Annotate-Sample-Average (ASA): A New Distant Supervision Approach for Twitter Sentiment Analysis

22nd European Conference on Artificial Intelligence
The Hague, The Netherlands

Felipe Bravo-Marquez, Eibe Frank, and Bernhard Pfahringer

University of Waikato
Computer Science Department

31 August, 2016

THE UNIVERSITY OF WAIKATO
Te Whare Wānanga o Waikato

Social Media

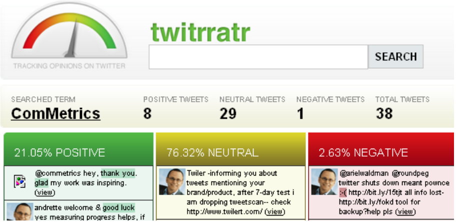
- Microblogging services are increasingly being adopted by people in order to access and publish information.
- **Twitter**: Massively used Microblogging platform where users post messages (a.k.a **tweets**) limited to 140 characters.
- Tweets use a unique **informal dialect** including many abbreviations, acronyms, misspelled words, hashtags, and emoticons, e.g., **lol**, **omg**, **hahaha**, **#hatemonday**, **:)** .

Sentiment Analysis and Social Media

- Twitter users tend to publish **personal opinions** regarding certain topics and news events.

    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 sentiment underlying these opinions has important applications in product **marketing** and **politics**.



Opinion Mining or Sentiment Analysis

- Application of **NLP** and **text mining** techniques to identify and extract subjective information from textual datasets.

Message-level Polarity Classification (MPC) of Tweets

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

- State-of-the-art solutions use **supervised** machine learning models trained from **manually** annotated examples [Kiritchenko et al., 2014].
- **The Label Sparsity Problem**: manual annotation is **labour-intensive** and **time-consuming**.

Distant Supervision

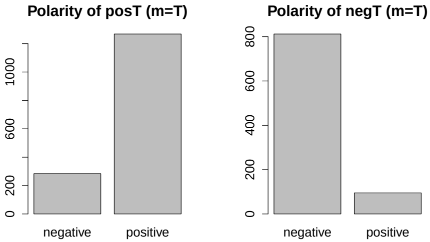
- The label sparsity problem can be addressed using **Distant Supervision**.
- Automatically **label** unlabelled data (**Twitter API**) using a heuristic method [Mintz et al., 2009].
- Unlabelled tweets are **cheap** to optain :)
- **Emoticon-Annotation Approach (EAA)**: tweets with positive **:)** or negative **:(** emoticons are labelled according to the polarity indicated by the emoticon [Read, 2005].
- The emoticon is **removed** from the content.
- Drawback: emoticons are **rarely** used in certain domains such as politics.

Lexicon-based Distant Supervision

- We propose a distant supervision method that builds **synthetically labelled tweets** based on **opinion lexicons** (we go beyond emoticons).
- An opinion lexicon $\mathcal{L}$ is a lists of terms labelled by sentiment.
- They are normally composed of positive and negative words such as **happy, wonderful** and **sad, bad**.
- Proposed method generate positive and negative training instances by **averaging** tweets containing words with the **same** polarity.
- Tweets are represented by **sparse feature vectors**.

Lexical Polarity Hypothesis

- A tweet containing a word with a certain polarity is more likely to express the **same polarity** than the **opposite** $p_d > 0.5$ (Bernoulli experiment).



- The opposite polarity may also be expressed due to the presence of **negation**, **sarcasm**, or other opinion words with the **opposite** polarity.

Why Averaging?

- Averaging multiple tweets with words with the same polarity **increases** the confidence of generating instances located in the **region** of the desired polarity.
- We assume that the average tweet will behave similarly to the **majority**.
- Probability that the **majority** of the tweets sampled from a collection of tweets with at least one word with the target polarity have the desired polarity:

$$P(M) = \sum_{i=\lfloor \frac{a}{2} \rfloor + 1}^{a} \binom{a}{i} p_d^i (1 - p_d)^{a-i}$$

	$p_d = 0.6$	$p_d = 0.7$	$p_d = 0.8$	$p_d = 0.9$
$a = 3$	0.648	0.784	0.896	0.972
$a = 5$	0.683	0.837	0.942	0.991
$a = 10$	0.633	0.850	0.967	0.998
$a = 50$	0.902	0.998	1	1
$a = 100$	0.973	1	1	1
$a = 500$	1	1	1	1
$a = 1000$	1	1	1	1

- $P(M) > p_d$, when $a \geq 3$ and $p_d \geq 0.5$. This is analogous to the **Condorcet's Jury Theorem**!!

Annotate-Sample-Average (ASA)

- **Annotation**: every time a word from $\mathcal{L}$ is found, the tweet is added to sets **posT** or **negT** (depending on the polarity).
- Tweets with both positive and negative words will be simultaneously added to both **posT** and **negT**.
- This will produce instances with better **generalisation** properties: Tweets are likely to contain words with the **opposite polarity**.
- **Sample**: randomly sample with replacement a tweets from either **posT** or **negT** for each generated instance.
- **Averaging**: average and label sampled feature vectors.
- We create **balanced training datasets** with size equal to 1% of the size of the source corpus ($20,000$ in our experiments).

Baselines

Emoticon-Annotation Approach (EAA)

- Labels tweets with positive or negative emoticons according to the emoticon's polarity after removing the emoticon from the message.
- Tweets containing both positive and negative emoticons are **discarded**.

Lexicon-annotation approach (LAA)

- Uses a given polarity lexicon $\mathcal{L}$.
- Tweets with at least one positive word and no negative word are labelled **positive**.
- Tweets with at least one negative word and no positive word are labelled **negative**.

We compare balance and unbalanced versions of each baseline!

Instances Generated by Distant Supervision Models

We use 10 collections of 2 million tweets as source corpora.

	Avg. Positive	(%)	Avg. Negative	(%)	Avg. Total	(%)
EAA	130,641	(6.5%)	21,537	(1.1%)	152,179	(7.6%)
EAA_B	21,537	(1.1%)	21,537	(1.1%)	43,074	(2.2%)
LAA	681,531	(34.1%)	294,177	(14.7%)	975,708	(48.8%)
LAA_B	294,177	(14.7%)	294,177	(14.7%)	588,354	(29.4%)
ASA	10,000	(0.5%)	10,000	(0.5%)	20,000	(1%)

Testing Datasets

	Positive	Negative	Total
6HumanCoded	1340	949	2289
Sanders	570	654	1224
SemEval	5232	2067	7299

Table : Message-level polarity classification datasets.

ASA results

	6HumanCoded		Sanders		SemEval	
EAA_U	0.576 ± 0.007	= - - -	0.506 ± 0.018	= - - -	0.591 ± 0.018	= - - -
EAA_B	0.735 ± 0.008	+ = + +	0.709 ± 0.018	+ = = =	0.711 ± 0.006	+ = - =
LAA_U	0.729 ± 0.004	+ - = +	0.711 ± 0.003	+ = = +	0.725 ± 0.002	+ + = +
LAA_B	0.719 ± 0.002	+ - - =	0.703 ± 0.004	+ = - =	0.712 ± 0.002	+ = - =
ASA ($a = 1$)	0.717 ± 0.007	+ - - =	0.691 ± 0.013	+ - - -	0.699 ± 0.008	+ - - -
ASA ($a = 5$)	0.755 ± 0.004	+ + + +	0.730 ± 0.008	+ + + +	0.735 ± 0.005	+ + + +
ASA ($a = 10$)	**0.761** ± 0.003	+ + + +	**0.735** ± 0.015	+ + + +	**0.742** ± 0.006	+ + + +
ASA ($a = 50$)	0.749 ± 0.004	+ + + +	0.673 ± 0.005	+ - - -	0.699 ± 0.009	+ - - -
ASA ($a = 100$)	0.717 ± 0.003	+ - - -	0.645 ± 0.006	+ - - -	0.664 ± 0.005	+ - - -
ASA ($a = 500$)	0.665 ± 0.002	+ - - -	0.621 ± 0.007	+ - - -	0.621 ± 0.004	+ - - -
ASA ($a = 1000$)	0.653 ± 0.003	+ - - -	0.619 ± 0.007	+ - - -	0.613 ± 0.002	+ - - -

Table : Macro-averaged F1 measure for different distant supervision models. Best results per column are given in bold.

Conclusions & Future Work

- ASA is a powerful distant supervison method that creates **compact and balanced** training datasets.
- It outperformed EAA and LAA!
- It could potentially used for domain-specific sentiment analysis.

Future Work

- Try ASA with other type of word-level labels: subjectivity, emotions.
- Design a mechanism for handling negations with ASA.
- Try non-linear representations with ASA such as paragraph embeddings.

Questions?

Thanks for your Attention!

Acknowledgements

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

References I

- Brown, P. F., Desouza, P. V., Mercer, R. L., Pietra, V. J. D., and Lai, J. C. (1992).
  Class-based n-gram models of natural language.
  Computational linguistics, 18(4):467–479.
- Esuli, A. and Sebastiani, F. (2005).
  Determining the semantic orientation of terms through gloss classification.
  In *Proceedings of the 14th ACM International Conference on Information and Knowledge Management*, CIKM '05, pages 617–624, New York, NY, USA. ACM.
- Esuli, A. and Sebastiani, F. (2006).
  Sentiwordnet: A publicly available lexical resource for opinion mining.
  In *In Proceedings of the 5th Conference on Language Resources and Evaluation (LREC'06*, pages 417–422.
- Harris, Z. (1954).
  Distributional structure.
  Word, 10(23):146–162.
- Hu, M. and Liu, B. (2004).
  Mining and summarizing customer reviews.
  In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '04, pages 168–177, New York, NY, USA. ACM.

References II

Kamps, J., Marx, M., Mokken, R. J., and De Rijke, M. (2004).
Using WordNet to Measure Semantic Orientation of Adjectives.
In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, volume 4, pages 1115–1118. European Language Resources Association (ELRA).

Kim, S.-M. and Hovy, E. (2004).
Determining the sentiment of opinions.
In *Proceedings of the 20th international conference on Computational Linguistics*, COLING '04, Stroudsburg, PA, USA. Association for Computational Linguistics.

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

Kouloumpis, E., Wilson, T., and Moore, J. (2011).
Twitter sentiment analysis: The good the bad and the omg!
ICWSM, 11:538–541.

Mintz, M., Bills, S., Snow, R., and Jurafsky, D. (2009).
Distant supervision for relation extraction without labeled data.
In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2 - Volume 2*, ACL '09, pages 1003–1011, Stroudsburg, PA, USA. Association for Computational Linguistics.

References III

Mohammad, S. M., Kiritchenko, S., and Zhu, X. (2013).
Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets.
Proceedings of the seventh international workshop on Semantic Evaluation Exercises (SemEval-2013).

Read, J. (2005).
Using emoticons to reduce dependency in machine learning techniques for sentiment classification.
In *Proceedings of the ACL Student Research Workshop*, ACLstudent '05, pages 43–48, Stroudsburg, PA, USA. Association for Computational Linguistics.

Thelwall, M., Buckley, K., and Paltoglou, G. (2012).
Sentiment strength detection for the social web.
JASIST, 63(1):163–173.

Turney, P. D. and Littman, M. L. (2003).
Measuring praise and criticism: Inference of semantic orientation from association.
ACM Transactions on Information Systems (TOIS), 21(4):315–346.

Zhou, Z., Zhang, X., and Sanderson, M. (2014).
Sentiment analysis on twitter through topic-based lexicon expansion.
In Wang, H. and Sharaf, M., editors, *Databases Theory and Applications*, volume 8506 of *Lecture Notes in Computer Science*, pages 98–109. Springer International Publishing.

References IV