Acquiring and Exploiting Lexical Knowledge for Twitter Sentiment Analysis

Felipe Bravo-Marquez
Chief Supervisor: Bernhard Pfahringer
Supervisor: Eibe Frank

Department of Computer Science, University of Waikato

17 July, 2017

THE UNIVERSITY OF WAIKATO
Te Whare Wānanga o Waikato

Message-level Polarity Classification (MPC)

1. Automatically classify a tweet to classes **positive**, **negative**, or **neutral**.

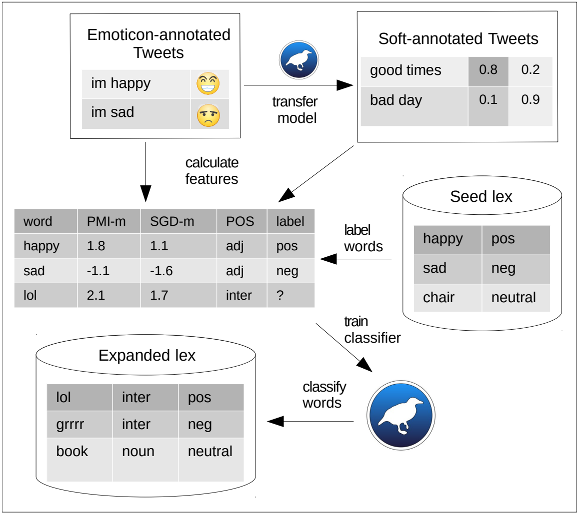
2. Challenge: Tweets use a unique **informal dialect** including many abbreviations, acronyms, misspelled words, hashtags, and emoticons, e.g., **lol**, **omg**, **hahaha**, **#hatemonday**, **#SweetAsBro**, **#yeahnah**, **:)** .
3. State-of-the-art solutions use **supervised** machine learning models trained from **manually** annotated examples [Kiritchenko et al., 2014].
4. **Label sparsity problem (LS)**: manual annotation is **labour-intensive** and **time-consuming**.

Research Problem

This thesis addresses the label sparsity problem for Twitter sentiment classification by automatically building **two type of resources**.

1. **Twitter-specific opinion lexicons**: we develop machine learning models to induce polarity lexicons from tweets.
2. **Synthetically labelled tweets**: we develop distant supervision methods based on **lexical knowledge** (we go beyond emoticons).

Word-sentiment Associations for Polarity Lexicon Induction

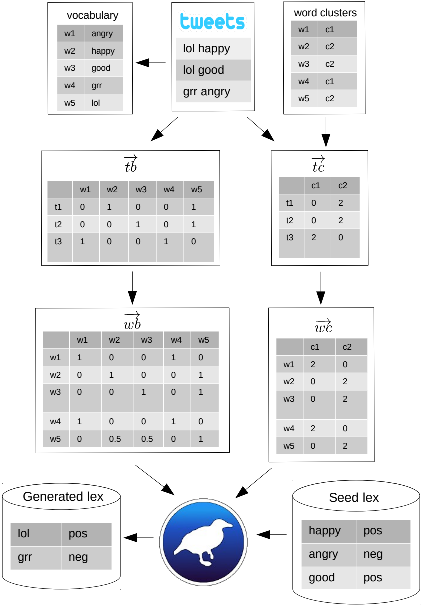


Word-level Classification Results using RBF SVMs

Weighted AUC		
Dataset	PMI-SO	ALL FEATURES
ED.EM	0.62 ± 0.02	$\mathbf{0.65} \pm \mathbf{0.02} +$
STS	0.64 ± 0.02	$\mathbf{0.66} \pm \mathbf{0.01} +$
ED.SL	0.63 ± 0.02	$\mathbf{0.65} \pm \mathbf{0.02} +$

Table: World-level classification performance.

Tweet-centroid Model for Lexicon Induction

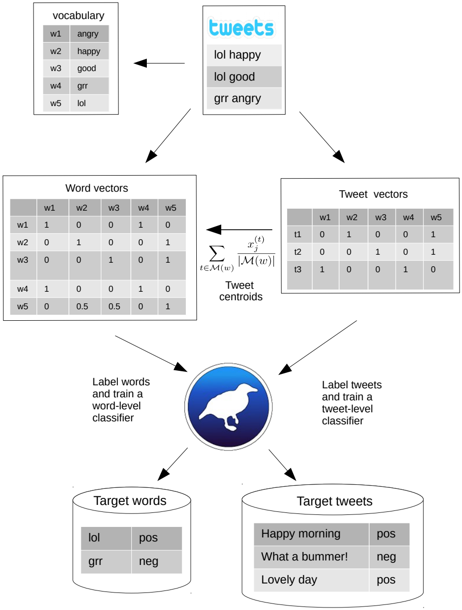


Message-level classification performance

AUC		
Dataset	Seed Lexicon	TCM Lexicon
Sanders	0.78 ± 0.04	$\mathbf{0.83} \pm \mathbf{0.04} +$
6-human	0.79 ± 0.03	$\mathbf{0.83} \pm \mathbf{0.02} +$
SemEval	0.78 ± 0.02	$\mathbf{0.84} \pm \mathbf{0.02} +$

Multi-Label Classification of Emotions with TCM

Transfer Learning with Tweet Centroids



Partitioned TCM

- We propose a modification of TCM for **increasing** the number of labelled instances it produces.
- The word-tweet set $\mathcal{M}(w)$ is **partitioned** into smaller disjoint subsets $\mathcal{M}(w)_1, \ldots \mathcal{M}(w)_z$ of a fixed size determined by p.
- We calculate one tweet centroid vector $\overrightarrow{w}$ for **each partition** labelled according to $\mathcal{L}$.

TCM for MPC

	6HumanCoded		Sanders		SemEval	
EAA	0.805 $\pm$ 0.005	= -	0.800 $\pm$ 0.017	= +	0.802 $\pm$ 0.006	= -
LAA	0.809 $\pm$ 0.001	+ =	0.778 $\pm$ 0.002	- =	0.814 $\pm$ 0.000	+ =
TCM	0.776 $\pm$ 0.004	- -	0.682 $\pm$ 0.024	- -	0.779 $\pm$ 0.008	- -
TCM (p=5)	0.834 $\pm$ 0.002	+ +	0.807 $\pm$ 0.008	= +	0.833 $\pm$ 0.002	+ +
TCM (p=10)	0.845 $\pm$ 0.003	+ +	**0.817** $\pm$ 0.006	+ +	0.841 $\pm$ 0.002	+ +
TCM (p=20)	**0.850** $\pm$ 0.003	+ +	0.815 $\pm$ 0.011	+ +	**0.844** $\pm$ 0.003	+ +
TCM (p=50)	0.844 $\pm$ 0.004	+ +	0.785 $\pm$ 0.010	- +	0.836 $\pm$ 0.004	+ +
TCM (p=100)	0.829 $\pm$ 0.003	+ +	0.752 $\pm$ 0.019	- -	0.821 $\pm$ 0.004	+ +

Table: Message-level Polarity Classification Results. Best results per column are given in bold.

Annotate-Sample-Average (ASA)

- What if we average random tweets containing **different words** with the same polarity?
- What if we can define the **number of instances** to generate?
- This could be useful for creating **compact and balanced** training datasets.

ASA algorithm

1. **Annotation**: every time a word from $\mathcal{L}$ is found, the tweet is added to sets **posT** or **negT** (depending on the polarity).
2. **Sample**: randomly sample with replacement a tweets from either **posT** or **negT** for each generated instance.
3. **Averaging**: average and label sampled feature vectors.

ASA results

	6HumanCoded		Sanders		SemEval	
EAA_U	0.805 ± 0.005	= = - -	0.800 ± 0.017	= = + +	0.802 ± 0.006	= + - -
EAA_B	0.809 ± 0.001	= = = =	0.795 ± 0.016	= = + +	0.798 ± 0.007	- = - -
LAA_U	0.809 ± 0.001	+ = = =	0.778 ± 0.002	- - = =	0.814 ± 0.000	+ + = =
LAA_B	0.809 ± 0.001	+ = = =	0.778 ± 0.003	- - = =	0.813 ± 0.001	+ + = =
ASA ($a = 1$, $m = F$)	0.793 ± 0.005	- - - -	0.762 ± 0.016	- - - -	0.787 ± 0.007	- - - -
ASA ($a = 5$, $m = F$)	0.837 ± 0.004	+ + + +	0.807 ± 0.010	= = + +	0.833 ± 0.003	+ + + +
ASA ($a = 10$, $m = F$)	**0.845** ± 0.001	+ + + +	**0.812** ± 0.015	+ + + +	**0.840** ± 0.003	+ + + +
ASA ($a = 50$, $m = F$)	0.815 ± 0.003	+ + + +	0.759 ± 0.006	- - - -	0.810 ± 0.004	+ + - -
ASA ($a = 100$, $m = F$)	0.781 ± 0.003	- - - -	0.720 ± 0.007	- - - -	0.779 ± 0.004	- - - -
ASA ($a = 500$, $m = F$)	0.723 ± 0.002	- - - -	0.670 ± 0.008	- - - -	0.729 ± 0.005	- - - -
ASA ($a = 1000$, $m = F$)	0.712 ± 0.002	- - - -	0.665 ± 0.007	- - - -	0.721 ± 0.005	- - - -

Table: AUC measure for different distant supervision models. Best results per column are given in bold.

Conclusions

- Twitter-specific polarity lexicons and lexicon-based distant supervision methods can successfully tackle the polarity classification of tweets when labels are **scarce**.
- We proposed two methods (Word Sentiment Associations and TCM) for building Twitter-specific **opinion lexicons** (acquisition of lexical knowledge).
- These methods could be used to create **domain-specific** lexicons.
- They could also be used to study the **dynamics** of opinion-words.
- Future work: try **non-linear representations** on TCM (Auto-Encoders or RBM).

Conclusions

- We proposed two **distant supervision methods** (TCM and ASA) that outperformed LAA and EAA for MPC.
- TCM is a **unified model** for message-level and word-level sentiment classification.
- Future work: subjectivity, emotions, handle negations, non-linear representations and deep networks.

Questions?

Thanks for your Attention!

Acknowledgements

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

THE UNIVERSITY OF WAIKATO
Te Whare Wānanga o Waikato

References I

Kiritchenko, S., Zhu, X., and Mohammad, S. M. (2014).
Sentiment analysis of short informal texts.
Journal of Artificial Intelligence Research, 50:723–762.