Determining Word–Emotion Associations from Tweets by Multi-Label Classification

Felipe Bravo-Marquez*, Eibe Frank*, Saif M. Mohammad† and Bernhard Pfahringer*
*Department of Computer Science, The University of Waikato
Hamilton, New Zealand
†National Research Council Canada
Ottawa, ON, Canada

***Abstract*—The automatic detection of emotions in Twitter posts is a challenging task due to the informal nature of the language used in this platform. In this paper, we propose a methodology for expanding the NRC word-emotion association lexicon for the language used in Twitter. We perform this expansion using multi-label classification of words and compare different word-level features extracted from unlabelled tweets such as unigrams, Brown clusters, POS tags, and word2vec embeddings. The results show that the expanded lexicon achieves major improvements over the original lexicon when classifying tweets into emotional categories. In contrast to previous work, our methodology does not depend on tweets annotated with emotional hashtags, thus enabling the identification of emotional words from any domain-specific collection using unlabelled tweets.**

I. Introduction

Analysing the emotions expressed in Twitter has important applications in the study of public opinion. Word-emotion association lexicons, which are lists of terms annotated according to emotional categories, are widely used resources for analysing emotions in textual passages. The NRC word-emotion association lexicon (NRC-10)[1] [1] is a well-known lexical resource for emotion analysis created by crowdsourcing via Mechanical Turk. It contains 14,182 distinct English words manually annotated according to ten non-exclusive binary categories including the eight emotions from Plutchik's wheel of emotions [2]: joy, sadness, anger, surprise, fear, disgust, trust and anticipation; and two sentiment classes: positive and negative. For example, the word **achieved** is mapped into the categories anticipation, joy, trust, and positive, and the word **exile** is mapped into anger, fear, sadness, and negative. There are 7,714 words that are not associated with any affective category and can be considered neutral, such as powder and corn. NRC-10 does not cover informal expressions commonly used in social media such as hashtags, slang words and misspelled words, and consequently suffers from limitations when analysing emotions from microblogging messages such as tweets.

In this paper, we study how to automatically expand NRC-10 with the words found in a corpus of unlabelled tweets. The expansion is performed using multi-label classification techniques. These techniques assign instances to multiple non-exclusive classes such as the ones provided by NRC-10. We represent words as feature vectors drawn from the contexts in which the words occur in a corpus. We experiment with two such approaches:

1) The word-centroid model [3], which creates word-vectors from tweet-level attributes (e.g., unigrams and Brown clusters) by averaging all the tweets in which the target word appears.
2) Word embeddings [4], which are low-dimensional continuous dense word vectors trained from document corpora.

The words from NRC-10 that occur in the corpus are labelled according to the emotional categories provided by the lexicon. The feature vectors for the words along with these affect labels are used for learning a word-level multi-label affect classifier. As some categories from NRC-10 correlate with each other, we explore multi-label classification techniques that exploit label co-occurrence such as classifier chains [5]. The fitted multi-label classification model is then used to classify the remaining unlabelled words into emotions.

To summarise, this paper proposes a method to automatically expand a hand-annotated emotion lexicon using a corpus of unlabelled tweets and a multi-label classifier. We empirically show which combinations of word-level features and learning techniques are most effective for this task and show that all our expanded lexicons produce substantial improvements over NRC-10 alone when classifying tweets into emotions. To the best of our knowledge, this is the first emotion lexicon expansion model for tweets in which a word-level multi-label classifier is trained using features calculated from unlabelled corpora.

This article is organised as follows. In Section II, we provide a review of existing work on lexicon expansion for emotions. In Section III, we describe the proposed methodology for expanding the NRC-10 lexicon. In Section IV, we present the experiments we conducted to evaluate the proposed approach and discuss results. The main findings and conclusions are discussed in Section V.

II. Related Work

Most previous work on Twitter lexicon expansion focuses on two polarity classes rather than multi-dimensional emotions. A common approach for polarity lexicon expansion is to compute associations between words and message-level

[1]http://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm

polarity labels using Pointwise Mutual Information (PMI). The message labels are derived by strong sentiment cues such as emoticons or hashtags [6] to avoid the expensive costs of data annotation. Another approach exploited in Amir et al. [7] and Tang et al. [8] represents words from a corpus of Tweets by low-dimensional word-embeddings that are classified into sentiment classes by training single-label classifiers with labels provided by a seed polarity lexicon. Our work goes beyond these methods because it tackles a multi-label classification problem (emotion classification), and compares word-embeddings with other word-level features.

In Mohammad and Kiritchenko [9], the authors collected around $50,000$ tweets annotated with hashtags corresponding to the six Ekman emotions: #anger, #disgust, #fear, #happy, #sadness, and #surprise and created a Twitter-specific emotion-association lexicon using PMI associations analogously to Mohammad et al. [6]. There are two limitations when depending on emotion-annotated tweets based on hashtags: 1) words that do not co-occur with those emotion-oriented hashtags will be excluded, and 2) there are many domains in which hashtags are not frequently used to express emotions, and hence, this approach would be unsuitable for creating domain-specific emotion lexicons for those domains. In contrast, our approach takes a target corpus of unlabelled tweets from any domain and a seed lexicon to perform the expansion.

III. Multi-label Classification of Words into Emotions

In this section, we describe our methodology for classifying Twitter words into emotions. The first step is to tokenise and extract word-level features from a target corpus of ten million unlabelled tweets written in English taken from the Edinburgh corpus (ED) [10], which is a general purpose collection of tweets. We use two models for extracting word-level features: 1) the word-centroid model [3], and 2) the skip-gram model [4]. In the word-centroid model, the tweets from the target corpus are represented by tweet-level feature vectors that are transferred to the word-level by averaging, for each word, all the tweet-level vectors in which the word occurs. The tweet-level features that we average for each word are:

1) Word unigrams (UNI): a vector space model based on counting the frequency of unigrams.
2) Brown clusters (BWN): we tag the tweet according to Brown clusters of words [11] to form a low-dimensional vector space in which the frequency of each word-cluster is counted.
3) POS n-grams (POS): the tweet is POS-tagged and the frequency of each POS unigram and bigram is counted.
4) Distant Polarity (DP): two features consisting of the positive and negative probabilities returned by a logistic regression model trained from a distant supervision corpus of 1.6. million tweets labelled with positive and negative emoticons [12] using unigrams as features.

The tokenisation process, the POS tags, and the Brown clusters are taken from the **TweetNLP** project[2].

We also use the negative sampling method for training skip-gram word-embeddings (W2V) from the target corpus that is implemented in **word2vec**[3]. In this method, a neural network with one hidden layer is trained for predicting the words surrounding a center word, within a window that is shifted along the target corpus.

The NRC-10 words that occur in the target corpus are labelled according to the corresponding emotions and their feature vectors are used for training a multi-label classifier. We use three multi-label classification techniques:

1) Binary Relevance (BR), in which a separate binary classifier is trained per label.
2) Classifier Chains (CC) [5], in which inter-label dependencies are exploited by cascading the predictions for each binary classifier as additional features along a random permutation of labels.
3) Bayesian Classifier Chains (BCC) [13], in which a Bayesian network that represents dependency relations between the labels is learned from the data and used to build a classifier chain based on these dependencies.

The resulting classifiers are used to classify the remaining unlabelled words into emotions.

IV. Evaluation

The proposed approach is evaluated both intrinsically and extrinsically as described in the sub-sections below.

A. Intrinsic Evaluation

We start with an intrinsic evaluation comparing the micro-averaged F1 measure obtained for the ten affective labels by different combinations of features and classifiers. These experiments are carried out using MEKA[4], a toolbox for multi-label classification. In order to obtain association scores for each label we use an L_2-regularised logistic regression from LIBLINEAR[5], with the regularisation parameter C set to 1.0, as the base learner in the different models.

All NRC-10 words that occur at least fifty times in the target corpus are used in our experiments. There were $10,137$ such words (902 are associated with anger, 694 with anticipation, $1,101$ with fear, 579 with joy, 885 with sadness, 432 with surprise, 981 with trust, $2,314$ with the negative sentiment, and $1,818$ with the positive sentiment).

Before training the word embeddings (W2V) from the target corpus of ten million tweets, we tune the window size and dimensionality of the skip-gram model by conducting a grid-search process in which we train a binary relevance word-level multi-label classifier on the NRC-10 words with 2-fold cross-validation for each parameter configuration. This process is performed over a collection of 1 million tweets independent from the target corpus using the micro averaged F1 measure

[2]http://www.ark.cs.cmu.edu/TweetNLP/
[3]https://code.google.com/p/word2vec/
[4]http://meka.sourceforge.net/
[5]http://www.csie.ntu.edu.tw/~cjlin/liblinear/

Fig. 1. A visualisation for the expanded emotion lexicon.

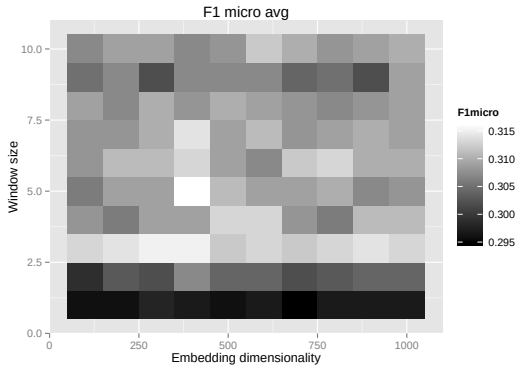

Fig. 2. Emotion classification results obtained using word embeddings of different dimensionalities, generated from various window sizes. Maximum F1 is achieved for 400 by 5.

Classifier	BR	CC	BCC
UNI (Baseline)	0.389 ± 0.03	0.371 ± 0.03	0.378 ± 0.03
UNI-BWN	0.410 ± 0.03 +	0.400 ± 0.03 +	0.407 ± 0.03 +
UNI-BWN-POS	0.411 ± 0.03 +	0.405 ± 0.02 +	0.407 ± 0.03 +
UNI-BWN-POS-DP	0.433 ± 0.03 +	0.427 ± 0.03 +	0.432 ± 0.03 +
UNI-BWN-POS-DP-W2V	0.477 ± 0.03 +	0.474 ± 0.03 +	0.478 ± 0.03 +
W2V	0.473 ± 0.03 +	0.469 ± 0.03 +	0.472 ± 0.03 +
W2V-BWN	0.468 ± 0.03 +	0.469 ± 0.03 +	0.47 ± 0.03 +
W2V-BWN-POS	0.465 ± 0.03 +	0.466 ± 0.03 +	0.466 ± 0.02 +
W2V-BWN-POS-DP	0.474 ± 0.03 +	0.473 ± 0.03 +	0.475 ± 0.03 +
W2V-DP	**0.479** ± 0.03 +	**0.476** ± 0.03 +	**0.479** ± 0.03 +

TABLE I
WORD-LEVEL MULTI-LABEL CLASSIFICATION MICRO-AVERAGED F1 RESULTS. BEST RESULTS PER COLUMN ARE SHOWN IN BOLD. THE SYMBOL + CORRESPONDS TO STATISTICALLY SIGNIFICANT IMPROVEMENTS WITH RESPECT TO THE BASELINE.

as performance metric. As shown in the heatmap in Figure 2, the optimum parameters are a window size of 5 and the number of dimensions set to 400. We used this parameter configuration for training the W2V features from the target corpus. From the figure, we can observe that embeddings built using windows smaller than two are not good for capturing emotional information.

The word-level multi-label classification results for the micro-averaged F1 measure obtained by training the three multi-label classification schemes BR, CC[6], and BCC using 10-fold cross-validation are shown in Table I. We compare word-level vectors by concatenating different combinations of the features presented in Section III: UNI, BWN, POS, DP, and W2V. The unigram feature-space (UNI) is used as the baseline and is compared with the other feature spaces using a corrected resampled paired *t*-test with an α level of 0.05 [14].

From the table we can see that distributional features that go beyond word counts, such as BWN, and DP, produce statistically significant improvements over using unigrams alone. On the other hand, W2V alone obtains a better performance than the other features and is only slightly improved when combined with certain features such as DP. This suggests that low-dimensional embeddings trained from unlabelled tweets capture stronger information for emotion classification than word-level features derived by the word-centroid model. Although these features can produce a competitive representation they do not add much value to W2V. Regarding the multi-label classification techniques, there are no observable benefits of methods that exploit label inter-dependencies such as CC and BCC over BR.

The trained multi-label classifiers are used to create Twitter-specific word-emotion associations by classifying the $42,900$ unlabelled words from the corpus into 10-dimensional affect vectors. A word cloud of the expanded lexicon that combines all the features trained with BCC is shown in Figure 1. The word sizes are proportional to the estimated probabilities associated with the corresponding emotions.

B. Extrinsic Evaluation

We conduct an extrinsic evaluation by studying the usefulness of the expanded lexicons for classifying Twitter messages

[6]Ensembles of classification chains were also evaluated, with no evidence of improvement over a single chain.

into emotion categories. We use the Twitter Emotion Corpus [15], which has $21,051$ tweets labelled by a single-label multi-class emotional label using hashtags. The number of tweets per class is $3,849$ for surprise, $3,830$ for sadness, $8,240$ for joy, 761 for disgust, $2,816$ for fear, and $1,555$ for anger. Using 10-fold cross-validation, we compare a one-vs-all logistic regression that uses attributes calculated from NRC-10 alone (the baseline), with the performance obtained by a classifier trained with attributes derived from NRC-10 and the expanded lexicon.

The comparisons are carried out using again the corrected resampled paired *t*-test. We calculate ten numerical features from NRC-10 by counting the number of words in a tweet matching each emotion category, and another ten features from the expanded lexicon, calculated as the sum of the corresponding affect probabilities for the matched words, obtained from the multi-label word-level model. Therefore, tweets are represented by ten features in the baseline (NRC-10 alone), and by twenty features for each expanded lexicon (with one lexicon for each multi-label classifier considered above). The kappa statistic and weighted area under the ROC curve (AUC) for all the logistic regression models trained with different expanded lexicons is given in Table II.

Lexicon	Kappa			AUC		
NRC-10 (alone)	0.077			0.633		
NRC-10+Expanded	BR	CC	BCC	BR	CC	BCC
UNI	0.191 +	0.201 +	0.198 +	0.711 +	0.714 +	0.713 +
UNI-BWN	0.174 +	0.178 +	0.176 +	0.708 +	0.712 +	0.711 +
UNI-BWN-POS	0.175 +	0.177 +	0.178 +	0.708 +	0.711 +	0.710 +
UNI-BWN-POS-DP	0.180 +	0.183 +	0.184 +	0.713 +	0.715 +	0.714 +
UNI-BWN-POS-DP-W2V	0.187 +	0.197 +	0.183 +	0.712 +	0.714 +	0.713 +
W2V	**0.223** +	**0.226** +	**0.226** +	0.720 +	**0.723** +	**0.723** +
W2V-BWN	0.199 +	0.201 +	0.197 +	0.713 +	0.715 +	0.715 +
W2V-BWN-POS	0.195 +	0.201 +	0.196 +	0.710 +	0.713 +	0.712 +
W2V-BWN-POS-DP	0.199 +	0.204 +	0.199 +	0.714 +	0.715 +	0.715 +
W2V-DP	**0.223** +	0.223 +	**0.226** +	**0.722** +	**0.723** +	**0.723** +

TABLE II
MESSAGE-LEVEL CLASSIFICATION RESULTS OVER THE HASHTAG EMOTION CORPUS. BEST RESULTS PER COLUMN ARE GIVEN IN BOLD.

All the expanded lexicons are statistically significantly better than using NRC-10 alone. Note that all these improvements are substantial in all cases. Similarly to the intrinsic results, we observe that the lexicons created using W2V alone and W2V-DP are the strongest ones. Another interesting result is that lexicons created with multi-label classifiers that exploit label correlations such as CC and BCC are slightly better than the ones created using BR in most cases.

V. CONCLUSIONS

In this work, we have proposed a methodology for expanding a multi-label emotion lexicon based on a collection of unlabelled tweets using multi-label classification. We have shown that all the produced lexicons achieve substantial improvements over the seed lexicon for classifying tweets into emotions[7]. The results indicate that low-dimensional word-embeddings are better than distributional word-level features obtained by averaging tweet-level features. This is aligned with recent findings in NLP showing that representations learned from unlabelled data using neural networks outperform representations obtained from hand-crafted features [16].

In contrast to earlier work on creating a lexicon of emotion words for Twitter [9], which is restricted to tweets annotated with emotional hashtags, our method can learn emotional words from any collection of unannotated tweets. Hence, our approach can be used, without any additional labelling effort, for creating domain-specific emotion lexicons based on unlabelled tweets collected from the target domain, such as politics and sports.

[7]The expanded lexicons as well as the word vectors used to build them are available for download at http://www.cs.waikato.ac.nz/ml/sa/lex.html#emolextwitter.

REFERENCES

[1] S. M. Mohammad and P. D. Turney, "Crowdsourcing a word-emotion association lexicon." *Computational Intelligence*, vol. 29, no. 3, pp. 436–465, 2013.

[2] R. Plutchik, "The nature of emotions," *American Scientist*, vol. 89, no. 4, pp. 344–350, 2001.

[3] F. Bravo-Marquez, E. Frank, and B. Pfahringer, "From unlabelled tweets to twitter-specific opinion words," in *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*. New York, NY, USA: ACM, 2015, pp. 743–746.

[4] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in Neural Information Processing Systems 26*, C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger, Eds. Curran Associates, Inc., 2013, pp. 3111–3119.

[5] J. Read, B. Pfahringer, G. Holmes, and E. Frank, "Classifier chains for multi-label classification," *Machine learning*, vol. 85, no. 3, pp. 333–359, 2011.

[6] S. M. Mohammad, S. Kiritchenko, and X. Zhu, "NRC-canada: Building the state-of-the-art in sentiment analysis of tweets," in *Proceedings of the seventh international workshop on Semantic Evaluation Exercises*, 2013, pp. 321–327.

[7] S. Amir, W. Ling, R. Astudillo, B. Martins, M. J. Silva, and I. Trancoso, "Inesc-id: A regression model for large scale twitter sentiment lexicon induction," in *Proceedings of the 9th International Workshop on Semantic Evaluation*. Denver, Colorado: Association for Computational Linguistics, June 2015, pp. 613–618.

[8] D. Tang, F. Wei, B. Qin, M. Zhou, and T. Liu, "Building large-scale twitter-specific sentiment lexicon : A representation learning approach," in *Proceedings of 25th International Conference on Computational Linguistics, August 23-29, 2014, Dublin, Ireland*, 2014, pp. 172–182.

[9] S. M. Mohammad and S. Kiritchenko, "Using hashtags to capture fine emotion categories from tweets," *Computational Intelligence*, vol. 31, no. 2, pp. 301–326, 2015.

[10] S. Petrović, M. Osborne, and V. Lavrenko, "The Edinburgh Twitter corpus," in *Proceedings of the NAACL HLT 2010 Workshop on Computational Linguistics in a World of Social Media*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2010, pp. 25–26.

[11] P. F. Brown, P. V. Desouza, R. L. Mercer, V. J. D. Pietra, and J. C. Lai, "Class-based n-gram models of natural language," *Computational linguistics*, vol. 18, no. 4, pp. 467–479, 1992.

[12] A. Go, R. Bhayani, and L. Huang, "Twitter sentiment classification using distant supervision," *CS224N Project Report, Stanford*, 2009.

[13] J. H. Zaragoza, L. E. Sucar, E. F. Morales, C. Bielza, and P. Larrañaga, "Bayesian chain classifiers for multidimensional classification," in *Proceedings of the 22nd International Joint Conference on Artificial Intelligence, Barcelona, Catalonia, Spain, July 16-22, 2011*. AAAI Press, 2011, pp. 2192–2197.

[14] C. Nadeau and Y. Bengio, "Inference for the generalization error," *Machine Learning*, vol. 52, no. 3, pp. 239–281, 2003.

[15] S. M. Mohammad, "#Emotional tweets," in *Proceedings of the Sixth International Workshop on Semantic Evaluation*. Montréal, Canada: Association for Computational Linguistics, 7-8 June 2012, pp. 246–255.

[16] M. Baroni, G. Dinu, and G. Kruszewski, "Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors." in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, 2014, pp. 238–247.