Multimodal Emotion Classification

Anurag Illendula
Department of Mathematics
Indian Institute of Technology Kharagpur
Kharagpur, India
aianurag09@iitkgp.ac.in

Amit Sheth
Kno.e.sis Center
Wright State University
Dayton, Ohio
amit@knoesis.org

ABSTRACT

Most NLP and Computer Vision tasks are limited to scarcity of labelled data. In social media emotion classification and other related tasks, hashtags have been used as indicators to label data. With the rapid increase in emoji usage of social media, emojis are used as an additional feature for major social NLP tasks. However, this is less explored in case of multimedia posts on social media where posts are composed of both image and text. At the same time, w.e have seen a surge in the interest to incorporate domain knowledge to improve machine understanding of text. In this paper, we investigate whether domain knowledge for emoji can improve the accuracy of emotion classification task. We exploit the importance of different modalities from social media post for emotion classification task using state-of-the-art deep learning architectures. Our experiments demonstrate that the three modalities (text, emoji and images) encode different information to express emotion and therefore can complement each other. Our results also demonstrate that emoji sense depends on the textual context, and emoji combined with text encodes better information than considered separately. The highest accuracy of 71.98% is achieved with a training data of 550k posts.

CCS CONCEPTS

• **Human-centered computing → Human computer interaction (HCI)**; *Social media*;

KEYWORDS

Emoji Understanding, Emotion Classification, Multimodal Analysis

ACM Reference Format:
Anurag Illendula and Amit Sheth. 2019. Multimodal Emotion Classification. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3308560.3316549

1 INTRODUCTION

Emotion is any conscious experience characterized by intense mental activity and a certain degree of pleasure or displeasure. It primarily reflects all aspects of our daily lives, playing a vital role in our decision-making and relationships. In recent years, there have been a growing interest in the development of technologies to recognize emotional states of individuals. Due to the escalating use of social media, emotion-rich content is being generated at an increasing rate, encouraging research on automatic emoji classification techniques. Social media posts are mainly composed of images and captions. Each of the modalities has very distinct statistical properties and fusing these modalities helps us learn useful representations of the data [3]. Emotion recognition is a process that uses low-level signal cues to predict high-level emotion labels. With the rapid increase in usage of emojis, researchers started using them as labels to train classification models [6]. A survey conducted by secondary school students suggested that the use of emoticons can help reinforce the meaning of the message [1]. Researchers found that emoticons when used in conjunction with a written message, can help to increase the "intensity" of its intended meaning [5].

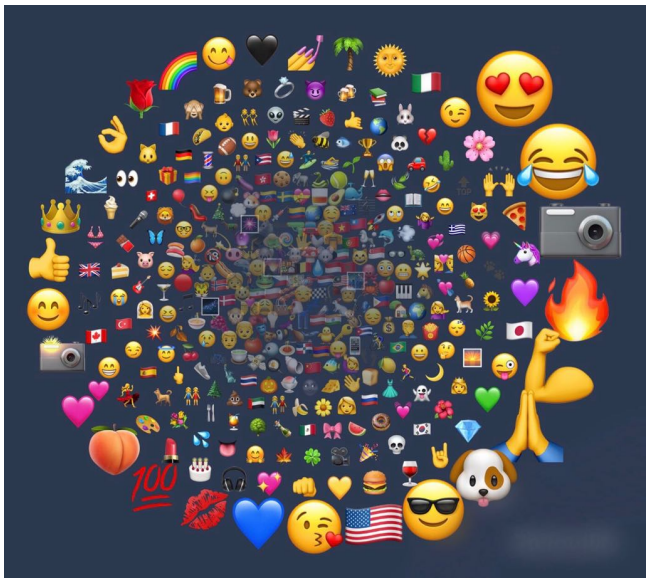

Figure 1: Spiral representing usage of different Emojis on Instagram. The image is copied from https://bit.ly/2W3ks5u, Article by Stefan Pettersson

Emojis are being used for the visual depictions of human emotions [23]. Emotions help us to determine the interactions among human beings. The context of emotions specifically brings out the complex and bizarre social communication. These social communications are identified as the judgment of other personś mood based on his emoji usage [23]. According to a study made by Rajhi et al. [23], the real-time use of emojis can detect the human emotions in different scenes, lighting conditions as well as angles in real time. Studies have shown that emojis when embedded with text to express emotion make the tone and tenor of the message

[1] https://bit.ly/2NyyiIp

WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA

ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3316549

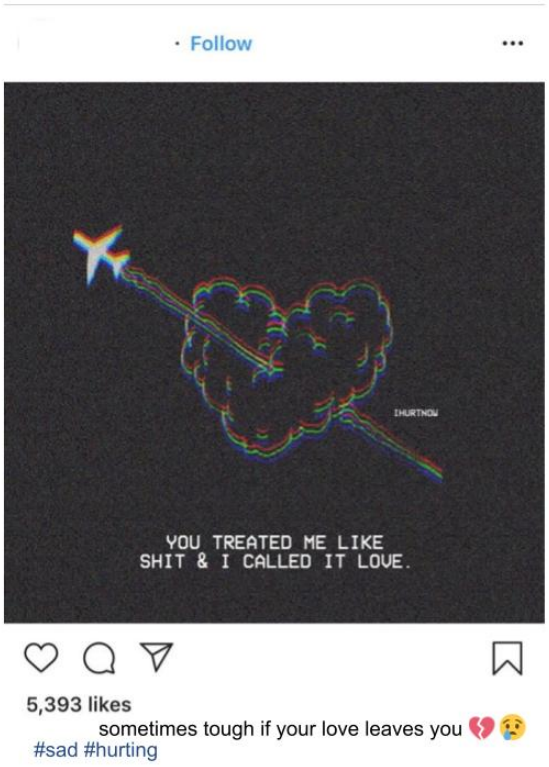

Figure 2: Example of an Instagram post which belongs to "Sad" emotional category in our dataset

clearer. This further helps in reducing or eliminating the chances of misunderstanding, often associated with plain text messages[2]. Recent study proved that co-occurrence helps users to express their sentiment more effectively[3] [13].

Psychological studies conducted in the early '80s provide us strong evidence that human emotion is closely related to the visual content. Images can both express and affect peopleś emotions. Hence it is intriguing and important to understand how emotions are conveyed and how they are implied by the visual content of images. With this as a reference, many computer scientists have been working to relate and learn different visual features from images to classify emotional intent. Convolutional Neural Networks (CNNs) have served as the baselines for major Image processing tasks. These deep CNNs combine the high and low-level features and classify images in an end-to-end multi layer fashion.

Earlier most researchers working in the field of social NLP have used either textual features or visual features, but there are hardly any instances where researchers have combined both these features. Recent works by Barbieri et al.'s [3], Illendula et al.'s [12] on multimodal emoji prediction and Apostolova et al.âĂŹs [1] work on information extraction fusing visual and textual features have shown that combining both modalities helps in improving the accuracies. While a high percentage of social media posts are composed of both images and caption, the researchers have not looked at the multimodal aspect for emotion classification. Consider the post in Firgure 2 where a user is sad and posts an image when a person close to him leaves him. The image represents a disturbed heart and has a textual description "sometimes tough if your love leaves you #sad #hurting" conveys a sad emotion. Similarly the emoji used 💔, 😢 conveys the emotion of being depressed. We hypothesize that all the modalities from a social media post including visual, textual, and emoji features, contribute to predicting the emotion of the user. Consequently, we seek to learn the importance of different modalities towards emotion prediction task. We first discuss relevant research in Section 2. The development of our dataset and present preliminary data analysis experiments which illustrate the importance to study the usage of emojis in different emotional contexts is described in Section 3. We present our model and approach of multimodal emotional classification in Section 4. The results from different experiments and relevant discussion are in Section 5 and Section 6, respectively. We conclude with some interesting findings and our future plan in Section 7.

2 RELATED WORK

Most NLP tasks are limited to the scarcity of labeled data. Earlier many researchers have used manual annotation technique to evaluate their models, but this requires much understanding of the emotional content of all the expressions, is time-consuming, requires much effort and may differ according to one's perspective. This creates a misinterpretation of emotion and effects the accuracies of respective tasks. Hence in most social NLP related tasks namely sentiment analysis and other emotion classification tasks hashtags were used as features for automatically labeling data to corresponding categories [6, 30]. However, the rapid increase in the usage of emojis on social media helped researchers use emoticons as features for data labeling. Felbo et al. [6] has introduced a transfer learning approach for emotion classification, sentiment analysis, sarcasm identification through emoji prediction on the text.

Using emoji knowledge for sentiment analysis, emotion classification and related tasks is not a new idea. Researchers have started using Emojinet [31] to learn embeddings for sentiment analysis task and have achieved better accuracies than the previous state of the art emoji embeddings [32]. There have been many approaches which use emoji as a feature to classify sentiment on social media using emoji as a feature, Illendula et al. [13] has used emoji co-occurrence has a feature to learn sentiment features of emojis. Also, emojis have also been a very important feature to classify emotional content; previous research has always manually specified which emotional category each hashtag or emoji belongs to [22]. Prior research work has used theories of emotion such as Ekman's six basic emotions, Pluchtik's eight basic emotions, and other psychology works [21, 26, 28]. Shaver et al.'s [26] work which is the most used and most cited work helped to automatically label tweet without any human intervention to take care of biases and misinterpretation to corresponding emotional label and develop an emotion classification model for text [30].

Content-Based Information Retrieval (CBIR) is the historical line of research in multimedia tasks. This task usually deals with retrieving images in the dataset that are most similar to the query image. Bag-of-Words representation [34] has seen a sustained line

[2]https://bit.ly/2AV1sNA
[3]https://bit.ly/2QW91sx

Table 1: Emotion words used for collecting tweets and the statistics of instagram posts in our dataset after filtration

Emotion	# of Hash Tags	Examples of Hash Tags	# of Instagram Posts with Emoji	# of Instagram Posts without Emoji	Number of Instagram Posts	Percentage of Posts with Emoji
Anger	23	aggravation, irritation, agilation, anger	59458	28549	88007	67.56
Fear	22	shock, fear, fright, terror, panic	44511	28078	63644	61.32
Joy	36	excited, happy elated, proud	55566	29947	85513	64.98
Love	17	affection, love, loving, fondness	54156	31596	89486	56.38
Sad	36	sorrow, unhappy, depressing, lonely	54156	31596	85752	63.16
Surprise	5	amazement, surprise, astonishment	45640	17230	72870	62.25
Thankful	2	thankfulness, thankful	42870	31308	74178	57.80
Overall	141		352649	205746	558395	63.16

of research in this task which has been effective up to million sized image datasets. This Information Retrieval task has eventually led to research in image classification which is one of the fundamental challenges in Computer Vision. Convolutional Neural Networks (CNNs) [16] have shown promising results in image classification. Research on shortcut connections has been an emerging topic since the development of multilayer perceptrons which has shown promising results for Image processing tasks. Generally, these multiple layers have been connected using shortcut connections using gated functions [10]. In image classification, depth of the network, i.e., number of layers within the network is of crucial importance as noted by Simonyan et al. [27]. Increasing the depth can have an adverse effect on the image classification task. Most notably, vanishing/exploding gradients problem [7] and the degradation problem [8] are of significant importance. These problems are overcome by the introduction of shortcut connections and residual representations introduced in Residual Networks [9] which won the 1st place in ImageNet 2015 Image classification competition [4]. Residual Networks and its extensions which consist of many residual units have shown to achieve state-of-the-art accuracy for image classification tasks on datasets such as ImageNet [24].

Knowledge graphs have been proved to be on the important addition to data for training machine learning and deep learning model architectures for various Natural language and Image Processing tasks. Recent research in NLP Emojinet has improved the accuracies of emoji similarity [32], learning emoji embeddings, emoji sense disambiguation [31]. Emojinet has also improved the accuracies of emoji prediction in case of images [12]. In the case of image processing tasks, knowledge graphs have also improved the task of video captioning [29] and other related tasks [19].

In this paper, we present an emotion classification approach using techniques in deep learning leveraging the three modalities from social media posts namely textual, visual and emoji features. We use state of the art approaches for learning features from three modalities namely ResNet model architecture, attention mechanism, Bag of words model for visual, textual and image features respectively. We used two different word embeddings namely FastText (word embedding model which is capable of capturing sub word information) [14] model trained on the set of the processed captions (we explain the pre-processing steps involved in Section 3) and pretrained fast text model trained on Wiki corpus [20]. We use the bag of words model developed by Wijeratne et al. [32] to learn the emoji embeddings using the emoji knowledge concepts extracted from Emojinet. We report the results observed considering different emoji knowledge concepts from EmojiNet namely emoji names, emoji senses, emoji sense definitions, and three different word embedding models. We also discuss the importance of different modalities towards emotion classification and report our results and observations in further sections.

3 DATASET DEVELOPMENT

3.1 Instagram Dataset

Instagram is one of the major platforms where people share multimedia posts composed of images and text in the form of description. Hence Instagram provides us an emotion-rich content to analyze different user emotions using machine learning and deep learning techniques. We have used the same approach as Wang et al. [30] to collect labeled data. Following Shaver et al. [26] they collected set of possible hashtags for six different emotions (for example words amazement, surprise, astonishment convey the emotion "Surprise"). This approach to collect emotion-labeled posts has been proved to be effective to label text. Shaver et al. have constructed an emotion-related tree where the first layer represents the parent category, i.e., the six different emotion, the second layer consists of 25 emotion-related keywords categorized into six different parent emotions. We have added a seventh emotional category different from the above mentioned emotional categories named "Thankful". We then checked posts having hashtags #gratified, #humbled, #blessed, #thankful on instagram, because in general these key words mean thankful according to english literature. But there are

[4] https://bit.ly/2y4J8Cz

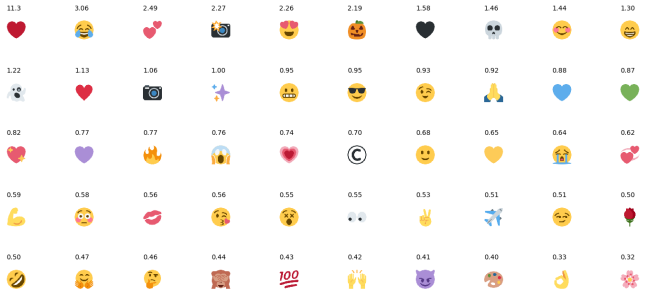

Figure 3: Percentage of number of posts having different types of emojis embedded in their post description.

hardly any posts having a hashtag #gratified (around 4k posts in total and no posts in our analysis period). We then checked posts having hashtags #humbled and #blessed but several posts convey a Joy emotion rather than Thankful emotion. Prior research of study of emotion related key words convey that # blessed could be considered as happy emotion but not necessarily Thankful or Joy or Love emotion is particular [18]. Hence we considered only the key words reported by Wang et al. for the seventh emotional category namely #thankful and #thankfulness for Thankful emotion. We then followed the pre-processing steps for emotion-related keywords by adding lexical variants of a word and removing ambiguous words. Table 1 gives a complete idea of the set of hashtags along with the number of posts collected from Instagram. We then used Instalooter[5] API which is an open source tool to collect public posts from Instagram. Instalooter API gives us access to HashTaglooter method similar to Twitter's hashtag search which helps us to collect multimedia posts having a particular hashtag. For example, we collected the posts having "#amazement" and labeled the post to Surprise emotional category. Similarly, we extracted the emotion-labeled multimedia posts for other emotion labels.

3.2 Pre Processing

In this section, we explain the different pre-processing steps to ensure the quality of our Instagram dataset. In total, we have collected about 1.1 million posts of different languages from Instagram using instalooter API posted between 1st December 2018 to 20th December 2018, a duration of 20 days. The preprocessing of this corpus consisted of a set of filtering, followed by annotation, as discussed next.

Filtering:

- Our current approach is limited to English language only, so we filtered the dataset for the posts that used English language only.
- We also ensured that the multimedia posts contain only images but not videos. Instagram also allows users to post multiple images in the same post, so we have filtered them and also considered posts containing one image but not multiple images.
- The next step of filtration involved removal of posts which have multiple hashtags belonging to different emotional content. For instance, user's have used surprise and joy in the same post, and we couldn't categorize this post to a single emotional category.
- We also made sure that we remove all the emotion-related hashtags, for instance, users may have used #amazement and #surprise in the same post. Hence we removed all the emotion-related hashtags which are present in our lexicon.
- We then removed posts having URL's embedded in the textual description section. We hypothesize that the content obtained from the URLs are likely to be important for understanding the emotional content of the post.

Annotation and Text Processing:

- Each post was annotated first by removing the emotion-related keywords from the textual description with the corresponding emotion label.
- Suppose, if the description of the posts has #bliss, then we removed the emotion-related hashtag and then labeled the post to "Joy" emotional category, if a hashtag is not an emotion-related keyword we converted the hashtag to corresponding word ("#car" to "car").
- We also followed some text pre-processing steps like removing characters repeated more than two times in a word (e.g., looooool to lol).
- We replaced user mentions to "USER" to anonymize users.

3.3 Explorational Data Analysis

Leveraging emoji as an additional feature has improved accuracies of various socialNLP tasks. Examples include improved user classification based on marijuana usage [17], better twitter street gang member identification [2], and higher accuracy in sarcasm and emotion identification on Twitter [6]. A recent study by Scott Ayres on emoji usage on the Instagram platform reports that 31% of all image posts and 29% of all video posts contained emojis [6]. Hence, we have analyzed the usage of emojis in each emotional category. Table 1 represents the number of posts having emojis embedded in the textual description out of the complete set of posts under each emotional category. We observed that about 63.16% of posts contain emojis. In many cases, users use them to express their emotion which cannot be conveyed in the form of text. Hence we explore the use of domain knowledge of emojis from EmojiNet[7] in our multimodal emotion classification task and reported our results in further sections.

Figure 3 shows the percentage of posts having different emojis. We have found that users used more than 1800 different types of emojis in their posts, the emoji ❤️ which is the most frequently used emoji, is used in only 11.3% posts, which is about $1/9^{th}$ of the total number of posts. We have then looked into the use of emojis to express a particular emotion. Table 2 reports the five most frequently used emojis to express a particular emotion. Consequently we looked at the sense forms of these emojis from EmojinNet, and the results looked convincing. For instance, the most frequently used emojis under the emotional category "Sad", 💔, 😔, 😭, 😢, 🥀

[5]https://bit.ly/2RGBYxF

[6]https://bit.ly/2E7rO1L

[7]https://bit.ly/2HiSRIq

Table 2: The five most frequently used emojis to express a particular emotion

Emotion \ Emojis	1	2	3	4	5
Anger	🖤	🔥	✨	😠	💔
Fear	🎃	👻	💀	😱	😬
Joy	❤️	😍	😊	💜	🎵
Love	❤️	😍	💕	😘	💙
Sad	💔	😔	😭	😢	🥀
Surprise	😲	😍	❤️	😊	✨
Thankful	❤️	💕	🙏	😊	💗

have sense forms which are closely related to sad or depressed. Similarly, we have studied emoji usage in other emotional categories and observed that the emojis corresponding to each emotional category have sense forms related to the emotional category. This supports our hypothesis that emojis play a vital role in relating to users' emotion.

4 PRE TRAINING

We collected set of all user captions from each post and followed pre-processing steps and trained a fastText word embedding model. We chose fasttext over other conventional word embedding models due to its capability of capturing sub-word information which plays a significant role in social NLP tasks. Also fasttext word embedding model has proved to give greater accuracies compared to other word embedding models over various NLP tasks such as sentiment analysis, emotion detection, sarcasm identification [14], and emoji prediction [12].

As we have observed that the emotional intent of the emoji used and the emotion expressed by the user in the post are parallel, we tried to use EmojiNet which lists 21,000 emoji sense definitions and 4,618 emoji sense forms to 2,389 emojis. Emoji sense forms list the possible sense forms related to each emoji and emoji sense definitions explaining the context of use of emojis [31]. We chose to use the Bag of words model developed by Wijeratne et al. [32] to learn the emoji representations in the same vector space as words.

5 MODELS

In this section, we present and motivate the models that we used to learn features from an Instagram post composed of a picture and the associated comment for emotion classification task.

5.1 ResNets

ResNet architecture [9], composed of convolution layers is currently the state of the art for image classification tasks achieving high accuracies[24]. Before the introduction of ResNet's, deep CNNs were used for most image processing tasks. ResNet is a feed forward CNN that utilizes two or more convolution layers [25]. It has been observed that the accuracy of image classification majorly depends on the depth of the network, i.e., the number of layers which can be related to the number of layers within the network [7]. Hence

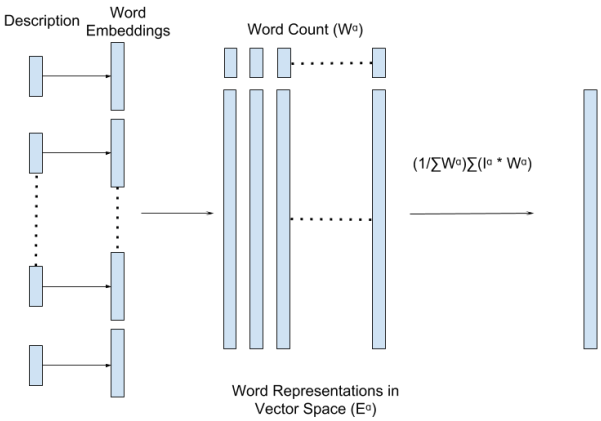

Figure 4: Bag Of Words model to learn emoji representation in the similar vector spaces as words.

we used different types of ResNet model architectures, namely ResNet-101 and ResNet-152 to check the classification accuracy in different cases. ResNet model architecture was first tested on the ImageNet dataset where the number of classes for classification was 1000. Since our task of emotion classification requires us to learn features from both text and image, we use the ResNet model to learn a feature vector from an image. We use the implementation of ResNet by Kotikalapudi which is open-sourced and available on github[8]. Table 3 reports the precision, recall, macro F1 score observed using different ResNet model architectures.

5.2 Bag Of Words Model

As discussed earlier there are different word embedding models which learn rich representations of words in vocabulary using neural networks. The neural network takes a large n-dimensional vector for each word (where n is the number of distinct words in the vocabulary), then learns a transformation of the vector in a low dimensional vector space. There have been many word embedding models to learn word representations including GloVe, skip-gram and CBOW. Similarly, there are different approaches to learn emoji embeddings that use co-occurrence feature [13], skip-gram word embedding architecture [4], and semantic knowledge of emojis from EmojiNet [32]. Since EmojiNet gives us access to the emoji sense forms, we make use of the embedding model developed by Wijeratne et al. [32] to learn emoji embeddings. We also check the accuracy of our approach using other emoji embeddings. Wijeratne et al. [32] replaced the word vectors of all words in the emoji definition and formed a 300-dimensional vector performing vector average. Also, the vector mean (or average) adjusts for word embedding bias that could take place due to certain emoji definitions having considerably more words than others has been noted by Kenter et al. [15].

We make use of different knowledge concepts, namely sense forms, sense definitions, emoji names from EmojiNet and use the bag-of-words model (Figure 4), fastText trained word embeddings to

[8]https://bit.ly/2W69fRC

learn the emoji representations. Since we use the word embeddings to learn emoji representation, we could say that both emojis and words are embedded on a similar vector space. We define two types of knowledge embeddings termed as Emoji_Embeddings _Definitions, Emoji_Embeddings SenseLabels learned using emoji sense definitions and emoji sense forms extracted from EmojiNet respectively. Then we evaluate our model using these embeddings as external knowledge concepts.

Emoji_Embeddings _Definitions: Emoji Definitions are the textual descriptions that relate to the context of use of particular emoji. The emoji embedding from the set of descriptions is calculated using the bag-of-words model shown in Figure 4. For example, consider the emoji 😎, EmojiNet lists "One of the temperate seasons, Summer is the warmest of the four temperate seasons, falling between spring and autumn.", "The period or season of summer." and so forth as emoji definitions for the emoji (😎).

Emoji_Embeddings _SenseLabels: Emoji Sense labels are the list of different senses what emoji mean in different contexts. The emoji embedding from the set of Sense labels is calculated following the bag-of-words model shown in Figure 4. For example "amusing", "swagger" and so forth are the sense labels listed by EmojiNet for the emoji (😎).

Let C_i represent the word count of word W_i, E_i represent the word embedding of word W_i; then the emoji embedding can be calculated as :

$$Emoji\_Embedding = \frac{\sum E_i * C_i}{\sum C_i}$$

5.3 Attention Model

Recurrent neural networks (RNNs) are a class of neural networks that take sequential data or time series data as input and compute hidden state vector at each time step, and these networks make use of the entire history of inputs to compute hidden state vector. A LSTM network is a special class of RNNs which has a memory cell and three gating units. These three gating units – input gate, forget gate, and output gate, allow the model to control what information to add, what information to use from previous history, and what information to use to output from current memory cell, respectively. Each gate, implemented as a logistic function σ, takes input vector and outputs a value between 0 and 1.

Bi-directional LSTM's are special kind of LSTM's which combine two LSTM running in forward and backward directions. The hidden state vector $\vec{h}_t$ is computed by the forward LSTM, $\overleftarrow{h}_t$ represent the hidden state vector computed by the backward LSTM. Then we compute the hidden state vector h_t concatenating both the hidden state vector, $h_t = [\vec{h}_t : \overleftarrow{h}_t]$. These networks have been shown to be efficient for a wide range of NLP tasks and have improved the accuracy over the LSTM networks because the current hidden state vector is computed using the past and future information. The importance of a word is highly context-dependent, i.e., the same word can have a different degree of importance in different context. To incorporate this perspective, we add an attention layer [33] on top of this encoded bidirectional LSTM which helps the model decide the importance of each word for the emotion classification task. For instance, words like "lovely" or "extraordinary" are likely to add more weight to the emotion carried in the text. The attention layer helps the model learn the importance of each word using attention scores.

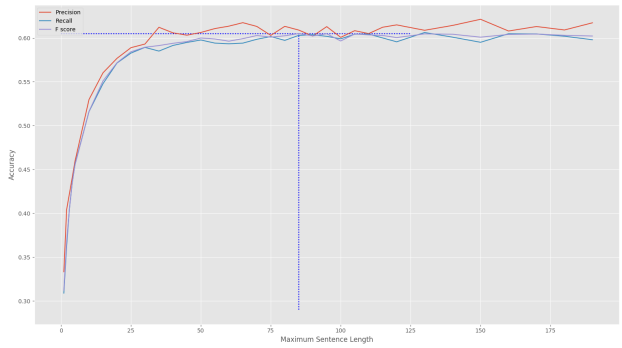

Figure 5: Variation of Accuracy measures vs the input sentence length

Table 3: Accuracies of emotion classification task using visual features from images

Resnet Architecture	Accuracy	Precision	F score
Resnet-101	25.2	25.8	25.4
Resnet-152	**29.6**	**29.8**	**29.7**

6 EXPERIMENTS

In this section, we report the accuracies of different experiments which we performed. We first check the classification accuracy when each modality is used independently, then we combine these modalities and check the classification accuracy of multimodal features towards the emotion classification task. All models are trained using the Keras library and run on Theano background on a cuda GPU, using Adam's gradient descent optimizer. For all experiments, models are trained using 60% of the dataset, validated on 20% of the dataset and tested on the remaining 20% of the dataset.

6.1 Visual Features

We have used the 152-layered residual network which is the current state of the art model architecture for image classification. Also the state of the art model architectures for emoji prediction task for images have used ResNet architecture [3, 12]. Here we trained a 152-layered residual network with learning rate as 0.0001 and stopped the training when there is no further increase in accuracy on the validation set. However, we also checked the classification accuracy using other ResNet model architectures.

6.2 Emoji features

As seen in Table 2, emojis play an integral part in predicting user's emotions on social media. As discussed earlier, we use the Bag of Words model developed by Wijeratne et al. [32] to learn emoji embeddings using emoji knowledge concepts extracted from EmojiNet. We learn four different emoji embeddings using emoji names, emoji

Table 4: Accuracies of emotion classification task using emoji features from caption and fasttext trained word embeddings on the set of descriptions

Knowledge Concept Embeddings	Precision	Recall	Macro F1 Score
Fasttext trained Emoji Embeddings	20.13	19.87	19.43
Emoji Names	27.55	26.67	25.83
Emoji Sense Forms	**34.81**	**30.82**	**32.83**
Emoji Sense Definitions	29.16	27.06	27.38
Processed Emoji Sense Definitions	30.39	28.56	29.66

sense forms, emoji sense definitions, and processed emoji sense definitions. Processed emoji sense definitions are the sense definitions formed after following pre-processing steps, namely removal of stop words, lemmatizing words on the emoji sense definitions extracted from EmojiNet. We consider the vector average of the emoji embeddings of emojis emb2dded in each post as the emoji feature vector for each post. For example if a post is embedded with the emoji set (💔, 💔, 😔), then we take $Feature\_Vector$ = 1/3(vec(💔) + vec(💔) + vec(😔)). Next we train a neural network using these feature vectors as input and encoded emotion vector as labels.

6.3 Textual Features

It has been proved that textual information is more appropriate than visual features and increases the accuracy of emoji prediction in case of multimodal information [3]. Hence we tried to check the importance of textual features for emotion classification task. We use the attention mechanism to learn the importance of words towards emotion classification task. Instagram allows users to post a caption of maximum 2200 characters which translates to approximately 300 words [9]. It would be challenging to train a Bi-LSTM with such long input sentence length, we experimented on different input sentence lengths to check the variation of accuracy of Bi-LSTM on sentence length. Figure 5 gives the clear picture of variation of accuracy measures namely precision, recall and F score with input sentence length to the Bi-LSTM using word embeddings trained using fasttext on the set of captions. We have observed that maximum accuracy for classification is achieved when the input sentence length is nearly 80 words per caption and thereafter there is no significant rise in the accuracy of the model. We also reported our accuracies of the emotion classification task using the bag of words model (which is also a extensively used approach for long input sentences) for text using different word embeddings, and we have observed that attention mechanism has achieved better accuracies than the bag of words model. This can be due to the reason that bag of words model gives equal importance to all the words which is not the case with attention mechanism. Table 5 and Table 6 report the accuracies for emotion classification task using different word embeddings and using different model architectures attention mechanism and bag of words model respectively.

[9]https://bit.ly/2RAg2zu

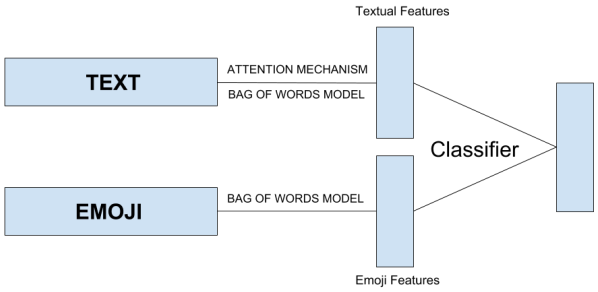

Figure 6: Multimodal emotion classification approach using caption of a post.

6.4 Combining both Textual and Emoji features

Here we present two different model architectures: one which considers complete caption (emojis embedded in between text) as sequential input and fed to a single input layer, while the other considers text and emoji as different input features and are fed to two different input layers as illustrated in Figure 6.

6.4.1 Bag Of Words approach for whole caption. Here we consider complete caption as the input to the classification model, and learn the feature vector by calculating vector average of the embeddings of entities, i.e., word embeddings of words and emoji embeddings of emojis. Since the emoji embeddings are learned using the word embeddings, this supports the addition of word embeddings and emoji embeddings to calculate vector average. Table 8 reports the accuracies of emotion classification task using bag of words model with different emoji knowledge concepts and different word embedding models.

6.4.2 Attention mechanism for whole caption. According to the study by Kevin Cohn which says emojis can be used as language on social media[10] it is noted that the sense of emoji depends on the context of use [31]. Hence Bi-LSTM which have been the SOTA for sequential data would capture the sense of the emoji in the textual context with the help of recurrent units. Hence we train an attention model to check the accuracy where complete caption (emoji+text) is fed to the input layer of a bi-lstm. Table 7 reports the accuracies of emotion classification task using attention mechanism on whole caption.

6.4.3 Considering emoji and text as different features. Recent research by Wijeratne et al. has proved that sense of emoji depends on the textual context [31]. To prove this assertion in the case of emotion classification we use caption without emojis (only text) and emojis embedded in the caption as different features to train classification model. Here we feed the textual input to the attention layer, feature vector learned using emoji embedded in caption to different input layers. We then merge the outputs of attention layer and emoji input layer and add a softmax layer on top of this for classification. The classification accuracies for this model architecture are reported in Table 9. Figure 6 illustrates the model architecture

[10]https://bit.ly/2UbmTRG

Table 5: Accuracies of emotion classification task using textual features from text and attention mechanism at different caption length

Caption Length	Word Embedding Model	Precision	Recall	Macro F1 Score
20	Fasttext trained on Post Descriptions	57.67	57.14	57.17
20	Pretrained Fasttext on Wiki Corpus	41.84	38.92	40.24
40	Fasttext trained on Posts Descriptions	60.59	59.15	59.39
40	Pretrained Fasttext on Wiki Corpus	46.83	42.39	43.20
80	Fasttext trained on Posts Descriptions	61.33	59.72	60.25
80	Pretrained Fasttext on Wiki Corpus	45.95	44.88	43.05

Table 6: Accuracies of emotion classification task using textual features from text and Bag Of Words model

Word Embedding Model	Precision	Recall	F1 Score
Fasttext trained on Posts Descriptions	56.57	55.82	54.50
Pretrained Fasttext on Wiki Corpus	44.26	43.57	43.47

in this case, where we learn independent features and then merge the features.

6.5 Combining textual, visual and emoji features

Barbieri et al. [3] showed that there would be an increase in prediction accuracies if both visual and textual features are combined. We investigated the same in the context of multimodal emotion classification using SOTA model architectures, i.e., ResNet-152 for learning image features, attention mechanism to learn features from text and merge these features to form a hidden layer. We then use a softmax layer on top of this for classification. We report the classification accuracies using this model architecture in Table 10. Figure 7 shows the model architecture.

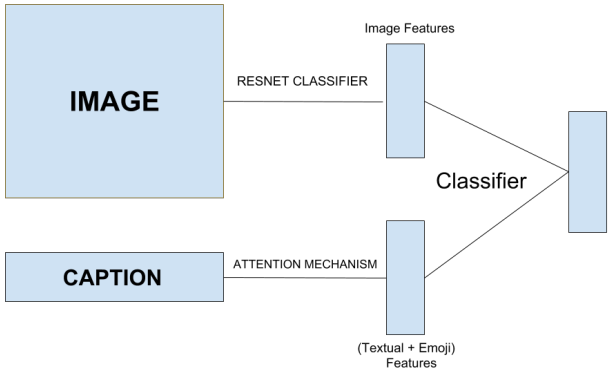

Figure 7: Multimodal emotion classification approach using both caption and image

7 DISCUSSION

We presented extensive experiments to check the accuracies of different approaches and importance of different modalities towards emotion classification. Researchers have identified the importance to study emoji for social NLP tasks considering its usage on social media platforms. Hence we considered emojis as an addition feature to learn the emotional intent from a social media post. We have first looked at how important visual features are to study user emotions. We used ResNet architecture developed by He et al. [8] to learn features from images. We had observed a very low F score of about 30% using ResNet-152 model architecture which is currently the state of the art for image classification. This low accuracy can be explained by the presence of captioned photos which may confuse the model while learning image features for emotion classification [11].

We then looked at the importance of textual features for emotion classification. Recent research has identified the importance of textual features over visual features for task involving social media data [3]. The same has been observed in the case of emotion classification. A caption may contain emoji or text, hence we looked at the importance of these modalities for emotions classification individually. Unlike Twitter where users are restricted to a caption of character length 300, instagram allows users to use a caption of length 2200 which can give room to about 350 words. In our dataset, we have observed that the maximum characters in the caption is nearly 1948, maximum word count of a caption is 423 words, and the average number of words in the caption is about 44 words. We then looked at the variation of accuracies with input sentence length. We observed that the maximum classification accuracy at a input sentence length of 80 words/caption. Further increasing the sentence length did not affect the classification accuracy.

We then looked at the importance of the third modality– emoji features, we used the different emoji knowledge concepts extracted

Table 7: Accuracies of emotion classification task using attention mechanism combining text and emoji as input for Bi-LSTM

Word Embedding Model	Knowledge Concepts	Precision	Recall	Macro F1 Score
Fasttext on Post Descriptions	Emoji Names	66.20	64.82	65.23
	Emoji Senses	65.40	64.96	64.95
	Emoji Sense Definitions	66.04	64.99	65.27
	Processed Emoji Sense Definitions	65.94	64.58	65.04
Pretrained Fasttext on Wiki Corpus	Emoji Names	53.99	52.14	52.42
	Emoji Senses	54.95	51.21	51.99
	Emoji Sense Definitions	52.34	51.85	51.83
	Processed Emoji Sense Definitions	53.72	51.83	52.25

Table 8: Accuracies of emotion classification task using Bag of words technique combining text and emoji as input to learn description embedding using fasttext trained word embeddings on the set of captions

Word Embedding Model	Knowledge Concepts	Precision	Recall	Macro F1 Score
Fasttext on Post Descriptions	Emoji Names	60.51	59.76	59.74
	Emoji Senses	61.29	59.79	60.19
	Emoji Sense Definitions	61.11	60.13	60.50
	Processed Emoji Sense Definitions	60.48	60.03	60.02
Pretrained Fasttext on Wiki Corpus	Emoji Names	54.28	53.85	53.85
	Emoji Senses	53.87	53.48	53.19
	Emoji Sense Definitions	54.20	53.15	53.02
	Processed Emoji Sense Definitions	54.87	53.16	53.51

Table 9: Accuracies of emotion classification task considering text (without emoji) and emoji features as separate input layers

Mechanism for Text	Knowledge Concepts	Precision	Recall	Macro F1 Score
Bag Of Words Approach	Emoji Names	61.38	60.45	60.43
	Emoji Senses	61.94	59.98	60.60
	Emoji Sense Definitions	61.00	60.11	60.19
	Processed Emoji Sense Definitions	61.30	60.63	60.88
Attention Mechanism	Emoji Names	64.68	61.66	62.54
	Emoji Senses	64.71	61.53	62.34
	Emoji Sense Definitions	64.96	61.42	62.12
	Processed Emoji Sense Definitions	63.69	61.54	61.64

Table 10: Accuracies of emotion classification task using attention mechanism combining text and emoji as input for Bi-LSTM

Knowledge Concepts	Precision	Recall	Macro F1 Score
Emoji Names	70.23	69.25	69.42
Emoji Senses	73.79	70.25	71.98
Emoji Sense Definitions	71.56	69.98	70.49
Processed Emoji Sense Definitions	72.23	70.26	70.78

from EmojiNet and trained a neural network. We used the vector average of the emojis embeddings used in each post as the feature vector and input of the neural network. Considering the use of ❤ in different contexts, the emoji ❤ is the most frequently used emoji to express Love or Joy. Hence using only emoji features for emotion classification is not a good idea and the same is observed in the results. The classification accuracy using emoji features is about 32.83% which is very low compared to textual features.

We then looked at the importance of emojis in addition to text for the emotion classification task. We used three different model architectures., The first was the bag of words model for the whole

caption (emojis embedded in between text). The second being attention mechanism with whole caption as input (emojis embedded in between text), where we consider an embedding layer consisting of both textual and emoji embeddings to train an attention model. In the third architecture, emojis and text are considered as two different features and fed through different input layers. The accuracies have been found to be better if emojis embedded with text is considered as sequential input to the attention model. This can be related to a study by Kevin Cohn which says that emojis on social media are being used as a language[11] and hence a Bi-LSTM model would give better accuracies for sequential data. For example consider the caption "My ❤ for life!!", here ❤ is used in context of "love". If both text and emoji are considered as different inputs, this would confuse the model since ❤ can be used in different contexts (as seen in Table 2), this decreases the classification accuracy. This is the reason resulting in high accuracy when both text + emoji is considered as sequential input to attention model.

Finally, we combined the three modalities – textual, visual, and emoji. Table 10 reports the classification accuracies observed using different emoji knowledge concepts wherethe caption (text + emoji) is sent through attention mechanism and image is sent through ResNet-152 model. We then merge the outputs of these layers and train a softmax on top of this for classification. This combination of all the modalities results in better classification accuracy.

8 CONCLUSION AND FUTURE WORK

In this paper, we explored the usage of different emojis in different emotional contexts and the importance of different modalities towards emotion classification task. We have presented a multimodal emotion classification approach which makes use of all modalities – emoji, textual and visual features. We have further shown that combining all modalities can outperform state of the art unimodal approaches (based on only on textual or visual or emoji contents). We also observed better accuracy for the emotion classification task when the caption (emoji and text) is considered as sequential input compared to accuracy when used as different input features. As a future work and with Felbo et al. [6] as reference, we plan to work on building transfer learning approaches using pre-trained classifiers to learn the emotional features from visual and textual contents towards emotion classification task. We also plan to evaluate our approaches using human annotated test set.

REFERENCES

[1] Emilia Apostolova and Noriko Tomuro. 2014. Combining visual and textual features for information extraction from online flyers. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 1924–1929.

[2] Lakshika Balasuriya, Sanjaya Wijeratne, Derek Doran, and Amit Sheth. 2016. Finding street gang members on twitter. In *Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. IEEE Press, 685–692.

[3] F Barbieri, M Ballesteros, F Ronzano, and H Saggion. 2018. Multimodal emoji prediction. *arXiv preprint arXiv:1803.02392* (2018).

[4] F Barbieri, F Ronzano, and H Saggion. 2016. What does this Emoji Mean? A Vector Space Skip-Gram Model for Twitter Emojis.. In *LREC*.

[5] Daantje Derks, Arjan ER Bos, and Jasper Von Grumbkow. 2008. Emoticons and online message interpretation. *Social Science Computer Review* 26, 3 (2008), 379–388.

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

[7] X Glorot and Y Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics (Proceedings of Machine Learning Research)*, Yee Whye Teh and Mike Titterington (Eds.), Vol. 9. PMLR, Chia Laguna Resort, Sardinia, Italy. http://proceedings.mlr.press/v9/glorot10a.html

[8] K He, X Zhang, S Ren, and J Sun. 2015. Deep Residual Learning for Image Recognition. *arXiv preprint arXiv:1512.03385* (2015).

[9] K He, X Zhang, S Ren, and J Sun. 2016. Identity Mappings in Deep Residual Networks. *arXiv preprint arXiv:1603.05027* (2016).

[10] S Hochreiter and J Schmidhuber. 1997. Long short-term memory. *Neural computation* (1997).

[11] Yuheng Hu, Lydia Manikonda, Subbarao Kambhampati, et al. 2014. What We Instagram: A First Analysis of Instagram Photo Content and User Types.. In *Icwsm*.

[12] Anurag Illendula, Kv Manohar, and Manish Reddy Yedulla. 2018. Which Emoji Talks Best for My Picture?. In *2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*. IEEE, 514–519.

[13] Anurag Illendula and Manish Reddy Yedulla. 2018. Learning Emoji Embeddings using Emoji Co-occurrence Network Graph. *arXiv preprint arXiv:1806.07785* (2018).

[14] A Joulin, E Grave, P Bojanowski, M Douze, Hé Jégou, and T Mikolov. 2016. FastText.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651* (2016).

[15] T Kenter, A Borisov, and M de Rijke. 2016. Siamese cbow: Optimizing word embeddings for sentence representations. *arXiv preprint arXiv:1606.04640* (2016).

[16] A Krizhevsky, I Sutskever, and G E. Hinton. 2017. ImageNet classification with deep convolutional neural networks. *Commun. ACM* 60, 6 (2017).

[17] U Kursuncu, M Gaur, U Lokala, A Illendula, K Thirunarayan, R Daniulaityte, A Sheth, and I B Arpinar. 2018. " What's ur type?" Contextualized Classification of User Types in Marijuana-related Communications using Compositional Multiview Embedding. *arXiv preprint arXiv:1806.06813* (2018).

[18] Jasy Suet Yan Liew and Howard R Turtle. 2016. Exploring fine-grained emotion detection in tweets. In *Proceedings of the NAACL Student Research Workshop*. 73–80.

[19] Kenneth Marino, Ruslan Salakhutdinov, and Abhinav Gupta. 2016. The more you know: Using knowledge graphs for image classification. *arXiv preprint arXiv:1612.04844* (2016).

[20] T Mikolov, E Grave, P Bojanowski, C Puhrsch, and A Joulin. 2018. Advances in Pre-Training Distributed Word Representations. In *LREC*.

[21] Saif M Mohammad. 2012. # Emotional tweets. In *Proceedings of 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*. Association for Computational Linguistics, 246–255.

[22] Petra Kralj Novak, Jasmina Smailović, Borut Sluban, and Igor Mozetič. 2015. Sentiment of emojis. *PloS one* 10, 12 (2015), e0144296.

[23] Mohammed Rajhi. 2017. Emotional Recognition Using Facial Expression by Emoji in Real Time. (2017).

[24] O Russakovsky, J Deng, Z Huang, A C. Berg, and L Fei-Fei. 2013. Detecting avocados to zucchinis: what have we done, and where are we going?. In *ICCV*.

[25] Pierre Sermanet and Yann LeCun. 2011. Traffic sign recognition with multi-scale convolutional networks. In *Neural Networks (IJCNN), The 2011 International Joint Conference on*. IEEE, 2809–2813.

[26] Phillip Shaver, Judith Schwartz, Donald Kirson, and Cary O'connor. 1987. Emotion knowledge: Further exploration of a prototype approach. *Journal of personality and social psychology* 52, 6 (1987), 1061.

[27] K. Simonyan and A. Zisserman. 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition. *CoRR* abs/1409.1556 (2014).

[28] Jared Suttles and Nancy Ide. 2013. Distant supervision for emotion classification with discrete binary values. In *International Conference on Intelligent Text Processing and Computational Linguistics*. Springer, 121–136.

[29] S Venugopalan, L A Hendricks, R Mooney, and K Saenko. 2016. Improving lstm-based video description with linguistic knowledge mined from text. *arXiv preprint arXiv:1604.01729* (2016).

[30] Wenbo Wang, Lu Chen, Krishnaprasad Thirunarayan, and Amit P Sheth. 2012. Harnessing twitter" big data" for automatic emotion identification. In *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom)*. IEEE, 587–592.

[31] S Wijeratne, L Balasuriya, A Sheth, and D Doran. 2017. EmojiNet: An Open Service and API for Emoji Sense Discovery. In *ICWSM*. Montreal, Canada.

[32] S Wijeratne, L Balasuriya, A Sheth, and D Doran. 2017. A semantics-based measure of emoji similarity. In *WI*. https://doi.org/10.1145/3106426.3106490

[33] Z Yang, D Yang, C Dyer, X He, A Smola, and E Hovy. 2016. Hierarchical attention networks for document classification. In *NAACL-HLT*.

[11] https://bit.ly/2UbmTRG

[34] Y Zhang, R Jin, and Z Zhou. 2010. Understanding bag-of-words model: a statistical framework. *International Journal of Machine Learning and Cybernetics* (2010).