, based on adaptive estimates of lower-order moments, was used for the optimization of the neural network. The adam optimizer makes use of the first and the second moments of the gradients for the weights. This helps it to achieves good results fast. The default parameters used for the optimizer are: $\alpha = 0.001, \beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 1e - 08$

where α is the learning rate, β_1 is the exponential decay rate for the first moment estimates, β_2 is the exponential decay rate for the second-moment estimates and ϵ is a constant to prevent division by zero.

3.2.3. Training the Convolutional Neural Networks

109 emojis pertaining to major emotions were downloaded as images(png) from [4]. These pictures were hand classified into 9 major polarities which range from -4 to 4. (-4,-3,-2,-1,0,1,2,3,4). Using these images as the training set and another set of similar but slightly different images of emojis as the testing set, the convolutional neural network was trained. A separate folder, named AllEmojis, consisting of all the images of emojis was created. Then in a csv file, two columns were created, where one column contained the UNICODE of these emojis and the other column consisted of the path of these images in the AllEmojis folder, corresponding to the emoji UNICODE. Two convolutional layers were present in the neural network. For both the layers the rectified linear unit function was used as the activation function. The optimizer used here was the adam optimizer, similar to the one used for the ANN. For the output layer, the sigmoid activation function was used with 9 units, each unit corresponding to the polarity.

3.2.4. Hand-classifying emoji polarity for our procedure

Emojis were hand-classified with polarities ranging from -4 to 4 and were represented in a dataframe.

3.3. Prediction

By using the above trained classifiers and neural networks, the polarities of the emojis in the tweets were predicted.

3.3.1. Predicting the polarity with machine learning classifiers

A tweet consists of several emojis, all these emojis were extracted into a list and were fed to the machine learning classifier with the highest accuracy. The average of all the polarities of the emojis consisting in a tweet was taken and was assigned as the combined polarity of all the emojis in a single tweet.

3.3.2. Predicting the polarity with artificial neural networks

A tweet consists of several emojis, all these emojis were extracted into a list and were fed to the artificial neural network with the highest accuracy. The average of all the polarities of the emojis consisting in a tweet was taken and was assigned as the combined polarity of all the emojis in a single tweet.

3.3.3. Predicting the polarity with convolutional neural networks

The previously extracted emojis from the tweets were mapped to their image file in the AllEmojis folder using the csv file. The convolutional neural networks was then fed with these image paths in order to predict the polarity. A tweet consists of several emojis, all these emojis were extracted into a list and were fed to the CNN separately. The average of all the polarities of the emojis consisting in a tweet was taken and was assigned as the combined polarity of all the emojis in a single tweet.

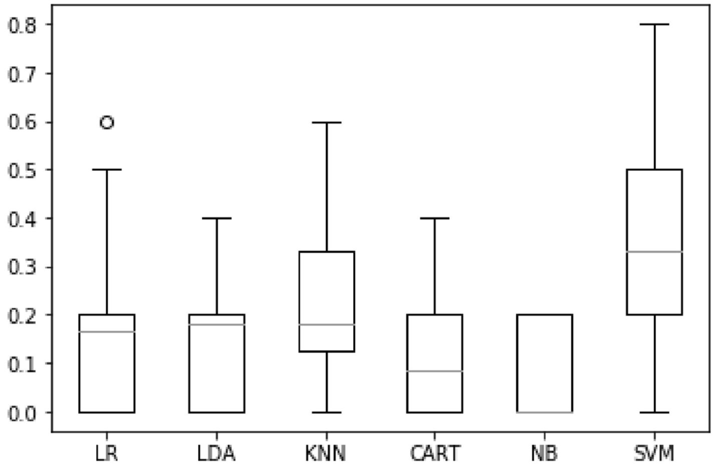

Fig. 1. Comparison of machine learning classifiers.

3.4. Text mining

A standard and a basic text mining classifier using natural language processing has been used to obtain the polarity of the text in the tweet. The textblob library on the python platform was used to mine the text. For a single tweet the text is mined and the polarity of the text is stored in a separate column alongside the emoji polarity.

3.5. Aggregation

Once the polarities are obtained, both, the text polarity and the emoji polarity are first normalized into a range between -1 and 1. They are then aggregated by taking the average of the values. This will represent the final polarity of the tweet

3.6. Final Classification

After the final polarity of a certain tweet is obtained, the tweet is then classified into positive, negative or neutral. If the polarity is positive then the tweet is positive. If the polarity is negative then the tweet is also negative. A tweet is considered neutral if the polarity is zero

4. Analysis

Based on the accuracy that was acquired from the machine learning classifiers, artificial neural networks and the convolutional neural networks, an analysis was done separately in order to determine the best way to predict the polarity of the emojis.

4.1. Machine Learning Classifier Analysis

All the classifiers were analyzed on the python platform using a 10-fold cross validation algorithm to find the best model to polarize emojis. Due to the lack of the number of parameters that an emoji contains, these machine learning classifiers tend to perform on a mediocre level. Since the only parameter that the algorithms can learn from is the emoji code, the accuracy the give is poor. Fig 1 explains the poor performance of the Machine Learning Algorithms on polarizing the emojis.

Table 1. Accuracy values of the Machine Learning Classifiers.

Classifier	Accuracy
Logistic Regression	16.5%
Linear Discriminant Analysis	13.5%
k-Nearest Neighbors	21.5%
Classification And Regression Trees	12.5%
Naive Bayes	6.83%
Support Vector Machine	36.33%

The exact accuracy values of the Machine Learning classifiers are given Table 1.

From Fig 1 and Table 1 one can easily deduce that the major machine learning classifiers are not suitable for classifying the polarity of an emoji. Therefore, deeper learning algorithms were introduced to increase the accuracy of the prediction.

4.2. Deep learning models and search-based classifier analysis

Two different neural network models and a basic search-based classifier were formed for the classification. Table 2 depicts the accuracy for the three models. As our classifier does a basic search-and-assign task for an emoji, the accuracy is always a hundred percent.

Table 2. Accuracy values of the Deep Learning models and the Search-based classifier.

Classifier	Accuracy
Artificial Neural Network Model	48.95%
Convolutional Neural Network Model	97.43%
A Simple Search-based Classifier	100%

Evidently, it is not advisable to use the Machine Learning classifiers or the ANN model for polarizing emojis but CNN and the mentioned classifier can be used to do that. However, even though the CNN gives us an excellent accuracy value, the time taken to train the convolutional neural network model is very high. Moreover, the process that is involved in training the CNN is also a bit tedious. This is one of the major drawbacks of the convolutional neural networks. Therefore the best way is to go for a basic assignment of polarities from a hand-classified list of emojis.

4.3. A major finding in the detection of sarcasm

By analysing the results for around 7000 tweets through the CNN and mentioned search-based classifier, a small yet impactful finding was discovered. The finding is illustrated with a following example:
Example: One of the tweets in the dataset read the following on the latest iphone release:

”Eagerly waiting for Apple to drop the new iPhone like I can afford it.”

The tweet also included two laughing emojis at the end. This tweet can be evidently stated as a **sarcastic tweet**.
Finding: The normalized text polarity for the above text was negative whereas the normalized emoji polarity for the emojis was positive. For this tweet there was a difference in the emoji polarity and text polarity. For such tweets there was a major difference in the emoji and the text polarity. Whenever such a difference in the two polarities striked in a tweet, that tweet was found to be a sarcastic tweet. In this way, sarcasm can be detected. The only drawback is that this works only with tweets having emojis and it cannot be detected by using plain text.

5. Conclusion and Future scope

The polarities of the emojis were obtained by training the standard machine learning classifiers. But due to the poor accuracy of the classifiers, deeper algorithms were required. A procedure was discovered to use the convolutional neural networks to determine the polarity of a given emoji. This procedure linked the unicodes of the emojis to their corresponding png format image.Another procedure was also introduced in which the emojis were polarised manually.Various tweets were obtained from the Twitter platform on a specific topic or product. The polarities for all the emojis in a given tweet was found through the above mentioned methods. A normalized polarity was acquired for each tweet and the text in the tweet was mined using a standard natural language processing method. The analysis of these procedures led to the path of detecting sarcasm. The detection of sarcasm in tweets using emojis is a very simple yet impacful finding that has not been found by anyone. These results also reflect on the issues in mining emojis. The main problem for mining emojis is the attributes that the emojis carry. As all the ML classifiers and the ANN ran on a single attribute, the result was a failure. As the images had many attributes that the CNN could grab and learn, the accuracy shot up to 97%. The impact and importance of emojis in expressing the opinions that people have, is going to increase at a rapid rate. In order to make the most out of their opinions, there is a need to find more technical attributes for emojis that will help the ML classifiers and the ANN to perform better in polarizing them. This is an open question worthy of pursuing in the future.

References

[1] Amalanathan, A., Anouncia, S.M., 2015. Social network users content personalization based on emoticons. Indian Journal of Science and Technology 8, 23.

[2] Baccianella, S., Esuli, A., Sebastiani, F., 2010. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining., in: LREC, pp. 2200–2204.

[3] Bravo-Marquez, F., Mendoza, M., Poblete, B., 2013. Combining strengths, emotions and polarities for boosting twitter sentiment analysis, in: Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining, ACM. pp. 2–3.

[4] Contributors, E.I., . Emojis in png format. URL: www.emojiisland.com.

[5] Cui, A., Zhang, M., Liu, Y., Ma, S., 2011. Emotion tokens: Bridging the gap among multilingual twitter sentiment analysis. Information retrieval technology , 238–249.

[6] Felbo, B., Mislove, A., Søgaard, A., Rahwan, I., Lehmann, S., 2017. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. arXiv preprint arXiv:1708.00524 .

[7] Gokulakrishnan, B., Priyanthan, P., Ragavan, T., Prasath, N., Perera, A., 2012. Opinion mining and sentiment analysis on a twitter data stream, in: Advances in ICT for emerging regions (ICTer), 2012 International Conference on, IEEE. pp. 182–188.

[8] Khan, F.H., Bashir, S., Qamar, U., 2014. Tom: Twitter opinion mining framework using hybrid classification scheme. Decision Support Systems 57, 245–257.

[9] Kingma, D., Ba, J., 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 .

[10] Mitchell, T.M., 1997. Artificial neural networks. Machine learning 45, 81–127.

[11] Pang, B., Lee, L., et al., 2008. Opinion mining and sentiment analysis. Foundations and Trends® in Information Retrieval 2, 1–135.

[12] Schalkoff, R.J., 1997. Artificial neural networks. volume 1. McGraw-Hill New York.

[13] Shiha, M., Ayvaz, S., 2017. The effects of emoji in sentiment analysis. Int. J. Comput. Electr. Eng.(IJCEE.) 9, 360–369.

[14] Taboada, M., Brooke, J., Tofiloski, M., Voll, K., Stede, M., 2011. Lexicon-based methods for sentiment analysis. Computational linguistics 37, 267–307.

[15] Turney, P.D., 2002. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews, in: Proceedings of the 40th annual meeting on association for computational linguistics, Association for Computational Linguistics. pp. 417–424.

[16] Wang, X., Wei, F., Liu, X., Zhou, M., Zhang, M., 2011. Topic sentiment analysis in twitter: a graph-based hashtag sentiment classification approach, in: Proceedings of the 20th ACM international conference on Information and knowledge management, ACM. pp. 1031–1040.

[17] Wilson, T., Wiebe, J., Hoffmann, P., 2005. Recognizing contextual polarity in phrase-level sentiment analysis, in: Proceedings of the conference on human language technology and empirical methods in natural language processing, Association for Computational Linguistics. pp. 347–354.