



EMOJI IDENTIFICATION AND PREDICTION IN HEBREW POLITICAL CORPUS

Chaya Liebeskind

Department of Computer Science,
Jerusalem College of Technology,
Lev Academic Center, Jerusalem,
Israel

liebchaya@gmail.com

ABSTRACT

Aim/Purpose	Any system that aims to address the task of modeling social media communication need to deal with the usage of emojis. Efficient prediction of the most likely emoji given the text of a message may help to improve different NLP tasks.
Background	We explore two tasks: emoji identification and emoji prediction. While emoji prediction is a classification task of predicting the emojis that appear in a given text message, emoji identification is the complementary preceding task of determining if a given text message includes emojis.
Methodology	We adopt a supervised Machine Learning (ML) approach. We compare two text representation approaches, i.e., n-grams and character n-grams and analyze the contribution of additional metadata features to the classification.
Contribution	The task of emoji identification is novel. We extend the definition of the emoji prediction task by allowing to use not only the textual content but also metadata analysis.
Findings	Metadata improve the classification accuracy in the task of emoji identification. In the task of emoji prediction it is better to apply feature selection.
Recommendations for Practitioners	In many of the cases the classifier decision seems fitter to the comment content than the emoji that was chosen by the commentator. The classifier may be useful for emoji suggestion.
Recommendation for Researchers	Explore character-based representations rather than word-based representations in the case of morphologically rich languages.
Impact on Society	Improve the modeling of social media communication.
Future Research	We plan to address the multi-label setting of the emoji prediction task and to investigate the deep learning approach for both of our classification tasks.

Accepted by Executive Review by Eli Cohen | Received: February 14, 2019 | Revised: March 8, March 26, April 2, 2019 | Accepted: June 9, 2019

Cite as: Liebeskind, L. (2019). Emoji identification and prediction in Hebrew political corpus. *Issues in Informing Science and Information Technology*, 16, 343-359. <https://doi.org/10.28945/4372>

(CC BY-NC 4.0) This article is licensed to you under a [Creative Commons Attribution-NonCommercial 4.0 International License](https://creativecommons.org/licenses/by-nc/4.0/). When you copy and redistribute this paper in full or in part, you need to provide proper attribution to it to ensure that others can later locate this work (and to ensure that others do not accuse you of plagiarism). You may (and we encourage you to) adapt, remix, transform, and build upon the material for any non-commercial purposes. This license does not permit you to use this material for commercial purposes.

Keywords emoji prediction, machine learning, social media, supervised learning

INTRODUCTION

The use of emojis in social media has increased in recent years. Cambridge dictionary defines emoji as “a digital image that is added to a message in electronic communication in order to express a particular idea or feeling.” Any algorithm aimed at modeling social media online communication must deal with the use of emojis.

Novak, Smailović, Sluban, and Mozetič (2015) and Barbieri, Ballesteros, and Saggion (2017) claimed that effective forecast of the most probable emoji given the short message text could contribute to the improvement of various Natural Language Processing (NLP) tasks, such as information retrieval, social media content generation, sentiment analysis and emotion recognition. (Barbieri et al., 2017) presented the emoji prediction task in Twitter and demonstrated that their neural network model can exceed humans.

Driven by the promising results of Barbieri et al. (2017), the first shared task on multilingual emoji prediction was proposed by Barbieri, Camacho-Collados, et al. (2018). Earlier results on the idiosyncrasy of emoji use in several languages inspired Barbieri, Camacho-Collados, et al. (2018) to focus on emoji prediction for two languages, English and Spanish. Further researchers followed and examined the task in other languages, including Italian (Ronzano, Barbieri, Wahyu Pamungkas, Patti, & Chiusaroli, 2018), Japanese (Tomihira, Otsuka, Yamashita, & Satoh, 2018), Hindi, Bengali and Telugu (Choudhary, Singh, Bindlish, & Shrivastava, 2018; Choudhary, Singh, Rao, & Shrivastava, 2018). We are interested in Hebrew emoji prediction. Hebrew has a highly productive morphology and has hardly been investigated before.

As the majority of Hebrew speakers are living in Israel, we inspected the statistics of social media usage in Israel in comparison to the usage of social media usage in the United States of America (USA) (see Figure 1, generated by <http://gs.statcounter.com/social-media-stats/>). While in the USA Facebook is leading with 47.82% of the users and Twitter is in the third place with 9.07% of the users, in Israel Facebook is leading with 75.47% of the users but Twitter is in the fourth place with only 4.31% of the users. Thus, even though most of the previous work used Twitter data, we used another social network, Facebook, to collect enough data.

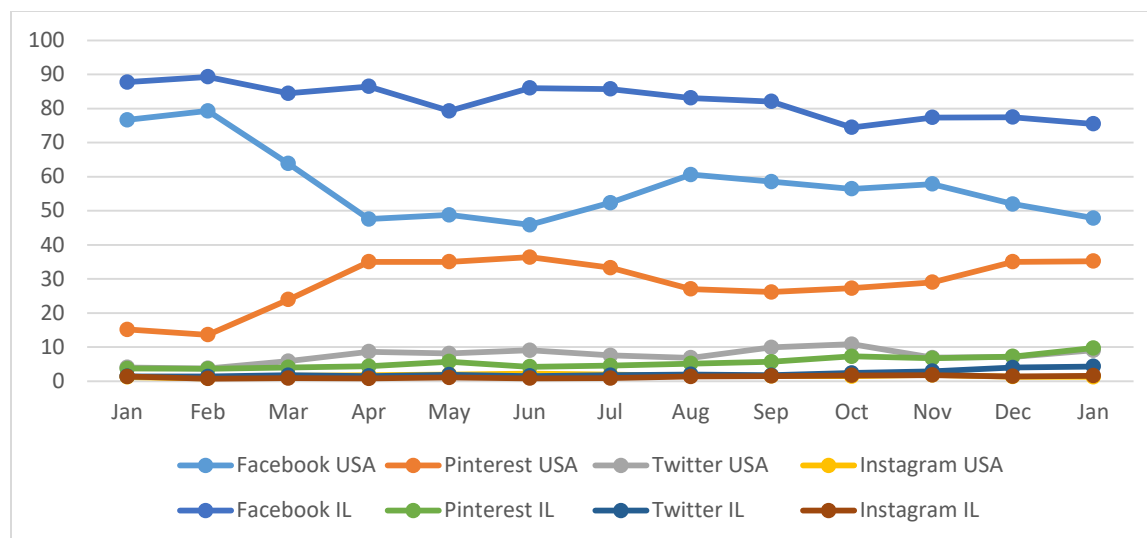


Figure 1: Social Media Stats Israel vs. United State of America, Jan 2018 - Jan 2019

In this paper, we first describe emoji usage in our political Facebook dataset. Then, we explore two tasks: emoji identification and emoji prediction. While emoji prediction is a classification task of pre-

dicting the emojis that appear in a given text message, emoji identification is the complementary preceding task of determining if a given text message includes emojis. To perform these tasks, we adopt a supervised Machine Learning (ML) approach. We compare two text representation approaches, i.e., n-grams and character n-grams and analyze the contribution of additional metadata features to the classification.

Our metadata features include metadata both on the post and on the comment. We demonstrate that, in the task of emoji identification, the metadata features improve the classification accuracy. However, in the task of emoji prediction, the contribution of the metadata features on the comment is minor and the advantage of applying feature selection is greater. We also show that the best character n-grams representation for the emoji prediction task outperforms the FastText baseline (Joulin, Grave, Bojanowski, & Mikolov, 2017), which is often on par with the accuracy of deep learning classifiers.

The rest of this paper is organized as follows: the next section introduces the necessary background on emoji prediction. The third section presents the emoji identification and prediction tasks, datasets, the representation methods, and the metadata features. The fourth section introduces the experimental setting, the experimental results and a deep analysis of the best method. The final section summarizes the main findings and suggests future directions.

BACKGROUND

Recently, modeling emoji semantics has been an area of active research. Distributional embeddings has been the most widely used approach to investigate the meaning of emojis. Two types of data were utilized in order to learn emoji embeddings, a large dataset of text and words describing the emoji. The first type of data was utilized by Barbieri, Ronzano, and Saggion (2016). They trained emoji embeddings from a massive Twitter dataset of over one hundred million English tweets using the skip-gram neural embedding model (Mikolov, Chen, Corrado, & Dean, 2013). They assessed their pre-trained emoji representations on two assignments: a pair similarity and relatedness, and clustering. They demonstrated that the representations enhance accuracy on both assignments. An identical emoji representation technique was followed to compare the meaning and usage of emojis across two Spanish cities: Barcelona and Madrid (Barbieri, Espinosa-Anke, & Saggion, 2016), and across various languages (Barbieri, Kruszewski, Ronzano, & Saggion, 2016).

Pohl, Domin, and Rohs (2017) investigated a similar neural embedding model for the emoji entry task. Since search is a critical problem of emoji entry, emoji keyboards need to be optimized for search. To support users' search, they suggested to place related emojis close to each other. The emoji embeddings enabled them to compute the level of similarity between two emojis. They demonstrated that the model has good performance in recognizing specific relationships between emojis.

The second type of data was utilized by Eisner, Rocktäschel, Augenstein, Bosnjak, and Riedel (2016) who claimed that Barbieri, Ronzano, et al.'s (2016) method can not learn vigorous representations for rare emojis. They used the emoji descriptions in the Unicode emoji standard to straightforwardly estimate the representation of emojis. They compared their representation to the emoji embeddings trained by Barbieri, Ronzano, et al. (2016) on the task of Twitter sentiment analysis and found that their representation generally has a better performance.

Wijeratne, Balasuriya, Sheth, and Doran (2017) also used the second type of data. To improve emoji embedding models, they included more words by using longer emoji definitions. Utilizing the data in EmojiNet (Wijeratne, Balasuriya, Sheth, & Doran, 2016), they used three different forms to represent the meaning of an emoji: emoji descriptions, emoji sense labels, and the emoji sense definitions. They demonstrated that their models beat the past best-performing emoji embedding model of (Eisner et al., 2016) on the sentiment analysis task. Additionally, they introduced a new openly accessible dataset called EmoSim508 (Wijeratne et al., 2016), which assigns human-annotated semantic similarity scores to a set of 508 carefully selected emoji pairs.

Presently, neural networks with word embeddings representations have been used to model the semantics of emojis. The word embeddings representations are generated by either word2vec models (Mikolov et al., 2013) or gradient descent based learning algorithms.

Some variants of Recurrent Neural Network (RNN) based on Long Short-Term Memory (LSTM) networks have been investigated and promising results are reported on the task of emoji prediction. A hierarchical LSTM model was proposed by Xie, Liu, Yan, and Sun (2016) to construct better dialogue representations by encoding the contextual information located in conversations. Their method significantly outperforms other LSTM models and a baseline Logistic Regression classifier with bag-of-words feature representation and tf-idf weights on the task of emoji recommendation in multi-turn dialogue systems.

Bidirectional LSTM (BLSTM) networks (Graves & Schmidhuber, 2005; Hochreiter & Schmidhuber, 1997) were employed by Barbieri et al. (2017) for emoji prediction. They investigated two types of embedding representations: word embeddings and character-based continuous-space vector embedding (Ling et al., 2015). They demonstrated that in this task, the BLSTM networks outperform a baseline of bag-of-words, a baseline based on semantic vectors, and human annotators.

Felbo, Mislove, Søgaard, Rahwan, and Lehmann (2017) used another variant of LSTM networks (Hochreiter & Schmidhuber, 1997; Sutskever, Vinyals, & Le, 2014) to predict sentiment, emotions and irony using pre-trained emoji prediction classifier. Their pre-trained model includes a mechanism of attention (Yang et al., 2016) to increase their sensitivity to individual words during prediction. On 8 benchmark datasets, their model obtained state-of-the-art performance. Based on Felbo et al.'s (2017) network design, Barbieri, Anke, Camacho-Collados, Schockaert, and Saggion (2018) implemented a label-wise attention mechanism suitable for underrepresented emojis. They demonstrated an increase in performance over Felbo et al.'s (2017) model and the efficient FastText (Joulin et al., 2017) classification algorithm.

LSTM networks were also utilized to incorporate temporal information in emoji prediction systems. Barbieri, Marujo, Karuturi, Brendel, and Saggion (2018) showed that some emojis are used differently depending on the time of the year by comparing emoji embeddings trained on a corpus of different seasons. Two types of embedding are extracted by their method: character BLSTM (Ling et al., 2015) and word embeddings. Then the two representations are concatenated (as in Barbieri et al., 2017) and passed on to a word LSTM and word attention units. They demonstrated that their method exceeds state-of-the-art methods.

Neural models have been used to address the multi-label setting of the emoji prediction task as well. An automatic recommendation system based on user message analysis and real emoji usage was developed by Guibon, Ochs, and Bellot (2018). They showed that a Random Forest multi-label classifier with a bag-of-words/characters representation and calculated features outperforms (Barbieri et al., 2017) BLSTM networks.

Wu, Wu, Wu, Huang, and Xie (2018) also focused on multi-label classification and proposed a hierarchical neural model with attention mechanism. The model includes three modules, a character encoder for learning hidden word representations using Convolutional Neural Networks (CNNs), a word encoder for learning sentence representations using a combination of CNN and LSTM, and an emoji classifier for predicting tweet emojis. Their approach is superior to a number of baselines, including K independent bag-of-word Support Vector Machine (SVM) models, CNN as a word encoder, and a hierarchical model with LSTM in both character and word encoders, as well as humans in this task.

A new SemEval task, i.e. the multilingual emoji prediction task, has recently been introduced by Barbieri, Camacho-Collados, et al. (2018). The task was divided into two subtasks dealing with the emoji prediction associated with English and Spanish tweets respectively. The tweets which contain one of 20 most frequently occurring emojis in the Twitter data were selected for each subtask. The

task datasets only contain tweets with a single emoji so that the challenge can be viewed as a single-label classification problem. The task required to predict the emoji using the textual content of the text message exclusively. In total, 49 teams took part in the English subtask and 22 teams attended the Spanish subtask. While many of the participating teams preferred neural architectures, especially LSTMs and CNNs, the most effective system (Çöltekin & Rama, 2018) on both English and Spanish datasets used a SVM classifier with bag-of-n-grams features (both characters and words). The task organizers suggested extending the problem of emoji semantics modeling by incorporating more and more diverse languages.

One step has been taken in this direction. In the context of the Evalita 2018 evaluation campaign (Caselli, Novielli, Viviana Patti, & Rosso, 2018), Ronzano et al. (2018) proposed the shared task also for the Italian language (ITAMoji). Five teams have submitted twelve runs at ITAMoji. In this task, systems that employ neural network architecture generally achieved good performance, particularly when relying on the BLSTM model.

The first attempt to focus on Japanese was the work of Tomihira et al. (2018). They collected Twitter's Japanese tweets and suggested a new model that learns from phrases. They explored multiple models based on CNN and RNN's Encoder-Decoder model. Contrary to Zhao & Zeng (n.d.) who demonstrated that the CNN model has higher classification accuracy than the RNN model, in their case study the Encoder-Decoder model with attention had better performance than the CNN model in accuracy and F1 score.

The study of creating a corpus for multilingual sentiment analysis and emoji prediction in Hindi, Bengali and Telugu (Choudhary, Singh, Rao, et al., 2018) was another step in this direction. They tackled resource-poor languages because such discourse is available on resource-rich languages, such as English and Spanish, while resource-poor languages are largely disregarded. A twin BLSTM RNNs model to learn emoji-based representations of resource-poor languages was introduced by Choudhary, Singh, Bindlish, et al. (2018). They jointly train the resource-poor languages (Hindi and Telugu) with resource-rich languages (English and Spanish) in a common emoji space by using a similarity metric based on the emojis present in sentences from both languages. They showed that their model exceeds the state-of-the-art emoji prediction approaches based on distributional semantics, semantic rules, lexicon lists and deep neural network representations without shared parameters.

We also study emoji prediction in a resource-poor language, Hebrew, in this research. Recently, Liebeskind and Liebeskind (2019) made a first attempt to focus on Hebrew. They explored n-grams representations, character n-grams representations and four dimension reduction methods for emoji prediction. They showed that the common Word Embedding dimension reduction method is not optimal and that the character n-grams representations outperform all the other representations. In this research, we extend their prediction task by allowing meta-data analysis. In addition, we investigate the novel complementary preceding task of emoji identification.

EMOJI IDENTIFICATION AND PREDICTION

TASKS

Emoji Identification

Emoji identification is a binary classification task of determining if a given text message includes emojis by relying on either textual content or meta-data analysis.

Emoji Prediction






Emoji prediction is defined as a classification task of predicting the emojis that appear in a given text message by relying exclusively on the textual content of that message (Barbieri et al., 2017; Barbieri,

Camacho-Collados, et al., 2018; Ronzano et al., 2018). We extend this definition by allowing to use not only the textual content but also meta-data analysis.

In practice, we remove emojis from the messages' text and use them as labels. This task can therefore be seen as a problem of multi-label classification.

Following previous works on emoji prediction (Barbieri et al., 2017; Ronzano et al., 2018), only messages with a single emoji were selected, so that the challenge can be directed as a single-label classification problem, detailed examples from our Hebrew dataset are shown in Table 1 (to facilitate readability, we used a transliteration of Hebrew using Roman characters; the letters used, in Hebrew lexicographic order, are abgdhwzxtiklmns'pcqršt).

Table 1. Examples from our Hebrew dataset.

#	Comment	Label
1	חזק וברוך xzq wbrwk be strong and blessed	
2	אל תפסיק לחלום al tpsiq lxlwm do not stop dreaming	
3	לחיים lxim cheers!	
4	כאלה מקסימים יחד kalh mqsimim ixh these are lovely together	
5	תברכו tbwrkw you will be blessed	

DATASETS

We have constructed two datasets for the tasks of emoji identification and prediction using a political dataset by Liebeskind, Liebeskind, and HaCohen-Kerner (2017), Liebeskind and Nahon (2018), and Liebeskind, Nahon, HaCohen-Kerner, and Manor (2017). All posts of Members of Knesset (MKs) between 2014-2016 ($n=130$ MKs, $m=33,537$ posts) have been downloaded via Facebook Graph API. The data included also the comments to these posts ($n=5.37$ M comments posted by 702,396 commentators).

We have analyzed the commentators' emoji use. There are 786 types of emojis and 98,865 of the comments include at least one of them. There are 50,243 comments with a single emoji. Emojis are used by 41,789 of the commentators.

For the task of emoji identification, we used 80,276 comments which include at least one emoji and less than 6 emojis as positive examples. We extracted negative examples by randomly selecting an equal number of comments without emojis. For each class, we randomly selected 90% of the data for training, and the remaining 10% for the test set. Our training and test sets include 160,552 and 16,057, respectively.

As described in the previous section, the prediction task dataset should contain only comments with a single emoji. However, we include comments with multiple emoji appearances, such as כל הכבוד לגיבור שלנו 🙌👏👏👏 (kl hkbwd lgibwr šlnw – well done our hero), to increase the number of comments. After limiting our dataset to comments that contain only one **type** of emojis, we resulted with 78,147 comments that include 593 emoji types.

For the emoji prediction task, we selected the comments that include one of the twenty most frequently occurring emojis in the dataset we have described. For each emoji, we randomly selected 90% of the data for training, and the remaining 10% for the test set. Table 2 shows the emoji distribution of the comments in our dataset.

Table 2. Emoji distribution of the comments in our dataset.

#	Emoji	Training set	Test set
1	👍	11282	1254
2	❤️	5190	577
3	👏	3785	421
4	🙏	3355	373
5	😂	2797	311
6	😊	2361	263
7	❤️	1692	188
8	🌹	1608	179
9	😬	1456	162
10	😘	1307	146
11	👉	1199	134
12	😏	1191	133
13	👑	1152	128
14	😊	1099	123
15	💙	1075	120
16	😊	1070	119
17	😄	1031	115
18	😘	1002	112
19	✌️	895	100
20	😞	864	97

METHOD

In this research, we adopt a supervised Machine Learning (ML) approach for emoji identification and prediction. The first step in a classifier training is to determine which text characteristics are relevant and how those features are coded.

First, we explored two types of text representations:

1. N-grams representation - An n-gram is a contiguous sequence of n words. Each of the n-grams in the comment is considered as a feature. The score of the feature is the n-gram tf-idf. For document of any length, the n-grams representation is a high-dimensional sparse representation. But for short texts, as for comments, where most words occur only one time, the sparsity problem is much more critical.
2. Character n-grams representation - Character n-grams are strings of length n. For example, the character 3-grams of the string “identification” would be: “ide”, “den”, “ent”, “nti”, “tif”, “if”, “fic”, “ica”, “cat”, “ati”, “tio”, and “ion”. Each of the character n-grams of the comment is considered as a feature and scored by its tf-idf. Because there are far fewer character combinations than n-gram combinations, character n-grams representation overcomes the problem of sparse data that arises when using n-grams representation. It nevertheless produces a considerably large feature set. Due to the tendency of noise and incorrect spellings to have less impact on substring patterns than on n-gram patterns, character n-gram features can be quite effective for short informal text classification (Aisopos, Papadakis, Tserpes, & Varvarigou, 2012; Raaijmakers & Kraaij, 2008).

Since previous works on short Hebrew text classification showed that lemmatization using a Part-of-Speech (PoS) tagger do not improve the performance (Liebeskind, Nahon, et al., 2017; Mughaz, Fuchs, & Bouhnik, 2018), we did not lemmatize the comments.

Next, we detail how the special metadata of Facebook is encoded as features. We encoded metadata on both the post and the comment.

1. Metadata on the post: number of characters, number of words, normalized number of punctuations, normalized number of emojis. Using Facebook API, we extracted additional four Facebook depended features: the number of “reactions” that the post got, i.e., like, love, haha, wow, angry, and sad, the number of “shares”, the number of comments that the post got, and the post type, i.e., photo, link, status, video, and event. We also encoded two features on the writer of the post: MK identifier and MK gender.
2. Metadata on the comment: number of characters, number of words, temporal information, i.e., hour, day, and month of the comment publication. Following (Liebeskind, Nahon, et al., 2017), we extracted additional three Facebook depended features: the number of “likes” that the comment got, the number of comments on the comment, and a Boolean feature, which indicates whether the commentator also “liked” the status. Another two features that we defined are the number of occurrences of the MK writer of the post and the number of occurrences of other MKs, either aliens or rivals of the post writer.

For the identification task, we combined both types of metadata, assuming that some of the metadata might trigger the commentator to include an emoji in the comment. For example, post with an emoji or popular post with many “likes”. For the prediction task, we combined only the metadata on the comment.

EVALUATION

EVALUATION SETTING

While the dataset of the positive and negative comments was used for the emoji identification task, the emoji prediction was performed on the comments that include one of the twenty emojis that occur most frequently in the Facebook data we collected.

For classification, Scikit-learn (<https://scikit-learn.org/stable/index.html>) machine learning python module (Pedregosa et al., 2011) was used. We used the training set for learning the model and the test set for estimating the classification performance.

In our experiments, four commonly used classification measures were used to compare the performance of our algorithms: precision, recall, F1, and accuracy. The scores are macro-averaged; we first calculate the measure for each label/emoji and then take the average of these scores.

RESULTS

In our experiments, we combined the features in a supervised classification framework using the Logistic Regression ML methods. Since this classifier achieves good performance with reasonable run time, it was empirically selected out of eight classifiers, i.e., Bernulli Naive Bayes, Decision Tree, Logistic Regression, Random Forest, Multilayered Perceptron, Support Vector Classification, Adaboost, and Bagging.

Emoji identification

A large feature set was generated by the text representations. Therefore, we filtered out features that have less than 15 appearances or appear in more than 5% of the comments in our dataset, limiting the total number of features to 17,000. We also removed comments without any text.

Table 3 presents the results of the two types of text representation: n-grams representations (unigram, bigram, and trigram) and character n-grams (character 2-grams, 3-grams, 4-grams, and 5-grams) for the emoji identification task. The unigram representation significantly outperforms all the other n-grams representations. However, the best representation is the character 3-grams.

Table 3. Logistic Regression results for the n-grams and character n-grams representations for the emoji identification task.

Representation	Precision	Recall	F1	Accuracy
Char 2-grams	0.6623	0.6613	0.6614	0.6629
Char 3-grams	0.6896	0.6873	0.6874	0.6894
Char 4-grams	0.6867	0.6843	0.6844	0.6865
Char 5-grams	0.6849	0.6825	0.6826	0.6847
Unigrams	0.6843	0.6804	0.6802	0.6832
Bigrams	0.6522	0.6391	0.6346	0.6450
Trigrams	0.6381	0.5736	0.5252	0.5876

Next, we chose the best text representation and combined it with the metadata feature set. Table 4 presents the results of the classifier using the metadata feature set alone and in combination with the best text representation. In addition, we tried to filter out non-relevant features using the chi-square

feature selection method. The results of the combined feature set after the feature selection are also presented in the table.

Table 4. Logistic Regression results for the combined feature set.

Representation	Precision	Recall	F1	Accuracy
Char 3-grams	0.6896	0.6873	0.6874	0.6894
Metadata feature set	0.25	0.5	0.3333	0.5
Char 3-grams + feature selection	0.6851	0.6823	0.6823	0.6846
Combined feature set	0.7001	0.6984	0.6987	0.7001
Combined feature set + feature selection	0.6959	0.6941	0.6943	0.6959

The combined representation significantly increases the accuracy of the character 3-grams representation. The feature selection method does not improve the character 3-grams representation. Moreover, when applied on the combined feature set, it decreases the combined representation performance.

We note that with the metadata feature set the Logistic Regression classifier classified all the comments as negative. The Bernulli Naive Bayes achieved 0.6041 accuracy and 0.6018 F1. However, it did not perform well with the other representations.

Analysis

We used the information obtained by the chi-square feature selection method to better understand which features have more influence on the classification accuracy. We selected 50% of the features from the representation that was generated by the metadata feature set. Out of the 10 selected features, 5 were encoded from information on the post, i.e., type, normalized number of punctuations, number of characters, MK identifier, and 5 from information on the comment, i.e., publication day of the week, number of comments, number of “likes”, number of characters, number of occurrences of the MK writer of the post.

In Table 5, we complete our analysis by presenting the confusion matrix of the best classification results. Each column of the matrix represents the instances in a predicted class while each row represents the instances in an actual class. Table 5 shows that most of the classification errors were due to incorrect classification of positive comments as negative (2,536) and vice versa (2,087).

Table 5. Confusion matrix of the combined feature set.

	Positive	Negative
Positive	4,857	2,536
Negative	2,087	5,940

Emoji prediction

A large feature set was generated by the text representations. Thus, we filtered out features that have less than 8 appearances or appear in more than 5% of the comments in our dataset. We also removed fake comments with more than 15 emojis.

Table 6 presents the results of the two types of text representation: n-grams representations (unigram, bigram, and trigram) and character n-grams (character 2-grams, 3-grams, 4-grams, and 5-grams) for the emoji prediction task.

Table 6. Logistic Regression results for the n-grams and character n-grams representations for the emoji prediction task.

Representation	Precision	Recall	F1	Accuracy
Char 2-grams	0.3519	0.1647	0.1729	0.3604
Char 3-grams	0.4661	0.1910	0.2121	0.3676
Char 4-grams	0.4802	0.1854	0.2058	0.3665
Char 5-grams	0.4880	0.1805	0.2013	0.3588
Unigrams	0.4401	0.1725	0.1896	0.3525
Bigrams	0.4275	0.1364	0.1518	0.3055
Trigrams	0.3808	0.1027	0.1065	0.2689

The unigram representation significantly outperforms all the other n-grams representations. However, the best representations are the character 3-grams and character 4-grams. The F1 advantage of the character 3-grams representation over the unigram representation is statistically significant at the 0.01 level according to the two-sided Wilcoxon signed-rank test (Wilcoxon, 1945). This discovery is interesting since there are three letters in the Hebrew root, which is the most basic form of the word, to which other parts, such as affixes, can be added.

To improve the results, we combined the comment metadata feature set with the character 3-grams and character 4-grams representations. Additionally, we tried to filter out non-relevant features using the chi-square feature selection method. Table 7 shows the results of the improved representations.

Table 7. Logistic Regression results for the improved character n-grams representations.

Representation	Precision	Recall	F1	Accuracy
Char 3-grams	0.4661	0.1910	0.2121	0.3676
Char 3-grams - Combined	0.4030	0.1835	0.1909	0.3782
Char 3-grams - Feature Selection	0.4517	0.1957	0.2181	0.3859
Char 4-grams	0.4802	0.1854	0.2058	0.3665
Char 4-grams - Combined	0.3845	0.1751	0.1823	0.3719
Char 4-grams - Feature Selection	0.4914	0.1872	0.2099	0.3773

The combined representation slightly increases the accuracy of the character 3-grams representation but decrease its F1. The feature selection method removes the metadata features, yet it outperforms the character 3-grams representation. The same is true for the character 4-grams representation

We also compared our results to a common FastText baseline (Joulin et al., 2017), which is often in line with the accuracy of deep learning classifiers. We obtained an accuracy of 0.3701, and precision, recall and a F1 scores of 0.2217, 0.1494, 0.1483, respectively. We observed that the character 3-grams representation outperforms the FastText baseline and its F1 advantage is statistically significant at the 0.01 level.

In Table 8, we show the precision, recall, and F1 measure of the character 3-grams with feature selection representation for each emoji, sorted by the emoji frequency. With the exception of the 🍷 emoji, frequent emojis have higher F1 scores. Other emojis with good performance are: 👑, 💙, 😊 and 😞. The accuracy is generally higher than the recall. Only part of the emojis features were learned by the classifier.

Table 8. The precision, recall, and F1 measure of the character 3-grams with feature selection representation for each emoji.

#	Emoji	Precision	Recall	F1
1	👍	0.32	0.86	0.46
2	❤️	0.63	0.72	0.67
3	👏	0.39	0.12	0.19
4	🙏	0.43	0.4	0.41
5	😂	0.39	0.33	0.36
6	😊	0.33	0.1	0.15
7	❤️	0.58	0.11	0.18
8	🌹	0.27	0.08	0.12
9	😊	0.04	0.01	0.02
10	😍	0.71	0.04	0.07
11	👉	0.48	0.11	0.17
12	😜	0.22	0.03	0.05
13	👑	0.57	0.34	0.43
14	😊	1	0.06	0.11
15	💙	0.65	0.2	0.3
16	😊	0	0	0
17	😊	1	0.16	0.28
18	😘	0.25	0.01	0.02
19	👉	0.32	0.06	0.1
20	😞	0.45	0.17	0.25

Analysis

To better understand the challenges of the emoji prediction task, we analyzed the classification errors of the character 3-grams with feature selection representation. In Figure 2, we present the classification confusion matrix.

Most classification errors occurred because the comments were wrongly classified into the most common emoji 👍. In order to better understand the reasons for the classification errors, in Table 9, we present the meaning of some emojis according to the Emojipedia (<https://emojipedia.org/>). Example #1 and #2 share a similar meaning. Therefore, it is not surprising that 75% of the comments with 🙌 were classified as 👍.

The meaning of examples #3 and #4 may be related in the context of sickness. A commentator might feel sadness and pray for a quick recovery. Thus, 12% of the comments with 😞 were misclassified as 🙏.

The various heart emojis express a closely related meaning as demonstrated in Examples #5, #6, and #7. It is hard to tell the differences between them. The most common heart emoji is ❤️. 18% of the comments with 💙 and 17% of the comments with ❤️ were misclassified as ❤️. Classification errors were not common in the opposite direction.

Finally, we analyzed 150 misclassified comments and observed that in many of the cases both the classifier decision and the emoji that was chosen by the commentator seem to fit well the comment content. For example, the emojis: 🌹 and 😊 for the comment **תג שג** (xg šmx - happy holiday). The results of recent studies have explained this in inconsistent interpretation of Emoji characters (Barbieri et al., 2017; Miller, Kluver, Thebault-Spieker, Terveen, & Hecht, 2017).

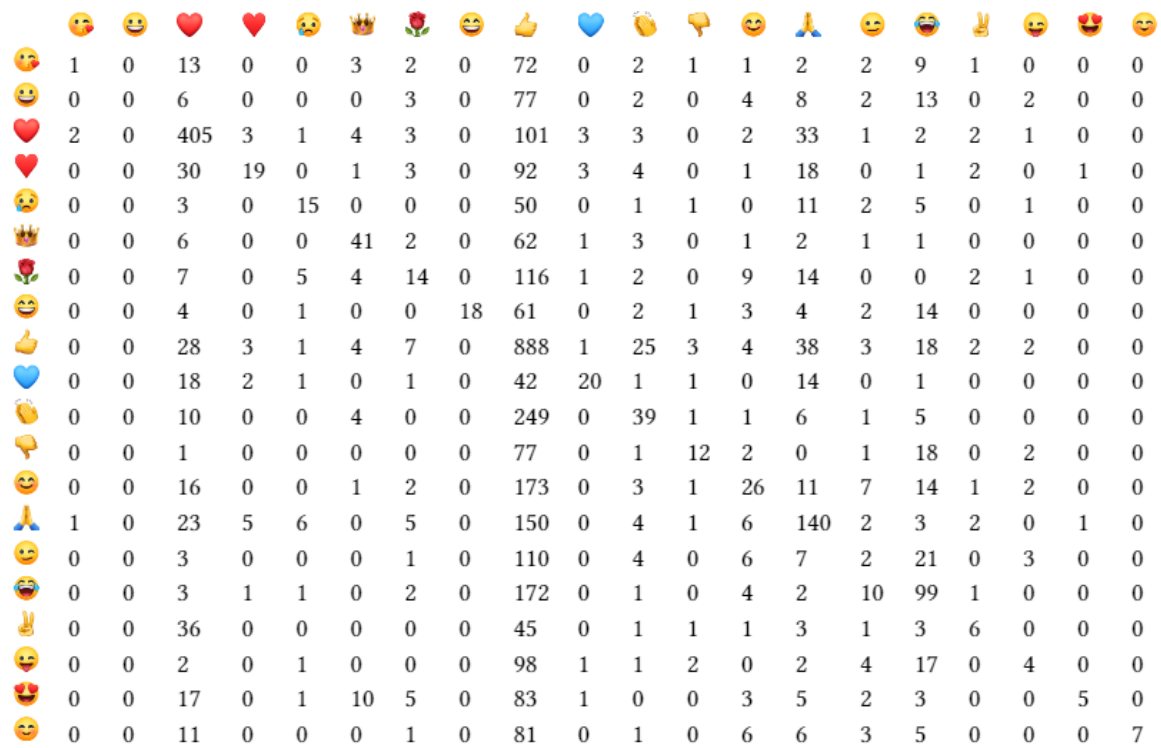









Figure 2. Confusion matrix of the character 3-grams with feature selection representation

Table 9. The meaning of some emojis according to the Emojipedia.

#	Emoji	Emoji name	Emoji meaning
1		Thumbs Up	A thumbs-up gesture indicating approval.
2		Clapping Hands	Two hands clapping emoji, which when used multiple times can be used as a round of applause.
3		Folded Hands	Two hands placed firmly together, meaning <i>please</i> or <i>thank you</i> in Japanese culture. A common alternative use for this emoji is for prayer, using the same gesture as praying hands.
4		Crying Face	A yellow face with raised eyebrows and a slight frown, shedding a single, blue tear from one eye down its cheek. May convey a moderate degree of sadness or pain, usually less intensely than 😭 Loudly Crying Face.
5		Red Heart	A classic love heart emoji, used for expressions of love. Displayed in various shades of red on most platforms. A similar emoji exists for the heart suit in a deck of playing cards.
6		Heart Suit	A heart symbol emoji, which is used in card games for the hearts suit. Generally shown in red, despite the name.
7		Blue Heart	A blue heart can symbolize a deep and stable love. Trust, harmony, peace and loyalty (from the description of https://www.emoji.com/view/emoji/36/symbols/blue-heart)

In this paper, we simplified the setting of prediction to classification with a single label. In real life, however, it is a task of multi-label prediction.

CONCLUSIONS AND FUTURE WORK

We explored two tasks: emoji identification and emoji prediction as a single-label classification problem. By applying ML methods, we investigated two text representation approaches, i.e., n-grams and character n-grams. We also explored the contribution of additional metadata features on both the post and the comment to the classification.

We showed that while the metadata features improve the classification accuracy in the task of emoji identification, in the task of emoji prediction, the contribution of the metadata is minor. We demonstrated that in the case of emoji prediction it is better to apply feature selection and showed that the best character n-grams representation for the emoji prediction task significantly outperforms the efficient FastText algorithm.

In practice, a text message can contain more than one emoji. So we plan to deal with the emoji prediction task multi-label setting. In addition, we plan to investigate the deep learning approach for our classification tasks as it has been shown to be effective for the emoji prediction task (Barbieri et al., 2017; Choudhary, Singh, Bindlish, et al., 2018; Felbo et al., 2017; Tomihira et al., 2018; Wu et al., 2018).

REFERENCES

- Aisopos, F., Papadakis, G., Tserpes, K., & Varvarigou, T. (2012). Content vs. context for sentiment analysis: A comparative analysis over microblogs. In *Proceedings of the 23rd ACM Conference on Hypertext and Social Media* (pp. 187–196). ACM. <https://doi.org/10.1145/2309996.2310028>
- Barbieri, F., Anke, L. E., Camacho-Collados, J., Schockaert, S., & Saggion, H. (2018). Interpretable emoji prediction via label-wise attention LSTMs. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (pp. 4766–4771).
- Barbieri, F., Ballesteros, M., & Saggion, H. (2017). Are emojis predictable? In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers* (pp. 105–111). <https://doi.org/10.18653/v1/e17-2017>
- Barbieri, F., Camacho-Collados, J., Ronzano, F., Anke, L. E., Ballesteros, M., Basile, V., ... Saggion, H. (2018). SemEval 2018 Task 2: Multilingual emoji prediction. In *Proceedings of The 12th International Workshop on Semantic Evaluation* (pp. 24–33). <https://doi.org/10.18653/v1/s18-1003>
- Barbieri, F., Espinosa-Anke, L., & Saggion, H. (2016). Revealing patterns of Twitter emoji usage in Barcelona and Madrid. *Frontiers in Artificial Intelligence and Applications. 2016;(Artificial Intelligence Research and Development) 288*, 239-244.
- Barbieri, F., Kruszewski, G., Ronzano, F., & Saggion, H. (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* (pp. 531–535). ACM. <https://doi.org/10.1145/2964284.2967278>
- Barbieri, F., Marujo, L., Karuturi, P., Brendel, W., & Saggion, H. (2018). Exploring emoji usage and prediction through a temporal variation lens. *arXiv Preprint arXiv:1805.00731*.
- Barbieri, F., Ronzano, F., & Saggion, H. (2016). What does this emoji mean? A vector space skip-gram model for twitter emojis. In *LREC*.
- Caselli, T., Novielli, N., Viviana Patti, & Rosso, P. (2018). EVALITA 2018: Overview of the 6th evaluation campaign of natural language processing and speech tools for Italian. In Tommaso Caselli Nicole Novielli, Viviana Patti & P. Rosso (Eds.), *Proceedings of Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018)* (Vol. 2263, pp. 1–6). <https://doi.org/10.4000/books.aaccademia.1924>
- Choudhary, N., Singh, R., Bindlish, I., & Shrivastava, M. (2018). Contrastive learning of emoji-based representations for resource-poor languages. *arXiv Preprint arXiv:1804.01855*.
- Choudhary, N., Singh, R., Rao, V. A., & Shrivastava, M. (2018). Twitter corpus of resource-scarce languages for sentiment analysis and multilingual emoji prediction. In *Proceedings of the 27th International Conference on Computational Linguistics* (pp. 1570–1577).
- Çöltekin, Ç., & Rama, T. (2018). Tübingen-Oslo at SemEval-2018 Task 2: SVMs perform better than RNNs in emoji prediction. In *Proceedings of The 12th International Workshop on Semantic Evaluation* (pp. 34–38). <https://doi.org/10.18653/v1/s18-1004>
- Eisner, B., Rocktäschel, T., Augenstein, I., Bosnjak, M., & Riedel, S. (2016). emoji2vec: Learning *Emoji Representations from their Description*. In *Proceedings of The Fourth International Workshop on Natural Language Processing for Social Media* (pp. 48–54). <https://doi.org/10.18653/v1/w16-6208>
- Felbo, B., Mislove, A., Sogaard, A., Rahwan, I., & Lehmann, S. (2017). Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* (pp. 1615–1625). <https://doi.org/10.18653/v1/d17-1169>

- Graves, A., & Schmidhuber, J. (2005). Framewise phoneme classification with bidirectional LSTM networks. In *Neural Networks, 2005. IJCNN'05. Proceedings. 2005 IEEE International Joint conference on* (Vol. 4, pp. 2047–2052). IEEE. <https://doi.org/10.1109/ijcnn.2005.1556215>
- Guibon, G., Ochs, M., & Bellot, P. (2018). Emoji recommendation in private instant messages. In *Proceedings of the 33rd Annual ACM Symposium on Applied Computing* (pp. 1821–1823). ACM. <https://doi.org/10.1145/3167132.3167430>
- Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735–1780. <https://doi.org/10.1162/neco.1997.9.8.1735>
- Joulin, A., Grave, E., Bojanowski, P., & Mikolov, T. (2017). Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers* (Vol. 2, pp. 427–431). <https://doi.org/10.18653/v1/e17-2068>
- Liebeskind, C., & Liebeskind, S. (2019). Emoji prediction for Hebrew political domain. In *Proceedings of the 2nd International Workshop on Emoji Understanding and Applications in Social Media (Emoji2019)*. San Francisco, CA.
- Liebeskind, C., Liebeskind, S., & HaCohen-Kerner, Y. (2017). Comment relevance classification in Facebook. In *18th International Conference on Computational Linguistics and Intelligent Text Processing*. Budapest, Hungary. https://doi.org/10.1007/978-3-319-77116-8_18
- Liebeskind, C., & Nahon, K. (2018). Challenges in applying machine learning methods: Studying political interactions on social networks. In J. Szymański & Y. Velegrakis (Eds.), *Semantic keyword-based search on structured data sources* (pp. 136–141). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-74497-1_13
- Liebeskind, C., Nahon, K., HaCohen-Kerner, Y., & Manor, Y. (2017). Comparing *Sentiment Analysis Models to Classify Attitudes of Political Comments on Facebook*. *Polibits*, 55, 17–23.
- Ling, W., Dyer, C., Black, A. W., Trancoso, I., Fernandez, R., Amir, S., ... Luis, T. (2015). Finding function in form: Compositional character models for open vocabulary word representation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* (pp. 1520–1530). <https://doi.org/10.18653/v1/d15-1176>
- Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv Preprint arXiv:1301.3781*.
- Miller, H., Kluver, D., Thebault-Spieker, J., Terveen, L., & Hecht, B. (2017). Understanding emoji ambiguity in context: The role of text in emoji-related miscommunication. In *Eleventh International AAAI Conference on Web and Social Media*.
- Mughaz, D., Fuchs, T., & Bouhnik, D. (2018). Automatic opinion extraction from short Hebrew texts using machine learning techniques. *Computación Y Sistemas*, 22(4). <https://doi.org/10.13053/cys-22-4-3071>
- Novak, P. K., Smailović, J., Sluban, B., & Mozetič, I. (2015). Sentiment of emojis. *PLoS One*, 10(12), e0144296. <https://doi.org/10.1371/journal.pone.0144296>
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... others. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12(Oct), 2825–2830. <https://doi.org/10.3389/jmlr.2014.00014>
- Pohl, H., Domin, C., & Rohs, M. (2017). Beyond just text: Semantic emoji similarity modeling to support expressive communication. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 24(1), 6. <https://doi.org/10.1145/3039685>
- Raaïmakers, S., & Kraaij, W. (2008). A shallow approach to subjectivity classification. In *ICWSM*.
- Ronzano, F., Barbieri, F., Wahyu Pamungkas, E., Patti, V., Chiusaroli, F., & others. (2018). Overview of the EVALITA 2018 Italian Emoji Prediction (ITAMoji) Task. In *6th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop, EVALITA 2018* (Vol. 2263, pp. 1–9). CEUR-WS. <https://doi.org/10.4000/books.aaccademia.1924>
- Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems* (pp. 3104–3112).

- Tomihira, T., Otsuka, A., Yamashita, A., & Satoh, T. (2018). What does your tweet emotion mean? Neural emoji prediction for sentiment analysis. In *Proceedings of the 20th International Conference on Information Integration and Web-based Applications & Services* (pp. 289–296). ACM. <https://doi.org/10.1145/3282373.3282406>
- Wijeratne, S., Balasuriya, L., Sheth, A., & Doran, D. (2016). Emojinet: Building a machine readable sense inventory for emoji. In *International Conference on Social Informatics* (pp. 527–541). Springer. https://doi.org/10.1007/978-3-319-47880-7_33
- Wijeratne, S., Balasuriya, L., Sheth, A., & Doran, D. (2017). A semantics-based measure of emoji similarity. *arXiv Preprint arXiv:1707.04653*. <https://doi.org/10.1145/3106426.3106490>
- Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6), 80–83. <https://doi.org/10.2307/3001968>
- Wu, C., Wu, F., Wu, S., Huang, Y., & Xie, X. (2018). Tweet emoji prediction using hierarchical model with attention. In *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers* (pp. 1337–1344). ACM. <https://doi.org/10.1145/3267305.3274181>
- Xie, R., Liu, Z., Yan, R., & Sun, M. (2016). Neural emoji recommendation in dialogue systems. *arXiv Preprint arXiv:1612.04609*.
- Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., & Hovy, E. (2016). Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 1480–1489). <https://doi.org/10.18653/v1/n16-1174>
- Zhao, L., & Zeng, C. (n.d.). *Using neural networks to predict emoji usage from twitter data*. CA.

BIOGRAPHY

Dr. Chaya Liebeskind is a lecturer and researcher in the Department of Computer Science at the Jerusalem College of Technology. Her research interests span both Natural Language Processing and data mining. Much of her recent work has been on applying Machine Learning methods to study political interactions on social networks.