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

2016 IEEE/WIC/ACM International Conference on Web Intelligence
Omaha, Nebraska, USA

Felipe Bravo-Marquez[1], Saif M. Mohammad[2], Eibe Frank[1], and Bernhard Pfahringer[1]

[1]University of Waikato,Computer Science Department
Hamilton, New Zealand
[2]National Research Council Canada
Ottawa, ON, Canada

October 15, 2016

#Emotional Tweets

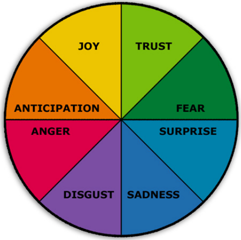
- Posts in Twitter or **tweets** are provided **freely and voluntarily** by users.

    1. Hey @Apple, pretty much all your products are amazing. You blow minds every time you launch a new gizmo. That said, your hold music is crap.
    2. #windows sucks... I want #imac so bad!!! why is it so damn expensive :(@apple please give me free imac and I will love you :D

- Analysing the emotions behind those messages has important applications in product **marketing**, **politics**, and even for **stock market analysis** [Bollen et al., 2011].

The NRC Emotion Lexicon

- A well known lexical resource for **automatically analysing** emotions from textual data is the **NRC word-emotion association lexicon** (NRC-10) [Mohammad and Turney, 2013]
- It contains more than 14,000 English words **manually annotated** according to ten non-exclusive **emotional** and **polarity categories**.
- Examples: **achieved** is mapped to **anticipation, joy, and trust**, and **exile** is mapped into **anger, fear, and sadness**.



- NRC-10 does not cover **informal expressions** used in Twitter.
- It suffers from **limitations** for analysing emotions from tweets.

Proposal

NRC Lex

angry	ang,fear
happy	joy,trust
good	joy,trust
upset	ang,sad

Tweets

t1	lol happy
t2	lol good
t3	grr angry
t4	grr upset

Words features

	f1	f2	f3	f4	f5
angry	1	0	0	1	0
happy	0	0	1	0	1
good	0	0	1	0	1
grr	0.5	0	0	1	0
lol	0	0.5	0.5	0	1
upset	1	0	0	1	0

Multi-label classification

Expanded lex

lol	joy,trust
grr	ang,fear

Word-level features

Attributes based on averaging tweet-level features:

1. **Word unigrams (UNI)**: based on an unigram frequency count.
2. **Brown clusters (BWN)**: in which the tweet is tagged according to low-dimensional Brown clusters of words.
3. **POS n-grams (POS)**: in which the frequency of each POS unigram and bigram is counted.
4. **Distant Polarity (DP)**: in which a logistic regression model is trained from a corpus of tweets with positive **:)** and negative **:(emoticons** and applied to the tweet.

Attributes based on dense Embeddings:

1. **Word2Vec Embeddings (W2V)**: skip-gram embeddings trained from a corpus of tweets.

Multi-Label Classification Models

1. **Binary Relevance (BR)**: in which one **separated binary classifier** is trained per label.
2. **Classifier Chains (CC)** [Read et al., 2011]: in which the predictions for each binary classifier are **cascaded** as additional features along a random permutation of labels.
3. **Bayesian Classifier Chains (BCC)** [Zaragoza et al., 2011]: in which a Bayesian network that represents **dependency relations** between the labels is learned and used to build a **classifier chain**.

Intrinsic Evaluation

- We compare the **micro-averaged F1** for the **ten affective** categories on the labelled words using **10-fold cross-validation**.
- We use **logistic regression** as the **base learner** in the different models.
- We compare different combinations of **features** and **classifiers**.

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

- W2V-embeddings produce the **strongest** features!
- There are **no clear differences** between multi-label models!

Expanded Lexicon

Extrinsic Evaluation

- We conduct an **extrinsic evaluation** by studying the usefulness of the expanded lexicons for classifying Twitter messages annotated with **emotional hashtags**.
- We compare a **logistic regression** that uses **NRC-10 alone** with another one using NRC-10 and **the expanded lexicon**.

Lexicon	Kappa			AUC		
NRC-10 (alone)	0.0769			0.633		
NRC-10+Expanded	BR	CC	BCC	BR	CC	BCC
UNI	0.1912	0.2006	0.1977	0.711	0.714	0.713
UNI-BWN	0.174	0.1783	0.176	0.708	0.712	0.711
UNI-BWN-POS	0.1753	0.1767	0.1776	0.708	0.711	0.710
UNI-BWN-POS-DP	0.1803	0.1829	0.1835	0.713	0.715	0.714
UNI-BWN-POS-DP-W2V	0.1871	0.1966	0.1832	0.712	0.714	0.713
W2V	**0.2234**	**0.2256**	0.2256	0.720	**0.723**	**0.723**
W2V-BWN	0.1988	0.2007	0.1974	0.713	0.715	0.715
W2V-BWN-POS	0.195	0.2012	0.1956	0.710	0.713	0.712
W2V-BWN-POS-DP	0.1994	0.2041	0.1992	0.714	0.715	0.715
W2V-DP	0.2228	0.2234	**0.2263**	**0.722**	**0.723**	**0.723**

- All the expanded lexicons are **substantially better** than using NRC-10 alone.
- Lexicons created with CC and BCC are slightly better than the ones created using BR in most cases.

Conclusions

- The results obtained 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.
- This method could be used for creating **domain specific** emotion lexicons for elections or sport competitions.

Questions?

Thanks for your Attention!

Acknowledgments

- University of Waikato Doctoral Scholarship
- Machine Learning Group at the University of Waikato

THE UNIVERSITY OF WAIKATO
Te Whare Wānanga o Waikato

References I

Bollen, J., Mao, H., and Zeng, X. (2011).
Twitter mood predicts the stock market.
Journal of Computational Science, 2(1):1–8.

Mohammad, S. and Turney, P. D. (2013).
Crowdsourcing a word-emotion association lexicon.
Computational Intelligence, 29(3):436–465.

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

Read, J., Pfahringer, B., Holmes, G., and Frank, E. (2011).
Classifier chains for multi-label classification.
Machine learning, 85(3):333–359.

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