Computer Speech and Language 73 (2022) 101330

Contents lists available at ScienceDirect

Computer Speech & Language

journal homepage: www.elsevier.com/locate/csl

A multi-label emoji classification method using balanced pointwise mutual information-based feature selection

Zahra Ahanin[a], Maizatul Akmar Ismail[a,*]

[a] *Department of Information Systems, Faculty of Computer Science and Information Technology, University of Malaya, Kuala Lumpur, 50603 Malaysia*

ARTICLE INFO

Keywords:
Multi-label classification
Emotion classification
Emoji classification
Emoji lexicon
Sentiment analysis
Pointwise Mutual Information
Natural Language Processing

ABSTRACT

The availability of social media such as twitter allows users to express their feeling, emotions and opinions toward a topic. Emojis are graphic symbols that are regarded as the new generation of emoticons and an effective way of conveying feelings and emotions in social media. With the surging popularity of Emojis, the researchers in the area of Emotion Classification strive to understand the emotion correlated to each Emoji. Two of the most the successful approaches in emoji analysis rely on: 1) official Unicode description and 2) manually built emoji lexicons. Since the use of emoji is socially determined, the former approach is not aligned with intended semantic and usage, which leads researchers to opt for emoji lexicons. To overcome problem of lexicon-based approach, we proposed a method to classify emojis automatically. Therefore, we present a modified Pointwise Mutual Information (PMI) method, called Balanced Pointwise Mutual Information-Based (B-PMI), to develop a balanced weighted emoji classification based on the semantic similarity. Further, deep neural network is used to represent emoji in vector form (emoji embedding) to extend the pre-trained word embeddings. We carefully evaluated the proposed method in multiple twitter datasets that are employed in sentiment and emotion classification using machine learning (ML) and deep learning (DL) approaches. In both approaches, extending word embedding with the proposed emoji embedding improved results. The DL-based approach achieved the highest f1-score of 70.01% for sentiment classification, and accuracy score of 56.36% for emotion classification. ML-based approach obtained accuracy score of 52.17% in emotion classification.

1. Introduction

Social media are web-based online tools that are now an important part of people’s lives, enabling them to use various platforms to communicate and share their personal opinion on a variety of topics. The shared information not only conveys literal information, but also shows one’s perception or emotional attitudes towards the information. Thus, evaluating the content on social media platforms is essential to understand peoples’ emotions (El-Naggar et al., 2017). The availability of large amount of user-generated textual data by various users in diverse types of social media (e.g., forum, weblogs, and social networks) makes the process of recognizing emotions and extracting logical emotional patterns from such unstructured data a critical task to perform (Al-Moslmi et al., 2015; Dashtipour et al., 2016; Wikarsa and Thahir, 2016). Sentiment analysis is used to overcome this problem. Sentiment analysis is the computational

* Corresponding author.
E-mail address: maizatul@um.edu.my (M.A. Ismail).

https://doi.org/10.1016/j.csl.2021.101330
Received 15 February 2021; Received in revised form 2 September 2021; Accepted 20 November 2021
Available online 27 November 2021

study of people's attitudes, emotions, and opinions in regard to different topic or events (Al-Moslmi et al., 2017). Sentiment analysis is the process of detecting polarity (negative, positive) in text, which is too general for some tasks such as decision making (Feldman, 2013; Saif et al., 2012). Thus, the efforts have shifted toward determining sophisticated and more fine-grained affective feelings, such as emotions (happy, sad, joy, etc.) in text, which is known as emotion analysis. The importance of emotional state on human communication, decision making, and social behavior coupled with the complexity in expressing and discerning emotion in language, lead the researchers in emotion analysis field to develop techniques to analyze and understand people's state of mind, emotion and feelings (Gambino and Calvo, 2019; Zhang et al., 2018). Twitter as one of the biggest social network platforms is enriched with diverse opinions and emotions that allow users to add expression in form of text, hashtag, emoji or gif. Such forms of expressions are shown to be significantly more engaging than just text (Bakhshi et al., 2016). Beside hashtag, which has shown to be indicative of emotion in tweet (Ahmad et al., 2019), using Emojis are an effective way of conveying feelings and emotions in social media. They are encoded in Unicode and are incorporated into Unicode Standard which indicates the possible long-time usage compared to application-specific smileys or stickers. According to the Unicode Character Encoding Stability Policies "Once a character is encoded, it will not be moved or removed."[1] New Emojis are continuously offered to the social media platform to add meaning and nuance to the text and enrich digital interactions. Hence, emojis are important especially in domains that deal with emotions such as: marketing, psychology, and politics. It gives organizations the ability to monitor users in different social media platforms and allows them to act accordingly. Analyzing emoji is different than analyzing text owing to their unique characteristics, such as their lack of a clear phonetic interpretation and their visual nature (Pohl et al., 2017).

According to the extensive research of psychologists, there are two main approaches for emotion modeling including categorical approach and dimensional approach (Alswaidan and Menai, 2020). In contrast to the dimensional approach which refers to broad emotional experiences (example pleasant, unpleasant), the categorical approach suggests basic emotions (for example: joy, disgust, and anger) with distinct expressions that are universally recognized. In this paper, since we refer to emotions as discrete categories, we utilize categorical approach. The most commonly used categorical emotion models in Natural Language Processing (NLP) tasks are theory of Ekman (Ekman and Friesen, 1972) with 6 basic emotions, and Plutchik's Wheel of Emotion (Plutchik and Kellerman, 1980) with 8 primary emotions. More number of distinct emotions require deeper insights, and it is difficult even for human annotators to distinguish emotions, such as anger and disgust, because of the similarities between them (Aman and Szpakowicz, 2007).

Humans use the power of logical, and linguistic reasoning to understand the intent and emotion that can be conveyed in a message. However, computers do not naturally have the ability to understand words or emojis as human does. In the field of NLP, each term (word or emoji) is represented to the computer in the form of numeric vectors. Word Embedding is one of the most popular concepts for vector representation. Bag of Words (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF) are two forms of word embeddings (Salton and Buckley, 1988). BoW is an algorithm that builds a vocabulary based on the unique words in the corpus and takes their number of occurrences to create vectors. The core idea is that similar documents include similar words which is used for applications like search, document classification, and topic modeling. In TF-IDF each word in the vocabulary is given a weight based on their importance. For example, higher weight is assigned to words occurring in only a few documents, and the word that appears in more documents may have less value. Removing the less valuable words makes the model building less complex by reducing the input dimensions. These two algorithms result in a vector with lots of zero scores, which is called a sparse vector representation. Sparse vectors require more memory and computational power.

Despite the success of BoW and TF-IDF in the field of NLP, Word Embedding based on neural network models such as GloVe, and Word2Vec have proven to be more effective and shown better results where understanding the context of words is concerned. In these models, each word is represented by a vector, instead of a number (wordcount or weight). These vectors, called dense vector representations, retain most of the linguistic information and semantics of the terms present in the sentence by grouping commonly co-occurring items, that share common contexts, together in the representation space. Word vectors with similar meanings have vector representations that are positioned close to one another in the embedding space. There are publicly available word embeddings which are generated as a result of training these models on very large amounts of text data. Such pre-trained word embeddings can be reused in various NLP tasks such as sentiment classification, emotion classification, depression detection.

The emotions in text are detected and analyzed using various techniques, such as machine learning approaches (supervised, and unsupervised learning methods), keyword-based, and lexicon-based approaches (Chaturvedi et al., 2016). While machine learning approaches rely on classifiers, such as Support Vector Machine (SVM) and Naïve Bayes, the lexicon-based approaches are developed based on the word-emotion lexicons. Two of the well-known emotion lexicons are WordNet-Affect (Strapparava and Valitutti, 2004), which used seed words to generate the same emotions for all WordNet synonyms, and the NRC word-emotion lexicon (Mohammad and Turney, 2013) in which an emotion is assigned to more than ten thousands English words by manual annotation. The vocabularies attributed to the aforementioned emotion lexicons are static and formal, making it challenging to apply to social media context that requires dynamic and informal lexicon for emotion detection. To address this limitation, researchers have applied various unsupervised techniques, where developing emotion lexicons rely on associations between words and emotions by finding their probabilities of appearing together. These techniques are applied on manually labeled dataset or weakly labeled emotion corpus. The core idea of weakly labeling, also known as distant supervision, is using hashtag for pseudo-labeling in the tweets. In the study by (Mohammad, 2012) the researchers applied Pointwise Mutual Information (PMI) to learn a word-emotion lexicon that led to better results compared to WordNet-Affect lexicon in an emotion classification task. In this approach, weakly labeled corpus was extracted from tweets that

[1] http://unicode.org/policies/stability_policy.html

contained any of Ekman's emotion hashtags. PMI measures the relationship between two variables by comparing the probability of observing two variables together with the probability of observing two variables independently. There are several researches that have discussed the effectiveness of PMI in capturing associations between words and emotions (Bandhakavi et al., 2017; Sintsova and Pu, 2016), quantifying implicit associations among nodes in community detection (Luo et al., 2020), and recognizing word synonyms in web search engine (Turney, 2001).

The aim of our research is to develop a method to automatically categorize emoji into fine-grained emotion classes using partially annotated data, which is a combination of annotated and unlabeled data. As it was mentioned, new emojis are introduced every year to the social media platform which makes it a difficult task to manually update and categorize emojis. Thus, there is a need to automatically categorize the emojis. To address this challenge, we used emotions from existing labelled data as well as hashtags as initial labelers, and then a modified PMI method is introduced to discover association between emotional terms and Emoji based on their co-occurrences. The success of proposed method owes to two main novelties:

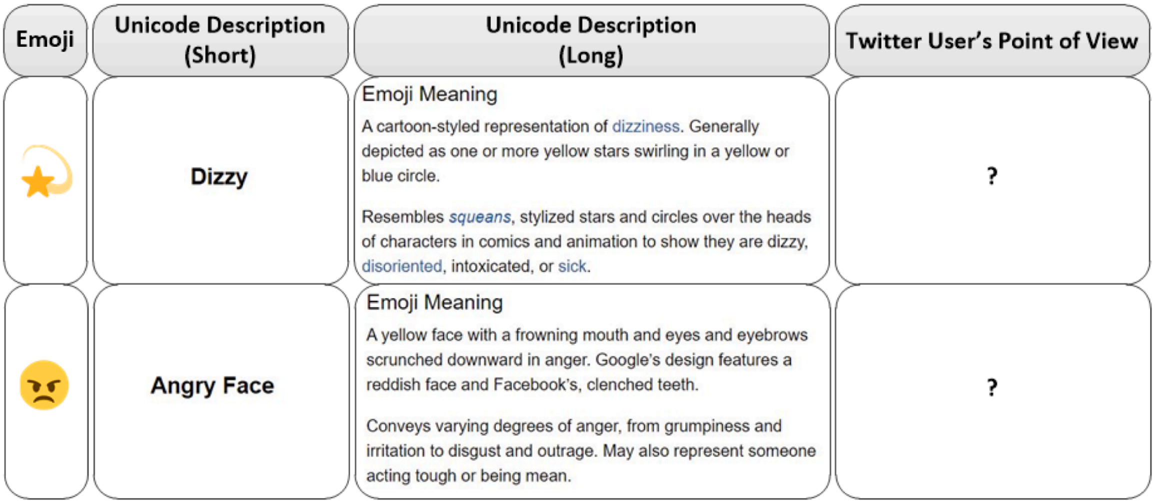
- Firstly, we use sophisticated emotion categories (11 emotions instead of 6 basic emotions) and perform multi-label emotion classification by discovering the semantic similarity between emoji and emotion categories. Classification into a greater number of distinct emotion categories is more difficult due to the similarities between emotions.
- Secondly, we suggest a novel Balanced Weighted PMI algorithm (B-PMI) that considers the imbalance of emotion classes in dataset, which allows to compensate for the PMI biasness toward less frequent emotion categories. As a result, an emoji lexicon is built that provides more emotion related features to enhance the performance of emotion classification. Concretely, applying pre-trained model to develop emoji embedding achieved higher classification accuracy.

The proposed method is validated on tweets in the field of sentiment analysis (negative, positive, neutral) and emotion classification (containing 11 emotion categories). The results show that the novelties increased the performance in both keyword-based and word embedding approaches using machine learning and deep learning methods. We compared our method with pre-trained word embeddings on two different datasets, which achieves higher micro-F1 scores. According to the performed experiments withing the domain of tweets, the practicality of this approach is confirmed.

This article is organized as follows: related works are presented in the Section 2. Section 3 presents the proposed B-PMI method that helps to classify emoji, assign labels to emoji, and develop emoji embedding. Section 4 describes a deep learning-based approach for sentiment analysis and multi-label emotion classification. Section 5 presents evaluation metrics and datasets. Section 6 discusses the results and finding obtained from the experiments. In the final section, conclusion and future work are discussed.

2. Emotion indicators in tweet

Text is a virtually unlimited resource on internet which is usually fused with various forms of expressions such as emoticon, emoji and gif. Using emoticons as indicators of emotion in tweets is discussed in earlier works, such as the study by Go et al. (2009), that eliminated time consuming task of manually annotating tweets by automatically creating training data. They suggested positive emoticons, like:), indicates tweet is positive, and tweets with negative emoticons, like:(, and are negative. Yu et al. (2019) developed a system to extract emoticons in Chinese context and utilized kinesics model to divide emoticons into semantic areas (eye, nose, etc.) and classified them into one of seven affect categories (e.g. negative leader, negative follower, etc.). In comparison to other models that used n-grams or dictionary, this study extracted feature vector from emoticons and then used machine learning methods such as Naïve Bayes (NB), Random Forest (RF), Support Vector Machines (SVM), and Logistic Regression (LR), as well as heuristics method to classify



Emoji	Unicode Description (Short)	Unicode Description (Long)	Twitter User's Point of View
💫	Dizzy	Emoji Meaning A cartoon-styled representation of dizziness. Generally depicted as one or more yellow stars swirling in a yellow or blue circle. Resembles *squeans*, stylized stars and circles over the heads of characters in comics and animation to show they are dizzy, disoriented, intoxicated, or sick.	?
😠	Angry Face	Emoji Meaning A yellow face with a frowning mouth and eyes and eyebrows scrunched downward in anger. Google's design features a reddish face and Facebook's, clenched teeth. Conveys varying degrees of anger, from grumpiness and irritation to disgust and outrage. May also represent someone acting tough or being mean.	?

Fig. 1. Sample Unicode Description for Emoji.

the emoticons, in which heuristics method outperformed machine learning methods.

2.1. Determining emotions based on emojis

Emoji as new generation of emoticons has been studied in recent researches. Jiang and Wilson (2018) explored the use of emoji in degree of misinformation in the original post. The result showed significant negative correlations between eight cluster of emoji and veracity of original tweets, as people usually used them when commenting on posts with low veracity. Raza et al. (2019) proposed a scoring approach to get the semantic orientation (positive, negative, neutral) of frequently used emoji by using a manually compiled list of Emoji. Vora et al. (2017) replaced the emojis with their textual meaning based on Unicode Consortium's emoji definitions (Fig. 1). Similarly, in the study by Hauthal et al. (2019), emojis are categorized based on their Unicode names, and the synonyms associated to their description, which results in categorizing 86 emojis to one of the six emotional categories by Ekman (Ekman and Friesen, 1972). AlMahmoud and AlKhalifa (2018) calculated the sentiment of the emoji by classifying emojis into positive, negative, and neutral. Despite the fact that some emojis clearly represent the emotion, other emojis may not carry a clear meaning or emotions. For the sentiment-ambiguous emoji which are difficult to classify, they surveyed 146 people in order to determine sentiment of each emoji. Fernández-Gavilanes et al. (2018) proposed an unsupervised sentiment analysis approach to create emoji lexica based on their Unicode description.

One of the remarkable studies in emoji classification is the study by Pohl et al. (2017) that proposed an emoji similarity model that automatically creates emoji-pairs and optimize emoji ordering to retrieve a fitting emoji based on the emoji or text user entered in the applications. They developed two models: one is based on the emoji annotations or tags which has poor coverage for many emoji, and the other model is based on semantic information on emoji which does not require manual annotated data, scale better for larger emoji pairs, captured more nuanced relationships of emoji despite having more noise. In their study they implement skip-grams which is a form of Word2vec (Mikolov et al., 2013). Similarly, Urabe et al. (2021) proposed a deep learning based method to improve the emoticon recommendation system using pre-trained embedding models such as Word2Vec. Eisner et al. (2016) presented emoji embedding approach to automatically interpreting the emotional content of an emoji from their Unicode description. Considering embedding models, there have been only a few studies on emoji embedding and all of them have been done recently (Eisner et al., 2016; Guibon et al., 2018). Many researches used existing pre-trained word embeddings such as Stanford's GloVe (Pennington et al., 2014), which is based on words co-occurrences, Google's Word2vec, and the pre-trained word embedding trained on 550 million English tweets (Baziotis et al., 2018) which used Word2Vec algorithm with skip-gram model. However, the available pre-trained word embeddings do not have emoji or support very few numbers of emojis.

Therefore, according to the studies, there are two main approaches to interpret the emoji in tweet:

1) Keyword-based approach in which each emoji is assigned a label that are used for expressing them, such as using their Unicode description.[2] However, it is not guaranteed that their popular usage aligns with their description (Wood and Ruder, 2016).

   Another way is using an emotion label such as negative, positive, happy, or disgust which can be assigned to each emoji manually or automatically. In most of the existing researches, emojis are often categorized in one emotional class (Fernández-Gavilanes et al., 2018), though emojis can express more than one emotion and can be considered as multi-label classification problem. Categorizing emojis manually, is prone to misinterpretations and may omit important details regarding usage. In the process of emotion classification using keyword-based approach, the emoji is replaced with the associated label.

1) word embedding approach that includes word representations in finite dimensional vector space; Word embeddings can be developed from scratch which requires large dataset or can be obtained from available pre-trained word embeddings. The emoji in existing pre-trained word embeddings are usually limited and does not include new emoji. Thus, developing an emoji classification method based on semantic similarity between emoji and emotional words can potentially improve emotion classification while requiring a smaller number of data. Learning the emotion class of more emojis and extending pre-trained word embeddings, results in providing more emotional related features in emotion classification and therefore, improve its discriminative power in the emotion classification.

2.2. Emotion classification in tweets

Co-occurrence based measures for word association such as Pointwise Mutual Information (PMI) (Church and Hanks, 1990), log likelihood ratio (LLR) (Dunning, 1993), and Dice (Dice, 1945) are widely used to measure the strength of association between pair of words. The wide use of word co-occurrence statistics for measuring semantic similarity is due to the popular assumption that words that are mentioned together more frequently are more likely to be conceptually related. Sintsova sand Pu (2016) proposed a distant learning method that focused on the problem of imbalance in emotion distribution, as well as detecting neutral tweets. The method includes creating emotion lexicons by using a list of descriptive emotional terms for each emotion category and then used PMI to determine emotion associations of new terms according to frequent occurrence of the new term with the given emotional terms. This

[2] emojipedia.org

method eliminates the need for manually annotated data. introduced a novel classifier namely as Balanced Weighted Voting (BWV), that categorized the tweets in the area of sport event into 20-category emotion model and balanced the weight between emotion categories that appeared more often than other emotion categories.

The study by Cheng et al. (2017) defined a model for emotion cause detection using multiple-user structure (i.e., Messages are written by multiple users instead of single user) in order to analyze psychology of a group of users who often interact with each other. In this study, they developed an emotion cause corpora and then used SVM and LSTM to examine the features extracted from texts, in which SVM significantly outperformed LSTM. Although the input of both SVM and LSTM is a word sequence, the mechanism differs in each model. In SVM the input is word bags such as tf-idf, but LSTM tries to model dependencies between words. According to the authors, due to using informal texts in tweets without proper grammar, LSTM is not powerful enough to learn from their proposed method. The method improved emotion cause detection with accuracy between 61.3 ~ 70.5 in comparison with majority-based-base line with accuracy of 67.1

In the study by Zhou et al. (2019) the researchers, proposed an emotional supervised model (nCG-ESM), that uses sequence-to-sequence (Seq2Seq) model to generate responses with emotional diversity, including five specified (Angry, Disgust, Happy, Like, Sad) or three unspecified emotions, that can be adapted and extended to different scenarios. They adopted a Bidirectional Long Short-Term Memory (Bi-LSTM) to produce an emotion distribution for each dialog sentence, and later used cosine distance between on-hot vector of emotion word, and emotion distribution. The model showed poor results for Angry and Disgust due to lack of training data. The study by Naskar et al. (2020) investigated the emotion changes of users over time in Twitter. This study adapted 16-state Russell's circumflex model of Affect (Feldman Barrett and Russell, 1998) and determined the valence and arousal using dictionary-based approach. The study by Jabreel and Moreno (2019) focused on the multi-label classification using binary relevance (Tsoumakas and Katakis, 2007) method. In the binary relevance method, a multi-label problem is transformed into multiple binary problems, one problem for each label. They proposed a transformation method to get a single binary dataset instead of multiple independent binary datasets. In addition to binary relevance method, label powerset method and classifier chain method have been used to treat the multi-label classification problem.

2.3. Theory of emotion

Two main emotion models, which are often used in NLP tasks are theory of Ekman (Ekman and Friesen, 1972), and Plutchik's Wheel of Emotion (Plutchik and Kellerman, 1980). In this study, we used Plutchik's theory of emotion which has 8 primary emotions, with addition of love, optimism, and pessimism. Therefore, anger, anticipation, disgust, fear, joy, love, optimism, pessimism, sadness, surprise, and trust are the 11 emotion classes which we used in this study. Plutchik's theory of emotion covers the 6 basic emotions in theory of Ekman, and it includes more fine-grained emotions that help to capture the nuances of human emotion. This is the reason for choosing Plutchik's theory of emotion.

2.4. Identify emojis that shows emotion

A list of 156 emojis is compiled based on the frequently used emojis in the tweet dataset, as well as commonly used emojis based on the real-time statistics in emojitracker.[3] We only included tweets that contain any of our specified list of 156 emojis. More information about dataset is given in Section 5.

3. Proposal

In this study, the problem of emoji categorization is formulated as a multi-label emotion classification task since we study fine-grained emotion categories and each emoji might belong to more than one emotion category.

Given an emoji set $X = \{emj_1,\ emj_2, \ldots,\ emj_K\}$ and emotion categories $E\ = \{e_1,\ e_2, \ldots,\ e_{|E|}\}$, the proposed method (Fig. 2) aims to categorize each emoji based on the semantic similarity. The Balanced PMI helps to determine the probability of emoji occurring together with emotion categories, and detects the weight W between each emoji and the emotion categories in order to assign labels $Y = \{y_1,\ y_2, \ldots,\ y_n\}$. Thus, we model the relationship between emoji and emotion category in which, each emoji can be assigned to one or more emotion labels $Y_{emj}\ \ = \{e_{i_k}\} \in E$. We later used the resultant emoji with the mapped emotion category, to produce emoji lexicon and emoji embedding to be used in the task of sentiment analysis and emotion classification.

3.1. Method input

The demonstration of our proposed method is given in Fig. 2. The labeled dataset (Table 1) includes the content of tweet and associated labels (0 means not associated, 1 means associated). We transformed the labeled dataset to only include the associated labels (Table 2). Besides, we used twitter API[4] to crawl English (published in USA) tweets labeled with any of the 11 emotional hashtags. The tweets were filtered to include at least one emoji from the list of 156 emojis.

Therefore, input contains a set of labeled tweets (Table 2) and a set of tweets with emotion hashtags (Table 3). As mentioned in

[3] emojitracker.com.

[4] https://developer.twitter.com/en/docs

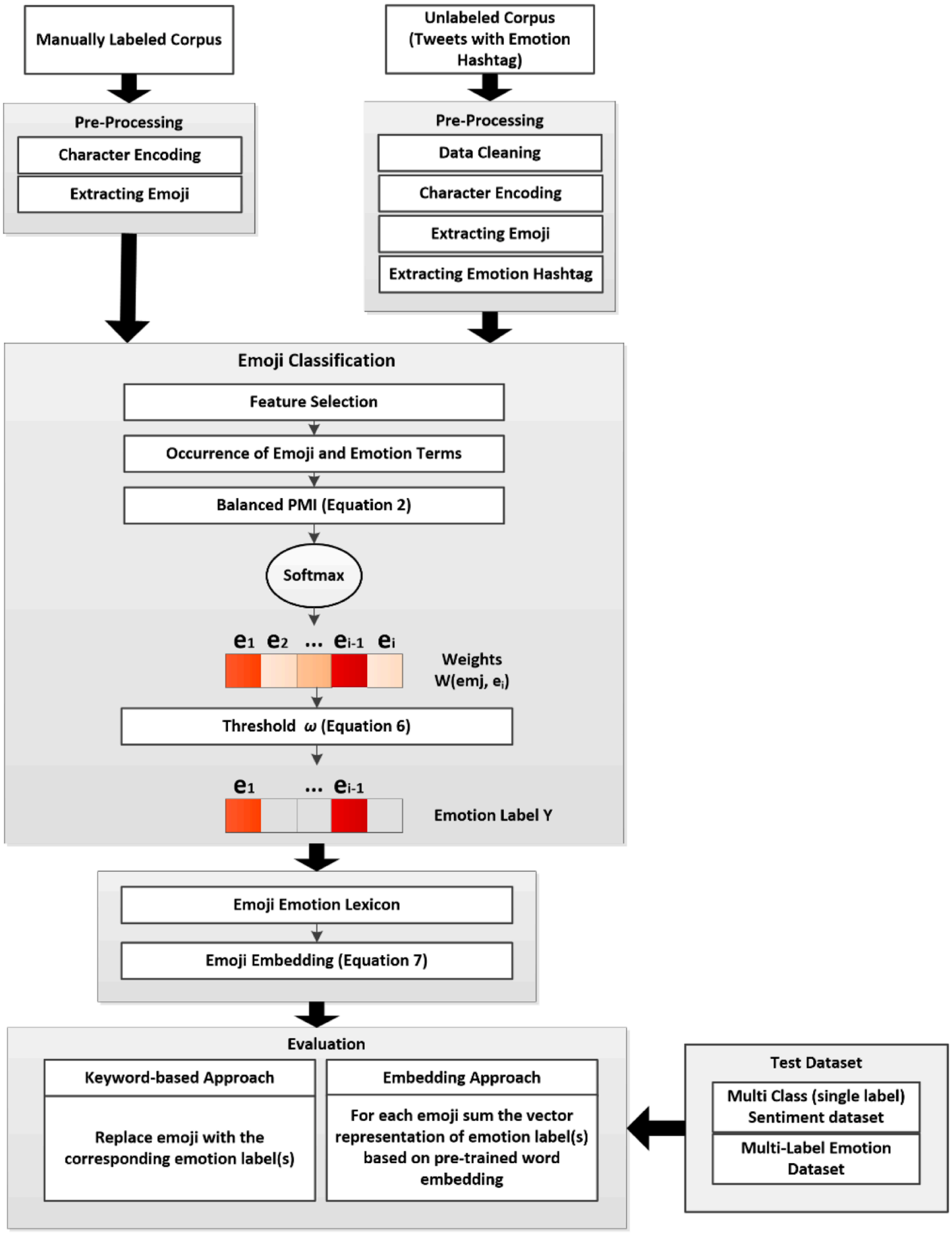

Manually Labeled Corpus
Pre-Processing
Character Encoding
Extracting Emoji
Unlabeled Corpus (Tweets with Emotion Hashtag)
Pre-Processing
Data Cleaning
Character Encoding
Extracting Emoji
Extracting Emotion Hashtag
Emoji Classification
Feature Selection
Occurrence of Emoji and Emotion Terms
Balanced PMI (Equation 2)
Softmax
e1 e2 ... ei-1 ei
Weights W(emj, ei)
Threshold ω (Equation 6)
e1 ... ei-1
Emotion Label Y
Emoji Emotion Lexicon
Emoji Embedding (Equation 7)
Evaluation
Keyword-based Approach
Replace emoji with the corresponding emotion label(s)
Embedding Approach
For each emoji sum the vector representation of emotion label(s) based on pre-trained word embedding
Test Dataset
Multi Class (single label) Sentiment dataset
Multi-Label Emotion Dataset

Fig. 2. An illustration of the proposed system that maps emojis to emotion categories to produce emoji lexicon; the darker shaded areas shows the higher weights and therefore, stronger association between emoji and emotion category.

Section 2.3, we chose a set of standard and representative emotions by Plutchik's theory of emotion (Plutchik and Kellerman, 1980) that consists of eight basic emotions with addition of love, optimism, and pessimism.

The size of total tweets that contain any of 156 emojis is shown in Table 4. As is evident from the table, this dataset suffers from imbalanced data for the emotion classes. The emotion *joy* is more dominant compared to any of the other Plutchik's emotions like *pessimism* or *trust*.

3.2. Pre-processing

Text pre-processing includes extraction of emoji and emotional hashtags from tweets which are fed into feature selection. In the first step, UTF-8 character encoding is performed to store an emoji as a unique sequence of bytes. Then the emoji is extracted from tweet corpus. In the unlabeled dataset, emoji and emotional hashtags are extracted. We used unigram as features to determine the number of co-occurrences of each term (term refers to emotion categories and emoji) and retain the terms that appeared M or more

Table 1
Sample of labelled dataset.

Id	Tweet	Anger	Anticipation	Disgust	Fear	Joy	Love	Optimism	Pessimism	Sadness	Surprise	Trust
1	@user yes ❤❤; cheering homecoming game!	0	0	0	0	1	1	1	0	0	0	0
2	@user bts' trilogy MV is my all-time fav 🙌 quite gloomy but beautiful as well✨	0	0	0	0	1	1	0	0	0	0	0
3	I've been disconnected whilst on holiday 😃 but I don't move house until the 1st October 😂	1	0	0	0	0	0	0	0	0	0	0
4	How can l rule my mind!!!!!! It's hilarious that you can't 😃 😂 💜	0	0	0	1	1	0	0	1	1	0	0

Table 2
Sample of labelled data after transforming the dataset.

Id	Tweet	Tweet label
1	@user yes ❤❤; cheering homecoming game!	Joy, Love, Optimism
2	@user bts' trilogy MV is my all time fav 💜 quite gloomy but beautiful as well✨	Joy, Love
3	I've been disconnected whilst on holiday 💜 but I don't move house until the 1st October 💜	Anger
4	How can l rule my mind!!!!!! It's hilarious that you can't 💜 😢 😢	Fear, Joy, Sadness, Pessimism

times in the corpus. In this study M is equal to 8.

3.3. Weighted emoji lexicons

In PMI (Eq. (1)) (Church and Hanks, 1990), P(x,y) is the ratio between number of observed co-occurrences of x and y (joint probability) and the size of the corpus (N). P(x,y) is divided by the individual probability of x and y, P(x) and P(y), which are estimated by counting the number of observations of x and y in a corpus divided by N (Church and Hanks, 1990).

$$PMI: \ Log\frac{P(x,y)}{P(x)P(y)} = Log\frac{\frac{f(x,y)}{N}}{\frac{f(x)}{N}\frac{f(y)}{N}} = Log\frac{f(x,y)N}{f(x)f(y)} \tag{1}$$

This approach is biased the less frequent emoji or emotion categories to have higher weight (Jurafsky and Martin, 2019; Levy et al., 2015). This limitation leads us to propose a balance weighted emoji categorization method. We categorized Twitter-specific emoji by using a modified pointwise mutual information (PMI) method. In this method, emotion weights (W) of emoji is learnt by measuring the semantic similarity between emoji (emj) and emotion categories (e_i) using a rebalancing coefficient value.

3.4. Balanced weighted emoji lexicon and emoji categories

In order to balance the distribution of emotions and boost the score of frequent pairs, we modified PMI to multiply a rebalancing coefficient (Eq. (2)). Therefore, the emotion weight of emoji is computed as:

Table 3
A partial list of tweets with emotion hashtags.

Id	Tweet	Tweet label
1	Happiness depends on your mindset and attitude ❤ #focus #trust	Trust
2	Prince Rogers Nelson Missed, And Loved Rih Gone But Not Forgotten!!!!!! 😢 😢 💫 🙌 😫 😫 💫 💁 #sad	Sadness
3	Happy Human Birthing Day to this beautiful young lady! - 12 cycles of the earth babe. Honored you chose me to walk this planet w ya! ~Daddy ❤ 💁 😂#joy #happy #loveandlight	Joy

$$Balanced\ PMI\ (B-PMI):\ W(x,y)=Log\left(\frac{f(x,y)}{f(x)}\frac{1}{P(x)\ P(y)}+\frac{f(x,y)}{f(y)}\frac{1}{P(x)\ P(y)}\right) \tag{2}$$

In this formula, x models the occurrence of emoji *emj*, and y models the occurrence of an emotion class *e*. Therefore:

$$Balanced\ PMI\ (B-PMI):\ W\ (emj,e)=Log\left(\frac{f(emj,e)}{f(emj)}\frac{1}{P(emj)\ P(e)}+\frac{f(emj,e)}{f(e)}\frac{1}{P(emj)\ P(e)}\right) \tag{3}$$

Where *f(e)* is the number of observations of an emotion class, and *f(emj)* is the number of observations of emoji. Using this formula, we consider the association between each emoji and emotion class relative to each other. In this case, to compensate for the PMI score toward more frequent emotion categories without overestimating their weights:

$$instead\ of\ using:\ \frac{f(x,y)}{N}$$

$$we\ used\ \frac{f(x,y)}{f(x)}\ and\ \frac{f(x,y)}{f(y)}.$$

The higher value of $W(emj,e)$ indicates the stronger association between pairs of emoji and emotion category.

The output of this step is a set of weighted emojis *(L)* compromises of emojis with mapped emotion categories that represents the strength of relationship based on their semantic similarity. Then it is normalized into probability distribution (Eq. (4)) using Softmax, a log-linear classification model which is based on the multinomial distribution (Table 5). Therefore, the weight ranges between 0 and 1, and the sum of all the weights will be equal to one (Eq. (5)).

$$W(emj,\ e_i)=\frac{e^{W\ (emj,e_i)}}{\sum_{e_j\in W}e^{W(emj,e_j)}} \tag{4}$$

Where

$$\sum_{e_i\ \in E} W\ (emj,\ e_i)=1 \tag{5}$$

In the set of weighted emojis *(L)*, we specified a threshold value ω in order to obtain the multi-label output (Table 6). The weight of emotions W_{e_i} that are lower than ω is set to zero. Thus, for each emoji *emj* those emotions with weights W_{e_i} greater than ω are assigned as labels (Eq. (6)).

$$Y_{emj}\ =\{e_{i_k}\}=\begin{cases} W_{e_i},\ W_{e_i}\geq\omega \\ 0,\ W_{e_i}<\omega \end{cases},\ e_i\in E \tag{6}$$

The specific value for the threshold (*ω*) value was set manually. The threshold was determined based on midrange which is the

Table 4
Total distribution of different emotion classes in manually labelled tweets, and tweets with emotional hashtags that include any of 156 emojis.

	Anger	Anticipation	Disgust	Fear	Joy	Love	Optimism	Pessimism	Sadness	Surprise	Trust	Total
# of instances	365	205	305	272	4554	1168	2970	108	1320	177	146	11,590
% of instances	3.14	1.76	2.63	2.34	39.29	10.07	25.62	0.93	11.38	1.52	1.25	100

Table 5
Sample of Weighted Emoji Lexicon.

description	anger	anticipation	disgust	fear	joy	love	optimism	pessimism	sadness	surprise	trust
😫	0.4474	0.0134	0.4319	0.0134	0.0134	0.0134	0.0134	0.0134	0.0134	0.0134	0.0134
😫	0.3406	0.0135	0.3273	0.0135	0.0135	0.0135	0.0135	0.0135	0.2238	0.0135	0.0135

Table 6
Sample of Emoji Emotion Lexicon.

description	Emotion
😫	Anger, Disgust
😫	Anger, Disgust, Sadness

mean of the highest and lowest values. In this study, the threshold is set to 0.2.

Fig. 3 compares the meaning of emoji provided by Unicode description and point of view of twitter user. Based on Unicode description 💫 means "dizzy", but from user's point of view, this emoji brings joy and optimism. It recommends that the description of emoji does not always reflect the actual usage of emoji in twitter.

We compared the results of PMI and B-PMI for two emoji to better understand the influence of rebalancing coefficient. It is assumed that x and y are different random variables, that create bigrams when these two variables appear together. P(x) refers to possibility of emoji occurring as the first word of bigram and P(y) refers to possibility of occurrence of y as second word of bigram. The difference between B-PMI and PMI is that, while PMI uses the ratio of bigram co-occurrence to the total number of bigrams, B-PMI adapts relative frequency where the ratio of bigram is divided by the total occurrence of the unigram. Hence, f(x,y) is divided by f(x) to calculate P(x, y).

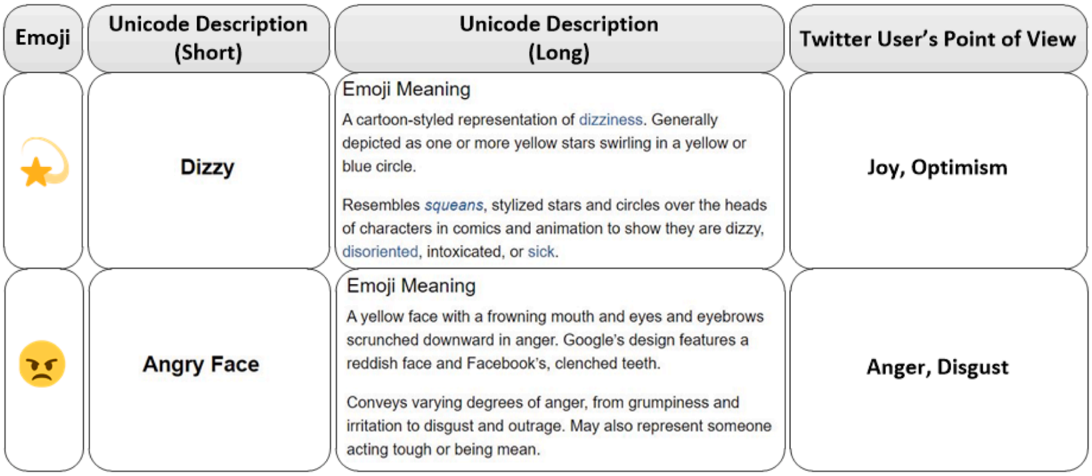
According to the results at the Table 7 and Table 8, B-PMI mapped the emoji to emotion classes in accordance with their frequency to address the class imbalance. Thus, B-PMI categorized *blue heart* emoji in *trust* and *anticipation* emotion classes, while based on PMI this emoji is more indicative of *anticipation* and *love*. Similarly, based on B-PMI *two hearts* emoji symbolizes *joy, optimism,* and *love* emotions, while PMI categorized this emoji as *joy* emotion class.

3.5. Word2vec embeddings

In this paper, we utilized pre-trained word embedding by Baziotis et al. (2018), which is based on Word2vec model (Mikolov et al., 2013). The reason why we used this pre-trained word embedding was because it was trained on twitter corpus, and Word2vec model is applied to capture the semantic similarity of the texts.

Word2vec is a two-layer neural network that learns word associations and semantic similarity from natural language inputs and represents each word with a fixed-size vector.

Word2vec uses continuous bag-of-words (CBOW) or continuous skip-gram (Skip-gram) to construct distributed representation of a word. The first model predicts the representation of target word by using the words that appeared in the context window. In contrast, Skip-gram model (Fig. 4) uses the representation of target word to predict the context.



Emoji	Unicode Description (Short)	Unicode Description (Long)	Twitter User's Point of View
💫	Dizzy	Emoji Meaning A cartoon-styled representation of dizziness. Generally depicted as one or more yellow stars swirling in a yellow or blue circle. Resembles *squeans*, stylized stars and circles over the heads of characters in comics and animation to show they are dizzy, disoriented, intoxicated, or sick.	Joy, Optimism
😠	Angry Face	Emoji Meaning A yellow face with a frowning mouth and eyes and eyebrows scrunched downward in anger. Google's design features a reddish face and Facebook's, clenched teeth. Conveys varying degrees of anger, from grumpiness and irritation to disgust and outrage. May also represent someone acting tough or being mean.	Anger, Disgust

Fig. 3. Meaning of emoji in Unicode description vs emoji from Twitter user's point of view.

Table 7
Results for B-PMI and PMI: blue heart emoji; s stands for softmax.

emotion	emoji	f(e)	f(emj)	f(emj,e)	B-PMI	PMI	s_B-PMI	s_PMI
anger	blueheart	365	267	0	0.00	0.00	0.00	0.05
anticipation	blueheart	205	267	20	**6.00**	**1.85**	**0.41**	**0.32**
disgust	blueheart	305	267	0	0.00	0.00	0.00	0.05
fear	blueheart	272	267	0	0.00	0.00	0.00	0.05
joy	blueheart	4554	267	135	4.04	0.13	0.06	0.06
love	blueheart	1168	267	78	5.00	**1.30**	0.15	**0.18**
optimism	blueheart	297	267	10	4.75	0.31	0.12	0.07
pessimism	blueheart	108	267	0	0.00	0.00	0.00	0.05
sadness	blueheart	1320	267	9	2.69	−1.99	0.01	0.01
surprise	blueheart	177	267	0	0.00	0.00	0.00	0.05
trust	blueheart	146	267	7	**5.50**	0.82	**0.25**	0.11

Table 8
Results for B-PMI and PMI: two hearts emoji.

emotion	emoji	f(e)	f(emj)	f(emj,e)	B-PMI	PMI	s_B_PMI	s_PMI
anger	twohearts	365	408	0	0.00	0.00	0.01	0.09
anticipation	twohearts	205	408	0	0.00	0.00	0.01	0.09
disgust	twohearts	305	408	0	0.00	0.00	0.01	0.09
fear	twohearts	272	408	0	0.00	0.00	0.01	0.09
joy	twohearts	4554	408	326	**4.10**	**0.79**	**0.31**	**0.19**
love	twohearts	1168	408	51	**3.82**	0.08	**0.23**	0.10
optimism	twohearts	297	408	12	**4.31**	−0.04	**0.38**	0.09
pessimism	twohearts	108	408	0	0.00	0.00	0.01	0.09
sadness	twohearts	1320	408	10	2.04	−2.45	0.04	0.01
surprise	twohearts	177	408	0	0.00	0.00	0.01	0.09
trust	twohearts	146	408	0	0.00	0.00	0.01	0.09

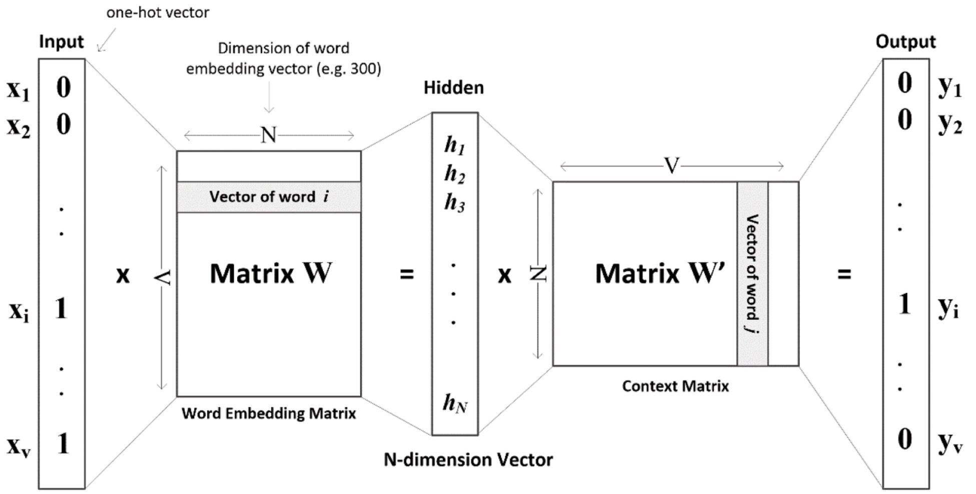


Fig. 4. Word2vec (Skip-gram) model with one hidden layer. The figure was created with reference to Orkphol and Yang (2019).

In the Skip-gram model, input vector x $(x_1, x_2, \ldots, x_V)$ and the output y $(y_1, y_2, \ldots, y_V)$ are one-hot encoded vector of size V. Given the vocabulary size V (total unique words), the model aims to learn word embedding vector of size N (N is the dimension of word embedding*)*. Matrix W of size $V \times N$ includes the embedding vector of the input word (target word). A hidden layer h is a multiplication matrix between one-hot vector x and matrix W. Context matrix W' is optimized to predict the surrounding words for the target word. The multiplication of the hidden layer and the word context matrix W' produces the output one-hot encoded vector y.

Word2vec model should be trained to learn the weights W and W' that minimizes the loss function (e.g. Negative Sampling). Then, the word embedding matrix W is used to obtain word vectors.

3.6. Emoji embedding

We further used the developed emoji lexicon to build an emoji embedding. For every emoji and the sequence of emotional terms $e_1 \ldots e_N$ describing that emoji, we take sum of the individual emotional term vectors as in pre-trained embedding (details are written in Section 3.5):

$$v_{emj} = \sum_{k=1}^{N} e_k \tag{7}$$

Where v_{emj} is the vector representation of the emoji.

Since we deployed the pre-trained word embedding with 300 dimensions, the resultant emoji embedding (v_{emj}) has 300 dimensions.

4. Experimental setup

We evaluated the effectiveness of our proposed B-PMI-based emoji lexicon and emoji embedding by utilizing machine learning and deep learning methods in the tasks of sentiment classification and emotion classification. For machine learning method, we performed a Logistic Regression (LR) algorithm, considering unigram as features, with a binary relevance method to treat the multi-label problem.

In order to perform deep learning method, a Bidirectional Long Short-Term Memory (Bi-LSTM) (Hochreiter, 1998) is used. This Deep Learning model is trained with the Keras library based on Tensorflow (Erickson et al., 2017) in Python. Fig. 5 demonstrates the system's architecture. This system includes three main parts which are embedding layer, encoding layer and attention layer.

Pre-processing includes the following steps:

- Spell correction: The "ekphrasis" library has been used to check the spelling.
- Term normalization: The "ekphrasis" library has been used to detect and replace "url", "email", "percent", "money", "phone", and "user" (e.g. url is replaced with the token URL, @user is replaced with token USER)
- Segmentation on hashtags: The "ekphrasis" library recognizes hashtags (e.g. #ilovemyteam: I love my team)
- Lowercasing: To convert the tweet to lowercase.
- Twitter abbreviation: A list of common twitter abbreviations is provided (e.g. tbh: to be honest, idk: i do not know).
- Tokenization
- Lemmatization: Lemmatization is used to map a word to its root form.

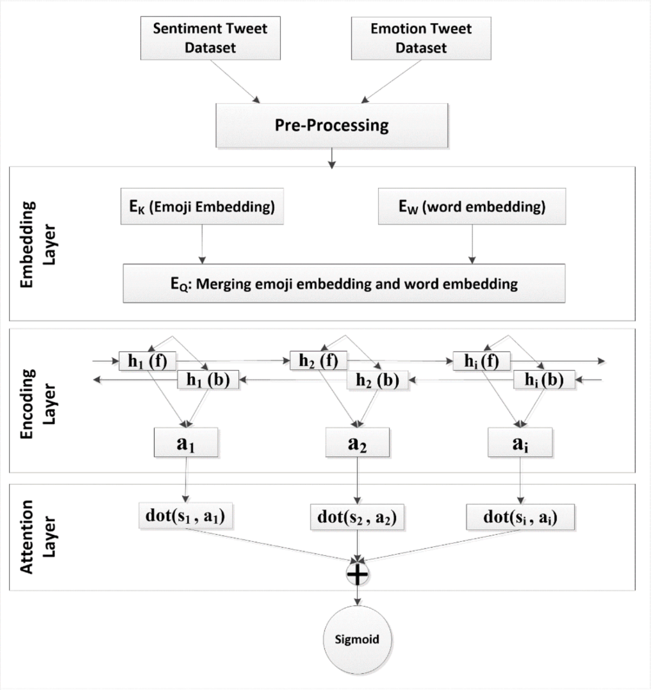


Fig. 5. The utilized Bi-LSTM system (f: forward, b: backward).

Algorithm 1: Combining emoji embedding and word embedding

```
Input: $E_P$: pre-trained word embedding, $E_K$: emoji embedding, D: tweet datasets
Output: Word embedding that includes Emoji and Word: $E_Q$
$E_Q = \{\}$; $E_W = \{\}$;
foreach $w \in D$ do
    if $w \in X$
        $E_Q = E_Q \cup E_k(w)$;
    else if $w \in W$
        $E_w(w) = E_P(w)$;
        $E_Q = E_Q \cup E_w(w)$;
End
End
Return $E_Q$
```

Each tweet includes word and emoji which $W = \{w_1, w_2, \ldots, w_l\}$ is the set of the words in a tweet and $X = \{\mathrm{emj}_1, \mathrm{emj}_2, \ldots, emj_k\}$ is the set of emojis in the tweet. The embedding layer aims to represent each word w_i and emoji emj_k by a vector v_{wi}. Let E_P be the pre-trained word embedding (Baziotis et al., 2018), then the word embedding E_w includes all words in the tweet dataset which have a vector presentation in pre-trained word embedding $E_W(w_i)\ \forall\ w_i \in W, w_i \in\ E_P$. The emoji embedding E_K includes the vector representation of the emojis obtained in Section 3.6. In the Algorithm 1, the emoji embedding (E_K) is combined with the word embedding (E_W) to create a new word embedding (E_Q).

Then, the obtained word representations from embedding layer is fed into encoding layer. Encoding layer consists of LSTM cell/units with x_t as the input vector at time t and output vector h_t. There are three gates: input gate (i_t), forget gate (f_t), and output gate (o_t) in which input vector x_t and previous output h_{t-1} in multiplied to input weight matrix (W) and output weight matrix (U) respectively. A Sigmoid activation function σ is employed on each gate to convert the values to probabilities between 0 and 1, which decides how much information to keep or forget. Afterwards, for the next cell state c_t and output h_t, *tanh* activation function τ is applied. The equations in each cell is given below:

$$f_t = \sigma\left(W_f x_t + U_f h_{t-1} + b_f\right) \tag{8}$$

$$i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i) \tag{9}$$

$$o_t = \sigma(W_o x_t + U_o h_{t-1} + b_o) \tag{10}$$

$$c_t = f_t\ c_{t-1} + i_t \tau(W_c x_t + U_c h_{t-1} + b_c) \tag{11}$$

$$h_t = o_t \tau(c_t) \tag{12}$$

Bi-LSTM concatenates a sequence of forward hidden state as well as sequence of backward hidden states as given bellow:

$$h_t = \left[\overrightarrow{h_t};\ \overleftarrow{h_t}\right] \tag{13}$$

Lastly, the word attention mechanism (Vaswani et al., 2017; Yang et al., 2016) is used to extract most relevant words.

$$s_i = \sum_t a_{it} h_{it} \tag{14}$$

Where:

Table 9
Bi-LSTM Hyperparameters.

Parameter	Value
Embedding Layer	E_K: Dimensions: 300 E_W: Dimensions: 300 E_Q: Dimensions: 300
Encoding Layer	RNN Cell: Bi-LSTM Layers: 2 Hidden size: 64 × 2 = 128 RNN Dropout: 0.3
Output	Activation: ReLU Sigmoid: 11 units (emotion classification: 11 emotion classes) or Softmax: 3 units (sentiment classification: positive, negative, neutral)

$$a_{it} = \frac{\exp\left(u_{it}^{\mathrm{T}} u_w\right)}{\sum_t \exp\left(u_{it}^{\mathrm{T}} u_w\right)} \tag{15}$$

$$u_{it} = \tanh(W_w h_{it} + b_w) \tag{16}$$

u_{it} is hidden representation of h_{it}, which is the result of non-linear activation function (*tanh*) on the Bi-LSTM output. To calculate the attention similarity score of a word, we measure the similarity of the context vector u_w with u_{it}. Then a Softmax function is used to normalize the importance weight a_{it}. Finally, the outputs of attention layer are summed and sent through Sigmoid operation to get the probability distribution of classes for the task of multi-label tweet classification. For the binary classification in sentiment analysis where the output is a single label, Softmax function is performed.

The system was trained using Adam (Kingma and Ba, 2014), with dropout of 0.3, and a mini-batch size of 32 to minimize the binary cross-entropy loss function (Table 9).

5. *Evaluation methodology*

We provided details of the ground-truth data used for evaluating the obtained emoji lexicon and emoji embedding. We evaluated the proposed method on two datasets for the emotion classification (11 emotion classes) and sentiment analysis (negative, positive, neutral) tasks on tweets. The result of proposed B-PMI-based method is compared with available methods based on lexicon (keyword-based) and pre-trained word embeddings (embedding-based). We used keyword-based approach to replace the emojis with the words that express them such as their description or emotion words.

5.1. *Benchmark dataset*

5.1.1. Emotion dataset. We used SemEval-2018 Task 1: E-C (Mohammad et al., 2018), a multi-label emotion dataset composed of English tweets that were manually labeled in 11 emotion classes. In this dataset the data for the emotion classes in not balanced, in which the number of examples of one class is substantially more than examples of other classes. Fig. 6 represents the occurrence of tweets per emotional class.

5.1.2. Sentiment dataset. We used dataset prepared by (Novak et al., 2015) that manually labeled tweets in one of the 3 classes: positive, negative, and neutral (Table 10). Statistics of the dataset is reported in Table 11.

5.2. *Evaluation metrics*

We used multiple evaluation metrics suitable for multi-class (single-label) classification as well as multi-label classification for both tasks of emotion classification and sentiment analysis.

The evaluation of the predictive performance for multi-label learning systems needs a special approach which is more complicated than multi-class (single-label) learning system. In the experiments, various evaluation measures (Gibaja and Ventura, 2015; Tsoumakas and Katakis, 2007) have been used (Table 12, Table 13).

6. *Evaluation results*

We conducted experiments on two datasets and the results of the proposed algorithm on the benchmark datasets is presented in this

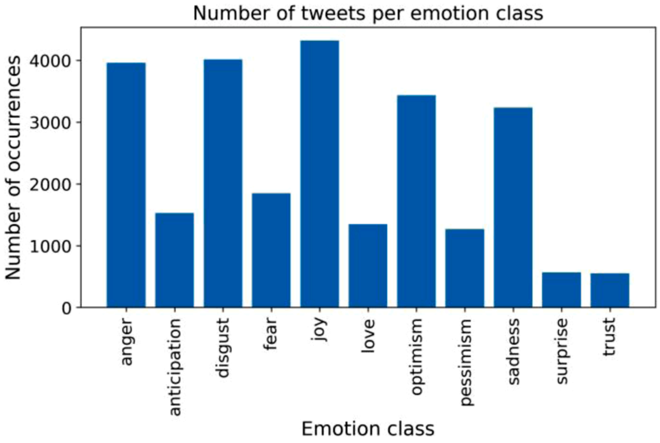


Fig. 6. Tweet occurrences per emotion class in Multi-label dataset.

Table 10
Distribution of tweets in each class in sentiment dataset.

Class	# of tweets
positive	3929
negative	2564
neutral	2009

Table 11
Statistics of the experimental datasets.

Dataset	Single-label / Multi-label	Reference	Labels	Type	#Tweets	#tweets with emoji*
SemEval-2018 Task 1: E-C	Multi-label	(Mohammad et al., 2018)	Anger, Anticipation, Disgust, Fear, Joy, Love, Optimism, Pessimism, sadness, surprise, trust	Train + validation	7506	878
SemEval-2018 Task 1: E-C	Multi-label	(Mohammad et al., 2018)	Anger, Anticipation, Disgust, Fear, Joy, Love, Optimism, Pessimism, sadness, surprise, trust	Test	3184	762
Sentiment-tweet	Multi-class (Single-label)	(Novak et al., 2015)	Negative, positive, neutral	Train	51,679	6877
Sentiment-tweet	Multi-class (Single-label)	(Novak et al., 2015)	Negative, positive, neutral	Test	12,920	1625

*We only included tweets that has any of our specified list of 156 emoji.

Table 12
Definition of Confusion Matrix values.

Metric	Definition	Formula
TN	Negative classes which are correctly predicted as negative class	–
FN	Positive classes which are incorrectly predicted as negative class	–
TP	Positive classes which are correctly predicted as positive class	–
FP	Negative classes which are incorrectly predicted as positive class	–

Table 13
Definition of multi-class single-label and multi-label performance measure.

Metric	Definition	Formula
Precision	ratio of tweets that are correctly predicted to the total predicted tweets	$\frac{TP}{TP+FP}$
Recall	ratio of correctly predicted tweets to all tweets in actual class	$\frac{TP}{TP+FN}$
F1-Score	harmonic mean of precision and recall	$\frac{2 \times Precision \times Recall}{Precision + Recall}$
Precision$_{\text{micro}}$	For multi-label task, where E is the emotion classes	$\frac{\sum_{e\in E} TP}{\sum_{e\in E} TP + \sum_{e\in E} FP}$
Recall$_{\text{micro}}$	For multi-label task, where E is the emotion classes	$\frac{\sum_{e\in E} TP}{\sum_{e\in E} TP + \sum_{e\in E} FN}$
Precision$_{\text{e}}$	Precision of emotion class “e”, for multi-label task	$\frac{TP_e}{TP_e + FP_e}$
Recall$_{\text{e}}$	Recall of emotion class “e”, for multi-label task	$\frac{TP_e}{TP_e + FN_e}$
Micro F1	micro-averaged F-score aggregates the contributions of all classes to compute the average metric	$\frac{2 \times Precision_{micro} \times Recall_{micro}}{Precision_{micro} + Recall_{micro}}$
Macro F1	macro-averaged F-score compute harmonic mean of precision and recalls independently for each emotion class and then take the average (hence treating all classes equally)	$\frac{1}{\lvert E\rvert}\sum_{e\in E}\frac{2 \times Precision_e \times Recall_e}{Precision_e + Recall_e}$
Jaccard Similarity Index	A measure of similarity for the two sets of data which divides the number of correctly predicted labels by the union of predicted and true labels. G_t is the set of the gold labels for tweet t, P_t is the set of the predicted labels for tweet t, and T is the set of tweets.	$\frac{1}{\lvert T\rvert}\sum_{t\in T}\frac{G_t \cap P_t}{G_t \cup P_t}$
Hamming Loss	fraction of the wrong labels to the total number of labels. Smaller value of Hamming Loss indicates a better performance. N defines the number of tweets and L shows the number of emotion labels.	$\frac{1}{NL}\sum_{t=1}^{N}\sum_{j=1}^{L} XOR(G_{t,j}, P_{t,j})$

section. We reported the performance of the machine learning and deep learning classifiers with the baseline emoji labeling methods, and further show the significant effect of the resulted emoji lexicon and emoji embedding from the proposed emoji categorization method (B-PMI) on the performance of classifiers. We compared the proposed B-PMI-based approach with four other existing embeddings as follows:

- Pre-trained Embedding A (Baziotis et al., 2018),
- Pre-trained Embedding B (Pennington et al., 2014),
- Unicode emoji description: replacing emoji with its Unicode description (Vora et al., 2017) (Fig. 7),
- Description embedding: generating emoji embedding by taking sum of the word vectors of its Unicode description (Eisner et al., 2016) (Fig. 8),
- Without emoji: removing emoji from tweet.

6.1. *Sentiment analysis (Bi-LSTM with attention model + pre-trained word embedding)*

For sentiment analysis (negative, positive, neutral), we used the dataset provided by Novak et al. (2015) and compared the results of different existing embedding researches with the proposed method.

It is preferable to use micro-average in a multi-class classification in case of imbalanced dataset. Thus, we used micro-average to evaluate the performance of the classifier. The proposed emoji embedding is merged with the existing pre-trained word embedding (Algorithm 1) to extend the features in the existing embedding. We used word embedding A, since it was trained on a large twitter dataset.

As it can be seen in Table 14, the proposed method improved the micro-recall (67.07%) and micro-f1score (70.01%) with competitive, but lower micro-precision (73.21%). Removing emoji from tweets reduced performance of the classification.

6.2. *Emotion analysis*

Two different approaches are implemented to evaluate the proposed method for the task of multi-label emotion classification. As it was discussed in Section 5.1, we used the dataset that is provided in SemEval-2018 Task 1: E-C.

6.2.1. Binary relevance with logistic regression (considering unigram as features). In this experiment a machine learning algorithm is applied to classify the tweets. We used Binary Relevance with Logistic Regression algorithm to analyze the effect of the proposed emoji lexicon on multi-label classification of tweets. In this method we used unigram as feature and replaced emoji with its description (keyword-based approach). The proposed emoji lexicon based on PMI and B-PMI achieved 64.28% and 63.62% micro-average, respectively, in comparison with using emoji Unicode description (60.13%). Excluding emoji as emotion indicators in tweet dramatically reduced the micro-average result with 46.05%. Using B-PMI based approach slightly improved multi-label classification performance in comparing with PMI-based method which shows considering rebalancing factor could potentially increase the accuracy. The least value of Hamming Loss belongs to B-PMI-based approach that indicates fraction of wrong labels is less than other methods.

Table 15 and Fig. 9 show the benefits of having balanced weighted PMI (B-PMI) comparing to other methods. Fig. 9 breaks down the f1-score for each emotion category. The F1-score for anticipation, optimism and pessimism has significantly improved; however, trust recorded zero f1-score due to availability of very few tweets.

6.2.2. Bi-LSTM with attention model (considering word embedding as features). Similar to Section 6.1, we used Bi-LSTM with attention model for the multi-label emotion classification. As shown in Table 16, the proposed method outperformed other methods. Pre-trained embedding A shows better results comparing with pre-trained embedding B, since it is specifically trained on twitter corpus.

Fig. 10 show improvements in the f1-score with B-PMI-based emoji embedding compared to other methods. F1-score noticeably increased for anger, disgust, joy, love, optimism, pessimism, and sadness; however, trust recorded zero score due to availability of very few tweets.

Describing emojis according to their Unicode description, help in classification results compared to classification without emoji

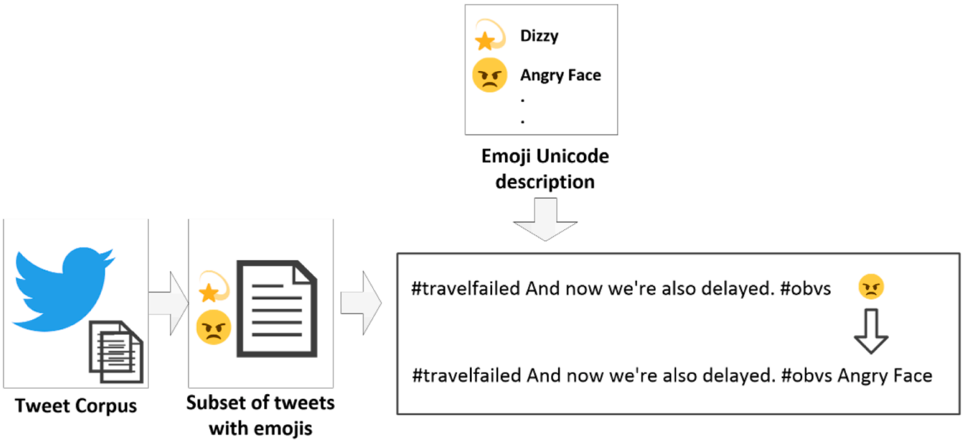

Dizzy
Angry Face
Emoji Unicode description
Tweet Corpus
Subset of tweets with emojis
#travelfailed And now we're also delayed. #obvs
#travelfailed And now we're also delayed. #obvs Angry Face

Fig. 7. Replacing emoji with its Unicode description.

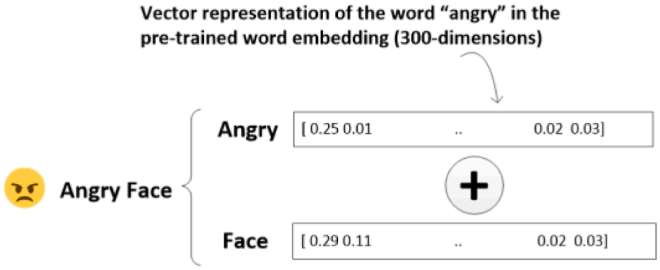

Vector representation of the word "angry" in the pre-trained word embedding (300-dimensions)
Angry Face
Angry
[0.25 0.01 .. 0.02 0.03]
+
Face
[0.29 0.11 .. 0.02 0.03]

Fig. 8. Description embedding which is deployed by taking the sum of word vectors of Emoji Unicode description.

Table 14
Results of sentiment classification considering word embedding as features (embedding approach).

		Micro Precision	Recall	F1-score
	Without Emoji + Word Embedding A	66.37%	52.38%	58.55%
	Unicode Emoji Description + Word Embedding A	68.14%	57.5%	62.37%
	Word Embedding B	68.54%	48.61%	56.88%
	Word Embedding A	**75.25%**	63.61%	68.94%
	Description embedding + Word Embedding A	71.34%	63.76%	67.34%
Proposed Method	B-PMI-based Emoji Embedding + Word Embedding A	73.21%	**67.07%**	**70.01%**

Table 15
Binary Relevance with Logistic Regression for multi-label emotion classification (using keyword-based approach to replace emoji).

	Proposed Method B-PMI-based Emoji Lexicon	PMI-based	Emoji Unicode Description	Without emoji
Micro	**64.28%**	63.62%	60.13%	46.05%
Macro	**44.99%**	44.18%	40.44%	28.72%
Hamming Loss	**16.69%**	16.85%	17.99%	18.92%
Jaccard	**52.17%**	51.32%	47.98%	31.69%

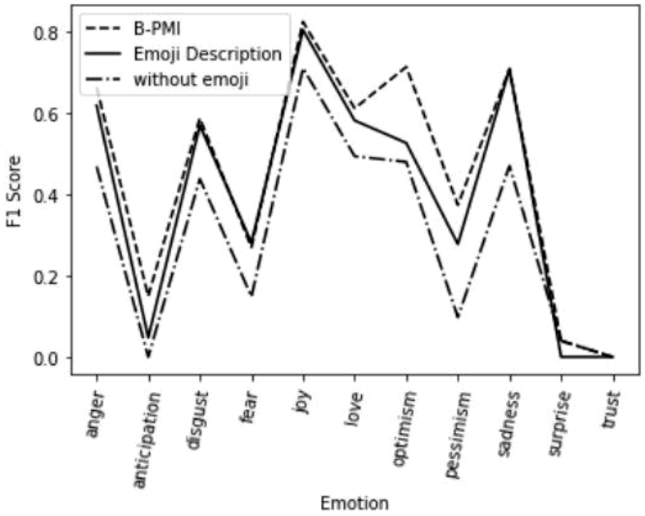


Fig. 9. F1-Score using Binary Relevance with Logistic Regression (keyword-based approach).

features. However, the results were not as high as the proposed B-PMI-based emoji embedding. Moreover, the proposed B-PMI-based emoji embedding showed better performance comparing to pre-trained word embedding A and word embedding B.

The results indicate that our method can successfully model the relationship between emoji and emotion classes. Mapping emoji to fine-grained emotions and using emoji specific embeddings can potentially improve classification results.

Table 16
Experiment results of multi-label emotion classification using word embedding and emoji embeddings (embedding approach).

	Proposed method B-PMI-based Emoji Embedding + Embedding A	Pre-trained Embedding A	Pre-trained embedding B	Description embedding + Embedding A	Unicode Emoji Description + Embedding A	Without emoji + Embedding A
Micro	**67.68%**	65.14%	63.35%	67.05%	64.34%	53.40%
Macro	**47.23%**	37.27%	44.44%	46.40%	44.49%	29.58%
Hamming Loss	**14.15%**	16.76%	16%	15.14%	16.78%	24.53%
Jaccard	**56.36%**	52.92%	51.81%	55.11%	53.27%	41.46%

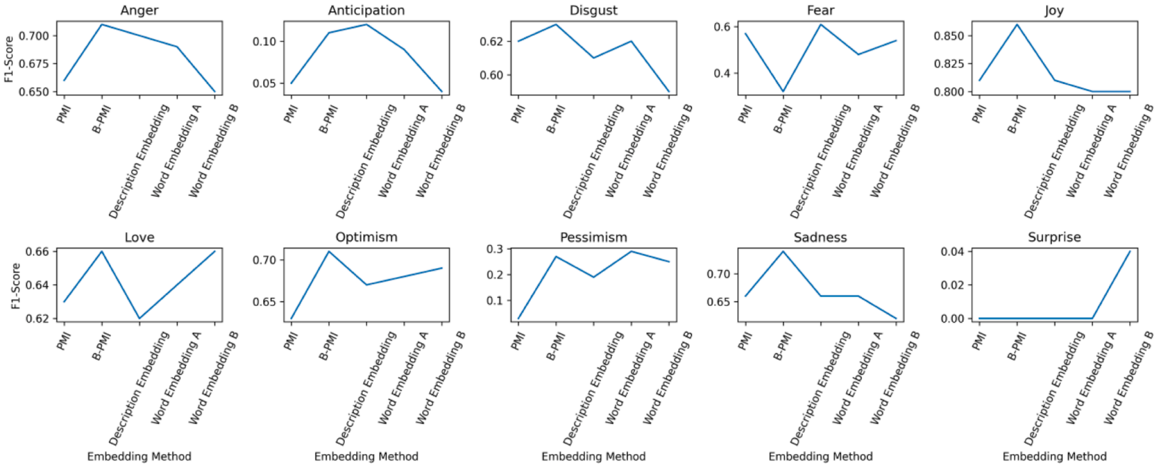


Fig. 10. Results of multi-label emotion classification per emotion class, using embedding approach (excluding trust).

The proposed emoji embedding can be used to extend existing pre-trained word embeddings since pre-trained word embeddings may not capture the emotion of emoji. The low score for trust (score of trust was zero) and surprise classes was due to very few numbers of tweets that contain emoji for these classes.

In order to get a better understanding of the performance of our method, we visualized the attention weights for each word in the tweet. The results of two example tweets are shown in Fig. 11 and Fig. 12. The color intensity corresponds to the weight assigned to each word. The weights indicate which words the model was paying attention to during the classification. As shown, high weight is given to the words and the emojis that are associated to emotions (e.g., "scariest" or "fear"). The emojis that are not present in the emoji embedding are less probable to be associated with emotions, since the existing word embeddings are trained on huge corpus containing noisy data, compared to the proposed emoji embedding that is specifically developed based on emotion labeled data. For example, the

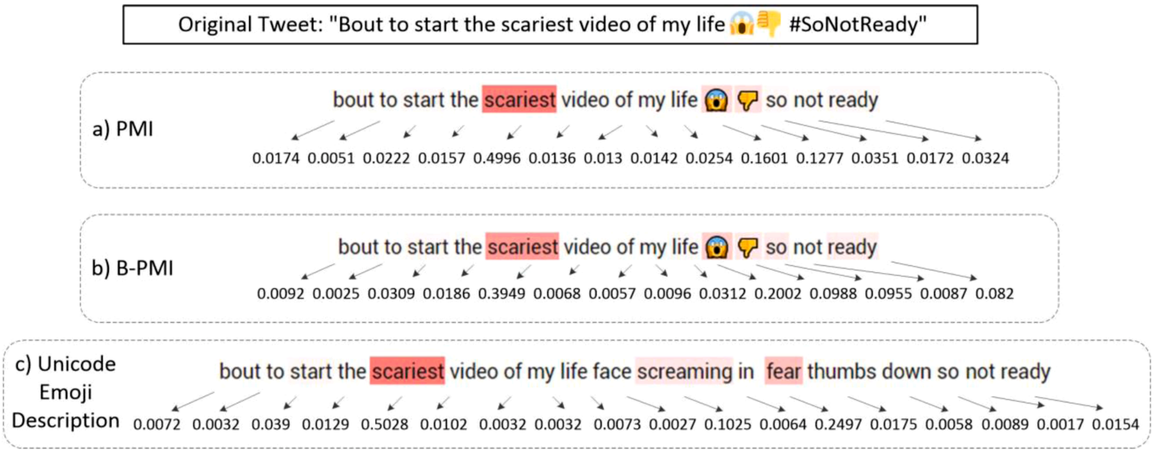


Fig. 11. Attention heat-map visualization example. Golden labels are {fear, pessimism}. Predicted labels based each method are: a){disgust, fear, sadness}, b){fear, sadness}, c){fear}.

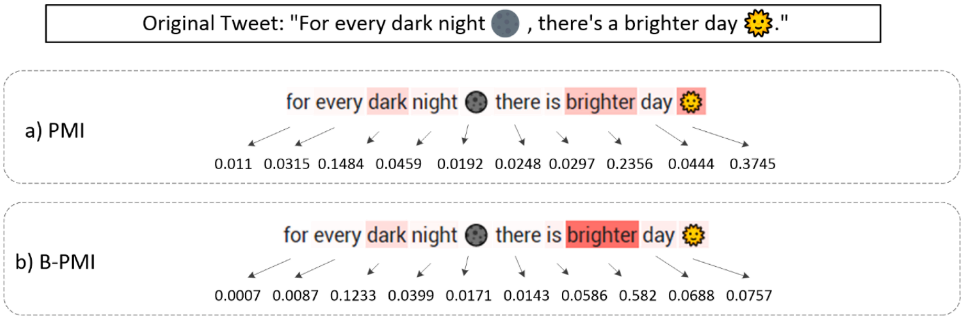

Original Tweet: "For every dark night 🌑 , there's a brighter day 🌞."
a) PMI
for every dark night 🌑 there is brighter day 🌞
0.011 0.0315 0.1484 0.0459 0.0192 0.0248 0.0297 0.2356 0.0444 0.3745
b) B-PMI
for every dark night 🌑 there is brighter day 🌞
0.0007 0.0087 0.1233 0.0399 0.0171 0.0143 0.0586 0.582 0.0688 0.0757

Fig. 12. Attention heat-map visualization example. Golden labels are {anticipation, optimism}. Predicted labels based each method are: a){joy, love, optimism}, b){joy, optimism}.

low value of weight for “moon” in Fig. 12 indicates its insignificance in the final prediction.

Therefore, the emoji features that are related to emotion are more likely to be important in the emotion classification. In some cases, the emotion of emoji may not be correctly assigned. For example, emoji of “person tipping hand 💁” is categorized as *disgust*, which associated the following sentence “Only I could be talking to a catfish on tinder 💁😊 glad I don’t use it seriously” with *disgust*, while the golden label of this sentence is *joy*.

7. Conclusion

This paper presented a method for categorizing emoji into one or more emotion classes. Instead of the using Unicode description of emoji, we proposed a method that categories emoji based on the semantic similarity between emoji and emotion classes. Considering the explosive growth of web 2.0 and the increasing use of emoji in textual content, it has become crucial to research emoji and emotion types and the approaches used for their analysis. Therefore, this paper aimed to develop a deep understanding of using emoji in emotion classification. With this in mind, we proposed model the relationship between emoji and emotion classes with a method called balanced weighted PMI (B-PMI) for dealing with imbalanced emotion classes. These improvements result in developing multi-label emoji classification and building an emoji lexicon and emoji embedding that improves the sentiment analysis and emotion classification in tweets. We validated the proposed method in two different datasets, for multi-class (positive, negative, neutral) and multi-label (11 emotion classes) classification. According to the report, the proposed method achieved the top accuracy results in comparison with the existing methods. The method obtains 3%−8% increase in micro F1-score in all datasets.

To the best of our knowledge, it is the first research that automatically classifies emoji into one or more emotion classes using a modified balanced PMI approach. Researchers and practitioners can use this method and categorize emoji based on their specific emotion classes to extend emotion related features for classification problem. Since new emoji are continuously offered to the social media platform, further research might be of interest to use n-gram features and analyze the emotion of emoji sequences.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Ahmad, S., Asghar, M.Z., Alotaibi, F.M., Awan, I., 2019. Detection and classification of social media-based extremist affiliations using sentiment analysis techniques. Human-centric Comput. Inf. Sci. 9, 24.

Al-Moslmi, T., Gaber, S., Al-Shabi, A., Albared, M., Omar, N., 2015. Feature selection methods effects on machine learning approaches in malay sentiment analysis. In: Proc. 1st ICRIL-Int. Conf. Inno. Sci. Technol.(IICIST), pp. 1–2.

Al-Moslmi, T., Omar, N., Abdullah, S., Albared, M., 2017. Approaches to cross-domain sentiment analysis: a systematic literature review. IEEE Access 5, 16173–16192.

AlMahmoud, H., AlKhalifa, S., 2018. TSim: a system for discovering similar users on Twitter. J. Big Data 5. https://doi.org/10.1186/s40537-018-0147-2 https://doi.org/.

Alswaidan, N., Menai, M.E.B., 2020. A survey of state-of-the-art approaches for emotion recognition in text. Knowl. Inf. Syst. https://doi.org/10.1007/s10115-020-01449-0 https://doi.org/.

Aman, S., Szpakowicz, S., 2007. Identifying expressions of emotion in text. In: International Conference on Text, Speech and Dialogue. Springer, pp. 196–205.

Bakhshi, S., Shamma, D.A., Kennedy, L., Song, Y., De Juan, P., Kaye, J., 2016. Fast, cheap, and good: why animated GIFs engage us. In: Proceedings of the 2016 Chi Conference on Human Factors in Computing Systems, pp. 575–586.

Bandhakavi, A., Wiratunga, N., Padmanabhan, D., Massie, S., 2017. Lexicon based feature extraction for emotion text classification. Pattern Recognit. Lett. 93, 133–142. https://doi.org/10.1016/j.patrec.2016.12.009 https://doi.org/.

Baziotis, C., Athanasiou, N., Chronopoulou, A., Kolovou, A., Paraskevopoulos, G., Ellinas, N., Narayanan, S., Potamianos, A., 2018. Ntua-slp at semeval-2018 task 1: predicting affective content in tweets with deep attentive rnns and transfer learning. arXiv Prepr. arXiv1804.06658.

Chaturvedi, I., Cambria, E., Poria, S., Bajpai, R., 2016. Bayesian deep convolution belief networks for subjectivity detection. In: 2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW). IEEE, pp. 916–923.

Cheng, X., Chen, Y., Cheng, B., Li, S., Zhou, G., 2017. An emotion cause corpus for Chinese microblogs with multiple-user structures. ACM Trans. Asian Low-Resource Lang. Inf. Process. 17 https://doi.org/10.1145/3132684 https://doi.org/.

Church, K., Hanks, P., 1990. Word association norms, mutual information, and lexicography. Comput. Linguist. 16, 22–29.

Dashtipour, K., Poria, S., Hussain, A., Cambria, E., Hawalah, A.Y.A., Gelbukh, A., Zhou, Q., 2016. Multilingual sentiment analysis: state of the art and independent comparison of techniques. Cognit. Comput. 8, 757–771.

Dice, L.R., 1945. Measures of the amount of ecologic association between species. Ecology 26, 297–302.

Dunning, T.E., 1993. Accurate methods for the statistics of surprise and coincidence. Comput. Linguist. 19, 61–74.

Eisner, B., Rocktäschel, T., Augenstein, I., Bošnjak, M., Riedel, S., 2016. Emoji2vec: learning emoji representations from their description. In: Conference on Empirical Methods in Natural Language Processing, p. 48.

Ekman, P., Friesen, W.V., 1972. Hand movements. J. Commun. 22, 353–374.

El-Naggar, N., El-Sonbaty, Y., Mohamad, A.E.-N., 2017. Sentiment analysis of modern standard Arabic and Egyptian dialectal Arabic tweets. In: 2017 Computing Conference, pp. 880–887. https://doi.org/10.1109/SAI.2017.8252198 https://doi.org/.

Erickson, B.J., Korfiatis, P., Akkus, Z., Kline, T., Philbrick, K., 2017. Toolkits and libraries for deep learning. J. Digit. Imaging 30, 400–405.

Feldman Barrett, L., Russell, J.A., 1998. Independence and bipolarity in the structure of current affect. J. Pers. Soc. Psychol. 74, 967.

Feldman, R., 2013. Techniques and applications for sentiment analysis. Commun. ACM 56, 82–89.

Fernández-Gavilanes, M., Juncal-Martínez, J., García-Méndez, S., Costa-Montenegro, E., González-Castaño, F.J., 2018. Creating emoji lexica from unsupervised sentiment analysis of their descriptions. Expert Syst. Appl. 103, 74–91. https://doi.org/10.1016/j.eswa.2018.02.043 https://doi.org/.

Gambino, O.J., Calvo, H., 2019. Predicting emotional reactions to news articles in social networks. Comput. Speech Lang. 58, 280–303. https://doi.org/10.1016/j.csl.2019.03.004 https://doi.org/https://doi.org/.

Gibaja, E., Ventura, S., 2015. A Tutorial on Multilabel Learning. ACM Comput. Surv. 47 https://doi.org/10.1145/2716262 https://doi.org/.

Go, A., Bhayani, R., Huang, L., 2009. Twitter sentiment classification using distant supervision. CS224N Proj. report, Stanford 1, 2009.

Guibon, G., Ochs, M., Bellot, P., 2018. From emoji usage to categorical emoji prediction. In: 19th International Conference on Computational Linguistics and Intelligent Text Processing (CICLING 2018).

Hauthal, E., Dirk, B., Alexander, D., 2019. Analyzing and visualizing emotional reactions expressed by emojis in location-based social media. ISPRS Int. J. Geo-Information 8. https://doi.org/10.3390/ijgi8030113 https://doi.org/.

Hochreiter, S., 1998. The vanishing gradient problem during learning recurrent neural nets and problem solutions. Int. J. Uncertainty, Fuzziness Knowledge-Based Syst. 6, 107–116.

Jabreel, M., Moreno, A., 2019. A deep learning-based approach for multi-label emotion classification in Tweets. Appl. Sci. 9 https://doi.org/10.3390/app9061123 https://doi.org/.

Jiang, S., Wilson, C., 2018. Linguistic signals under misinformation and fact-checking: evidence from user comments on social media. Proc. ACM Human-Computer Interact. 2 https://doi.org/10.1145/3274351 https://doi.org/.

Jurafsky, D., Martin, J.H., 2019. Speech and language processing (the 3nd edition draft).

Kingma, D.P., Ba, J., 2014. Adam: a method for stochastic optimization. arXiv Prepr. arXiv1412.6980.

Levy, O., Goldberg, Y., Dagan, I., 2015. Improving distributional similarity with lessons learned from word embeddings. Trans. Assoc. Comput. Linguist. 3, 211–225.

Luo, X., Liu, Z., Shang, M., Lou, J., Zhou, M., 2020. Highly-accurate community detection via pointwise mutual information-incorporated symmetric non-negative matrix factorization. IEEE Trans. Netw. Sci. Eng. 8, 463–476.

Mikolov, T., Chen, K., Corrado, G., Dean, J., 2013. Efficient estimation of word representations in vector space. arXiv Prepr. arXiv1301.3781.

Mohammad, S., 2012. # Emotional tweets. In: * SEM 2012: The First Joint Conference on Lexical and Computational Semantics–Volume 1: Proceedings of the Main Conference and the Shared Task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012), pp. 246–255.

Mohammad, S., Bravo-Marquez, F., Salameh, M., Kiritchenko, S., 2018. Semeval-2018 task 1: affect in tweets. In: Proceedings of the 12th International Workshop on Semantic Evaluation, pp. 1–17.

Mohammad, S.M., Turney, P.D., 2013. Crowdsourcing a word–emotion association lexicon. Comput. Intell. 29, 436–465.

Naskar, D., Singh, S.R., Kumar, D., Nandi, S., Rivaherrera, E.O.de la, 2020. Emotion dynamics of public opinions on Twitter. ACM Trans. Inf. Syst. 38 https://doi.org/10.1145/3379340 https://doi.org/.

Novak, P.K., Jasmina, S., Borut, S., Igor, M., 2015. Sentiment of emojis. PLoS ONE 10. https://doi.org/10.1371/journal.pone.0144296 https://doi.org/.

Orkphol, K., Yang, W., 2019. Word sense disambiguation using cosine similarity collaborates with Word2vec and WordNet. Futur. Internet 11, 114.

Pennington, J., Socher, R., Manning, C.D., 2014. Glove: global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1532–1543.

Plutchik, R., Kellerman, H., 1980. Emotion, theory, research, and Experience. Academic press.

Pohl, H., Domin, C., Rohs, M., 2017. Beyond just text: semantic emoji similarity modeling to support expressive communication. ACM Trans. Comput. Interact. 24 https://doi.org/10.1145/3039685 https://doi.org/.

Raza, A.A., Habib, A., Ashraf, J., Javed, M., 2019. Semantic orientation based decision making framework for big data analysis of sporadic news events. J. Grid Comput. 17, 367–383. https://doi.org/10.1007/s10723-018-9466-y https://doi.org/.

Saif, H., He, Y., Alani, H., 2012. Semantic sentiment analysis of twitter. In: International Semantic Web Conference. Springer, pp. 508–524.

Salton, G., Buckley, C., 1988. Term-weighting approaches in automatic text retrieval. Inf. Process. Manag. 24, 513–523.

Sintsova, V., Pu, P., 2016. Dystemo: distant supervision method for multi-category emotion recognition in tweets. ACM Trans. Intell. Syst. Technol. 8 https://doi.org/10.1145/2912147 https://doi.org/.

Strapparava, C., Valitutti, A., 2004. Wordnet affect: an affective extension of wordnet. In: Lrec. Lisbon, p. 40.

Tsoumakas, G., Katakis, I., 2007. Multi-label classification: an overview. Int. J. Data Warehous. Min. 3, 1–13.

Turney, P.D., 2001. Mining the web for synonyms: PMI-IR versus LSA on TOEFL. In: European Conference on Machine Learning. Springer, pp. 491–502.

Urabe, Y., Rzepka, R., Araki, K., 2021. Find right countenance for your input—Improving automatic emoticon recommendation system with distributed representations. Inf. Process. Manag. 58, 102414.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. In: Advances in Neural Information Processing Systems, pp. 5998–6008.

Vora, P., Khara, M., Kelkar, K., 2017. Classification of tweets based on emotions using word embedding and random forest classifiers. Int. J. Comput. Appl. 178, 1–7.

Wikarsa, L., Thahir, S.N., 2016. A text mining application of emotion classifications of Twitter's users using Naïve. In: Bayes method International Conference on Wireless & Telematics.

Wood, I., Ruder, S., 2016. Emoji as emotion tags for tweets, in: Proceedings of the Emotion and Sentiment Analysis Workshop LREC2016, Portorož, Slovenia. pp. 76–79.

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.

Yu, S., Zhu, H., Jiang, S., Zhang, Y., Xing, C., Chen, H., 2019. Emoticon analysis for Chinese social media and e-commerce: the azemo system. ACM Trans. Manag. Inf. Syst. 9 https://doi.org/10.1145/3309707 https://doi.org/.

Zhang, X., Li, W., Chen, X., Lu, S., 2018. MoodExplorer. In: Proc. ACM Interactive, Mobile, Wearable Ubiquitous Technol., 1, pp. 1–30. https://doi.org/10.1145/3161414 https://doi.org/.

Zhou, G., Fang, Y., Peng, Y., Lu, J., 2019. Neural conversation generation with auxiliary emotional supervised models. ACM Trans. Asian Low-Resource Lang. Inf. Process. 19 https://doi.org/10.1145/3344788 https://doi.org/.