I. J. Mathematical Sciences and Computing, **2022, 1, 37-43**
Published Online February 2022 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijmsc.2022.01.04

Emoji Prediction Using Emerging Machine Learning Classifiers for Text-based Communication

Sayan Saha
School of Computer Science and Engineering, Vellore Institute of Technology, Vellore, Tamil Nadu, India, 632014.
E-mail: sayan.saha1010@gmail.com

Kakelli Anil Kumar
Associate Professor, School of Computer Science and Engineering, Vellore Institute of Technology, Vellore, Tamil Nadu, India, 632014.
E-mail: anilsekumar@gmail.com

Received: 01 June 2021; Accepted: 18 July 2021; Published: 08 February 2022

Abstract: We aim to extract emotional components within statements to identify the emotional state of the writer and assigning emoji related to the emotion. Emojis have become a staple part of everyday text-based communication. It is normal and common to construct an entire response with the sole use of emoji. It comes as no surprise, therefore, that effort is being put into the automatic prediction and selection of emoji appropriate for a text message. Major companies like Apple and Google have made immense strides in this, and have already deployed such systems into production (for example, the Google Gboard). The proposed work is focused on the problem of automatic emoji selection for a given text message using machine learning classification algorithms to categorize the tone of a message which is further segregated through n-gram into one of seven distinct categories. Based on the output of the classifier, select one of the more appropriate emoji from a predefined list using natural language processing (NLP) and sentimental analysis techniques. The corpus is extracted from Twitter. The result is a boring text message made lively after being annotated with appropriate text messages

Index Terms: Google GBoard, Twitter, n-gram, Natural Language Processing (NLP), Sentimental Analysis, Machine Learning**,** Classifiers

1. Introduction

Creation of software that helps us to predict the most appropriate emoji for a given sentence predicting the emotion. We have used a Twitter corpus with 7480 statements and tweets and pre-processing the whole data and then apply classification methods to it for a better understanding of which classification technique when used will provide us with more accurate results [1]. After analyzing which one is giving us better training and test accuracy, select the classifier and try it. After the whole exercise, predict the emotion and sentiment of the tweet and assign an emoji to that tweet [2]. This whole feature would be a great addition in keypads where the software can suggest the use of a particular and applicable emoji to the user while he is typing out his sentence.

1.1. Motivation

According to most research papers that are out there; motivation for sentiment analysis is twofold. Nowadays all customers and product creators value "Client's feedback" that is given for all those services and products. Therefore, educationists are trying and giving good considerable inputs to the Sentimental Analysis industry. From a consumer perspective, it is important to know the opinion of the people around us when making a decision. Previously those small talks happened between close groups with friends and family but now with the dawn of the internet and its accompanying technologies, people are expressing their outlook and views in forums [5]. Since nowadays users have started to use the internet extensively so it has also led to an increase in blogs and forums where users can voice their opinions regarding the pros and cons of products. These views shape the future of a product or service. Hence the vendors and sellers have identified this opportunity to see and check reviews about their products and improve their service by looking into those statements [6]. Here the motivation also lies in providing users emoji while they type their text so that they don't have to search through all the emoticon lists.

Copyright © 2022 MECS

Emoji has been a way of expression for quite a long time so it makes sense to assign them to textual statements. Sentimental Analysis as an independent research area is still a work in progress. I wanted to work out something which helps people to identify the proper sentiment or mood of a statement including if it is sarcastic. That's why I came up with the idea of playing with emojis.

The main idea to bring out this paper was to use algorithms and techniques up from scratch and build a model on it. This will help to scale up the model for emoji prediction according to need and will give us the cushion of flexibility. The work where all the things were built from the scratch helping us to understand the minute intricacies of the actual problem statement. Our proposed work brings out a solution for the problem without using any APIs or inbuilt third-party software.

1.2. Background

There are many sources of sentiment analysis using text-based sentiment analysis and many of them focus on textual specific sentiment analysis. Such work can be found in the sentences labeled subjective and objective and then the application of scientific machine learning methods to the subjective parts. So that the polarity classification ignores words that are absurd or misleading. The problem of prediction of emoji constitutes multiple domains. Techniques from a variety of modern disciplines have been used to approach this problem. Sentiment analysis from Natural Language Processing, Classification from Machine Learning, word generation using LSTMs (Long Short Term Memory) / RNNs (Recurrent Neural Networks), etc. is some of the many areas of modern data science that are being leveraged in this domain [7, 8].

Sentimental analysis is being used in a variety of use cases from getting movie reviews through feature extraction from social media statements [14], getting customer feedback for products [15], and many more. My work will be to accurately identify the sentiment of the text and apply a suitable emoji justifying the tone of the text. One of the papers talks about textual extraction for sentimental analysis from the Twitter corpus [16]. In this paper, machine learning classifiers were first compared and then the best one was used to give the best prediction. Here they went ahead with the Neural Network classifier.

2. Literature Survey

A good number of research papers have focused on voice prediction accompanied by some extra features. One paper discusses the use of emoji entry through voice recognition [9]. It is mainly focused to aid visually impaired clients who can use their voice and the program will predict the correct emoji from the pool and announce it through Apple Voiceover. Here most of the functionalities were designed over pre-build APIs like the Google Search API. This was functioning as a web app with the usage of aria labels for UI and aria-live for the speech output feature, this paper was successful in addressing the problem of visually impaired people. A second paper also had something on the lines of the previous one [10]. It focused on solving the issue for deaf and hard of listening users in the online meetings by captioning the talks between the participants through concise emojis. Here too the work was based on pre-built APIs from Google like Google Speech to Text API, voice emotion recognition API. Another paper focuses on emoji's effects on the reading times and what effects it does have on the reader's mood [11]. Most of the papers addressed the basic core problems of sentimental analysis through more complex deep learning models, Convolutional NeuralNetwork (CNN), and attention-based Bidirectional Gated Recurrent Unit (BiGRU). One model uses the SLCABG model and pairs it up with the advantages of sentimental lexicons [12]. Good work is done in the field of sentimental analysis by including fuzzy sentiments by taking into account the elements of lexical, word positioning, and word-type and sentiment polarity [13].

3. Methodology

3.1. Project Description

The process is designed to be minimal, yet impactful. In this paper, we have proposed the usage of classification algorithms operating on features extracted from the text message (which will be converted to n-grams) to predict and classify the tone of the message into one of seven distinct.

Having successfully categorized a message, it is trivial to map it to a predefined list of emojis. Multiple emojis represent a single tone, thus allowing the possibility of annotating a single text message with multiple.

The grouping methods are utilizing:

- The first is the Support vector classifier (SVC). An SVM (Support Vector Machine) is technically a selective classifier that is formalized as an isolating hyperplane. As a result, the calculation yields an ideal hyperplane that arranges new models provided named preparing data (managed learning) [1].
- The second is Linear SVC. The whole point of Linear SVC (Linear Support Vector) is to accumulate all the data that can provide it with and giving us a "best fit" hyperplane which segregates or distinguishes

information. After having a hyperplane, feed some highlights into the classifier to figure out and see the expected class [2]. So here the classifier applies linear kernel function to perform classification, this helps it to perform better in the cases of large data. So when it comes to its comparison with SVC the Linear SVC takes additional parameters like penalty normalization (L1 and L2) and loss function.

- The third is the Decision Tree. The decision tree can be trained through the source part divided into further subsets which will be dependent on the trait test. In a recursive process known as recursive partitioning, this method is rehashed on each inferred subset. The recursion is complete when all of the subsets at a hub have the same estimate of the variable or the time when the component does not exceed the expectations again. The development of a Decision tree classifier does not require any space information or parameter setting, and in this way is proper for exploratory learning revelation [3].
- Fourth is Random Forest [4]. This classifier which is also known as random decision forests is an assembled learning process designed for segregation, regression, and some more tasks which work by making a large number of decision trees during the training and then doing the prediction of class which is mean prediction (regression) or mode of those classes (classification) belonging to those separate trees.

Understand SVC and Linear SVC both have the same parent features just Linear SVC comes with an extra parameter of Kernel-linear. And when it comes to random forest classifier it behaves like a superset of decision tree classifier. It does have a lot of decision trees in it. So it is slow in comparison to the decision tree which behaves well on linear data. The random forest does need more training. This paper will be showing a fun way of performing sentimental analysis. As mentioned above, first dissecting the sentences through n-grams then will look out for distinctive features which will constitute words or a group of words then train a classifier on those sentences to see and compare those results.

3.2. Objectives

Our first goal is to extract feature texts from the pickle dataset, which will help this model to learn. The dataset contains about 7500 instances of text categorized into seven categories:

- Joy
- Fear
- Anger
- Sadness
- Shame
- Guilt
- Disgust

3.3. Pre-processing

The following preprocessing steps are applied to a text before it is fed into the model:

- Stop word removal: This step removes common words that don't impart any meaning or context to a sentence but are present to make the sentence complete and grammatically accurate.
- Lemmatization: This step converts every word into its root form, thus reducing the number of features and increasing accuracy.

3.4. Feature Engineering

The text is converted into features by generating all the n-grams of the word for values of n ranging from 0 to 4 after converting the text into lowercase and ignoring anything which isn't alphanumeric or punctuation. The generated n-grams are used as features for the classifiers. In this work, we will first convert all the characters in lowercase and then divide the sentences through n-grams, how many times the n-grams appear in the base. We will then apply the four classifiers into data which we had already divided in 80-20 fashion. 80% will use for training and the remaining 20% for testing purposes. The results are estimated for finding the best and high performing classifier among four classifiers for the dataset.

The data was in the form of a pickle format, firstly converted the pickle file into a text file for able to feed it to code. The codebase will start with a utility function that will load data from the text file. This particular part will return a list after loading up the sentences from a text file. After that, eliminated all the stopwords from my data, used "nltk" library which will identify all the stopwords and after that, all the stopwords could be easily removed. Then it is converted the sentences into n-gram tokens, tokenizing all the sentences into words and appending them into a list. The code also has a create feature function which will create the vectors from the sentences given. So it is functioning as a Bag of Word model that takes in a lot of words and gives out a vectored form of data, giving out a representation of corpus. Then convert the labels into text data after which is defined all the seven categories to decided for the model. Our model will take in the sentence and predict one of these emojis for the sentences. Now it will be having two array

variables one holding the test input data and the other the labels to be assigned to the sentences. Start the modeling part with four classifiers namely Support Vector Classifier, Linear Support Vector Classifier, Random Forest Classifier, and Decision Tree Classifier. A test is created and training set using the "train_test_split" function to define the metrics against the tested models. Hence start modeling the data into these classifiers. Figure 2 shows the dictionary mapping that has been defined for the model, each emotion or sentiment is been given three emoji to be used randomly while assigning the emotion to textual data.

To the validation part of the code, a confusion matrix of all the predicted models is used to observe the total number of false positives, false negatives, true positives, and true negatives in a seven-by-seven matrix defining the confusion matrix. To visualize the matrix, a heat map is plotted as shown in figure 3. In the color bar as the opacity of the color increases the total true positives or the true negatives also increases. All the violet region shows the false negatives and false positives and the diagonal value shows the true positives and true negatives. Then applied to the chosen model which is Linear SVC in this case and applied it to the whole test data corpus while assigning the accurate emotions from the seven emotions listed to those sentences and grouping them up. After that, mapped the emojis to those emotions. From Figure 1, it can be observed that the whole corpus is divided into the identified emotions and apply emojis to these textual statements.

```
joy       (1. 0. 0. 0. 0. 0. 0.)  1084
anger     (0. 0. 1. 0. 0. 0. 0.)  1080
sadness   (0. 0. 0. 1. 0. 0. 0.)  1079
fear      (0. 1. 0. 0. 0. 0. 0.)  1078
disgust   (0. 0. 0. 0. 1. 0. 0.)  1057
guilt     (0. 0. 0. 0. 0. 0. 1.)  1057
shame     (0. 0. 0. 0. 0. 1. 0.)  1045
```

Fig.1. Segregation of whole text corpus according to the emotion features.

```
emoji_dict = {"joy":["😊", "😌", "😁"],
              "fear":["😱","😟", "😨"],
              "anger":["😡", "😤","😠"],
              "sadness":["😔", "😢", "😕"],
              "disgust":["😒", "😏", "🤮"],
              "shame":["😳","😳","😶"],
              "guilt":["😥", "😳", "🙄"]}
```

Fig.2. The emoji dictionary mapped to the emotion feature.

4. Results and Discussions

Our objective here is to apply the classifiers on the dataset and see which provides us with the best ratio. We've tried four different algorithms:

- SVC - Support Vector Classifier
- Linear SVC - Support Vector Classifier with a Linear Kernel function
- Decision Tree Classifier
- Random Forest Classifier (which is an ensemble method)

4.1. Error Metrics

Since this is a classification problem, the error metrics and performance evaluation measures being used are accuracies and the confusion matrix. This confusion matrix will give us a piece of abstract information regarding the performance of these four classifiers. Emoji has been a way of expression for quite a long time so it makes sense to assign them to textual statements. Sentimental Analysis as an independent research area. Our work helps to identify the proper sentiment or mood of a statement including if it is sarcastic. Figure 3 shows the accuracies in the testing data, and accounting for overfitting, chose the Linear SVC classifier as the classifier of choice for this problem. One thing that helps Linear SVC to put up a strong case is that it is designed to be faster while converging large data samples. Linear SVC has a linear kernel for the basis function.

Copyright © 2022 MECS

4.2. Results

Table 1. The accuracies of classifiers.

Classifiers	Training Accuracy	Testing Accuracy
SVC	0.9271390	0.4832888
Linear SVC	0.9981618	0.5681818
Decision Tree	0.9981618	0.5381016
Random Forest	0.9981618	0.4672460

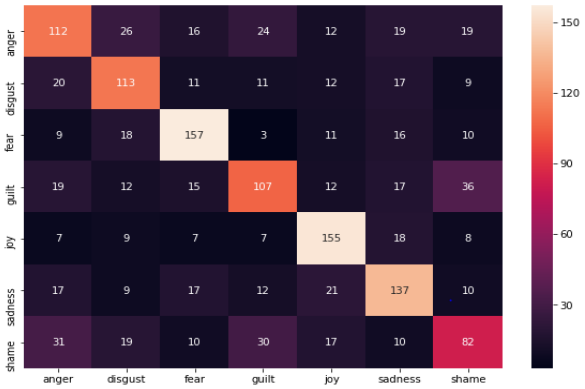

Fig.3. Heat map of the confusion matrix based on the accuracies and accounting for overfitting (using Linear SVC classifier)

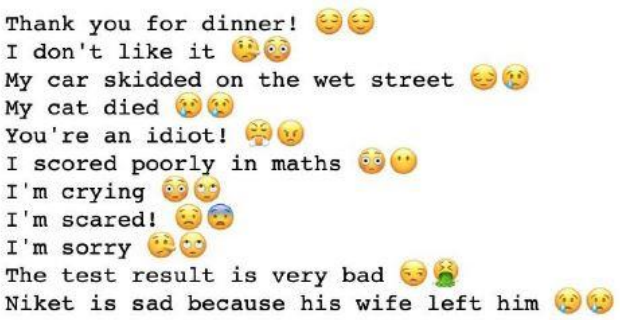

Fig.4. The results of the classifier on real-time data.

4.3. Result Description

After seeing the error metrics through accuracy scores that Linear SVC works comparatively better than the rest of the lot. After that, plot a confusion matrix that gave us the heat map. An ideal confusion matrix would have given us numbers in the diagonal part and null in the rest of the cells. Observe a pattern where anger is been classified with anger emoji most of the time. Though for the rest of the time it isn't that good. In the area of training accuracy, Linear SVC, Random Forest Classifier, and Decision Forest Classifier all got the highest and equal value whereas SVC got the least value. So now the choice pool is between the former three classifiers. Now coming to test accuracy Linear SVC got the highest value. Since in both the metrics of training and test accuracy Linear SVC is giving the best results I will proceed with the Linear SVC only. The accuracy of the classification techniques will be made by dividing the correct classification number of texts by the total number of texts using equation 1. The results of the classifier on the real-time data are shown in figure 4.

$$Accuracy = No.\ of\ correct\ classifications\ /Total\ No.\ of\ classifications \quad (1)$$

5. Conclusion

Emotion or sentimental analysis study is in its premature state and is still in its development stage because of the sheer difficulty in identifying and modeling sentiments. Researchers from all over the world are trying hard to dish out a full-proof emotion recognition system. The goal for the emotion recognition system will be to accurately judge the

Copyright © 2022 MECS

emotional state of the user and its degree of extent through the interfaces or some personalized systems. Here in this paper, assign emojis or emoticons to texts after doing extensive work on those classifiers to see which of those gives us the best result. Based on accuracy scores, choose Linear SVC through some overfitting still works better than the rest.

The point is that there is no one particular solution for this particular problem. The reason we got Linear SVC as the best classifier is dependent on many factors maybe if the data pool increases the accuracy game can shift towards Random Forest Classifier or Decision Tree Classifier. Since Linear SVC has more flexibility in the choice of loss functions and penalties it will mostly fair better than SVC with large data pools. Decision Tree works great with linear data. If the more data for training maybe Random Forest could give us better results. There are other classifiers too which can be a prospective classifiers. As talked about earlier making a project from scratch will help us to scale it easily. This classifier can be also used on other multilingual data sets too. So at last the conclusion comes out that Linear SVC works best for this dataset and situation.

6. Future Scope

There is a lot of scope in this paper which can be further researched and can be showed. As seen from the confusion matrix the results are not perfect. We can draw the accuracy ratings close to perfect by bringing more efficient classifiers or making it training on an even larger set of data. Paying more attention to text features like punctuations, sarcasm and all caps will give us better results. This system has a good chance to grow and can be used even for industrial purposes for customer feedback if the results can be polished further. This project can be tweaked according to the problem statement and can be applied by individuals and companies. Companies can apply this system on review forums to see and note valuable customer feedback.

From other papers, we can also note that emoji entries can be used for disabled people including visually impaired people and people having hearing problems [9, 10]. Another thing that can be added is third-party APIs to improve the efficiency including better feature extraction and prediction. Further one thing that can be done to improve the p[roject as a whole is to add extra features using voice extraction and voice announcement.

Acknowledgment

Firstly, I would like to thank my Professor Dr. Kakelli Anil Kumar from the Department of Analytics, SCOPE, Vellore Institute of Technology. I am privileged to work under his guidance. He always showered me with immense encouragement and invaluable guidance and took great interest in the progress of work. He always took out time for good discussions regarding our paper. I would also like to thank my university management Hon'ble Chancellor, Dean, HODs of the school of Computer Science and Engineering (SCOPE) for providing me with the opportunities and environment which helped me to write this paper. They provided me with all the required facilities and infrastructure which helped me out to complete this work successfully.

References

[1] Chunhua Zhang, Xiaojian Shao, Dewei Li, "Knowledge-based Support Vector Classification Based on C-SVC", Procedia Computer Science 17:1083-1090, December 2013.

[2] Daniel Dichiu, Irina Rancea, "Using Machine Learning Algorithms for Author Profiling In Social Media Notebook", PAN at CLEF 2016.

[3] S. R. Safavian, D. Landgrebe, "A survey of decision tree classifier methodology," in IEEE Transactions on Systems, Man, and Cybernetics, vol. 21, no. 3, pp. 660-674, doi: 10.1109/21.97458, 1991.

[4] Gerard Biau, "Analysis of a Random Forests Model", Journal of Machine Learning Research 13, 1063-1095 Submitted 10/10, Published 4/12, 2012.

[5] Nurendra Choudhary, Rajat Singh, Vijjini Anvesh Rao, Manish Shrivastava, Twitter corpus of Resource-Scarce Languages for Sentiment Analysis and Multilingual Emoji Prediction, 2018.

[6] Ronen Feldman, Techniques and applications for sentiment analysis, Communications of the ACM, April 2013.

[7] Jin Wang, Liang-Chih Yu, K. Robert Lai, Xuejie Zhang, Dimensional Sentiment Analysis Using a Regional CNN-LSTM Model, Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pages 225–230, Berlin, Germany, August 7-12, 2016.

[8] M. Thamarai, S P. Malarvizhi, "House Price Prediction Modeling Using Machine Learning", International Journal of Information Engineering and Electronic Business (IJIEEB), Vol.12, No.2, pp. 15-20, 2020. DOI: 10.5815/ijieeb.2020.02.03

[9] Mingrui "Ray" Zhang, Ruolin Wang, Xuhai Xu, Qisheng Li, Ather Sharif, Jacob O. Wobbrock, "Voicemoji: Emoji Entry Using Voice for Visually Impaired People", Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, May 2021, Article No.: 37Mingrui "Ray" Zhang, Ruolin Wang, Xuhai Xu, Qisheng Li, Ather Sharif, Jacob O. Wobbrock, "Voicemoji: Emoji Entry Using Voice for Visually Impaired People", Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, May 2021, Article No.: 37, https://dl.acm.org/doi/10.1145/3411764.3445338

[10] Kotaro Oomori, Akihisa Shitara, Tatsuya Minagawa, Sayan Sarcar, Yoichi Ochiai, "A Preliminary Study on Understanding Voice-only Online Meetings Using Emoji-based Captioning for Deaf or Hard of Hearing Users", The 22nd International ACM

SIGACCESS Conference on Computers and Accessibility, October 2020 Article No.: 54, https://doi.org/10.1145/3373625.3418032
[11] Alattar, Fetheya N., "Happy, Sad or Pizza: A Review of Emoji Effects on Reading Times and their Relation to Mood" (2021). University Honors Theses. Paper 1087. https://doi.org/10.15760/honors.1114
[12] L. Yang, Y. Li, J. Wang and R. S. Sherratt, "Sentiment Analysis for E-Commerce Product Reviews in Chinese Based on Sentiment Lexicon and Deep Learning," in *IEEE Access*, vol. 8, pp. 23522-23530, 2020, doi: 10.1109/ACCESS.2020.2969854.
[13] H. T. Phan, V. C. Tran, N. T. Nguyen and D. Hwang, "Improving the Performance of Sentiment Analysis of Tweets Containing Fuzzy Sentiment Using the Feature Ensemble Model," in *IEEE Access*, vol. 8, pp. 14630-14641, 2020, doi: 10.1109/ACCESS.2019.2963702.
[14] B. Narendra, K. Uday Sai, G. Rajesh, K. Hemanth, M. V. Chaitanya Teja, K. Deva Kumar, "Sentiment Analysis on Movie Reviews: A Comparative Study of Machine Learning Algorithms and Open Source Technologies", August 2016, MECS, DOI: 10.5815/ijisa.2016.08.08
[15] Anjali Dadhich, Blessy Thankachan, "Sentiment Analysis of Amazon Product Reviews Using Hybrid Rule-based Approach", April 2021, MECS, DOI: 10.5815/ijem.2021.02.04
[16] Golam Mostafa, Ikhtiar Ahmed, Masum Shah Junayed, "Investigation of Different Machine Learning Algorithms to Determine Human Sentiment Using Twitter Data", April 2021, MECS, DOI: 10.5815/ijitcs.2021.02.04

Authors' Profiles

Sayan Saha is from New Delhi, India. He is a B. Tech final year Computer Science and Engineering student from Vellore Institute of Technology, Vellore. He has interned with DRDO, India in 2019, Texas Instruments (NSIT) in 2018, and TATA Power in 2020 and is currently working in TeejLab Inc. as a Software and Services Developer. He has won hackathons organized by Google Developers and MLH. He has been also selected as an HPAIR delegate and has been part of the Schneider Go Green Challenge Greater India Team. He is interested in Blockchain, NLP, and Digital Forensics and wants to work in these fields in the future.

Dr. Kakelli Anil Kumar is an Associate Professor of the School of Computer Science and Engineering at the Vellore Institute of Technology (VIT), Vellore, TN, India. He earned his Ph.D. in Computer Science and Engineering from Jawaharlal Nehru Technological University (JNTUH) Hyderabad in 2017, and graduated in 2009 and under-graduated in 2003 from the same university. He started his teaching career in 2004 and worked as an Assistant Professor, and Associate Professor, and HOD in various reputed institutions of India. His current research includes wireless sensor networks, the internet of things (IoT), cybersecurity and digital forensics, Malware analysis, block-chain, and crypto-currency. He has published over 40 research articles in reputed peer-reviewed international journals and conferences.

How to cite this paper: Sayan Saha, Kakelli Anil Kumar," Emoji Prediction Using Emerging Machine Learning Classifiers for Text-based Communication ", International Journal of Mathematical Sciences and Computing(IJMSC), Vol.8, No.1, pp. 37-43, 2022. DOI: 10.5815/ijmsc.2022.01.04