ECNU at SemEval-2018 Task 1: Emotion Intensity Prediction Using Effective Features and Machine Learning Models

Huimin Xu[1], **Man Lan**[1,2*] **,Yuanbin Wu**[1,2]
[1]Department of Computer Science and Technology,
East China Normal University, Shanghai, P.R.China
[2]Shanghai Key Laboratory of Multidimensional Information Processing
51174506035@stu.ecnu.edu.cn, {mlan, ybwu}@cs.ecnu.edu.cn

Abstract

In this paper we describe our systems submitted to Semeval 2018 Task 1 "Affect in Tweet" (Mohammad et al., 2018). We participated in all subtasks of English tweets, including emotion intensity classification and quantification, valence intensity classification and quantification. In our systems, we extracted four types of features, including linguistic, sentiment lexicon, emotion lexicon and domain-specific features, then fed them to different regressors, finally combined the models to create an ensemble for the better performance. Officially released results showed that our system can be further extended.

1 Introduction

The Semeval 2018 Task 1 aims to automatically determine the intensity of emotions of the tweeters from their tweets, including five subtasks. That is, given a tweet and one of the four emotions (*anger*, *fear*, *joy*, *sadness*), the subtask 1 and 2 are to determine the intensity and classify the tweet into one of the four ordinal classes of intensity of the emotion respectively. Similarly, the subtask 3 and 4 determine the intensity and classify the tweet into one of seven ordinal classes of intensity of valance. Subtask 5 is a multi-label emotion classification task which classifies the tweets as neutral or no emotion or as one, or more, of eleven given emotions (*anger*, *anticipation*, *disgust*, *fear*, *joy*, *love*, *optimism*, *pessimism*, *sadness*, *surprise*, *trust*) that best represent the mental state of the tweeter. For each task, training and test datasets are divided into **English**, **Arabic**, and **Spanish** tweets. We participated in all subtasks of English tweets.

Traditional sentiment classification is a coarse-grained task in sentiment analysis which focuses on sentiment polarity classification of the whole sentence (*i.e.*, positive, negative, neutral, mixed). Semeval 2018 Task 1 subtask 5 takes basic human emotion proposed by Ekman (Ekman, 1999) into consideration, including *Anger*, *Anticipation*, *Disgust*, *Fear*, *Joy*, *Sadness*, *Surprise*, and *Trust*.

The difference between these subtasks lies in the emotion granularity and classification or quantification, so in our work, the similar method is adopted for five subtasks. We extracted a rich set of elaborately designed features. In addition to linguistic features, sentiment lexicon features and emotion lexicon features, we also extracted some domain specific features. Also, we conducted a series of experiments on different machine learning algorithms and ensemble methods to obtain the better performing for each subtask. For subask 5, we adopted multiple binary classification and constructed a model for each emotion.

2 System Description

We first performed data preprocessing, then extracted several types of features from tweets and constructed supervised models for this task.

2.1 Data Preprocessing

Firstly, all words are converted to lower case, URLs are replaced by "*url*", abbreviations, slangs and elongated words are transformed to their normal format. Then, emojis are replaced by corresponding emojis names by "Emoji Library"[1]. Finally, we use *Stanford CoreNLP tools* (Manning et al., 2014) for tokenization, POS tagging, named entity recognizing (NER) and parsing.

2.2 Feature Engineering

We extracted a set of features to construct supervised models for five subtasks, that is linguistic

[1] https://github.com/fvancesco/emoji/

Proceedings of the 12th International Workshop on Semantic Evaluation (SemEval-2018), pages 231–235
New Orleans, Louisiana, June 5–6, 2018. ©2018 Association for Computational Linguistics

features, sentiment lexicon features, emotion lexicon features and domain-specific features.

2.2.1 Linguistic Features

- ***Lemma unigram*** Considering there is similar emotion intensity expressed by "anger" and "angers", we choose word lemma unigram features from tweets rather than word unigram features.

- ***Negation*** Negation in a sentence often affects its sentiment orientation, and conveys its intensity of the sentiment. For example, a sentence with several negation words is more inclined to negative sentiment polarity. Following previous work (Zhang et al., 2015), we manually collected 29 negations[2] and designed two binary features. One is to indicate whether there is any negation in the tweet and the other is to record whether this tweet contains more than one negation.

- ***NER*** Given a tweet "*@JackHoward the Christmas episode genuinely had me in tears of laughter*", it has useful information like person name and festival which may convey tweeter's happiness. So we extracted 12 types of named entities (DURATION, SET, NUMBER, LOCATION, PERSON, ORGANIZATION, PERCENT, MISC, ORDINAL, TIME, DATE, MONEY) from the sentence and represented each type of named entity as a binary feature to check whether it appears in the sentence.

2.2.2 Sentiment Lexicon Features

Many tasks related to sentiment or emotion analysis depend upon affect, opinion, sentiment, sense and emotion lexicons. So we employ eight sentiment lexicons to capture the sentiment information of the given sentence. The eight sentiment lexicons are as follows: *Bing Liu lexicon*[3], *General Inquirer lexicon*[4], *IMDB*[5], *MPQA*[6], *NRC Emotion Sentiment Lexicon*[7], *AFINN*[8], *NRC Hashtag Sentiment Lexicon*[9], and *NRC Sentiment140 Lexicon*[10].

There is not a unified form among the eights lexicons. For example, Bing Liu lexicon use two values for each word to represent its sentiment scores which one for positive sentiment and the other for negative sentiment. In order to unify the form, we transformed the two scores into a one-dimensional value by subtracting negative emotion scores from positive emotion scores. Given a tweet, we calculated the following six scores:

- the ratio of positive words to all words.
- the ratio of negative words to all words.
- the maximum sentiment scores.
- the minimum sentiment scores.
- the sum of sentiment scores.
- the sentiment score of the last word in tweet.

2.2.3 Emotion Lexicon Features

Considering subtask 1, 2, 5 are related to emotion intensity prediction, subtask 3, 4 are related valence intensity prediction, three emotion lexicons and one valence lexion are adopted. That is NRC Hashtag Sentiment Lexicon (Mohammad and Kiritchenko, 2015), NRC Affect Intensity Lexicon (Mohammad, 2017), NRC Word-Emotion Association Lexicon (Bravo-Marquez et al., 2017) and ANEW-1999 Lexicon (Bradley and Lang, 1999). Given a tweet, we calculate three scores for each lexicon to construct emotion lexicon features: the maximum scores, the sum of scores, the number of words exist in lexicons.

2.2.4 Domain-specific Features

- ***Punctuation*** People often use exclamation mark(!) and question mark(?) to express surprise or emphasis. Therefore, we extract the following 6 features:
  - whether the tweet contains an exclamation mark.
  - whether the tweet contains more than one exclamation mark.
  - whether the tweet has a question mark.

[2] https://github.com/haierlord/resource

[3] http://www.cs.uic.edu/liub/FBS/sentiment-analysis.html#lexicon

[4] http://www.wjh.harvard.edu/inquirer/homecat.htm

[5] http://www.aclweb.org/anthology/S13-2067

[6] http://mpqa.cs.pitt.edu/

[7] http://www.saifmohammad.com/WebPages/lexicons.html

[8] http://www2.imm.dtu.dk/pubdb/views/publication details.php?id=6010

[9] http://www.umiacs.umd.edu/saif/WebDocs/NRC-Hashtag-Sentiment-Lexicon-v0.1.zip

[10] http://help.sentiment140.com/for-students/

  - whether the tweet contains more than one question mark.
  - whether the tweet contains both exclamation marks and question marks.
  - whether the last token of this tweet is an exclamation or question mark.

- ***Bag-of-Hashtags*** Hashtags reflect emotion orientation of tweets directly, so we constructed a vocabulary of hashtags appearing in the training set and development set, then adopted the bag-of-hashtags method for each tweet.

- ***Emoticon*** We collected 67 emoticons from Internet[11], including 34 positive emoticons and 33 negative emoticons, then designed the following 4 binary features:

  - to record whether the positive and negative emoticons are present in the tweet, respectively (1 for yes, 0 for no).
  - to record whether the last token is a positive or a negative emoticon.

- ***Intensity Words*** Some words appeared more frequently in tweets with higher intensity, some words has higher score in emotion lexicons, these words may contain information that express strong emotion intensity. So we extracted this type words in two ways:

  - Pick up words whose emotion score is greater than threshold from emotion lexicons.
  - Calculate the probability of each word appearing at different intensity for subtask 2 and 4, then pick up words whose probability greater than threshold(*i.e.*, 0.5).

  Finally, for each word in intensity words list, we use a binary feature to check whether it appears in the given tweet.

2.3 Learning Algorithms

We explore six algorithms as follows: Logistic Regression (LR) and Support Vector Regression (SVR) implemented in *Liblinear*[12], Bagging Regressor (BR), AdaBoost Regressor (ABR) and Gradient Boosting Regressor (GBR) implemented in *scikit-learn tools*[13] and XGBoost Regressor (XGB)[14]. All these algorithms are used with default parameters.

3 Experiments

3.1 Dataset

The statistics of the English datasets provided by Semeval 2018 Task 1 are shown in Table 1 and 2. How the English data created is described in (Mohammad and Kiritchenko, 2018).

Datasets		**anger**	**fear**	**joy**	**sadness**
train		1,701	2,252	1,616	1,533
dev		388	689	290	397
test	subtask 1	17,939	17,923	18,042	17,912
	subtask 2	1,002	986	1,105	975

Table 1: The statistics of data sets for subtask 1 and 2.

Subtask	**train**	**dev**	**test**
3	1,181	449	17,874
4	1,181	449	937
5	6,838	886	3,259

Table 2: The statistics of data sets for subtask 3, 4, 5.

3.2 Evaluation Metric

To evaluate the performance of different systems, the official evaluation measure *Pearson Correlation Coefficient* with the Gold ratings/labels is adopted for the first four subtasks. The correlation scores across all four emotions will be averaged (macro-average) to determine the final system performance.

As for the last subtask, systems are evaluated by calculating multi-label accuracy namely *Jaccard index*, the formula are follow:

$$Accuracy = \frac{1}{|T|} \sum_{t \in T} \frac{|G_t \cap P_t|}{|G_t \cup P_t|}$$

where G_t is the set of the gold labels for tweet t, P_t is the set of the predicted labels for tweet t, and T is the set of tweets.

3.3 Experiments on Training and Test Data

Firstly, we performed a series of experiments in order to explore the effectiveness of each feature type. Table 3 lists the performance contributed by

[11] https://github.com/haierlord/resource/blob/master/Emoticon.txt

[12] https://www.csie.ntu.edu.tw/ cjlin/liblinear/

[13] http://scikit-learn.org/stable/

[14] https://github.com/dmlc/xgboost

Features	macro-avg	anger	fear	joy	sadness
Linguistic	0.393	0.398	0.402	0.485	0.286
.+SentiLexi	0.594(+20.1%)	0.606	0.532	0.634	0.603
.+EmoLexi	0.635(+4.1%)	0.689	0.632	0.612	0.606
.+domain	**0.657**(+2.2%)	**0.691**	**0.658**	**0.642**	**0.638**

Table 3: Performance of different features on development set for subtask 1. ".+" means to add current features to the previous feature set. The numbers in the brackets are the performance increments compared with the previous results.

Algorithm	macro-avg	anger	fear	joy	sadness
BR	0.602	0.609	0.618	0.584	0.597
XGBOOST	0.628	0.663	0.656	0.576	0.618
ABR	0.635	0.664	0.666	0.573	0.637
SVR	0.657	0.691	0.658	0.642	0.638
GBR	0.667	0.694	0.675	0.630	0.668
XGBOOST+ABR+SVR+GBR	**0.680**	**0.715**	**0.689**	**0.647**	**0.670**

Table 4: Performance of different learning algorithm on development set for subtask 1.

Subtask	System	macro-avg	anger	fear	joy	sadness
1	rank 1	0.799 (1)	0.827 (1)	0.779 (1)	0.792 (1)	0.798 (1)
	our system	0.695(14)	0.713(15)	0.677(18)	0.693(16)	0.697(14)
	baseline	0.520(36)	0.526(33)	0.525(34)	0.575(33)	0.453(36)
2	rank 1	0.695 (1)	0.706 (1)	0.637 (1)	0.720 (2)	0.717 (1)
	our system	0.531(16)	0.565(13)	0.441(21)	0.581(15)	0.536(20)
	baseline	0.394(26)	0.382(27)	0.355(26)	0.469(26)	0.370(29)

Table 5: Performance of our system, top-ranked system and baseline on test set for subtask 1, 2. SVM and unigrams are adopted in baseline. The numbers in the brackets are the official rankings.

System	Subtask3	Subtask4	Subtask5
rank 1	0.873 (1)	0.836 (1)	0.588 (1)
our system	0.813(14)	0.686(17)	0.501(11)
baseline	0.585(28)	0.509(24)	0.442(19)

Table 6: Performance of our system, top-ranked system and baseline on test set for subtask 3, 4, 5. SVM and unigrams are adopted in baseline. The numbers in the brackets are the official rankings.

different features on development set with Support Vector Regression algorithm for subtask 1. We find that:

(1) All feature types make contribution to the performance of emotion intensity prediction and their combination achieves the best performance.

(2) Linguistic features act as baseline and have shown poor performance for emotion intensity prediction. However, we find the system performance drops once we remove the Linguistic features.

(3) Sentiment lexicon features make a considerable contribution to the performance, which indicates that sentiment lexicon features are beneficial not only in traditional sentiment polarity analysis tasks, but also in emotion intensity prediction tasks.

(4) Beside, we find that the system performance only drops by 0.2% if we remove intensity words features. This indicates that these intensity words fail to distinguish emotion intensity. The reason may be that their function have overlap with sentiment and emotion lexicon features.

Also, we explored the performance of different learning algorithms. Table 4 shows the results of different algorithms for subtask 1 based on all features described before. From table 4, we find that GBR outperforms other single algorithm, and the ensemble model are superior to the models using single algorithm. The ensemble model use the four algorithms to build the ensemble regression models, which averages the output scores of all regression algorithm.

Therefore, the system configurations for test data are: using all features for five subtasks, ensemble model for subtask 1 and 3, Logistic Regression for subtask 2, 4 and 5.

Based on the system configurations described above, we train separate model for each subtask and evaluate them against the test set in SemEval 2018 Task 1. Table 5 and Table 6 shows the results with ranks on test set for subtask 1 to 5. Compared with the top ranked systems, there is much room for improvement in our work. First, the biggest issue is that we only used hand-craft features but ignoring deep learning method. Second, we find that our system achieves greater performance on test set compared with the development set, the possible reason might be the different data distribution held between them.

4 Conclusion

In this paper, we extracted several traditional NLP, sentiment lexicon, emotion lexicon and domain specific features from tweets, adopted supervised machine learning algorithms to perform emotion intensity prediction. The system performance ranks above average. In future work, we consider to use deep learning method to model sentence with the aid of sentiment word vectors.

Acknowledgements

This work is is supported by the Science and Technology Commission of Shanghai Municipality Grant (No. 15ZR1410700) and the open project of Shanghai Key Laboratory of Trustworthy Computing (No.07dz22304201604).

References

Margaret M. Bradley and Peter J. Lang. 1999. Affective norms for english words (anew): Instruction manual and affective ratings. *Journal Royal Microscopical Society*, 88(1):630–634.

Felipe Bravo-Marquez, Eibe Frank, Saif M. Mohammad, and Bernhard Pfahringer. 2017. Determining word-emotion associations from tweets by multi-label classification. In *Ieee/wic/acm International Conference on Web Intelligence*, pages 536–539.

Paul Ekman. 1999. Basic emotions. *Handbook of Cognition and Emotion*, 99(1):45–60.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David Mcclosky. 2014. The stanford corenlp natural language processing toolkit. In *Meeting of the Association for Computational Linguistics: System Demonstrations*.

Saif M. Mohammad. 2017. Word affect intensities.

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.

Saif M Mohammad and Svetlana Kiritchenko. 2015. Using hashtags to capture fine emotion categories from tweets. *Computational Intelligence*, 31(2):301–326.

Saif M. Mohammad and Svetlana Kiritchenko. 2018. Understanding emotions: A dataset of tweets to study interactions between affect categories. In *Proceedings of the 11th Edition of the Language Resources and Evaluation Conference*, Miyazaki, Japan.

Zhihua Zhang, Guoshun Wu, and Man Lan. 2015. Ecnu: Multi-level sentiment analysis on twitter using traditional linguistic features and word embedding features. In *International Workshop on Semantic Evaluation*, pages 561–567.