Jordanian Journal of Computers and Information Technology (JJCIT), Vol. 05, No. 02, August 2019.

CONVOLUTIONAL NEURAL NETWORK MULTI-EMOTION CLASSIFIERS

S. S. Ibrahiem[1], S. S. Ismail [2], K. A. Bahnasy[3] and M. M. Aref [4]

(Received: 19-Apr.-2019, Revised: 10-Jun.-2019, Accepted: 26-Jun.-2019)

ABSTRACT

Natural languages are universal and flexible, but cannot exist without ambiguity. Having more than one attitude and meaning in the same phrase context is the main cause for word or phrase ambiguity. Most previous work on emotion analysis has only covered single-label classification and neglected the presence of multiple emotion labels in one instance. This paper presents multi-emotion classification in Twitter based on Convolutional Neural Networks (CNNs). The applied features are emotion lexicons, word embeddings and frequency distribution. The proposed networks performance is evaluated using state-of-the-art classification algorithms, achieving a hamming score range from 0.46 to 0.52 on the challenging SemEval2018 Task E-c.

KEYWORDS

Emotion classification, Multi-label classification, Convolutional neural network, Twitter.

1. INTRODUCTION

Online social media, such as Twitter, can communicate people's facts, opinions and emotions on different kinds of topics in short texts. Analyzing the emotions expressed in these texts has attracted researchers in the natural language processing research field. Emotion analysis is the task of determining the feeling or attitude towards a target or topic. It has a wide range of applications in politics, public health, commerce and business. Many real-world problems can be modeled by multi-label classification systems, like emotion analysis, since one tweet instance can imply more than one emotion. Traditional emotion analysis methods are single-label classification-based, while multi-label classification (MLC) recently attracts researchers' interest due to its applicability to a wide range of domains [1]-[3]. One of the most common used approaches is problem transformation methods that transform a multi-label dataset into a single-label dataset, so that existing single-label classifiers can be applied to multi-label datasets. Problem transformation approach replaces each multi-label instance with a single-class label for each class-label occurring. Binary relevance is the most common used method in the problem transformation approach; it works by decomposing the multi-label learning task into a number of independent binary learning tasks [2]-[4]. It suffers from directly modeling correlations that may exist among labels.

This paper is going to propose multi-emotion classification in Twitter based on Convolutional Neural Networks (CNNs). It performs two experiments to compare the proposed architecture with the state-of-art emotion classification approaches. The experimental results achieved a high classification accuracy and outperformed the state-of-the art approaches. The paper is organized in sections as follows. Section 2 discusses the research background and related work. Then, Section 3 discusses the applied datasets and lexicons. Section 4 discusses the proposed emotion detection approach and its evaluation. Section 5 concludes the paper.

2. RELATED WORK

Traditional approaches to sentiment analysis and emotion detection were categorized into lexicon-based methods (Y. Cao et al. [5], L. Flekova et al. [6], D. M. El-Din et al. [7], S. Mohamed et al. [8]), supervised machine learning methods (E. Cambria et al. [9], Y. Wang et al. [10], P. Sobhani et al. [11], N. Majumder et al. [12], R. Oramas et al. [13], M. Suhasini et al. [14]) and hybrid methods (S.

1. S. S. Ibrahiem is PhD researcher with Department of Computer Science, Ain Shams University, Cairo, Egypt. Email: soha.elshafey@cis.asu.edu.eg
2. S. S. Ismail is Computer Science Teacher, Ain Shams University, Cairo, Egypt. Email: sallysaad@cis.asu.edu.eg
3. K. A. Bahnasy is Information System Professor, Ain Shams University, Cairo, Egypt. Email: khaled.bahnasy@oi.edu.eg
4. M. M. Aref is Computer Science Professor, Ain Shams University, Cairo, Egypt. Email: mostafa.aref@cis.asu.edu.eg

Mohamed et al. [15], X. Sun et al. [16], B. Gaind et al. [17]).

Y. Cao et al. [5] focused on the task of disambiguating polarity-ambiguous words and the task was reduced to sentiment classification of aspects, which referred to sentiment expectation instead of semantic orientation. In order to disambiguate polarity-ambiguous words, the research constructed the aspect-and polarity-ambiguous lexicon using a mutual bootstrapping algorithm. So, the sentiment of polarity-ambiguous words in context could be decided collaboratively by the sentiment expectation of the aspect-and polarity-ambiguous words' prior polarity. Training corpus was 6000 positive and negative reviews about computer and books, while testing corpus was 1000 reviews. The average F1-measures were 75% in books and 79% in computer reviews.

L. Flekova et al. [6] introduced a method to identify frequent bigrams where a word switches polarity and to find out which words were bipolar to the extent, so that it was better to have them removed from the polarity lexica. The introduced method demonstrated that the scores match human perception of polarity and bring improvement in the classification results using its enhanced context-aware method. It enhanced the assessment of lexicon-based sentiment detection algorithms and could be used to quantify ambiguous words. 1600 Facebook messages were annotated by positive and negative sentiments that were used to asset the lexicon's performance on different feature sets. The accuracy ranged from 66% to 76%.

D. M. El-Din et al. [7] proposed a new technique to analyze online reviews. It was called sentiment analysis of online papers (SAOOP). SAOOP was a new technique used for enhancing bag-of-words model, improving accuracy and performance. SAOOP was useful in increasing the understanding rate of review's sentences through higher language coverage cases. SAOOP introduced solutions for some sentiment analysis challenges and used them to achieve higher accuracy. Two datasets were used; real dataset which splits into two datasets with training set (1000 text reviews) and test set (5000 text reviews) and the verified data set (10.000 text reviews) which included more than 5.000 positive words dataset and 5.000 negative words. The accuracy of SAOOP was 82%.

S. Mohamed et al. [8] presented a data-driven study comparing the emotionality of metaphorical expressions with that of their literal counterparts. Its results indicated that metaphorical usages are, on average, significantly more emotional than literal ones. It also showed that this emotional content was not simply transferred from the source domain into the target, but rather is a result of meaning composition and interaction of the two domains in the metaphor. It used 1639 senses of 440 verbs in WordNet. The confidence was 95%.

E. Cambria et al. [9] introduced a vector space model which was built by means of random projection to allow reasoning by natural language concepts. The model allowed semantic features associated with concepts to be generalized and to be intuitively clustered according to their semantic and affective relatedness. Such an affective intuition enabled the inference of emotions and polarity conveyed by multi-word expressions, thus achieving efficient concept-level sentiment analysis. An affective common-sense knowledge is built by applying concept frequency - inverse opinion frequency (CF-IOF) on a 5,000- blogpost database extracted from LiveJournal1, that is category-and mood-labeled by users. Test dataset was 2000 manually tagged patient reviews associating to each a category service. F-measure ranged from 74% to 85.1% according to evaluated service.

Y. Wang et al. [10] proposed a constraint optimization framework to discover emotions from social media content of the users. This framework employed several novel constraints, such as emotion bindings, topic correlations, along with specialized features proposed by prior work and well-established emotion NRC[2] lexicons[3]. It proposed an efficient inference algorithm and reported promising empirical results on three diverse datasets. Another distinguishing feature of this model was that it solved multi-label classification problem and allowed a document to have multiple emotions. The evaluated datasets were SemEval[4] of 1250 news headlines with an average F-measure of 0.63,

[1] http://livejournal.com/
[2] National Research Canada (NRC)
[3] http://saifmohammad.com/WebPages/lexicons.html
[4] http://web.eecs.umich.edu/~mihalcea/downloads.htm

ISEAR[5] of 7666 sentences annotated by 1096 participants with different cultural backgrounds with an average F-measure of 0.74 and a Twitter dataset of 1800 tweets using the Twitter API with an average F-measure of 0.522.

P. Sobhani et al. [11] developed a simple stance detection system that outperforms all 19 teams that participated in a recent shared task competition on the same dataset (SemEval-2016 Task #6). It applied n-grams, NRC lexicons, word embeddings and support vector machine learning. The classification range was in favour or against classes. The automatic system evaluation F1-measure was 70.32%.

N. Majumder et al. [12] presented a method to extract personality traits from a stream of consciousness essays using a convolutional neural network (CNN). It has been trained on five different networks, all with the same architecture, for the five studied personality traits. Each network was a binary classifier that predicted the corresponding trait to be positive or negative. It developed a novel document modeling technique based on a CNN feature extractor. Namely, it had been fed sentences from the essays to convolution filters in order to obtain the sentence model in the form of n-gram and word embedding feature vectors. Each individual essay was represented by aggregating the vectors of its sentences. For final classification, the document vector was fed into a fully connected neural network with one hidden layer and the final softmax layer of two sizes, representing the yes and no classes. 50 epochs for training and tenfold cross-validation were used to evaluate the trained network. The network was evaluated to SVM and multi-layer perception learning and the accuracy ranged from 50% to 62%.

R. Oramas et al. [13] created a corpus of phrases (opinions) and categorized them into frustration, boring, excitement and engagement phrases. The corpus was tested using several tests with different classifiers: Multi-nomial Naive Bayes classifier, Support Vector Machine, Linear Support Vector Machine, Stochastic Gradient Descent Classifier and K-Nearest Neighbors classifier. The used dataset consisted of 851 opinions. The classifier with the highest score was Bernoulli Naive Bayes classifier with an accuracy of 76.77%.

M. Suhasini et al. [14] proposed a method which detected the emotion or mood of the tweets and classified the Twitter messages under appropriate emotional categories. The method used was a two-approach method. The approach used were the Rule-Based approach and the Machine Learning approach. First approach contributed in pre-processing, tagging, feature selection and knowledge base creation. Rule-based approach was used to classify the tweets under four class categories (Happy-Active, Happy-Inactive, Unhappy-Active and Unhappy-Inactive). The second approach was based on a supervised machine learning algorithm called Naïve Bayes, which requires labeled data. The rule-based approach was able to classify the tweets with an accuracy around 85% and with the machine learning approach the accuracy was around 88%.

S. Mohamed et al. [15] automatically annotated a set of 2012 US presidential election tweets for a number of attributes pertaining to sentiment, emotion, purpose and style by crowdsourcing. Overall, more than 100,000 crowdsourced responses were obtained for 13 questions on emotions, style and purpose. Additionally, it was shown through an analysis of these annotations that purpose, even though correlated with emotions, was significantly different. Finally, it was described how automatic classifiers had been developed, using features from state-of-the-art sentiment analysis systems, to predict emotion and purpose labels, respectively, in new unseen tweets. These experiments resulted in an accuracy of 56.84% for automatic systems on this new data.

X. Sun et al. [16] presented a method for extracting emotional elements containing emotional objects and emotional words and their tendencies from product reviews based on a mixed model. First, some conditional random fields were constructed to extract emotional elements, lead-in semantics and word meanings as features to improve the robustness of feature template and rules were used for hierarchical filtering of errors. Then, a support vector machine was constructed to classify the emotional tendencies of the fine-grained elements to achieve key information from product reviews. Deep semantic information was imported based on a neural network to improve the traditional bag-of-

[5] http://www.affective-sciences.org/researchmaterial

word model. Experimental results showed that the proposed model with deep features efficiently improved the F-measure 50-80%.

B. Gaind et al. [17] proposed two approaches to classify social media texts into six categories of emotion: Happiness, Sadness, Fear, Anger, Surprise and Disgust. First approach extracted emotion in the texts using natural language processing, like emoticons, part of speech, negations and grammatical analysis. Second approach was based on two machine learnings, which are support to vector machine and J48 classifiers. A large bag of words in English was created that expressed word emotions in addition to their intensities. The training accuracy of the support vector machine was 91.7% and the training accuracy of J48 classifier was 85.4% of 900 tweets.

Through the research studies that have been listed, it can be concluded that firstly there is still insufficient inadequacy of models to explore emotions from the texts of social media due to data size, text structure's context or emotional granularity [4]. Secondly, emotion analysis is modeled as a supervised multi-label classification problem, because one instance may contain one or more emotions from a standard emotion set. This paper deals with multi-emotion classification based on Convolutional Neural Networks (CNNs). This network can keep complementary information and will bring higher accuracy with the assistance of different feature configurations that could lead to possible directions of further improvement.

3. DATASETS & LEXICONS

The datasets and lexicons that are going to be used in this research are listed as follows.

SemEval[6]-2018 is a group of datasets that include an array of subtasks, where automatic systems have to infer the affectual state of a person from his/her tweets. One of its tasks is Emotion Classification (E-c), where a given tweet is classified as 'neutral or no emotion' or as one or more of eleven given emotions that best represent the mental state of the user [18].

The NRC Emotion Lexicon is a list of English words and their associations with eight basic emotions (anger, fear, anticipation, trust, surprise, sadness, joy and disgust) and two sentiments (negative and positive). The annotations were manually done by crowdsourcing [19]-[20].

NRC Hashtag Emotion Lexicon: it is an association of words with eight emotions (anger, fear, anticipation, trust, surprise, sadness, joy and disgust) generated automatically from tweets with emotion-word hashtags such as #happiness and #anger [21].

NRC VAD Lexicon: The NRC Valence, Arousal and Dominance (VAD) Lexicon includes a list of more than 20,000 English words and their valence, arousal and dominance scores. Valence is the positive and negative or pleasure and displeasure dimension [22]. Arousal is the excited and calm or active and passive dimension [22]. Dominance is the powerful and weak or 'have full control' and 'have no control' dimension [22].

4. EMOTION DETECTION APPROACH

The supervised machine learning approach involves two processes; training process where an algorithm learns from a labeled data and testing process or evaluation process where the algorithm makes predictions on a sample data. In this research, the input tweets are firstly pre-processed and normalized, where different types of processes are applied to denoise and filter important tokens in the tweets (see Figure 1). Secondly, the pre-processed tokens are transformed to feature vectors. These featured vectors exploit contextual and semantic relations between tokens and each target emotion space using the prementioned lexicons and datasets, in addition to word frequency distribution and word embeddings. The word frequency distribution calculates each token's occurrence in the given dataset SemEval-2018 towards a given emotion space [23]. The applied word embedding method is word2vec that works on grouping each emotion's representative tokens together in the same vector space [24]-[25]. Figure 2 illustrates the procedure of transforming the tweets' pre-processed tokens

[6] Semantic Evaluation Workshop, 2018.

into feature vectors in both training and testing phases. Finally, the produced tweets' feature vectors are trained and evaluated on the following proposed network.

- Emojis are translated, and URLs are removed from the input tweet.
- The tweets are tokenized.
- Negation is handled, so that context meaning and attitude are conserved.
- Long words are corrected without context loss.
- Any spaces and punctuation letters are filtered, so that it doesn't adversely affect the efficiency of the classification.
- Part of speech tagging is applied to tokens.
- Spelling correction is applied to each token. Any meaningless tokens are neglected.
- Tokens are stemmed from wordnet lexicon, according to their part of speech tag.
- Stop words are removed from tokens. It doesn't imply any emotion.

Figure 1. Pre-processing steps sequence.

```
Input:
        train_tweets_tokens=load_train_tweets_tokens()
        test_tweets_tokens=load_test_tweets_tokens()
Output:
        //Initialize empty output lists that represent the relevance of each tweet to a target emotion class
        anger_frequency,fear_frequency,joy_frequency,sad_frequency=[],[],[],[]
        anger_w2v,fear_w2v,joy_w2v,sad_w2v=[],[],[],[]
        anger_wrdlex,fear_wrdlex,joy_wrdlex,sad_wrdlex=[],[],[],[]
        anger_hshlex,fear_hshlex,joy_hshlex,sad_hshlex=[],[],[],[]
        valence,arousal,dominance=[],[],[]
-------------------------------------------------------------------------------
//frequency distribution calculation and vector space generation to each emotion class from training tokens only
for each emotion tokens in train_tweets_tokens do
        Compute Frequency distribution of tokens occurred in each emotion class
        output lists of most frequent training tokens in each emotion class
        Compute each emotion vector space from tokens occurred in each emotion tweets
        output each emotions' representative training tokens together in a same vector space
end for;
for each tweet in train_tweets_tokens do
        for each token in tweet:
                //frequency distribution
                Get Frequency distribution of the input token to each training emotion frequent tokens
                Sum up frequency distribution of tokens in each emotion class in different variables
                //word embeddings
                Get token similarity to each emotion training vector space
                Sum up tokens similarities in each emotion class in separate variables
                //NRC word and hash lexicons
                Search word lexicon if the token belongs to an emotion class, increment the word emotion counter of the tweet
                Search hash lexicon if the token belongs to an emotion class, increment the hash emotion counter of the tweet
                //NRC VAD Lexicon
                Search VAD lexicon for the token, get three dimensions of the token valence, arousal, and dominance
                Sum up tokens three dimensions
        end for;

        Calculate average of frequency distributions of the tweet tokens in each emotion class
        Save value of frequency distribution of tokens in each tweet towards each emotion class lists
        in anger_frequency, fear_frequency, joy_frequency, and sad_frequency

        Calculate average similarities of the tweet tokens in each emotion class
        Save similarities values of each tweet towards each emotion class in
        anger_w2v, fear_w2v, joy_w2v, and sad_w2v lists

        Calculate average word lexicon of each emotion counter
        Calculate average hash lexicon of each emotion counter
        Save each average counters each tweet towards each emotion class in
        anger_wrdlex, fear_wrdlex, joy_wrdlex ,sad_wrdlex  lists and
        anger_hshlex,fear_hshlex,joy_hshlex,sad_hshlex lists

        Calculate average dimensions of tweets tokens
        Save average dimensions of each tweet in separate lists valence,arousal, and dominance
end for;
```

Figure 2. Feature generation procedure steps.

4.1 Building and Training Models

Convolutional Neural Networks (CNNs) are one of the most successful network architectures in state-of-the-art Artificial Neural Network (ANN) algorithms. A CNN can learn relevant features from the input text at different levels like the human brain. Its basic components are listed as follows. The convolution is over an input pixel matrix. A kernel or filter slides over the input matrix creating an entry in the activation map for each window in the input matrix. Hereby, the weights of the filter are multiplied by the values in the window of the input matrix and the results are added up. The weights in the filter are subject to the learning process of the network and are shared over all windows of the convolution operation. In CNN architectures, a convolution layer is usually followed by a pooling layer. Pooling layers sub-sample their input and can be applied over the whole matrix or over windows. They significantly reduce the output dimensionality without losing much information [26].

In this paper, a Convolutional Neural Network (CNN) is built and trained to predict multi-emotion labeled tweets. The output emotion classes are anger, fear, joy and sadness. The SemEval2018 task e-c dataset is divided into training and testing datasets. The training dataset consists of 5000 labeled tweets, while the testing dataset consists of 1000 annotated tweets. Two experiments are applied on two feature configurations. Figures 3 and 5 show two different CNN network architectures according to the applied feature configuration. In the first experiment, the network has five input layers. The first four input layers are of size four, representing the tweet's features for a target emotion using frequency distribution, word-to-vector similarity, NRC word emotion lexicon and NRC hashtag lexicon. The fifth input layer is of size three, which represents the tweet's average valence-arousal-dominance feature using NRC VAD lexicon. The first four input convolutional layers have 4 filters and window size 2, followed by an average pooling layer with pool size 3. The fifth input layer has three filters and window size 2, followed by an average pooling size 2. Flatten layers work as a connection between convolutional layer output and the following dense layers. This classifier has 2 fully connected dense layers with sizes of 10 and 4 nodes, respectively. The output node represents the probability of likelihood of the input tweet vector to the four emotions: anger, fear, joy and sadness.

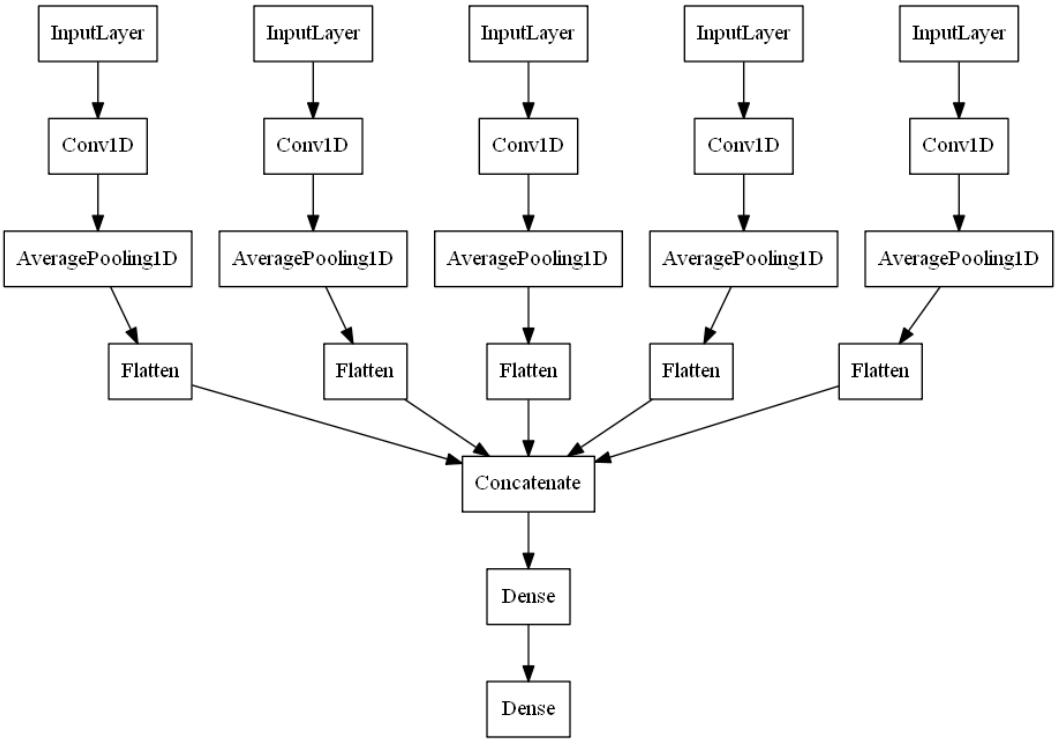

Figure 3. Large-feature CNN emotion classification architecture.

The training dataset is transformed to the defined feature vectors. These vectors are trained on the presented network to produce the emotion prediction model. The tools used are Keras [27], Python [28]-[29] and NLTK [30]. Figure 4 shows the CNN model accuracy and loss graphs during the training phase using a 0.34 cross-validation set. Cross-validation technique is used to evaluate

predictive models by partitioning the original sample into a training set to train the model and a test set to evaluate it. Figure 4 shows the progress of both accuracy and loss in the training phase during 70 epochs. The accuracy is greater than 85% and the loss is less than 0.35.

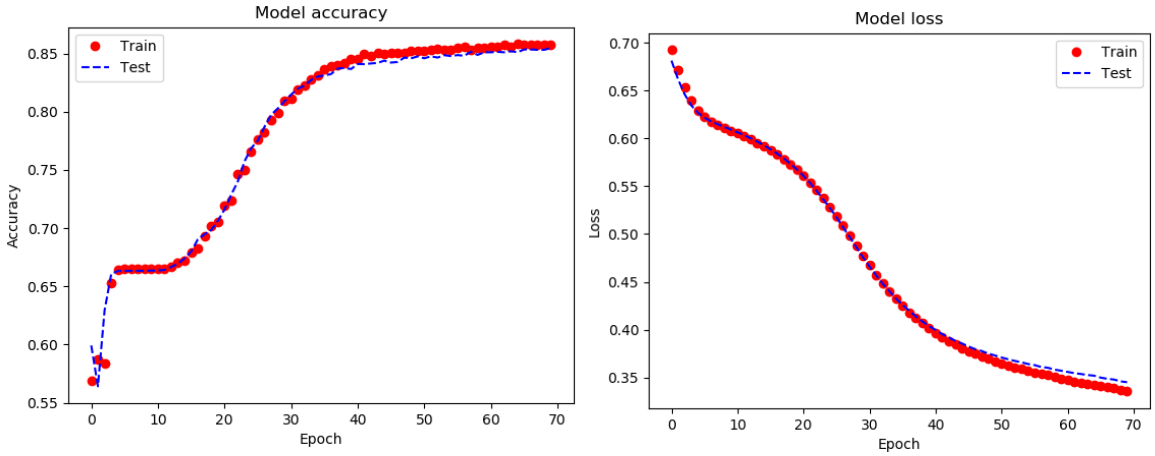

Figure 4. Training CNN accuracy and loss graphs for large feature configuration.

In the second experiment, another network's architecture is presented in Figure 5. This network works on small-feature configuration that includes frequency distribution, word-to-vector similarity and word emotion lexicon only. The network has four input layers. Each input layer represents an emotion space feature. The four input convolutional layers have 3 filters and window size 2, followed by an average pooling layer with pool size 3. There are 2 fully connected dense layers with sizes of 10 and 4 nodes, respectively.

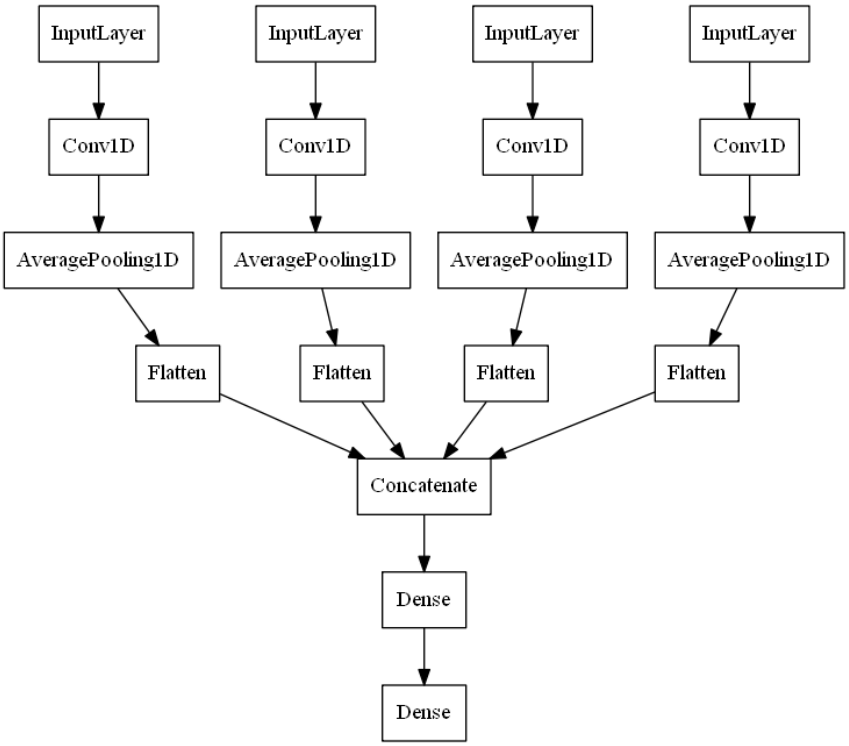

Figure 5. Small-feature CNN emotion classification architecture.

Figure 6 shows CNN model accuracy and loss graphs during the training phase. Figure 6 shows the progress of both accuracy and loss in the training phase in 70 epochs. The accuracy is greater than 85% and the loss is less than 0.35.

In these experiments, each convolutional layer has a four-or three-vector input according to feature set, Rectified Linear Units (ReLUs) [31] as activation functions and a batch size of 100. Filters are four

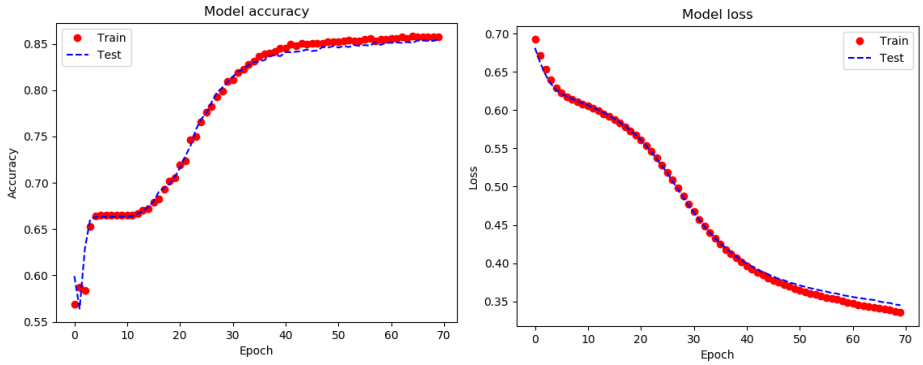

Figure 6. Training CNN accuracy and loss graphs for small-feature configuration.

and window is two, except the valence-arousal-dominance feature node with a three-vector input. Filters are three and window is two. The last decision layer is a sigmoid, so 0.5 is the threshold value to each output node. The model undergoes training through Adam optimizer [31] over shuffled mini-batches. The model stops the iterant processes of learning by a loss function binary cross-entropy [33]-[34].

4.2 Testing Models and Evaluation

The predefined training dataset consists of multi-labeled tweets. The output emotion classes are: anger, fear, joy and sadness. These tweets are trained on both convolutional neural networks in Figures 3 and 5. In the testing phase, the testing dataset consists of 1000 tweets from SemEval2018. They are multi-labeled in the same emotion range and never used in the training phase. Anger tweets are 392, fear tweets are 224, joy tweets are 402 and sadness tweets are 310. The proposed networks are evaluated by two different feature configurations with four standard machine learning algorithms; support vector machine, naïve Bayes, k nearest neighbour and multi-layer perceptron. Table 2 shows the precision (P), recall (R), f-measure (F) [35] values for each applied learning algorithm. Table 1 includes hamming score (HS), hamming loss (HL) and exact ratio (ER) [36] of the applied algorithms. First experiment evaluation has a mean of 0.756 and a standard deviation of 0.006. Second experiment evaluation has a mean of 0.745 and a standard deviation of 0.005.

Table 1. Hamming score, hamming loss and exact ratio on two feature configurations.

Techniques and feature configurations	HS	HL	ER
SVM			
Word2vec, FreqDist, NRC word lexicon	0.40	0.25	0.31
Word2vec, FreqDist, NRC word, hash, & VAD lexicons	0.46	0.23	0.37
KNN			
Word2vec, FreqDist, NRC word lexicon	0.50	0.26	0.37
Word2vec, FreqDist, NRC word, hash, & VAD lexicons	0.52	0.25	0.40
NB			
Word2vec, FreqDist, NRC word lexicon	0.38	0.28	0.26
Word2vec, FreqDist, NRC word, hash, & VAD lexicons	0.43	0.27	0.30
MLP			
Word2vec, FreqDist, NRC word lexicon	0.40	0.27	0.28
Word2vec, FreqDist, NRC word, hash, & VAD lexicons	0.44	0.25	0.31
Proposed CNN			
Word2vec, FreqDist, NRC word lexicon	0.46	0.24	0.35
Word2vec, FreqDist, NRC word, hash, & VAD lexicons	0.52	0.23	0.42

On the other hand, Figure 7 shows another performance measurement for classification problem using Receiver Operating Characteristics (ROC) curve and Area Under the Receiver Operating Characteristics (AUC) [37]-[38]. Large-feature configuration network has 0.78 AUC, while small-feature one has 0.75 AUC.

By utilizing extensive computational power, convolutional neural network processing has been proven to be a very powerful method by researchers in many fields, like computer vision and natural language processing. Applied experiments contribute in multi-label classification in natural language processing field. They showed more reliable results and higher overall accuracies compared to standard machine learning algorithms.

Table 2. Precision, recall and f-measure results on two feature configurations.

	Anger			**Fear**			**Joy**			**Sadness**		
	P	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
SVM												
Word2vec, FreqDist, NRC word lexicon	0.68	0.42	0.52	0.71	0.18	0.29	0.76	0.58	0.66	0.56	0.33	0.41
Word2vec, FreqDist, NRC word, hash, & VAD lexicons	0.74	0.46	0.57	0.77	0.23	0.35	0.81	0.67	0.73	0.61	0.34	0.44
K nearest neighbor												
Word2vec, FreqDist, NRC word lexicon	0.68	0.46	0.55	0.43	0.47	0.45	0.70	0.72	0.71	0.56	0.41	0.47
Word2vec, FreqDist, NRC word, hash, & VAD lexicons	0.68	0.48	0.57	0.49	0.41	0.44	0.70	0.77	0.74	0.58	0.38	0.46
Naïve Bayes												
Word2vec, FreqDist, NRC word lexicon	0.61	0.45	0.51	0.44	0.31	0.37	0.75	0.57	0.65	0.46	0.38	0.42
Word2vec, FreqDist, NRC word, hash, & VAD lexicons	0.63	0.46	0.53	0.46	0.42	0.44	0.78	0.63	0.70	0.50	0.44	0.47
MLP (three fully connected layers)												
Word2vec, FreqDist, NRC word lexicon	0.68	0.40	0.50	0.47	0.47	0.47	0.72	0.54	0.62	0.56	0.38	0.46
Word2vec, FreqDist, NRC word, hash, & VAD lexicons	0.68	0.44	0.54	0.56	0.50	0.53	0.74	0.56	0.64	0.59	0.42	0.49
Proposed CNN												
Word2vec, FreqDist, NRC word lexicon	0.69	0.52	0.59	0.64	0.26	0.37	0.72	0.67	0.69	0.60	0.31	0.41
Word2vec, FreqDist, NRC word, hash, & VAD lexicons	0.66	0.64	0.65	0.86	0.08	0.15	0.76	0.74	0.75	0.64	0.38	0.47

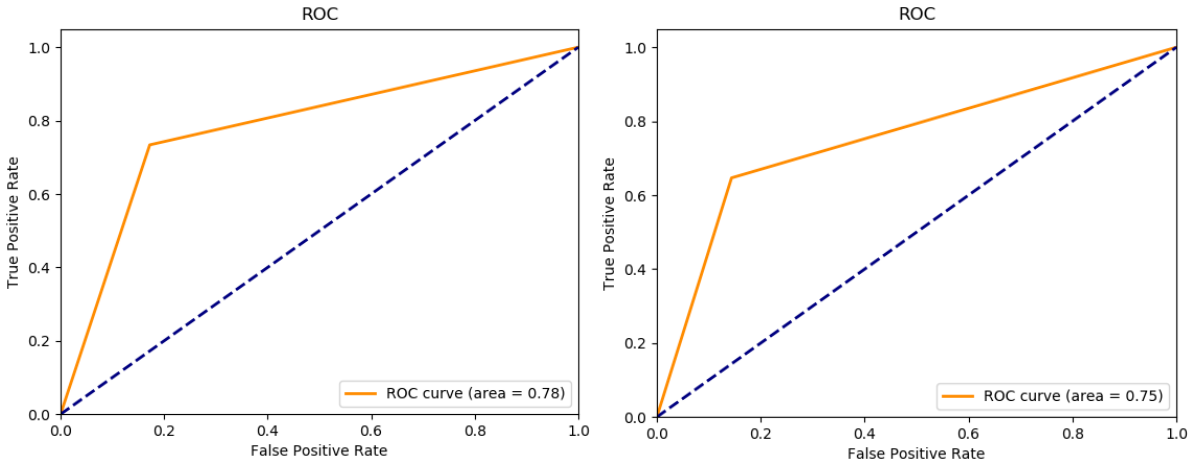

Figure 7. ROC curve of AUC calculation for large-feature configuration on the left-hand side and for small-feature configuration on the right-hand side.

5. CONCLUSION

This paper presented featured convolutional neural network architectures that applied multi-emotion classification in Twitter. It firstly discussed the related background about multi-label classification and emotion analysis. Secondly, it defined the annotated tweet datasets and lexicons that were used in pre-processing and feature extraction phases. Thirdly, it illustrated the architectures of the proposed convolutional neural networks and the applied experiments. Two experiments were applied using two different feature configurations. Fourthly, the evaluation metrics were illustrated to compare the CNN emotion classification models performance to the represented feature configurations and state-of-the-art classification algorithms. Python snapshots were shown to illustrate accuracy and loss performance during the training and testing phases. Finally, evaluation metrics were calculated and the proposed approach performance was evaluated. Tables 1 and 2 show the evaluation of the experimental results.

REFERENCES

[1] X. Quan, Q. Wang, Y. Zhang, L. Si and L. Wenyi, "Latent Discriminative Models for Social Emotion Detection with Emotional Dependency," ACM Transactions on Information Systems (TOIS), vol. 34 no. 1, p. 2, 2014.

[2] J. M. Nareshpalsingh and H. N. Modi, "Multi-label Classification Methods: A Comparative Study," International Research Journal of Engineering and Technology (IRJET), vol. 04, no. 12, December 2017.

[3] C. N. N. Kamath, S. S. Bukhari and A. Dengel, "Comparative Study between Traditional Machine Learning and Deep Learning Approaches for Text Classification," Proc. of the ACM Symposium Conference, pp. 1-11, 2018.

[4] M. Haggag, S. Fathy and N. Elhaggar, "Ontology-based Textual Emotion Detection," International Journal of Advanced Computer Science and Applications (IJACSA), vol. 6, no. 9, pp. 239- 246, 2015.

[5] Y. Cao, P. Zhang and A. Xiong, "Sentiment Analysis Based on Expanded Aspect-and Polarity-Ambiguous Word Lexicon," International Journal of Advanced Computer Science and Applications (IJACSA), vol. 6, no. 2, 2015.

[6] L. Flekova, E. Ruppert and D. P. Pietro, "Analyzing Domain Suitability of a Sentiment Lexicon by Identifying Distributionally Bipolar Words," Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (WASSA 2015), pp. 77–84, Lisboa, Portugal, September 2015.

[7] D. M. El-Din, H. M. O. Mokhtar and O. Ismael, "Online Paper Review Analysis," International Journal of Advanced Computer Science and Applications (IJACSA), vol. 6, no. 9, 2015.

[8] S. Mohammad, E. Shutova and P. Turney, "Metaphor As a Medium for Emotion: An Empirical Study," Proceedings of the Joint Conference on Lexical and Computational Semantics, pp. 23-33, Berlin, Germany, 2016.

[9] E. Cambria, J. Fu, F. Bisio and S. Poria, "AffectiveSpace 2: Enabling Affective Intuition for Concept-Level Sentiment Analysis," Proceedings of the 29th AAAI Conference on Artificial Intelligence, pp. 508-514, Austin, 2015.

[10] Y. Wang and A. Pal, "Detecting Emotions in Social Media: A Constrained Optimization Approach," Proceedings of the 24th International Joint Conference on Artificial Intelligence (IJCAI), pp. 996-1002, Buenos Aires, Argentina, 2015.

[11] P. Sobhani, S. M. Mohammad and S. Kiritchenko, "Detecting Stance in Tweets and Analyzing Its Interaction with Sentiment," Proceedings of the Joint Conference on Lexical and Computational Semantics, pp. 159–169, Berlin, Germany, August 2016.

[12] N. Majumder, S. Poria, A. Gelbukh and E. Cambria, "Deep Learning-based Document Modeling for Personality Detection From Text," IEEE Intelligent Systems, vol. 32, no. 2, pp. 74-79, 2017.

[13] R. Oramas, M. L. Barron-Estrada, R. Zatarain-Cabada and S. L. Ramírez-Ávila, "A Corpus for Sentiment Analysis and Emotion Recognition for a Learning Environment," Proc. of the 18th IEEE International Conference on Advanced Learning Technologies (ICALT), pp. 431-435, Mumbai, 2018.

[14] M. Suhasini and S. Badugu, "Two Step Approach for Emotion Detection on Twitter Data," International Journal of Computer Applications, vol. 179, no. 53, pp. 12 –19, June 2018.

[15] S. Mohammad, S. Kiritchenko, X. Zhu and J. Martin. "Sentiment, Emotion, Purpose and Style in Electoral Tweets," Information Processing and Management, vol. 51, no. 4, pp. 480–499, July 2015.

[16] X. Sun, C. Sun, C. Quan, F. Ren, F. Tian and K. Wang, "Fine-grained Emotion Analysis Based on Mixed Model for Product Review," International Journal of Networked and Distributed Computing, vol. 5, no. 1, pp. 1–11, January 2017.

[17] B. Gaind, V. Syal and S. Padgalwar, "Emotion Detection and Analysis on Social Media," Proceedings of the International Conference on Recent Trends in Computational Engineering and Technologies (ICTRCET'18), Bengaluru, India, May 2018.

[18] S. Mohammad, F. B. Marquez, M. Salameh and S. Kiritchenko, "Semeval-2018: Affect in Tweets," Proceedings of the International Workshop on Semantic Evaluation (SemEval-2018), New Orleans, LA, USA, June 2018.

[19] S. Mohammad and P. Turney, "Crowdsourcing a Word-Emotion Association Lexicon," Computational Intelligence, vol. 29, no. 3, pp. 436-465, 2013.

[20] S. Mohammad and P. Turney, "Emotions Evoked by Common Words and Phrases: Using Mechanical Turk to Create an Emotion Lexicon," Proceedings of the NAACL-HLT Workshop on Computational Approaches to Analysis and Generation of Emotion in Text, LA, California, June 2010.

[21] S. Mohammad, "#Emotional Tweets," The 1st Joint Conference on Lexical and Computational Semantics, vol. 1 (Proceedings of the Main Conference and the Shared Task) and vol. 2 (Proceedings of the 6th International Workshop on Semantic Evaluation (SemEval 2012)), Montr'eal, Canada, pp. 246-255, 7-8 June 2012.

[22] S. Mohammad, "Obtaining Reliable Human Ratings of Valence, Arousal and Dominance for 20,000 English Words," Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Melbourne, Australia, July 2018.

[23] V. A. Kharde and S. S. Sonawane, "Sentiment Analysis of Twitter Data: A Survey of Techniques," International Journal of Computer Applications, vol. 139, no. 11, pp. 5-15, April 2016.

[24] T. Mikolov, G. Corrado, K. Chen and J. Dean, "Efficient Estimation of Word Representations in Vector Space," Proceedings of the International Conference on Learning Representations (ICLR 2013), pp. 1–12, 2013.

[25] E. M. Alshari, A. Azman, S. Doraisamy, N. Mustapha and M. Alkeshr, "Improvement of Sentiment Analysis based on Clustering of Word2Vec Features," Proc. of the 28th International Workshop on Database and Expert System Applications, 2017.

[26] K. K. Lurz, Natural Language Processing in Artificial Neural Network Sentence Analysis in Medical Papers, Master Thesis, Department of Astronomy and Theoretical Physics, Lund University, June 11, 2018.

[27] F. Chollet, "keras," [Online], Available: GitHub. https://github.com/fchollet/keras, 2015.

[28] Python Software Foundation, [Online], Available: http://www.python.org, 2019.

[29] J. Perkins, Python 3 Text Processing with NLTK 3 Cookbook, Packt Publishing, 2014.

[30] N. Hardeniya, "NLTK Essentials Build Cool NLP and Machine Learning Applications Using NLTK and Other Python Libraries," July 2015.

[31] A. F. Agarap, "Deep Learning Using Rectified Linear Units (ReLUs)," arXiv:1803.08375, 2018.

[32] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," Proceedings of the 3rd International Conference on Learning Representations, 2014.

[33] S. Mannor, D. Peleg and R. Rubinstein, "The Cross-entropy Method for Classification," Proceedings of the 22nd International Conference on Machine Learning (ICML '05), pp. 561-568, Bonn, Germany, August 07 - 11, 2005.

[34] S. Baker and A. Korhonen, "Initializing Neural Networks for Hierarchical Multi-label Text Classification," Association for Computational Linguistics (BioNLP 2017), pp. 307-315, August 2017.

[35] D. Ganda and R. Buch, "A Survey on Multi-label Classification," Recent Trends in Programming Languages, vol. 5, no. 1, pp. 19-23, August 2018.

[36] S. R. Khade and S. R. Balwan. "Study and Analysis of Multi-label Classification Methods in Data Mining," International Journal of Computer Applications, vol. 159, no. 9, February 2017.

[37] J. A. Swets, "ROC Analysis Applied to the Evaluation of Medical Imaging Techniques," Invest. Radiol., vol. 14, no. 2, pp. 109-121, 1979.

[38] J. A. Hanley, "Receiver Operating Characteristic (ROC) Methodology: The State-of-the-Art," Crit. Rev. Diagn. Imaging, vol. 29, no. 3, pp. 307-335, 1989.

ملخص البحث:

تتسم اللغات الطبيعية بالكونية والمرونة، غير أنها لا يمكن أن توجد إلا أن تكون مصحوبة بشيء من الغموض. فوجود أكثر من منحنى وأكثر من معنى في سياق العبارة الواحدة هو السبب الأساسي في غموض الكلمات أو العبارات. لقد غطت غالبية الأبحاث السابقة المتعلقة بتحليل الأحاسيس التصنيف المبني على علامة مفردة فقط، في حين تم تجاهل وجود علامات إحساس متعددة في الكلمة أو العبارة ذاتها.

تعرض هذه الورقة تصنيفاً قائماً على تعدد الأحاسيس في نصوص تويتر (Twitter) يستخدم الشبكات العصبية الالتفافية. أما الخصائص المطبقة في هذا التصنيف فهي: معاجم الإحساس، ومضامين الكلمات، وتوزيعها التكراري. تم تقييم أداء الشبكات المقترحة عبر مقارنتها بأشهر خوارزميات التصنيف الشائعة. وأظهرت النتائج تفوق التقنية المقترحة عند تطبيقها على مجموعة من قواعد البيانات التي تمثل تحدّياً في هذا المجال، وتراوحت نتيجة المبالغة بين 0.46 و 0.52.