EPITA-ADAPT at SemEval-2019 Task 3: Using Deep Sentiment Analysis Models and Transfer Learning for Emotion Detection in Textual Conversations

Abdessalam Bouchekif [1] abdessalam.bouchekif@epita.fr
Praveen Joshi [2] praveen.joshi@mycit.ie
Latifa Bouchekif [3] latifa.bouchekif@univ-tlemcen.dz
Haithem Afli [2] haithem.afli@cit.ie

[1] LSE, EPITA Graduate School of Computer Science, France
[2] ADAPT Centre, Cork Institute of Technology, Cork, Ireland
[3] University Abou Bekr Belkaid-Tlemcen, Algeria

Abstract

Messaging platforms like WhatsApp, Facebook Messenger and Twitter have gained recently much popularity owing to their ability in connecting users in real-time. The content of these textual messages can be a useful resource for text mining to discover and unhide various aspects, including emotions. In this paper we present our submission for SemEval 2019 task 'EmoContext'. The task consists of classifying a given textual dialogue into one of four emotion classes : *Angry*, *Happy*, *Sad* and *Others*. Our proposed system is based on the combination of different deep neural networks techniques. In particular, we use Recurrent Neural Networks (LSTM, B-LSTM, GRU, B-GRU), Convolutional Neural Network (CNN) and Transfer Learning (TL) methods. Our final system, achieves an $F1\mu$ score of 74.51% on the subtask evaluation dataset.

1 Introduction

The world of text conversations has undergone drastic changes during the last few years. The generation and sharing of such information have become much easier than before with the advent of popular social media platforms such as Twitter, Facebook, *etc.* According to *Statistica* [1] WhatsApp is the most popular mobile messaging app in the world with one billion monthly active users. Facebook Messenger closely follows with 900 million monthly active users. The content of these messages can be useful resource for text mining to discover and unhide various aspects, including emotions (Chatterjee et al., 2019a).

Capturing and analysing these emotions from peoples conversations has raised growing interest within the scientific community in varied fields like cognitive and social psychology, signal processing and natural language processing (Gupta et al., 2017; Zhang et al., 2018; Majumder et al., 2018).

The goal of detecting emotions in SemEval2019 task3 described in (Chatterjee et al., 2019b) is to classify a given conversation into one of four classes - *happy*, *sad*, *angry* and *others*, that best represents the mental state of the users. This can be seen as a multiclass classification problem.

In this paper, we propose an approach to detect emotions like *happy*, *sad* or *angry* in textual messages using a combination of deep learning models. We apply, also, a transfer learning approach, from a model trained on a similar task consists on the prediction of the sentiment of the conversation, *i.e. positive*, *negative* or *neutral*. Then, the pre-trained model is re-used to classify the dialogue into one of four classes : *happy*, *sad*, *angry* and *others*.

The rest of the paper is organized as follows. Section 2 provides a brief literature review on emotion detection in textual datasets. The description of our proposed system is presented in Section 3. The experimental set-up and results are described in Section 4. Finally, a conclusion is given with a discussion of future works in Section 5.

2 Related Work

Emotions are closely related to sentiment with more analysis of the inferred polarity. For example, a negative sentiment can be caused by sadness or anger, while a positive sentiment can be caused by happiness or anticipation. Thus, following the way in sentiment analysis, many deep learning models are applied to detect emotions (Zhang et al., 2018; Poria et al., 2017).

(Zhou et al., 2018) proposed an Emotional Chatting Machine (ECM) that can generate appropriate responses grammatically relevant and emotionally consistent based on GRU. Their system

1. https://www.statista.com/topics/1145/internet-usage-worldwide/

Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval-2019), pages 215–219
Minneapolis, Minnesota, USA, June 6–7, 2019.

is modeling the emotion factor, using emotion category embedding, internal emotion memory, and external memory. A bilingual attention network model was proposed by (Wang et al., 2016) for code-switched emotion prediction. The authors used a LSTM model to construct a document level representation of each post, and an attention mechanism to capture the informative words from bilingual and monolingual contexts. (Abdul-Mageed and Ungar, 2017) built a large, automatically curated dataset for emotion detection using distant supervision and then used GRNNs to model fine-grained emotion. They extended the classification to model (Plutchik, 2001)'s 8 primary emotion dimensions.

(Felbo et al., 2017) show how millions of readily available emoji occurrences on Social Media can be used to pre-train models to learn a richer emotional representation. They transfer this knowledge to emotion, sarcasm and sentiment detection tasks using a new layer-wise fine-tuning method.
In (Daval-Frerot et al., 2018) the authors used a transfer learning approach, from a model trained on a similar task. They propose to pre-train a model to predict if a tweet is positive, negative or neutral by applying a B-LSTM on an external dataset. Then, they used the pre-trained model to classify a tweet according to the seven-point (range from 5 to +5 respectively from very negative to very positive) scale of positive and negative sentiment intensity.

3 Proposed System

Figure 1 provides a high-level overview of our system, which consists of three steps :

1. First step applies the basic text processing (Tokenisation, lemmatisation, filtering the noise from the raw text data, *etc*) and represents words in textual dialogue as vectors.
2. In the second step, we learn a model for each one of the emotions : *angry*, *happy* and *sad*.
3. The last step does the prediction based on the probabilities of our three models. The system classifies the dialogue in one of four classes (*angry*, *happy*, *sad* and *others*).

In this work, we consider each conversation as one single input, *i.e.* we didn't take into account the writing turn (utterance). Our decision was based on the fact that the language and the size of these conversations are similar to the standard user generated content type of data sets.

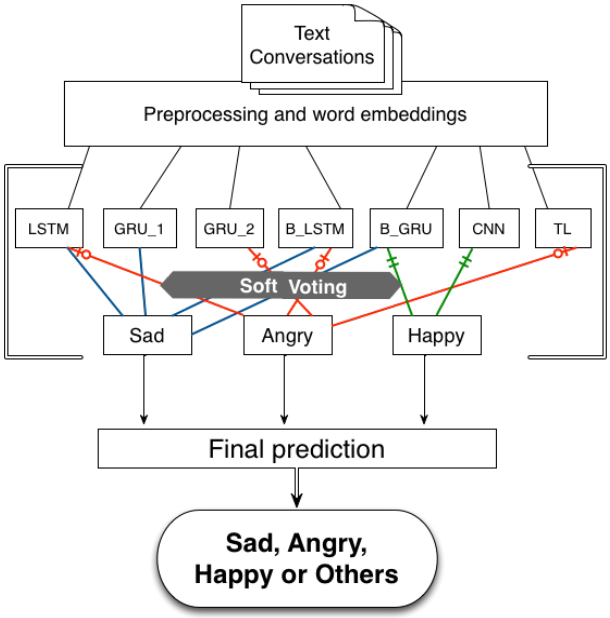

Figure 1 – Our proposed system architecture

3.1 Pre-processing and word representation

The textual dialogues are processed using *ekphrasis* [2] tool which allows performing the following tasks : tokenization, word normalization, word segmentation (for splitting hashtags) and spell correction (*i.e* replace a misspelled word with the most probable candidate word). All words are lowercased. E-mails, URLs and user handles are normalized. A detailed description of this tool is given in (Baziotis et al., 2017).
Each word in the text is represented by a vector of real numbers capturing the semantic meanings of words. We used datastories embeddings (Baziotis et al., 2017) trained on 330M english twitter messages posted from 12/2012 to 07/2016. The embeddings used in this work are 300 dimensional.

3.2 Neural Networks

With the recent advances in deep learning, the ability to analyse sentiments has considerably improved. Indeed, many experiments have used state-of-the-art systems to achieve high performance. For example, (Baziotis et al., 2017) use Bi-directional Long Short-Term Memory (B-LSTM) with attention mechanisms while (Deriu et al., 2016) use Convolutional Neural Networks (CNN). Both systems obtained the best performance at the the 2016 and 2017 SemEval 4-A task respecti-

2. https://github.com/cbaziotis/ekphrasis

vely. In this work, we use Long Short-Term Memory (LSTM), B-LSTM, Gated Recurrent Unit (GRU), Bidirectional-GRU (B-GRU) and CNN models and we, also, apply a Transfer Learning (TL) approach.

3.2.1 LSTM, GRU_1 and GRU_2 models

LSTM (Schuster and Paliwal, 1997) and GRU (Cho et al., 2014) are a special kind of RNN, capable of learning long-term dependencies. This ability comes from LSTM cells, that can decide which information to add and remove to the cell state regulated by gates (input gate, output gate and forget gate). The GRU cell is a simplified version of the LSTM cell.
LSTM and GRU_1 contain 2 layers of 128 neurons each. Similarly to (Baziotis et al., 2017), we added a Gaussian noise at the embedding layer with $\sigma = 0.3$ and dropout of 0.3 at LSTM/GRU layers. GRU_2 is similar to GRU_1, with some little differences where GRU_2 contains 3 layers of 100 neurons and dropout of 0.2 at the embedding layer.

3.2.2 B-LSTM and B-GRU models

The B-LSTM become a standard for deep sentiment analysis (Baziotis et al., 2017; Daval-Frerot et al., 2018; Moore and Rayson, 2017). B-LSTM and B-GRU consists of two LSTMs and two GRUs respectively in different directions running in parallel : the first forward network reads the input sequence from left to right and the second backward network reads the sequence from right to left. Each LSTM / GRU yields a hidden representation : $\vec{h}$ (left to right vector) and $\overleftarrow{h}$ (right-to-left vector) which are then combined to compute the output sequence. For our problem, capturing the context of words from both directions allows to better understand the textual conversations semantic. We use the same parameters of LSTM and GRU_1 models.

3.2.3 CNN model

The application of CNN models started with visual imagery. Many works apply CNN for sentiment analysis and obtained interesting results (Kim, 2014; Ouyang et al., 2015; Deriu et al., 2016). CNN is typically composed of three types of layers : convolution, pooling, and fully connected layers. Each neuron in the convolutional layer is connected only to a local region. In this work, we use multiple convolutional filters of sizes 3, 4 and 5.

3.2.4 TL-BLSTM model

In this work, we apply a TL, which allows to avoid learning from scratch. TL consists of transferring the knowledge learned on one task to a second related task. We start by training a first model to predict the sentiment of the dialogue : *positive*, *negative* or *neutral*. For this, we added a dense layer of 3 neurons to our B-LSTM (see subsection 3.2.2) . The model is learned using an external dataset [3] composed of 50333 tweets (7840 negatives, 19903 positives and 22590 neutrals). Then, the first model is re-used as the starting point to train a new model that classifies the dialogue into one of four classes : *happy*, *sad*, *angry* and *others*. For this, we remove the last layer of the pre-trained model and we add a fully-connected layer of 128 neurons followed by an output layer of 3 neurons (similar to our previous work (Daval-Frerot et al., 2018)).

4 Experimental settings and results

For each emotion, we use only the best models which maximizes the $F1\mu$ score. Table 2 presents the selected models for each emotion. Then, we combine these models using the soft voting approach considering only the probability of the selected emotion : *happy*, *sad* or *angry*.

For example, suppose that the CNN model gives 0.7, 0.1, 0.2 and 0.0 respectively for *happy*, *sad*, *angry* and *others*. And the GRU_1 model gives 0.6, 0.2, 0.0 and 0.2 for the same classes. We know, based on our results in table 2, that the CNN and GRU_1 are the best models for the emotion *happy*. So the soft voting combination is the average probability of 0.7 and 0.6, *i.e* 0.65.

In the last step we are applying a threshold after getting all prediction probabilities for our three classes. We choose threshold scores of 0.75 for angry and sad, and 0.67 for happy based on our experiments on the development set. These scores were high in order to avoid any possible confusion with the class *others*. This confusing can be caused by the fact that our training data is unbalanced.

Table 1 illustrates the performances of our system on *Dev* and *Test* sets. We can see an overall similarity in term of performance the system in both data sets. Indeed, the system achieves 74.4% and 74.5% on *Dev* and *test* sets respectively. The re-

3. https://github.com/cbaziotis/datastories-semeval2017-task4/tree/master/dataset/Subtask_A/downloaded

	Sad			Angry			Happy			Micro Average		
	P	*R*	*F1*	*P*	*R*	*F1*	*P*	*R*	*F1*	$P\mu$	$R\mu$	$F1\mu$
Dev	77.2	78.4	77.8	74.2	74.7	74.4	70.6	72.5	71.5	73.8	75.1	74.4
Test	82.5	75.6	78.9	73.9	75.8	74.8	70.8	70.1	70.4	75.2	73.8	74.5

Table 1 – performances of our system on *Dev* and *Test* corpora.

	models	$P\mu$	$R\mu$	$F1\mu$
Sad	B-GRU	73.7	79.6	76.5
	B-LSTM	72.6	78.8	75.6
	LSTM	80.1	71.2	75.4
	GRU_1	80.4	70.8	75.3
Angry	B-GRU	72.5	76.1	74.3
	GRU_2	69.5	78.1	73.6
	TL-BLSTM	70.1	76.5	73.1
	LSTM	72.2	71.4	71.8
Happy	CNN	69.4	68.6	69.6
	GRU_1	71.8	67.2	69.4

Table 2 – Models used for each emotion and their scores on test set.

sults shows a good performance for the *happy* and *Angry* and a better detection for the class *Sad* with a score around 78%.

Our final combination system, achieves an $F1\mu$ score of 74.51% on the test set gaining more than 3.5% compared to the best model before the combination which is B-GRU. This improvement comes from the fact that all systems used for the combination are using different methods that allows the diversification of their results and maximise the effect the soft voting.

Finally we can mention that after analysing our results, we have seen that most of errors came from the confusion between the class *Others* and the rest (*Happy*, *Angry* and *Sad*) which will be investigated in our future work.

5 Conclusion

In this paper, we propose to use a combination of different deep neural networks techniques including LSTM, B-LSTM, GRU, B-GRU, CNN and TL methods for the SemEval2019 task 3 of Emotions detection in textual conversations. Our system achieves a final $F1\mu$ of 74.51% on the subtask evaluation dataset.

As future work, we plan to develop an attention model to determine the importance of each part of the conversation (utterance) and its specific contribution to the emotion classification. We plan, also, to extend our work to other modalities such as audio emotions classification.

6 Acknowledgments

This research is partially supported by EPITA Systems Laboratory (www.lse.epita.fr) and Science Foundation Ireland through ADAPT Centre (Grant 13/RC/2106) (www.adaptcentre.ie).

References

Muhammad Abdul-Mageed and Lyle Ungar. 2017. Emonet : Fine-grained emotion detection with gated recurrent neural networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, page 718728, Vancouver, Canada.

Christos Baziotis, Nikos Pelekis, and Christos Doulkeridis. 2017. Datastories at semeval-2017 task 6 : Siamese LSTM with attention for humorous text comparison. In *Proceedings of the 11th International Workshop on Semantic Evaluation, SemEval@ACL 2017, Vancouver, Canada*.

Ankush Chatterjee, Umang Gupta, Manoj Kumar Chinnakotla, Radhakrishnan Srikanth, Michel Galley, and Puneet Agrawal. 2019a. Understanding emotions in text using deep learning and big data. *Computers in Human Behavior*, 93 :309–317.

Ankush Chatterjee, Kedhar Nath Narahari, Meghana Joshi, and Puneet Agrawal. 2019b. Semeval-2019 task 3 : Emocontext : Contextual emotion detection in text. In *Proceedings of The 13th International Workshop on Semantic Evaluation (SemEval-2019)*, Minneapolis, Minnesota.

Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1724–1734.

Guillaume Daval-Frerot, Abdessalam Bouchekif, and Anatole Moreau. 2018. Epita at semeval-2018 task 1 : Sentiment analysis using transfer learning approach. In *Proceedings of The 12th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT, New Orleans, Louisiana, June 5-6, 2018*, pages 151–155.

Jan Deriu, Maurice Gonzenbach, Fatih Uzdilli, Aurélien Lucchi, Valeria De Luca, and Martin Jaggi. 2016. Swisscheese at semeval-2016 task 4 : Sentiment classification using an ensemble of convolutional neural networks with distant supervision. In *Proceedings of the 10th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT, USA*.

Bjarke Felbo, Alan Mislove, Anders Sgaard, Iyad Rahwan, and Sune Lehmann. 2017. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, page 16151625.

Umang Gupta, Ankush Chatterjee, Radhakrishnan Srikanth, and Puneet Agrawal. 2017. A sentiment-and-semantics-based approach for emotion detection in textual conversations. In *Proceedings of Neu-IR 2017 SIGIR Workshop on Neural Information Retrieval (Neu-IR 17)*.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. *CoRR*, abs/1408.5882.

Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria. 2018. Dialoguernn : An attentive rnn for emotion detection in conversations. In *arXiv :1811.00405v3*.

Andrew Moore and Paul Rayson. 2017. Lancaster A at semeval-2017 task 5 : Evaluation metrics matter : predicting sentiment from financial news headlines. In *Proceedings of the 11th International Workshop on Semantic Evaluation, SemEval@ACL 2017, Canada*.

X. Ouyang, P. Zhou, C. H. Li, and L. Liu. 2015. Sentiment analysis using convolutional neural network. In *2015 IEEE International Conference on Computer and Information Technology ; Ubiquitous Computing and Communications ; Dependable, Autonomic and Secure Computing ; Pervasive Intelligence and Computing*, pages 2359–2364.

Robert Plutchik. 2001. The nature of emotions human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. In *American scientist*, volume 89(4), page 344350.

Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Mazumder, Amir Zadeh, and Louis-Philippe Morency. 2017. Context-dependent sentiment analysis in user-generated videos. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, page 873883, Vancouver, Canada.

Mike Schuster and Kuldip K. Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE Trans. Signal Processing*.

Zhongqing Wang, Yue Zhang, Sophia Yat Mei Lee, Shoushan Li, and Guodong Zhou. 2016. A bilingual attention network for code-switched emotion. In *Proceedings of the International Conference on Computational Linguistics (COLING 2016)*.

Lei Zhang, Shuai Wang, and Bing Liu. 2018. Deep learning for sentiment analysis : A survey. In *arXiv :1801.07883v2*.

Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. 2018. Emotional chatting machine : Emotional conversation generation with internal and external memory. In *Proceedings of The Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18))*.