Mutux at SemEval-2018 Task 1: Exploring Impacts of Context Information On Emotion Detection

Pan Du
DIRO, University of Montreal
pandu@iro.umontreal.ca

Jian-Yun Nie
DIRO, University of Montreal
nie@iro.umontreal.ca

Abstract

This paper describes MuTuX, our system that is designed for task 1-5a, emotion classification analysis of tweets on SemEval2018[1]. The system aims at exploring the potential of context information of terms for emotion analysis. A Recurrent Neural Network is adopted to capture the context information of terms in tweets. Only term features and the sequential relations are used in our system. The result submitted ranks 16th out of 35 systems on the task of emotion detection in English-language tweets.

1 Introduction

Emotion analysis on social media is attracting more and more reserach interests (Strapparava and Mihalcea, 2008; Balahur et al., 2011; Agrawal and An, 2012; Wang et al., 2012; Hasan et al., 2014a; Canales and Martínez-Barco, 2014) from industry and academia. Commercial applications such as product recommendation, online retailing, and marketing are turning their interests from traditional sentiment analysis to emotion analysis as well. Emotion analysis is generally taken as a multi-lable classification problem. Given a piece of text, such as a tweet, it assigns several lables such as depressed, sad, angry and so on to it (Mohammad et al., 2018) based on the meaning contained in the text.

Techniques related to emotion detection can be divided into lexicon-based approaches (Valitutti, 2004; Strapparava and Mihalcea, 2008; Balahur et al., 2011) and machine learning approaches (Hasan et al., 2014b; Wang et al., 2012; Roberts et al., 2012; Suttles and Ide, 2013). Lexicon-based approaches leverage lexical resources to detect emotions, the resources can be keywords (Hasan et al., 2014a), WordNet-Affect (Valitutti,

2004), ontologies (Balahur et al., 2011) and so on. Machine learning based approaches (Balabantaray et al., 2012) generally take emotion detection as a classification problem using SVM, neural network (Abdul-Mageed and Ungar, 2017; Bravo-Marquez et al., 2016), naieve bayes, Decision Tree, KNN and so on, or using certain unsupervised techniques such as LSA (Deerwester et al., 1990; Wang and Zheng, 2013; Gill et al., 2008), pLSA, NMF to transform the feature space into a more reasonable one before conducting classification. The main challenges of emotion analysis of tweets are the following:

1. Informal languages used on social media are not always obeying formal grammar, which makes traditional grammatical features less reliable for detecting emotions on social media.

2. New words are frequently created on social media, making it difficult to understand their emotional meaning even for a human being.

To address the challenges above, we use recurrent neural network to make use of terms, sequential information, and contextual information simultaneously for emotion detection. We believe that contextual information can partly solve the new-term problem and grammar-breach problem. To use recurrent neural network, a pre-trained embedding is used as our initial input of each term.

2 External Resource

We only used one external resource in our analysis, which is a pre-trained word embedding (Mikolov et al., 2013) word2vec provided by Google. It is trained on a part of the Google News dataset (about 100 billion words) and it contains 300-dimensional vectors for 3 million words and phrases.

---

[1] https://competitions.codalab.org/competitions/17751

3 System Description

To explore the limit of term features with RNN for emotion detection, we did not use various features other than term embedding. The system could be improved by using features like emojis or emoticons. We will conduct further analysis afterwards by addressing problems in combining different feature space.

As for the system used for SemEval18 task 1, the main steps, features used in the model are described in this section.

3.1 Preprocessing

Since the method heavily depends on terms appear in the text, the corpus is carefully pre-processed as described below.

- **Normalization** Each word in each tweet is converted into lowercase. Non-linguistic content such as URLs, emojis, emoticons, user names are removed (some important features such as emojis and emoticons will be explored in the future).

- **Tokenization** Each tweet is split into a word sequence. No stemming is applied since some special word forms may convey more apparent emotions than its original form.

- **Stop-words Removing** NLTK[2] toolkits is leveraged to remove stop words from tweets. Some other meaningless terms such as single characters, digits, their compositions and so on, are also eliminated.

3.2 Embedding Usage

Word embeddings are a widely used semantic presentation of words for almost any neural network based text analysis approach. A vector of real numbers is used for a single word to represent its distributional semantics in the embedding space. Since the space is generated by the language model, words that are functional similar in certain language are close with each other in the embedding space, for example, "cat" and "dog" could be similar in the space.

In this system, a tweet is represented by concatenating embeddings of the words in it.

3.3 Our Approach

The system submitted is based on a recurrent neural network approache, GRU, to be specific.

_____
[2] https://www.nltk.org/

3.3.1 The Basic Idea

Lexicons play the key role in lexicon-based approaches and bag-of-feature based machine learning approaches for emotion analysis. However, in addition to the emotion lexicons, we believe that linguistic characteristics may also contribute a lot to emotion analysis. For example, the context of the emotion lexicon such as negation could revert the emotion of the utterence if it is neglected. The sequence of the sentence terms also play an important role for understanding its meaning, hence important for uncovering the true emotions. Lack of newly created terms in vocabulary or grammatically incorrect utterances can also lead to poor performance of traditional emotion analysis approaches. By modeling long-term dependencies of terms inside a tweet, fusing the semantics of the terms and their contexts together with GRU, a recurrent neural network, we hope that above problems can be alleviated in the new space.

3.3.2 Problem Statement

We take emotion analysis as a multi-label classification problem in our system as usual. A tweet is represented as a sequence of terms

$$x_i = \{w_i^0, w_i^1, \cdots, w_i^M\} \tag{1}$$

where M is the length of the tweet x_i. Given a tweet x_i, the task aims at prediting the labels of it as y_i, where y_i is a d-dimensional Boolean vector $y_i \in B^d$, $d = 11$ in this case. Each dimension of vector y_{ij} indicates an emotion label of the 11 labels space _anger, anticipation, disgust, fear, joy, love, optimism, pessimism, sadness, suprise_, and _trust_ respectively. For example $y_{i0} = 1$ means emotion of _anger_ is detected in the tweet x_i, $y_{i9} = 0$ means emotion of _sadness_ does not appear in the tweet.

3.3.3 The model architecture

The architecture of the model is shown in Figure 1. The model is composed of 3 layers. The input of the network is the pre-trained embedding of each term occurs in a tweet. The sequential term embeddings are then turned into a tweet level representation by a classical GRU, the output of GRU is directly inputed into a linear perceptron layer, and maps the tweet representation into class representation directly by this layer. The output of the linear perceptrons are then processed by a sigmoid function to get the final predictions.
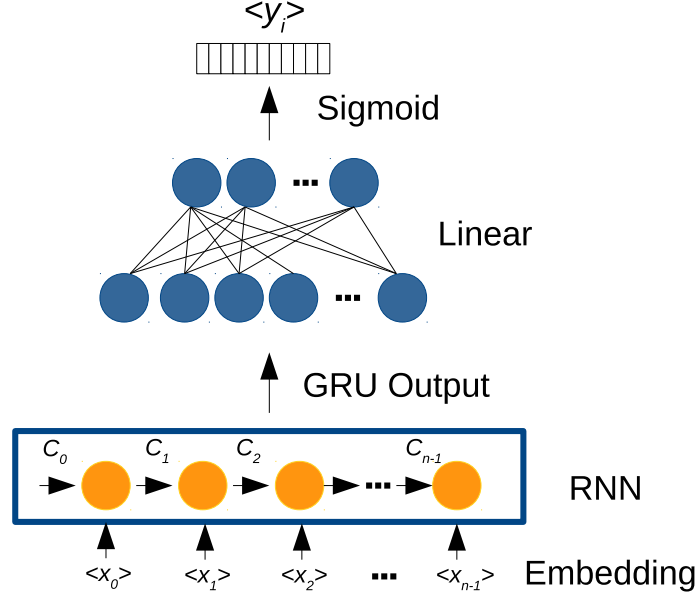
Figure 1: Overview of the architecture

3.3.4 The GRU Layer

The input of the GRU layer is the sequence of term embeddings of the tweet. We denote by $H = h_1, \ldots, h_n$ the input sequence of length n, where $h_i \in R^d$ is the term representation for the i-th token w_i. The new representation of the whole tweet r_i is then obtained from h_i via a GRU network:

$$r_i = (1 - z_i) \odot r_{i-1} + z_i \odot \tilde{r}_i \qquad (2)$$

where,

$$
\begin{aligned}
g_i &= \sigma(W_g r_{i-1} + U_g h_i) \\
z_i &= \sigma(W_z r_{i-1} + U_z h_i) \\
\tilde{r}_i &= \tanh(W_r(g_i \odot r_{i-1}) + U_r h_i).
\end{aligned}
\qquad (3)
$$

Here, g_i and z_i are reset and update gates respectively that control the information flow from the previous timestamp. W_g, U_g, W_z, U_z, W_r, and U_r are weight matrices to be learned for transforming r_{i-1} and h_i to gate units. By applying GRU on h_i, the tweet representation $r_i \in R^K$ encodes the context information and historical information simutaneously.

3.3.5 The Perceptron Layer

With the output of GRU, a vector $r_j \in R^K$ representing the overall information of a tweet, we use a perceptron layer together with a sigmoid activation function to map the tweet from feature space to label space $y_j \in R^L$, where L is dimension of the label space.

$$
\begin{aligned}
\hat{y}_j &= \sigma(W_p r_j) \\
\sigma(x) &= \frac{1}{1 + e^{-x}}
\end{aligned}
\qquad (4)
$$

The predicted label vector $\hat{y}_j$ of each tweet t_j is then compared with the true label vector y_j on training data to guide the training process with an appropriate loss function.

3.3.6 The Loss Function

Binary cross entropy loss can be used for multi-label classification problems, it is computed as the formula below:

$$
\begin{aligned}
loss&(p, q) \\
&= -\frac{1}{N} \sum_{i=1}^{N} H(p_n, q_n) \\
&= -\frac{1}{N} \sum_{i=1}^{N} [y_i \log \hat{y}_i + (1 - y_i) \log(1 - \hat{y}_i)]
\end{aligned}
\qquad (5)
$$

where y_i is the true label vector of the tweet x_i, and $\hat{y}_i$ is its predicted label vector.

4 Training

To train these models, we use the training data provided by SemEval18 task 1, which includes $6,839$

human labeled english tweets for Subtask 1. A data set of 887 labeled english tweets for development is also avaible, we leverage this set for validation. The trained model is then applied on the testing set with 3260 unlabeled tweets in it. A vocabulary is generated by extracting terms from all the training set, validating set and testing set to ensure its coverage.

The parameter configuration for the best system performance on validation set is defined as follows:

- **Hidden Dimension** The initial embedding of each term is 300 as we adopt the pretrained word embedding trained on part of the Google News dataset[3]. The hidden dimension of GRU is set to 200 when we get the best validation result.

- **Maximum Tweet Length** The length of each tweet is different, we regularize the length with a maximum limit of 30 meaningful terms after preprocessing steps. A tweet that is longer than that is trimmed, and shorter than that is populated with zero paddings.

- **Learning Rate** We adopt an Adam optimizer to train the model for the submitted system. The learning rate for the optimizer is set to 0.0001 when we get the best system performance on validation set.

- **Dropout Rate** Dropout operation is reported to have similar effects of boosting approches in neural network based models. A dropout operation is executed on the linear perceptron layer with a dropout rate of 0.4 when achieving the optimum.

- **Batch Size** The batch size settings also affact the performance of the proposed system. The optimum is obtained with a setting of 20 as the batch size.

The number of epochs is used for terminating the training process when optimum is obtained. Terminating condition depends on not only the values of the loss function, but also its transient performance on the validating dataset. Some random factors, such as the initial state of various random variables also show their impacts on it. In our experiments, the optimum is achieved at the 3-rd

epoch, it may vary with different intial states of other parameters.

Using the model parameters above, which produced the best performance on validating dataset, we predicted the labels of each tweet in the testing dataset. The evaluation results provided by SemEVal18 is described in the next section.

5 Results

Among all the 35 systems which participated in the task of emotion classification subtask of task 1 (Mohammad et al., 2018), our only submission is ranked 16-th on the evaluation metric of Accuracy, and 19-th on both metrics of micro-avg F1 and macro-avg F1, as is shown in Table 1. Our

Rank	System	Acc.	mi-F1	Ma-F1
1	cbaziotis	0.588	0.701	0.528
15	**mutux**	**0.473**	**0.591**	**0.446**
21	SVM	0.442	0.570	0.443
28	Random	0.185	0.307	0.285

Table 1: System Evaluation Results

model structure and feature space are designed as simple as possible intentionally, so that it can test the idea without distractive factors. As shown in above table, the system outperforms SVM-based approach consistently on all three different evaulation metrics.

6 Conclusion

We have presented a GRU-based multi-lable classifier to leverage context information and historical information for emotion analysis. It outperforms the unigram SVM model consistently on three evaluation metrics, even though only term features and a pre-trained word embedding are used. Some key factors such emojis, emoticons, emotion lexicons and multi-layer neural structures will be explored in the futrue for further analysis.

Acknowledgments

References

Muhammad Abdul-Mageed and Lyle Ungar. 2017. Emonet: Fine-grained emotion detection with gated

---

[3]https://code.google.com/archive/p/word2vec/

recurrent neural networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 718–728.

Ameeta Agrawal and Aijun An. 2012. Unsupervised emotion detection from text using semantic and syntactic relations. In *Proceedings of the The 2012 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology-Volume 01*, pages 346–353. IEEE Computer Society.

Rakesh C Balabantaray, Mudasir Mohammad, and Nibha Sharma. 2012. Multi-class twitter emotion classification: A new approach. *International Journal of Applied Information Systems*, 4(1):48–53.

Alexandra Balahur, Jesús M. Hermida, and Andrés Montoyo. 2011. Detecting implicit expressions of sentiment in text based on commonsense knowledge. In *Proceedings of the 2Nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis*, WASSA '11, pages 53–60, Stroudsburg, PA, USA. Association for Computational Linguistics.

F. Bravo-Marquez, E. Frank, S. M. Mohammad, and B. Pfahringer. 2016. Determining word-emotion associations from tweets by multi-label classification. In *2016 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*, pages 536–539.

Lea Canales and Patricio Martínez-Barco. 2014. Emotion detection from text: A survey. In *Proceedings of the Workshop on Natural Language Processing in the 5th Information Systems Research Working Days (JISIC)*, pages 37–43.

Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE*, 41(6):391–407.

Alastair J Gill, Robert M French, Darren Gergle, and Jon Oberlander. 2008. Identifying emotional characteristics from short blog texts. In *Proc. 30th Ann. Conf. Cognitive Science Soc., BC Love, K. McRae, and VM Sloutsky, eds*, pages 2237–2242.

Maryam Hasan, Emmanuel Agu, and Elke Rundensteiner. 2014a. Using hashtags as labels for supervised learning of emotions in twitter messages. In *ACM SIGKDD Workshop on Health Informatics, New York, USA*.

Maryam Hasan, Elke Rundensteiner, and Emmanuel Agu. 2014b. Emotex: Detecting emotions in twitter messages.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

Saif M. Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. Semeval-2018 Task 1: Affect in tweets. In *Proceedings of International Workshop on Semantic Evaluation (SemEval-2018)*, New Orleans, LA, USA.

Kirk Roberts, Michael A Roach, Joseph Johnson, Josh Guthrie, and Sanda M Harabagiu. 2012. Empatweet: Annotating and detecting emotions on twitter. In *LREC*, volume 12, pages 3806–3813. Citeseer.

Carlo Strapparava and Rada Mihalcea. 2008. Learning to identify emotions in text. In *Proceedings of the 2008 ACM Symposium on Applied Computing*, SAC '08, pages 1556–1560, New York, NY, USA. ACM.

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*, pages 121–136. Springer.

Ro Valitutti. 2004. Wordnet-affect: an affective extension of wordnet. In *In Proceedings of the 4th International Conference on Language Resources and Evaluation*, pages 1083–1086.

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)*, pages 587–592. IEEE.

Xuren Wang and Qiuhui Zheng. 2013. Text emotion classification research based on improved latent semantic analysis algorithm. In *Proceedings of the 2nd International Conference on Computer Science and Electronics Engineering (ICCSEE 2013)*, pages 210–213.