SINAI at SemEval-2019 Task 3: Using affective features for emotion classification in textual conversations

Flor Miriam Plaza-del-Arco, M. Dolores Molina-González,
M. Teresa Martín-Valdivia, L. Alfonso Ureña-López

Department of Computer Science, Advanced Studies Center in ICT (CEATIC)
Universidad de Jaén, Campus Las Lagunillas, 23071, Jaén, Spain
{fmplaza, mdmolina, maite, laurena}@ujaen.es

Abstract

Detecting emotions in textual conversation is a challenging problem in absence of nonverbal cues typically associated with emotion, like facial expression or voice modulations. However, more and more users are using message platforms such as WhatsApp or telegram. For this reason, it is important to develop systems capable of understanding human emotions in textual conversations. In this paper, we carried out different systems to analyze the emotions of textual dialogue from SemEval-2019 Task 3: EmoContext for English language. Our main contribution is the integration of emotional and sentimental features in the classification using the SVM algorithm.

1 Introduction

Emotions seem to govern our daily lives since most of our decisions are guided by our mood. They are complex and that is why they have been studied in many areas over time. Given the importance to develop systems to be able to mimic functioning of the human brain, emotions have attracted the attention in the field of affective computing (Thilmany, 2007).

To our knowledge, there are not many works that focus on studying how emotions are reflected verbally. However, studying emotions on text messaging platforms such as WhatsApp, Facebook Messenger or Telegram is important as more and more users are using them to share their experiences and emotions.

Currently, detecting emotions in instant messaging has multiple applications in different fields (Gupta et al., 2017; Yadollahi et al., 2017; Hakak et al., 2017), such as businesses intelligence to increase customer satisfaction knowing their preferences, social media to alert users if they are going to post an offensive tweet or psychology to detect some disorders like anorexia, anxiety or stress.

In this paper, we present the different systems we developed as part of our participation in SemEval-2019 Task 3: Contextual Emotion Detection in Text (EmoContext) (Chatterjee et al., 2019b). It is an emotion classification task. Given a textual dialogue along with two turns of context, its consists of classify the emotion of user utterance as one of the emotion classes: Happy, Sad, Angry or Others.

The rest of the paper is structured as follows. In Section 2 we explain the data used in our methods. Section 3 introduces the lexical resources used for this work. Section 4 presents the details of the proposed systems. In Section 5, we discuss the analysis and evaluation results for our system. We conclude in Section 6 with remarks and future work.

2 Data

To run our experiments, we used the English datasets provided by the organizers in SemEval19 Task 3 : EmoContext (Chatterjee et al., 2019b). The datasets containing 3-turn conversations along with their emotion class labels (Happy, Sad, Angry, Others) provided by human judges. The Turn 1 contains the first turn in the three turn conversation, written by User 1. The turn 2 contains the second turn, which is a reply to the first turn in conversation and it is written by User 2 and finally, the turn 3 contains the last turn, which is a reply to the second turn in the conversation, which is written by User 1.

During pre-evaluation period, we trained our models on the train set, and evaluated our different approaches on the dev set. During evaluation period, we trained our models on the train and dev sets, and tested the model on the test set. Table 1 shows the number of 3-turn conversations used in our experiments in English.

Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval-2019), pages 307–311
Minneapolis, Minnesota, USA, June 6–7, 2019.

Dataset	train	dev	test
Happy	4,243	142	284
Sad	5,463	125	250
Angry	5,506	150	298
Others	14,948	2,338	4,677
Total	30,160	2,755	5,509

Table 1: Number of 3-turn conversations per EmoContext dataset

3 Resources

For the development of the task, we used different lexicons that we explain in detail below.

NRC Affect Intensity Lexicon (Mohammad, 2017). It has almost 6,000 entries in English. Each of them has an intensity score associated to one of the following basic emotions: anger, fear, sadness and joy. The scores range from 0 to 1, where 1 indicates that the word has a high association to the emotion and 0 that the word has a low association to the emotion.

NRC Word-Emotion Association Lexicon (EmoLex) (Mohammad and Turney, 2010). This lexicon has a list of English words associated to one or more of the following emotions: anger, fear, anticipation, trust, surprise, sadness and joy.

VADER (Valence Aware Dictionary and sEntiment Reasoner) (Gilbert, 2014). The VADER sentiment lexicon is a rule-based sentiment analysis tool. This is sensitive both the polarity and the intensity of sentiments expressed in social media contexts, and is also generally applicable to sentiment analysis in other domains. VADER has been validated by multiple independent human judges. The tool returns four values: positive, negative, neutral and compound. The first three scores represent the proportion of text that falls in these categories. The compound score is computed by summing the valence scores of each word in the lexicon, adjusted according to the rules, and then normalized to be between -1 (most extreme negative) and +1 (most extreme positive).

4 System Description

In this section, we describe the systems developed for the EmoContext task. During our experiments, the scikit-learn machine learning in Python library (Pedregosa et al., 2011) was used for benchmarking.

4.1 Data Preprocessing

In first place, we preprocessed the corpus of conversations provided by the organizers. We applied the following preprocessing steps: the documents were tokenized using NLTK Tweet Tokenizer[1] and all letters were converted to lower-case.

4.2 Feature Extractor

Converting sentences into feature vectors is a focal task of supervised learning based sentiment analysis method. Therefore, the features we chose in our system can be divided into three parts: statistic features, morphological features and lexical features.

- **Statistic features**. We employed the feature that usually perform well in text classification: Term Frecuency (TF) taking into account unigrams and bigrams.
- **Morphological features**. We employed Part-of-speech tagging (PoS). For each sentence, we obtain a vector associated with the part of speech recognized in each word of the sentence.
- **Lexical features**. As we explained in Section 3, we used three lexicons obtained different features in the following way:
  1. **NRC Affect Intensity**. We checked the presence of lexicon terms in the sentence and then we computed the sum of the intensity value of the words of the sentence grouping them by the emotional category (*fear*, *sadness*, *anger* and *joy*). Therefore, we obtained a vector of four values (four emotions) for each sentence. Each value of intensity is normalized $\hat{i}_e$ applying the following equation:

$$\hat{i}_e = \frac{i_e}{\sum_e i_e}$$

Where e = {fear, sadness, anger, joy} and i_e is equal to value of intensity per emotion. Note that the components of the normalized vector add up to 1, and each of them is a positive number between 0 and 1.

[1] https://www.nltk.org/api/nltk.tokenize.html

2. **Emolex**. We identified the presence of lexicon terms in the sentence and we assigned 1 as confidence value (CV). Then, we summed the CV of the words whose emotion is the same obtaining a vector of emotions for each sentence. As a result, we obtained a vector of eight values (eight emotions). Each value $\hat{v}_e$ is normalized following the next equation:

$$\hat{v}_e = \frac{CV_e}{\sum_e CV_e}$$

Where e = {anger, fear, anticipation, trust, surprise, sadness and joy} and CV_e is equal to confidence value per emotion. Note that the components of the normalized vector add up to 1, and each of them is a positive number between 0 and 1.

3. **VaderSentiment**. We use the sentiment.vader module[2] provided by the Natural Language Toolkit (NLTK). With this module, we analyze each sentence and we obtained a vector of four scores: negative sentiment, positive sentiment, neutral sentiment and compound polarity.

4.3 Classifier

The concatenation of the features described before are applied for the classification using the SVM algorithm. We selected the Linear SVM formulation, known as C-SVC and the value of the C parameter was 1.0.

5 Experiments and analysis of results

During the pre-evaluation phase we carried out several experiments and the best experiments were taken into account for the evaluation phase. The architecture of the different systems can be seen in Figure 1 and are described below:

- **Basic system (BS)**. For this experiment, we have combined the 3-turn conversations of the corpus in a text string separated by spaces. For example, for turn 1: "Hahah i loved it" , turn 2: "Yay! Glad you loved it X3" and turn 3: "You always make us happy", the final sentence is "Hahah i loved it Yay! Glad you loved it X3 You always make us happy". Then, each sentence is represented as a vector of unigrams and bigrams choosing the TF weighting scheme and it is used as feature for the classification using the SVM algorithm.

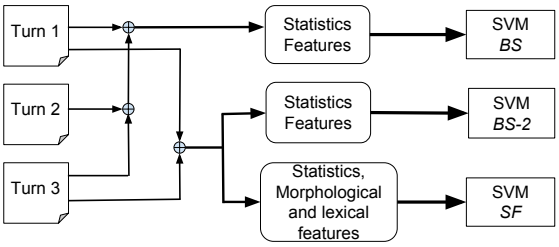

Figure 1: Systems architecture.

- **Basic system with turn 1 and 2 (BS-2)**. This experiment is similar to the previous one. However, we have only taken into account the first and last conversation turns because analyzing the training data, we realized that the second conversation turn is not useful for the classification as it does not provide representative information. For example, for turn 1: "Hahah i loved it" , turn 2: "Yay! Glad you loved it X3" and turn 3: "You always make us happy", the final sentence is "Hahah i loved it You always make us happy". We notice that the emotion is the same (happy) as if we consider the three turns.

- **System with features (SF)**. In this system, also we have only taken into account the first and last conversation turns. With these turns of conversations, we have tested several combinations with the lexical resources during the development phase and we chose the best combination for the evaluation phase. The best combination is the set of the vector of NRC (four values), the vector of Emolex (eight values) and the vector of VaderSentiment (four values) explained in Subsection 4.2. Therefore, the union of the best lexical features and the TF of the two conversation turns are used as features to perform the classification with the selected SVM algorithm.

The official competition metric to evaluate the

[2] https://www.nltk.org/_modules/nltk/sentiment/vader.html

Experiment	Sad			Angry			Happy			Micro - Avg		
	Prec	Rec	F1	Prec	Rec	F1	Prec	Rec	F1	Prec	Rec	F1
BS	0.49	0.73	0.58	0.54	0.79	0.64	0.43	0.73	0.54	0.48	0.75	0.59
BS-2	0.64	0.74	0.68	0.60	0.85	0.70	0.57	0.76	0.65	0.60	0.79	0.68
SF	0.63	0.81	0.71	0.62	0.87	0.72	0.57	0.77	0.66	0.61	0.82	0.7

Table 2: Results on the dev set.

Experiment	Sad			Angry			Happy			Micro - Avg		
	Prec	Rec	F1	Prec	Rec	F1	Prec	Rec	F1	Prec	Rec	F1
BS	0.59	0.74	0.66	0.53	0.77	0.66	0.5	0.64	0.56	0.56	0.72	**0.63**
BS-2	0.64	0.77	0.7	0.59	0.8	0.68	0.62	0.71	0.66	0.61	0.76	**0.68**
SF	0.61	0.78	0.69	0.58	0.81	0.68	0.61	0.70	0.65	0.6	0.76	**0.67**

Table 3: Results on the test set.

User name (ranking)	F1
leo1020 (1)	0.79
gautam_naik (60)	0.72
fmplaza (92)	**0.68**
emocontext_organizers (140)	0.59
waylensu (161)	0.0143

Table 4: System test results per user in EmoContext task.

systems in EmoContext task is the microaveraged F1 score (F1μ) for the three emotion classes (Happy, Sad and Angry). This metric is calculated between the real classes and the predicted classes. The results of our participation in the task can be seen in Tables 2 and 3.

In relation to our results, during the pre-evaluation phase and evaluation phase, we noticed that 1 and 3 conversation turns performed better the classification due to the reason that the 2 conversation turn is usually a contradiction or a question of the 1-turn. In Tables 2 and 3 we can observed that the BS-2 experiments outperformed the BS experiments. According to the classification per emotion, we may note different issues. On the one hand, the use of lexical features (SF experiment) improve about 2% of F1 with respect to the BS-2 experiment in the dev set. Nevertheless, this is not the case in the test set. On the other hand, the Happy emotion class perform worse than other emotion classes in both datasets, as it happens in other works (Chatterjee et al., 2019a; Gupta et al., 2017). Besides, if we observed the SF experiment in test set, we can see that the emotional features do not help to improve the classification because there are some words like "love" or "cool" whose assigned emotion is Happy class but in the 3-turn conversation of test set have been marked as Others class by the judges. Finally, in Table 4 we can observe our official position in the competition. We are ranked 92 out of 165 participating teams and our system outperforms the baseline system provided by the organizers of the task.

6 Conclusions and Future Work

In this paper, we present different systems to predict the emotion of user in a textual dialogue along with two turns of context. To carry out the task, we have developed three different systems. The first two are base systems, combining different turns of conversation and in the last system we decided to incorporate lexical features from sentiment and emotional resources.

In the future, we plan to continue working in emotion classification tasks because we have observed that the participation in this tasks is very high and this shows the interest by the scientific community in solving this type of tasks. Efforts will also be made to include more contextual information and to explore other multiple classifier methods.

Acknowledgments

This work has been partially supported by Fondo Europeo de Desarrollo Regional (FEDER) and REDES project (TIN2015-65136-C2-1-R) from the Spanish Government.

References

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.

CJ Hutto Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth International Conference on Weblogs and Social Media (ICWSM-14). Available at (20/04/16) http://comp. social. gatech. edu/papers/icwsm14. vader. hutto. pdf*.

Umang Gupta, Ankush Chatterjee, Radhakrishnan Srikanth, and Puneet Agrawal. 2017. A sentiment-and-semantics-based approach for emotion detection in textual conversations. *arXiv preprint arXiv:1707.06996*.

Nida Manzoor Hakak, Mohsin Mohd, Mahira Kirmani, and Mudasir Mohd. 2017. Emotion analysis: A survey. In *Computer, Communications and Electronics (Comptelix), 2017 International Conference on*, pages 397–402. IEEE.

Saif M Mohammad. 2017. Word affect intensities. *arXiv preprint arXiv:1704.08798*.

Saif M Mohammad and Peter D Turney. 2010. Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon. In *Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text*, pages 26–34. Association for Computational Linguistics.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830.

Jean Thilmany. 2007. The emotional robot: Cognitive computing and the quest for artificial intelligence. *EMBO reports*, 8(11):992–994.

Ali Yadollahi, Ameneh Gholipour Shahraki, and Osmar R Zaiane. 2017. Current state of text sentiment analysis from opinion to emotion mining. *ACM Computing Surveys (CSUR)*, 50(2):25.