University_of_Warwick: SENTIADAPTRON - A Domain Adaptable Sentiment Analyser for Tweets - Meets SemEval

Richard Townsend
University of Warwick
Richard.Townsend@warwick.ac.uk

Aaron Kalair
University of Warwick
Aaron.Kalair@warwick.ac.uk

Ojas Kulkarni
University of Warwick
Ojas.Kulkarni@warwick.ac.uk

Rob Procter
University of Warwick
Rob.Procter@warwick.ac.uk

Maria Liakata
University of Warwick
M.Liakata@warwick.ac.uk

Abstract

We give a brief overview of our system, SentiAdaptron, a domain-sensitive and domain adaptable system for twitter analysis in tweets, and discuss performance on SemEval (in both the constrained and unconstrained scenarios), as well as implications arising from comparing the intra- and inter- domain performance on our twitter corpus.

1 Introduction

A domain is broadly defined as a set of documents demonstrating a similar distribution of words and linguistic patterns. Task 9 of *SemEval* treats Twitter as a single domain with respect to sentiment analysis. However previous research has argued for the topic-specific treatment of sentiment given domain-specific nuances and the over-generality of current sentiment analysis systems with respect to applications in the social sciences (Thelwall and Buckley, 2013). Thelwal's method - manually extending a sentiment lexicon for a particular topic or domain - highlights that expression of sentiment varies from one domain to another. Rather than relying on the manual extension of lexica, we developed an approach to Twitter sentiment classification that is domain sensitive. To this effect we gathered tweets from three primary domains - financial news, political opinion, technology companies and their products - and trained our system on one domain while adapting to the other. Using this methodology we obtained both intrinsic as well as extrinsic evaluation of the system on real world applications with promising results. As our approach to sentiment analysis has been influenced by the task description of SemEval 2013 we decided to also evaluate the system on SemEval's data, since it provides a well established benchmark. In the following we briefly describe our system and corpus and discuss our approach for the SemEval submission.

	Positive	Negative	Neutral	Objective
SemEval	11610	6332	905	189887
Our corpus	10725	17837	3514	36904

(a) Contextual polarity.

	Positive	Negative	Neutral	Objective
SemEval	4215	1798	4082	1243
Our corpus	1090	1711	1191	-

(b) Message polarity.

Table 1: The distribution of sentiment classes between SemEval and our corpus at word and tweet level.

2 A Corpus of Three Domains: Source of Unconstrained Data

Our goal in developing SentiAdaptron was domain adaptive tweet-level classification. We decided to follow SemEval 2013 and collect both word-level as well as message level annotations. We prepared a corpus of 4000 tweets with a balanced coverage of financial, political and technology related tweets. Tweets were collected using keywords from domain-specific websites: the final list was chosen after evaluating each candidate keyword's popularity using a third-party service[1]. Each tweet is tagged with multiple candidate domains, based on a hierarchy of terms generated from the original keyword list and tweets are filtered using a clustering methodology based on the DBSCAN clustering algorithm to remove robotic and repetitive tweets. We performed domain disambiguation for annotation through keyword filtering and also by picking a number of synsets from WordNet and computing the tweet's mean semantic distance using the NLTK toolkit. Tweets which didn't contain

[1] http://topsy.com

any words associated with a score of less than 0.3 in SentiWordNet were removed, in a manner similar to Nakov et al (2013). After further manual relevance checks, the remaining tweets were submitted to Amazon's Mechanical Turk service for message and phrase-level annotation. We initially used the form demonstrated by Nakov et al., although we later redesigned it with a dramatic improvement in annotator performance and annotation quality. Each tweet was annotated by four workers and annotation sparsity at the phrase level was addressed by taking the majority of annotations following the precedence neutral > pos > neg > other. This is in contrast to the approach used by Nakov et al. for SemEval which intersects annotations. Annotations at the tweet level were aggregated using a majority vote. We found that using the proportion of positive, negative and neutral words in a tweet is a surprisingly robust feature for cross-domain classification, and boosted our performance when using bigrams.

3 Subjectivity Detection and Contextual Polarity Disambiguation (Subtask A)

Sentiment lexicons such as SentiWordNet (Esuli and Sebastiani, 2006), the NRC emotions lexicon (Mohammad and Turney, 2010), the MPQA lexicon (Wilson et al., 2005) and the Bing Liu Lexicon (Hu and Liu, 2004) have been used for determining whether a phrase should be labelled as positive, negative or neutral within a tweet or sentence (contextual polarity). However, lexical resources are by nature non-contextual and may not have good coverage over a given domain. We instead considered how to infer contextual polarity purely from the data available.

To address the problem of class imbalance in the tweets, we break the problem of contextual polarity detection into two stages: (i) we first determine whether a given word should be assigned a positive, negative or neutral annotation (*subjectivity detection*) and (ii) distinguish subjective tweets into positive, negative neutral.

3.1 Contextual Subjectivity Detection

Task A asks participants to predict the contextual subjectivity annotation of a text span at a given offset: our extrinsic applications don't have this fundamental structure, so we considered whether it was possible to automatically separate the content of input documents into those regions which should be assigned an annotation and those which should not. We considered a unigram and a bigram baseline using a naive Bayes classifier, which gave an F1 score of 0.640 and 0.520 respectively on our in domain data (under 10-fold cross-validation). We followed a number of approaches to subjectivity detection to try and improve on the baseline including sequential modelling using linear-chain Conditional Random Fields (CRFs) with CRFsuite (Okazaki, 2007) and lexical inference using semantically disambiguated WordNet (Miller, 1995) synsets in conjunction with their occurrence in a subjective context.

We found that the observed subjective proportion of a given word alongside its successor and predecessor[2] was a viable feature engineering scheme, which we call *neighbouring subjectivity proportions*. This gave the best subjectivity performance on our in-domain data when fed to a *voted perceptron* (Freund and Schapire, 1999), an ensemble approach which assigns particularly predictive iterations of an incrementally trained perceptron a greater weight when deciding the final classification, and offers wide-margin classification akin to support vector machines whilst also requiring less parameter exploration. We used the implementation provided by the WEKA machine-learning environment (Hall et al., 2009), which achieved an F1-score of 0.740 (again under 10-fold cross validation) for our in-domain data, but performance dropped to an F1-score of 0.323 on the SemEval 2013 training and development data. Table 1 indicates that the proportion of objective features in SemEval is much greater than that seen within our own corpus, likely due to the differences in the way we processed annotations (outlined in Section 2).

3.2 Contextual Polarity

We considered a naive Bayes unigram baseline, (similar approaches have proven popular with SemEval 2013 participants for Task A) and achieved an F1-measure of 0.662 when training using SemEval's 2014 training and development data and evaluating on SemEval's 2013 gold-standard annotations. However we could not detect the neutral class, and the test did not consider the objective class.

[2] As an example, if we observe 14 total occurrences of the word "heartbreaking", and 13 of them appear with a positive, negative or neutral label, the subjective proportion computed would be 13/14.

	F1-score				Precision			Recall		
	P	N	E	Overall	P	N	E	P	N	E
Task A (constrained)	85.5	49.0	9.84	67.3	91.1	41.4	3.8	80.6	60.2	16.7
Task A (unconstrained)	85.1	49.2	12.2	67.2	89.8	43.4	8.0	80.9	59.8	25.9
Finance	78.0	83.1	51.2	78.0	77.3	81.7	58.7	78.7	45.5	84.6
Politics	67.3	83.3	42.1	74.3	70.9	79.7	50.3	64.1	87.2	36.3
Technology	75.1	85.6	47.4	77.8	77.8	81.8	56.7	89.8	89.8	40.7

(a) Performance on word-level contextual annotation tasks.

	F1-score				Precision			Recall		
	P	N	E	Overall	P	N	E	P	N	E
Task B (constrained)	57.1	34.0	57.0	45.6	46.1	42.6	68.8	75.0	28.3	48.7
Task B (unconstrained)	57.2	33.0	57.7	45.1	46.2	36.1	72.1	74.9	30.4	47.9
Finance	78.7	78.9	64.6	73.8	77.9	79.5	64.8	78.3	79.5	64.5
Politics	80.8	75.0	61.9	72.3	78.4	74.8	63.7	83.3	75.1	60.2
Technology	73.2	78.2	58.0	70.1	69.7	76.3	62.9	76.9	80.2	53.7

(b) Performance on document-level annotation tasks.

Table 2: Classifier metrics obtained from 4-fold intra-domain cross-validation (using reference annotations) and results for subtasks A and B of SemEval task 9 (computed using reference scorer).

	Tech	Politics	Finance
Unigrams	0.53	0.35	0.53
Bigrams	0.38	0.31	0.52
Bigrams + SP	0.77	0.65	0.78
Unigrams + SP	0.68	0.62	0.76

Table 3: F1-scores achieved for each domain on our corpus (naive Bayes, 10-fold cross validation) using reference annotations with and without subjective (positive, negative, neutral) proportions (SP).

We improved on this baseline by combining unigrams with information from the wider context of the tweet. The algorithm first runs subjectivity detection on the entire document and then, for each word we need to classify (or otherwise each word detected as subjective), effectively generates two bags of words consisting of the subjective words before and after that word (we also included any adverbs as annotated by the Gimpel tagger (2011) in this bag to improve robustness). We output the word itself as a further feature, and use a random forest classifier (10 trees, $\log_2 N + 1$ features) to generate the annotation. We found this approach outperformed the other approaches we tried (including Naive Bayes and OneR) and also gave us better F1-scores on the neutral class. Results from this approach for our in-domain data and the SemEval 2014 data can be seen in Table 2a. The drop in performance from our in-domain data to SemEval 2014 can be explained by the different class distribution observed in SemEval (Table 1). Subjectivity detection was used to generate features for subtask B, but not subtask A, where the target subjective phrases are already given.

4 Message Polarity Classification (Subtask B)

We tried various different combinations of features to discover the best intra-domain classification approach for our corpus and found that the proportion of positive, negative and neutral words within a tweet boosted performance using bigram binary features (Table 3). This involves first running the contextual polarity detection component as described in Section 3 and feeding in the results as features (together with bigrams) into a naive Bayes classifier for tweet level sentiment detection. However, one of our hypotheses was that domain adaptation could help improve performance when moving from one domain to another, effectively allowing us to port our classifier from our own corpus to SemEval.

4.1 Cross-Domain Adaptation

Our research in domain adaptation uses and extends the technique described by Blitzer and Pereira (2007) called Structural Correspondence Learning (SCL), which derives a relationship (or correspondence) between features from two different domains. This is done via *pivot features* selected from the intersection of features from both domains which have been ranked according to mutual information. The technique then uses N pivot features from both the seen and unseen domains to learn a set of binary problems corresponding to whether a given pivot exists in a target document. A perceptron is then used to train each of the binary problems, giving a matrix of weights (where a weight represents covariance of non pivot features with pivot features). We extended Blitzer's technique to encompass the neutral class and gave a wider notion of domain than previously found in the literature. As an example Liu et al. (2013) use

Training domains	Test	Accuracy	Precision	Recall	F-measure
Tech & Finance	Politics	0.76	0.52	0.40	0.45
Tech & Politics	Finance	0.69	0.51	0.78	0.61
Finance & Politics	Tech	0.63	0.53	0.58	0.55

Table 4: Binary class metrics with structural correspondence learning on our own corpus.

			F-measure			**Precision**			**Recall**		
Train	**Test**	**Accuracy (SP)**	**SP**	**Baseline**(Bigrams)	**Loss**	**SP**	**Baseline (Bigrams)**	**Loss**	**SP**	**Baseline (Bigrams)**	**Loss**
Tech	**Politics**	69.74%	0.67	0.24	0.43	0.64	0.19	0.45	0.71	0.33	0.38
Tech	**Finance**	72.28%	0.78	0.20	0.58	0.74	0.15	0.59	0.83	0.33	0.50
Politics	**Tech**	72.40%	0.76	0.21	0.55	0.71	0.25	0.46	0.82	0.33	0.49
Politics	**Finance**	73.06%	0.79	0.22	0.57	0.77	0.16	0.61	0.82	0.33	0.49
Finance	**Politics**	68.39%	0.66	0.23	0.43	0.67	0.18	0.49	0.66	0.33	0.33
Finance	**Tech**	71.92%	0.76	0.22	0.54	0.71	0.16	0.55	0.82	0.33	0.49
Tech & Finance	**Politics**	69.17%	0.63	0.23	0.40	0.61	0.18	0.43	0.67	0.33	0.34
Tech & Politics	**Finance**	70.49%	0.63	0.20	0.43	0.61	0.15	0.46	0.64	0.33	0.31
Finance & Politics	**Tech**	72.88%	0.77	0.22	0.55	0.72	0.16	0.56	0.84	0.33	0.51

Table 5: Classifier metrics from training and testing on different domains, with and without proportions of positive, negative and neutral phrases from the source domain (Subjective Proportions SP).

an SVM-derived technique to adapt on domains containing terms relevant to *Google* and *Twitter*, which are both considered part of the technology domain in our corpus, whereas we attempted to adapt from technology topics to financial news and political opinions.

We found that the amount of mutual information in our three domains was very low and was practically zero for the three class version of our problem. The results for the binary version of the classifier generated poorer results (Table 4) than those produced by our back-up classifier (based on the naive Bayes bigrams and subjective phrase proportions from the source domain, see Table 5, last three rows). Therefore we generated our submission to SemEval 2014 based on bigrams and subjective proportions rather than SCL, since we found that the proportion of pos/neg/neutral phrases is a robust feature across domains (as long as it can be reliably predicted during the contextual polarity prediction stage, which was the case for our data). Our results for SemEval task B using the subjective phrase proportions can be found in Table 2b. Our unconstrained performance indicates that whilst this classifier provides reasonable cross domain performance for our own data (Table 5), it is very sensitive to the performance of subjectivity and contextual polarity detection, which is lower for SemEval than it is for our own corpus. Presumably the reason for this is the different assumptions in annotations in the two cases and the differences in the class distributions between SemEval and our own data. This meant that our performance was lower than systems that had specifically trained on the SemEval data.

5 Conclusions

Our goal was to demonstrate the potential of domain sensitivity and domain adaptability for sentiment analysis in tweets - a task which brings challenges defying the use of fixed lexica. We found that the proportions of positive, negative and neutral tweets are quite robust cross-domain features, although we do think that domain adaptation techniques such as Structural Correspondence Learning merit further investigation in the context of sentiment analysis for Twitter.

Access to the source code of this submission

The source code of the applications used to gather and prepare our corpus, conduct CRF-suite and structural correspondence learning, and the Java-based environment used to generate our final submission are available at `https://github.com/Sentimentron/Nebraska-public`[3] and `https://github.com/Sentimentron/PRJ9081`[4].

Acknowledgements

Warwick Research Development Fund grant RD13129 provided funding for crowdsourced annotations. We thank our partners at CUSP, NYU for enabling us to use Amazon Mechanical Turk for this process.

[3] `http://dx.doi.org/10.5281/zenodo.9906`

[4] `http://dx.doi.org/10.5281/zenodo.9904`

References

Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD*, volume 96, pages 226–231.

Andrea Esuli and Fabrizio Sebastiani. 2006. SentiWordNet: A publicly available lexical resource for opinion mining. In *Proceedings of LREC*, volume 6, pages 417–422.

Yoav Freund and Robert E Schapire. 1999. Large margin classification using the perceptron algorithm. *Machine learning*, 37(3):277–296.

Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A. Smith. 2011. Part-of-speech tagging for Twitter: annotation, features, and experiments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers - Volume 2*, HLT '11, pages 42–47.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. 2009. The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1):10–18.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177.

Mark Dredze John Blitzer and Fernando Pereira. 2007. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 440–447.

Shenghua Liu, Fuxin Li, Fangtao Li, Xueqi Cheng, and Huawei Shen. 2013. Adaptive co-training SVM for sentiment classification on tweets. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*, pages 2079–2088.

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

Saif M Mohammad and Peter D Turney. 2010. Emotions evoked by common words and phrases: Using Mechanical Turk to create an emotion lexicon. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 26–34.

Preslav Nakov, Sara Rosenthal, Zornitsa Kozareva, Veselin Stoyanov, Alan Ritter, and Theresa Wilson. 2013. SemEval-2013 Task 2: Sentiment Analysis in Twitter. pages 312–320, June.

Naoaki Okazaki. 2007. CRFsuite: a fast implementation of Conditional Random Fields (CRFs). [ONLINE] `http://www.chokkan.org/software/crfsuite/` (Retrieved: July 16, 2014).

Mike Thelwall and Kevan Buckley. 2013. Topic-based sentiment analysis for the social web: The role of mood and issue-related words. *Journal of the American Society for Information Science and Technology*, 64(8):1608–1617.

Theresa Wilson, Paul Hoffmann, Swapna Somasundaran, Jason Kessler, Janyce Wiebe, Yejin Choi, Claire Cardie, Ellen Riloff, and Siddharth Patwardhan. 2005. OpinionFinder: A system for subjectivity analysis. In *Proceedings of HLT/EMNLP on Interactive Demonstrations*, pages 34–35.