LT3 at SemEval-2018 Task 1: A classifier chain to detect emotions in tweets

Luna De Bruyne, Orphée De Clercq and Véronique Hoste
LT^3, Language and Translation Technology Team, Ghent University
Groot-Brittanniëlaan 45, 9000 Ghent, Belgium
{luna.debruyne, orphee.declercq, veronique.hoste}@ugent.be

Abstract

This paper presents an emotion classification system for English tweets, submitted for the SemEval shared task on Affect in Tweets, subtask 5: Detecting Emotions. The system combines lexicon, n-gram, style, syntactic and semantic features. For this multi-class multi-label problem, we created a classifier chain. This is an ensemble of eleven binary classifiers, one for each possible emotion category, where each model gets the predictions of the preceding models as additional features. The predicted labels are combined to get a multi-label representation of the predictions. Our system was ranked eleventh among thirty five participating teams, with a Jaccard accuracy of 52.0% and macro- and micro-average F1-scores of 49.3% and 64.0%, respectively.

1 Introduction

Most research in the domain of sentiment analysis focuses on the automatic prediction of polarity or valence in text, but also the detection of emotions has attracted growing interest in the last couple of years (Mohammad and Bravo-Marquez, 2017). Although emotion detection is a rather new research focus in NLP, the study of emotions has a long history in fields like psychology and neuroimaging. Many different frameworks exist, but the specific emotion approach, in which emotions are classified as specific discrete categories, predominates. In a lot of those approaches, some emotions are considered more basic than others, with Ekman's theory of six basic emotions (*joy, sadness, anger, fear, disgust*, and *surprise*) (Ekman, 1992) as the most well-known. Another popular theory is Plutchik's wheel of emotions (Plutchik, 1980), in which *joy, sadness, anger, fear, disgust, surprise, trust*, and *anticipation* are considered most basic.

Emotion analysis in NLP makes use of the frameworks developed by psychologists, mostly by employing categorical models of (basic) emotions. In traditional emotion classification tasks, a 'document' or sentence is classified under one or more emotion classes (or classified as neutral/no class when no emotions are present). Such emotion classification systems have been developed and tested on different kinds of data, including fairy tales (Alm et al., 2005), newspaper headlines (Strapparava and Mihalcea, 2007), chat messages (e.g. Holzman and Pottenger, 2003; Brooks et al., 2013), and tweets (e.g. Mohammad, 2012; Wang et al., 2012). The big advantage of using tweet datasets is the relative ease with which twitter data can be collected and the possibility of using hashtags as emotion labels (distant supervision approach).

For this paper, we used the data that was collected for the SemEval shared task on Affect in Tweets (Mohammad et al., 2018), a collection of tweets annotated for eleven emotions: *anger, anticipation, disgust, fear, joy, love, optimism, pessimism, sadness, surprise,* and *trust* (Mohammad and Kiritchenko, 2018). We participated in Subtask 5: Detecting Emotions (English emotion classification).

The remainder of this paper is structured as follows: in Section 2 we describe how we first analyzed the data in order to get more insight in the task. Section 3 discusses how the data was preprocessed and which information sources were extracted. Next, in Section 4 the actual experimental setup and results are discussed and we end this paper with a conclusion in Section 5.

2 Data analysis

We first analyzed the training data provided by the task organizers, which consisted of 6838 tweets. We found that *disgust, anger* and *joy* were present in the largest numbers (present in about 35 to 40%

Proceedings of the 12th International Workshop on Semantic Evaluation (SemEval-2018), pages 123–127
New Orleans, Louisiana, June 5–6, 2018.

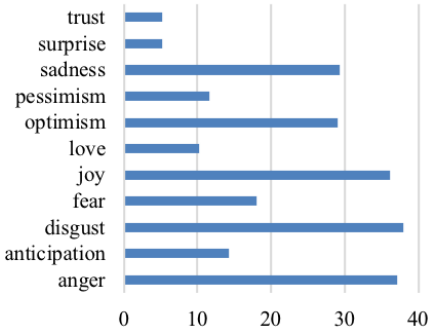

Figure 1: Proportion of training tweets in which the specified emotion is present (%).

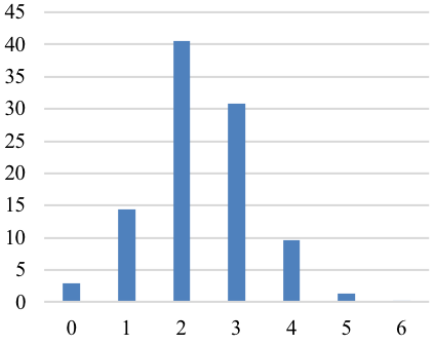

Figure 2: Proportion of training tweets in which a specific amount of emotion classes is present (%).

of the tweets), while *surprise* and *trust* only occur in around 5% of the tweets (Figure 1). Only three percent of the tweets was annotated as neutral.

As can be derived from Figure 2, most tweets contained two or three emotions (together 70%), and in only about 1% of the tweets five or more (max six) emotions were present. We also calculated the correlations and found ten emotion pairs that were moderately or highly correlated ($\|phi\| \geq 0.30$ for moderate correlation, $\|phi\| \geq 0.50$ for high correlation, according to Cohen's conventions on effect size (Cohen, 1988)). The correlated pairs are shown in Table 1 and suggest that the classification performance can be boosted when correlations between emotion categories are implemented in the model.

In order to get more insight into the data, we re-annotated a subset of 500 tweets from the training set. In Table 2, inter-annotator agreement (IAA) scores per emotion class between the gold labels and our annotations are presented. Except for *anger* and *joy* these scores are rather low. Overall, we assigned less emotion classes to a tweet than the official annotators. We often disagreed with the gold labels and had the feeling that the annotators of the official labels focused too much on lexical clues instead of keeping the context and the perspective of the writer of the tweet in mind. This leads us to presume that the threshold to assign an emotion label to a tweet when two out of seven annotators agreed (Mohammad and Kiritchenko, 2018) might have been a bit too generous.

Pair	phi
anger - joy	-0.44
anger - optim.	-0.37
disg. - optim.	-0.41
joy - disg.	-0.46
joy - sadn.	-0.33
surpr. - pessim.	-0.40

Pair	phi
anger - disg.	0.68
joy - love	0.40
joy - optim.	0.52
sadn. - pessim.	0.30

Table 1: Phi coefficients for moderate or high negative (left) and positive (right) correlations between emotion pairs.

Emotion	Kappa	Emotion	Kappa
Anger	0.678	Optimism	0.436
Anticipation	0.259	Pessimism	0.124
Disgust	0.132	Sadness	0.537
Fear	0.399	Surprise	0.276
Joy	0.717	Trust	0.367
Love	0.470		

Table 2: IAA (Kappa) per emotion class based on 500 re-annotated instances.

We further noticed that some tweets appeared twice in the data set, but not completely identically: we suspect that one of them was the original tweet with emotion hashtag and the other one with the hashtag removed. An example:

(1) a. *Whatever you decide to do make sure it makes you #happy.*

b. *Whatever you decide to do make sure it makes you .*

Since labels differed depending on the presence or absence of the emotion hashtag, we decided to keep both variants in our training set.

3 Preprocessing & Feature Extraction

3.1 Preprocessing

While we did not remove the 'almost identical' tweets from the data set, there were also some tweets in the training set that were completely identical but had been assigned other emotion labels. For those tweets, we took the majority class for each binary emotion category, and removed all other instances. This reduced our training set from 6838 to 6782 tweets. No duplicates were present in the development set, so the amount of

886 tweets was preserved. In the updated training set, as well as in the development and test set, all user names were replaced with the generic *@ID*.

All tweets were processed with Weka (Witten et al., 2016) using the Affective Tweets package (Mohammad and Bravo-Marquez, 2017), in order to extract lexicon and word embedding features. We used the default preprocessing settings for each filter. For the other features, we performed word and sentence tokenization (using NLTK), stemming (using spaCy), lowercasing, and POS-tagging (simple and detailed, corresponding to spaCy's `POS` and `Tag` function).

3.2 Feature extraction

For our supervised classification system, we employed features that measure different aspects of the tweet. These can be subsumed under five different categories: lexicon features (see Table 3 for an overview), n-gram features (binary, n equal to 3, 4 and 5 for characters and n equal to 1 or 2 for tokens), and various style, syntactic and semantic features (see Table 4).

Regarding the latter category, both features from traditional and distributional semantics were integrated. We first took the synset depth (distance to root) of all content words (calculated with WordNet (Miller, 1995)) and averaged the scores to get a mean synset depth for the tweet. Furthermore, we included two types of features from distributional semantics, namely word embeddings and word clusters. The word embeddings were extracted with Weka Affective Tweets, using pre-trained embeddings from 10 million tweets taken from the Edinburgh Twitter Corpus (Petrovic et al., 2010). For the word clusters, we downloaded a subset of around 1.5M tweets from the SemEval 2018 AIT DISC corpus (Mohammad et al., 2018). We first created word embeddings with word2vec using both skipgram and continuous bow and afterwards applied k-means clustering on the resulting word vectors. We experimented with various cluster sizes (800 of size 100, 1000 of size 100 and 800 of size 300). These clusters were implemented as binary features.

Lexicon	**Type**
MPQA	polarity
Bing Liu	polarity
AFINN	polarity
Sentiment140	polarity
NRC Hashtag Sentiment	polarity
NRC Word-Emotion	polarity + Plutchik emotions
NRC-10 Expanded	polarity + Plutchik emotions
NRC Hashtag Emotion	Plutchik emotions
SentiWordNet	polarity
Emoticons	polarity
Sentistrength	polarity sentiment strengths
Warinner et al. 2013	valence, arousal, dominance

Table 3: Lexicons used for feature extraction.

Style	**Syntax**	**Semantics**
avg word/sent. length	POS n-grams	synset depth
# words and sents	POS freq.	embeddings
# capitals	POS 1^{st} token	clusters
# punct. marks	presence imp.	
# non-standard words	presence fut.	
# connectives		

Table 4: Style, syntactic and semantic features.

4 Experiments & Results

4.1 Baseline & Binary Experiments

We trained different models on the training set and tested them on the development set, using scikit-learn (Pedregosa et al., 2011). For the baseline experiments, we used an SVM classifier with linear kernel (LinearSVC) and used the lexicon features from the Weka Affective Tweets package. The results for each binary classifier are shown in Table 5 (second column). Combining the predictions of these eleven binary classifiers resulted in a jaccard accuracy of 42.7%.

Before optimizing the separate classifiers, we took a more detailed look at the lexicon features and the clusters to assess whether it is beneficial to use only a part of the lexicons (e.g. only the emotion lexicons) or whether it is better to use all lexicons (even polarity lexicons). We found that the combination of all lexicons (including the valence-arousal-dominance lexicon of Warriner et al. (2013)) gave the highest performance. As regards the clusters, we tried all cluster types on each emotion category and picked the cluster that gave the highest performance on that particular category.

For every emotion category, we tested different classifiers on different combinations of features. The classifiers we used, were SVM, SGD (linear SVM with stochastic gradient descent learning), Logistic Regression, and Random Forest. Table 5 shows the F1-scores (in bold) on the positive class for the best performing classifiers and feature combinations, which are significantly higher than the baseline results. We joined the predictions of these optimized binary classifiers, and achieved a jaccard accuracy of 47.7%.

	BL	Optimized		
Emotion	**F1**	**Classifier**	**Features**	**F1**
Anger	*0.67*	SGD	all features except clusters	**0.73**
Anticip.	*0.00*	SGD	all features	**0.30**
Disgust	*0.56*	Log. R.	lexicons, embeddings, clusters	**0.67**
Fear	*0.62*	Log. R.	lexicons, embeddings, n-grams, clusters	**0.69**
Joy	*0.75*	Log. R.	lexicons, embeddings, n-grams, puncts, pos n-grams, pos frequencies, clusters	**0.80**
Love	*0.29*	Log. R.	all features	**0.55**
Optim.	*0.59*	SGD	all features	**0.68**
Pessim.	*0.04*	SGD	lexicons, embeddings, clusters	**0.20**
Sadness	*0.52*	Log. R.	all features	**0.59**
Surprise	*0.00*	SGD	all features except clusters	**0.35**
Trust	*0.00*	SGD	lexicons	**0.12**

Table 5: F1-scores on the positive class for the binary classifiers in the baseline (BL) setup (italics) and with the optimal classifier and feature sets (in bold)

.

4.2 Classifier Chain

Because the emotion categories are highly correlated (see Section 2), we envisaged to implement these relations in the model by using a classifier chain. We combined the best performing classifier per emotion category in a chain that passes predicted labels on to the next classifiers. We ordered the classifiers by performance on the positive class F1-score on the baseline (the emotion that is easiest to predict first, the emotion that is the most difficult to predict last). On the development set, this classifier chain approach led to a jaccard accuracy of 52.37%, which is significantly higher than the score without classifier chain (47.7%, see Section 4.1).

In our final model, the training and development data were joined, resulting in a combined training set of 7668 tweets. During the evaluation period, we achieved 52.0% jaccard accuracy, 64.0% micro-avg F1-score and 49.3% macro-avg F1-score on the held-out test set (see Table 6).

4.3 Discussion

As can be derived from Table 7 the number of false positives is rather low for all emotion classes (below 20% for most emotions). The model had most trouble with recognizing positive instances of *surprise*, *pessimism*, and *trust*, but also *love* and *anticipation* were more challenging. For these categories, the false negative rate was thus very high. We assume that these bad results are mostly due to a lack of sufficient training data for these categories.

Evaluation	**jaccard**	**micro F1**	**macro F1**
dev set	0.524	0.644	0.478
held-out test set	0.520	0.640	0.493

Table 6: Jaccard accuracy, micro averaged F1-score and macro averaged F1-score of the optimized model on the development and held-out test set.

		P				*P*	
	G	*0*	*1*		*G*	*0*	*1*
anger	*0*	0.72	0.28	optim.	*0*	0.65	0.35
	1	0.17	0.83		*1*	0.16	0.84
antic.	*0*	0.89	0.11	pess.	*0*	0.98	0.02
	1	0.68	0.32		*1*	0.86	0.14
disg.	*0*	0.75	0.25	sadn.	*0*	0.91	0.09
	1	0.21	0.79		*1*	0.46	0.54
fear	*0*	0.97	0.03	surpr.	*0*	>0.99	<0.01
	1	0.42	0.58		*1*	0.98	0.02
joy	*0*	0.89	0.11	trust	*0*	0.94	0.06
	1	0.20	0.80		*1*	0.82	0.18
love	*0*	0.95	0.05				
	1	0.52	0.48				

Table 7: Confusion matrices for the results on the held out test set. P = predicted labels; G = gold labels.

We evaluated all features by computing the ANOVA F-values, and extracted the hundred most predictive features for each emotion category. For all emotions, the top 100 features consisted exclusively of lexical information. In none of the emotion categories, style or syntactic features occurred in this top 100. However, features regarding labels of preceding classifiers belonged to the most predictive features for all emotions except for *optimism* and *surprise*.

5 Conclusion

Our emotion classification system for English tweets achieved 52.0% jaccard accuracy on the held-out test set. We started from binary classifiers which we optimized for each emotion category separately, and combined them in a classifier chain. We proved that passing on labels from previously predicted emotions categories improves the performance significantly. For future work, it would be interesting to investigate the model's performance on other datasets than twitter data.

References

Cecilia Ovesdotter Alm, Dan Roth, and Richard Sproat. 2005. Emotions from text: machine learning for text-based emotion prediction. In *Proceedings of the conference on human language technology and empirical methods in natural language processing*, pages 579–586. Association for Computational Linguistics.

Michael Brooks, Katie Kuksenok, Megan K Torkildson, Daniel Perry, John J Robinson, Taylor J Scott, Ona Anicello, Ariana Zukowski, Paul Harris, and Cecilia R Aragon. 2013. Statistical affect detection in collaborative chat. In *Proceedings of the 2013 conference on Computer supported cooperative work*, pages 317–328. ACM.

Jacob Cohen. 1988. Statistical power analysis for the behavioral sciences . hilsdale. *NJ: Lawrence Earlbaum Associates*, 2.

Paul Ekman. 1992. An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200.

Lars E Holzman and William M Pottenger. 2003. Classification of emotions in internet chat: An application of machine learning using speech phonemes. Technical report, Leigh University.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Saif Mohammad and Felipe Bravo-Marquez. 2017. Emotion intensities in tweets. In *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics, *SEM @ACM 2017, Vancouver, Canada, August 3-4, 2017*, pages 65–77.

Saif M Mohammad. 2012. # emotional tweets. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, pages 246–255. Association for Computational Linguistics.

Saif M. Mohammad and Felipe Bravo-Marquez. 2017. Wassa-2017 Shared Task on Emotion Intensity. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (WASSA-2017)*, Copenhagen, Denmark.

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. 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.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Sasa Petrovic, Miles Osborne, and Victor Lavrenko. 2010. The edinburgh twitter corpus. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Linguistics in a World of Social Media*, pages 25–26. Association for Computational Linguistics.

Robert Plutchik. 1980. A general psychoevolutionary theory of emotion. In Robert Plutchik and Henry Kellerman, editors, *Emotion: Theory, research, and experience: Vol. 1. Theories of emotion*, pages 3–33. Academic Press, New York.

Carlo Strapparava and Rada Mihalcea. 2007. Semeval-2007 task 14: Affective text. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 70–74. Association for Computational Linguistics.

Wenbo Wang, Lu Chen, Krishnaprasad Thirunarayan, and Amit P Sheth. 2012. Harnessing twitter "big data" for automatic emotion identification. In *Proceedings of the 2012 ASE/IEEE international conference on social computing and 2012 ASE/IEEE international conference on privacy, security, risk and trust, SOCIALCOM-PASSAT'12*, pages 587–592, Washington, DC. IEEE Computer Society.

Amy Beth Warriner, Victor Kuperman, and Marc Brysbaert. 2013. Norms of valence, arousal, and dominance for 13,915 english lemmas. *Behavior Research Methods*, 45(4):1191–1207.

Ian H Witten, Eibe Frank, Mark A Hall, and Christopher J Pal. 2016. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.