SemEval-2018: Task 1, 2 and 3: Emoji Recognition, Emoji Prediction and Irony Detection through GioVe model and 3 Machine Learning Classifiers

*Note: Project Final Github

Julio Ernesto Pérez Lara - 2003838
School of Computer Science and Electronic Engineering
University of Essex

***Abstract*—In this paper, we will describe a methodology to predict and recognise emoji in tweets as well as irony detection. The approach of this project is based on the classic bag-of-words model together with word embeddings techniques.**
For the classification process, the used algorithms were (i) Logistic Regression Classifier (LRC), Support Vector Machine (SVM) and Random Forest Classifier (RFC).

In any event, this system will be used and evaluated in the context of the SemEval 2018 project, especially in task 1, task 2 and task 3.

I. Introduction

In recent years, technology has markedly changed the way humans communicate, specifically due to social media like Twitter[1], Facebook[2], WhatsApp[3], among others. Such applications provide users with the ability to express their emotions and opinions not only with words, but also with digital pictures: emojis.

As described by [1], emojis were first used to emphasize conversations before becoming representations of specific emotions or ideas. As of today, emojis in conjunction with slang words and hashtags are used in almost every conversation in social media, particularly in Twitter. [2]. This complex context is arising a new challenge for the NLP community [3] and is attracting more and more research interest. [4]
For example, this phenomenon has been widely observed by [3] [4], and they found that automatic emoji recognition may reveal valuable information in customer service. in such cases, a client is conversing with an automatic chatbot; empowering the algorithm with the ability to detect user's emotion is a big step to the construction of an emotionally intelligence agent capable of generate an empathetic response.

In this report, a methodology to predict emojis and detect emotions is developed for the SemEval 2018 task 1, 2 and 3 [5], in which the goal is to achieve better results than TweetEval leader-board on test set [5].

This methodology is based on the bag-of-words model in conjunction with word embeddings (GloVe), and applying classification algorithms, such as (i) Logistic Regression Classifier (LRC), Support Vector Machine (SVM) and Random Forest Classifier (RFC).

The SemEval 2018-tasks consist in:

- **Task 1**: Emotion Recognition.
  This activity consists of recognizing the emotion evoked by a tweet. The dataset is called "Affects in Tweets" [5]
- **Task 2**: Emoji Prediction.
  This activity consists in, given a tweet, predicting its most likely emoji, which is based on the emoji prediction dataset [5]
- **Task 3**: Irony Detection
  This activity consists of recognizing whether a tweet includes ironic intents or not, based on the dataset [5].

This work is organized as follows: (II) Literature review, which explains some related works. (III) Methodology, which addresses the methodology applied in the 3 tasks [5]. (IV) Results, which presents the outcome of the project. (V) Discussion, which describes the main insights from the created models, (VI) Conclusion, which presents the final remarks as well as future works, and (VII) Plan, which provides a project planning view.

II. Literature review

Emoji meanings are very difficult to examine in detail not only due to the subjective nature of emoji semantics, but also because it is a relative novel research problem [6]. Similarly, the process of analyzing emojis could be very arduous if we consider the coexistence of many possible meanings for one digital symbol[4].

[1]https://twitter.com/.
[2]https://www.facebook.com/.
[3]https://www.whatsapp.com//.

[4]As explained by [6], the image that represents the same emoji can vary. For example, the emoji U+1F40F, which has different rendering by platform in Unicode v11 (up to February 2021): Apple , Google , Twitter , Facebook , Samsung , Windows .

For example, one emoji may be displayed differently across distinct devices or platforms [6], and perhaps, the same digital icon is used in different cultures and communities over the world who interpret the emoji individually [7]. Under these circumstances, it is quite complicated to apply traditional Natural Language Processing (NLP) techniques to analyze the small images.

Nevertheless, there certain research works that have described an emoji, either with a set of semantically close emojis, or by emoji pair co-occurrence in the same tweet [2]. For example, as described in [8], a collection of emoji definitions can enable a semantics-based measure of similarity though vector word embeddings. Similarly, more evidence of this approach can be also explained in [6], in which concludes that modeling emoji semantics via vector representations is a well-defined avenue of work.

In fact, the field of emoji vector evaluation has also experienced a notable growth as of recent. For instance, this contribution [6] mentions the introduction of: EmoTwi50 [9] and EmoSim508 [8], two new datasets with pairwise emoji similarity with human annotations.

Having considered the vector representation approach, it is also reasonable to look at two different proposals in regard to the analyze of emoji semantics. The first perspective includes models trained on Twitter data [9] [10], while the other collaborations have worked in the context of sentiment analysis; for example, in creating sentiment lexicons for emojis. [11] [12]

A. *Word Embeddings*

Originally, pictures were used instead of text as the source of emoji prediction. [13]. However, as stated by [1], several research papers focus on emoji prediction, and most of them use word embeddings in order to do multi-class emoji prediction. [3] [1] [4].

Briefly, word embeddings is nothing but the process of converting text data to numerical vectors, and it is used to capture not only the semantic of the emojis, but also their emotional content. Below are the popular and simple ”vectorization” techniques:

- Bag of words
- TF-IDF
- Word2vec
- GloVe embedding
- ELMO (Embeddings for Language models)

For more information about these methods, go to this post: Word embeddings in 2020 (Towards Data Science)

There appears to be an acceleration in the use of text vectorization techniques. For example, this paper [10] used simple embeddings techniques based on emoji description in the Unicode3 list and obtained 85% of accuracy, while the authors of this paper [14] trained a neural network model on Weibo4 to predict emojis and obtained more than 50% of accuracy in their algorithm. Similarly, this research [15] predicted 20 emojis in million of tweets using Long Short-Term Memory (LSTM) and obtained 65% of F1-Score.

III. Methodology

The methodology applied in the 3 tasks described in the introduction section will consist of two big stages, one based on the ”bag-of-words model” and another based on the word embeddings ”GloVe model” as shown in Figure 1

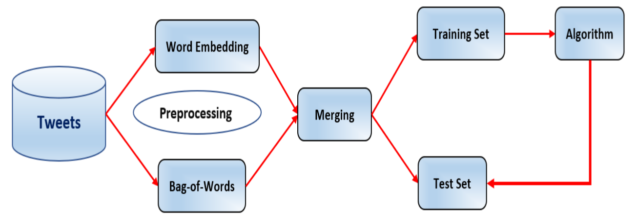

Fig. 1. Model used to solve task 1, 2 and 3

A. *Dataset*

The data for the 3 tasks consists of 500,000 tweets in English for training, 50,000 for trial and 50,000 for test.

As described by [5], the tweets were retrieved with the Twitter APIs from October 2015 to February 2017 and ”geolocalized” in United States. It is important to mention that this dataset includes tweets that contain one emoji out of twenty of the most frequent emoji.

IV. Results

In this section, the obtained results by the created model will be reported in accordance with the evaluation metrics proposed by [5] and mention in the TweetEval Leaberboard. [5]

Results will be divulged in table 1 for diverse models: (i) a model based on word embeddings (WE) and bag-of-words (BoW) with Logistic Regression Classifier (LRC); (ii) a model based on word embeddings (WE) and bag-of-words (BoW) with Support Vector Machine (SVM); a model based on word embeddings (WE) and bag-of-words (BoW) with Random Forest Classifier (RFC).

Each of this models will be replicated in each of the three tasks.

TABLE I
Results of SemEval 2018: task 1, 2 and 3.

	Model	*F1 Score*	*Precision*	*Recall*	*Accuracy*
Task 1	WE+BoW+LRC	?	?	?	?
	WE+BoW+SVM	?	?	?	?
	WE+BoW+RFC	?	?	?	?
Task 2	WE+BoW+LRC	?	?	?	?
	WE+BoW+SVM	?	?	?	?
	WE+BoW+RFC	?	?	?	?
Task 3	WE+BoW+LRC	?	?	?	?
	WE+BoW+SVM	?	?	?	?
	WE+BoW+RFC	?	?	?	?

This table show model’s performances

Finally, a comparison analysis between the 3 ML algorithms will be performed and all the remarks will be reported here. Additionally, this part of the report will describe the interpretation of the results and what they tell about the relative strenghts and weaknesses of the 3 alternative methods when applied to the given data (SemEval 2018 [5]).

V. Discussion

Initially, the models presented in this project will use text vectorization through word embeddings (GloVe) in order to represent each tweet by their polarity intensity. Therefore, the classification process will be made by using decision tree based algorithms with bagging techniques for better generalization to match the objective of macro F1-Score metric [5]. Finally, the implications, challenges and limitations of the results will be discussed here.

VI. Conclusion

In this project, a couple of NLP models based on word embeddings and bag-of-words for the SemEval 2018 task 1, 2 and 3 will be proposed [5]. The Machine Learning Classifiers that will be used to predict emojis and detect emotions are: (i) Logistic Regression Classifier (LRC), (ii) Support Vector Machine (SVM) and (iii) Random Forest Classifier (RFC).

As future works, it could be considered the idea to learn at least 2 more robust embeddings methods, such as Word2Vec or Doc2Vec [16] as well as Deep Neural Networks

VII. Plan

The project schedule is developed by analyzing the activity sequences, durations, resources, and requirements of this research. To illustrate this main approach, a Gantt Chart will be proposed to control the project planning. (See figure 2.0)

References

[1] G. Guibon, M. Ochs, and P. Bellot, "Lis at semeval-2018 task 2: Mixing word embeddings and bag of features for multilingual emoji prediction," in *Proceedings of The 12th International Workshop on Semantic Evaluation*, 2018, pp. 502–506.

[2] F. Barbieri, L. Marujo, P. Karuturi, W. Brendel, and H. Saggion, "Exploring emoji usage and prediction through a temporal variation lens," *arXiv preprint arXiv:1805.00731*, 2018.

[3] C. Huang, A. Trabelsi, and O. R. Zaïane, "Ana at semeval-2019 task 3: Contextual emotion detection in conversations through hierarchical lstms and bert," *arXiv preprint arXiv:1904.00132*, 2019.

[4] P. Du and J.-Y. Nie, "Mutux at semeval-2018 task 1: exploring impacts of context information on emotion detection," in *Proceedings of the 12th international workshop on semantic evaluation*, 2018, pp. 345–349.

[5] F. Barbieri, J. Camacho-Collados, L. Neves, and L. Espinosa-Anke, "Tweeteval: Unified benchmark and comparative evaluation for tweet classification," *arXiv preprint arXiv:2010.12421*, 2020.

[6] F. Barbieri, J. Camacho-Collados, F. Ronzano, L. E. Anke, M. Ballesteros, V. Basile, V. Patti, and H. Saggion, "Semeval 2018 task 2: Multilingual emoji prediction," in *Proceedings of The 12th International Workshop on Semantic Evaluation*, 2018, pp. 24–33.

[7] J. Ge and S. C. Herring, "Communicative functions of emoji sequences on sina weibo," *First Monday*, 2018.

[8] S. Wijeratne, L. Balasuriya, A. Sheth, and D. Doran, "A semantics-based measure of emoji similarity. in 2017 ieee/wic," in *ACM International Conference on Web Intelligence (WI), Leipzig, Germany. ACM, ACM*, 2017.

[9] F. Barbieri, F. Ronzano, and H. Saggion, "What does this emoji mean? a vector space skip-gram model for twitter emojis," in *Calzolari N, Choukri K, Declerck T, et al, editors. Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016); 2016 May 23-28; Portorož, Slovenia. Paris: European Language Resources Association (ELRA); 2016. p. 3967-72.* ELRA (European Language Resources Association), 2016.

[10] B. Eisner, T. Rocktäschel, I. Augenstein, M. Bošnjak, and S. Riedel, "emoji2vec: Learning emoji representations from their description," *arXiv preprint arXiv:1609.08359*, 2016.

[11] D. Rodrigues, M. Prada, R. Gaspar, M. V. Garrido, and D. Lopes, "Lisbon emoji and emoticon database (leed): Norms for emoji and emoticons in seven evaluative dimensions," *Behavior research methods*, vol. 50, no. 1, pp. 392–405, 2018.

[12] M. Kimura and M. Katsurai, "Automatic construction of an emoji sentiment lexicon," in *Proceedings of the 2017 IEEE/ACM international conference on advances in social networks analysis and mining 2017*, 2017, pp. 1033–1036.

[13] S. Cappallo, T. Mensink, and C. G. Snoek, "Image2emoji: Zero-shot emoji prediction for visual media," in *Proceedings of the 23rd ACM international conference on Multimedia*, 2015, pp. 1311–1314.

[14] R. Xie, Z. Liu, R. Yan, and M. Sun, "Neural emoji recommendation in dialogue systems," *arXiv preprint arXiv:1612.04609*, 2016.

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

[16] M. Abdullah and S. Shaikh, "Teamuncc at semeval-2018 task 1: Emotion detection in english and arabic tweets using deep learning," in *Proceedings of the 12th international workshop on semantic evaluation*, 2018, pp. 350–357.

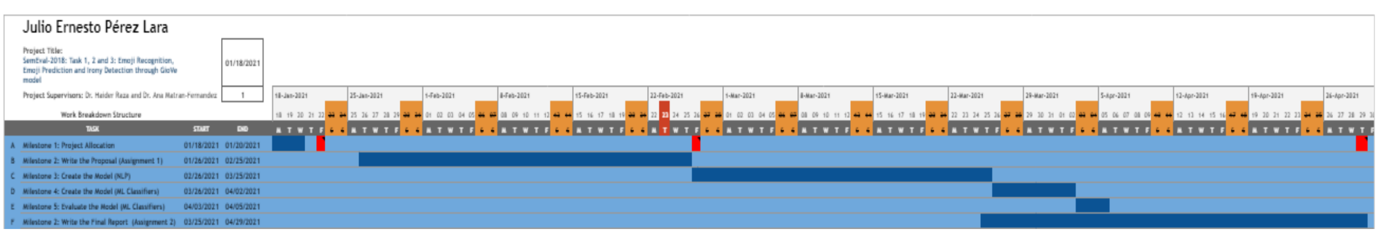

Fig. 2. Project Gantt Chart used to show task displayed against time.
Source: Own elaboration in Google Sheets.
For more details about the Gantt Chart, click here: Live Gantt Chart
Notes:
• On the left side of the sheet, there is a list of project tasks and milestones.
• On the top right, there is a calendar timeline.
• This Gantt Chart has been shared with other Essex students as part of the module CE903-7-SP (Group Project). As a result of this, it is probable that two or more students may also present this template