Cross-discourse Development of Supervised Sentiment Analysis in the Clinical Domain

Phillip Smith
School of Computer Science
University of Birmingham
pxs697@cs.bham.ac.uk

Mark Lee
School of Computer Science
University of Birmingham
M.G.Lee@cs.bham.ac.uk

Abstract

Current approaches to sentiment analysis assume that the sole discourse function of sentiment-bearing texts is expressivity. However, the persuasive discourse function also utilises expressive language. In this work, we present the results of training supervised classifiers on a new corpus of clinical texts that contain documents with an expressive discourse function, and we test the learned models on a subset of the same corpus containing persuasive texts. The results of this indicate that despite the difference in discourse function, the learned models perform favourably.

1 Introduction

Examining the role that discourse function holds is a critical part of an in-depth analysis into the capabilities of supervised sentiment classification techniques. However, it is a field that has not been comprehensively examined within the domain of sentiment analysis due to the lack of suitable cross-discourse corpora to train and test various machine learning methods upon.

In order to carry out such an investigation, this study will focus on the relationship between sentiment classification and two types of discourse function: *Expressive* and *Persuasive*. The expressive function denotes the feelings or attitudes of the author of a document. This is demonstrated in the following examples:

1. *"I didn't like the attitude of the nursing staff."*
2. *"The doctors treated me with such care."*

Intuitively, the associated polarity of each example is trivial to determine in these explicit examples. However, expressive statements do not operate in isolation of other respective discourse functions. As Biber (1988) notes, a persuasive statement incorporates elements of the expressive function in order to advise an external party of a proposed action that should be taken. The following example shows how persuasive statements make use of expressive functions:

1. *"The clumsy nurse who wrongly diagnosed me should be fired."*

The role of a persuasive statement is to incite an action in the target, dependent upon the intention that the author communicates. By using plain, sentiment-neutral language, the reader may misinterpret why the request for action is being given, and in the worst-case scenario not carry it out. Through the incorporation of expressive language, the weight of the persuasive statement is increased. This enables the speaker to emphasise the underlying sentiment of their statement, thereby increasing the likelihood of the intended action being undertaken, and their goals being accomplished. In the above example, the intention communicated by the author is the firing of the nurse. This in itself holds negative connotations, but through the use of the word 'clumsy', the negative sentiment of the statement becomes clearer to understand.

The inclusion of expressive aspects in the language of the persuasive discourse function, enables us to identify the sentiment of a persuasive comment. As there is this cross-over in the language of the two discourse functions, we can hypothesise that

Proceedings of the 3rd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis, pages 79–83,
Jeju, Republic of Korea, 12 July 2012. ©2012 Association for Computational Linguistics

if we train a supervised classifier on an expressive corpus, a learned model will be created that when applied to a corpus of persuasive documents, will classify these texts to an adequate standard.

As the corpus that we developed is in the clinical domain, it is worth noting the important role that sentiment analysis can play for health practitioners, which unfortunately has not received a great deal of attention. In assessing the effectiveness of treatments given by the health service for a condition which is curable, the results themselves indicate the effectiveness of such a process. However, for palliative treatments which merely alleviate the symptoms of an illness or relieve pain, it is vital to discover the extent to which these are effective. Feedback has progressed from the filling in of paper forms to the ability to give feedback through web pages and mobile phones. Text is stored in a highly accessible way, and is now able to be efficiently processed by sentiment classification algorithms to determine the opinions that patients are expressing. This in turn should enable health services to make informed decisions about the palliative care which they provide.

2 Patient Feedback Corpus

NHS Choices[1] is a website run by the National Health Service (NHS), which acts as an extensive knowledge base for any health-related queries. This website not only provides comprehensive articles about various ailments, but also gives the users of the site the option to rate and comment on the services that are provided to them at hospitals and GP surgeries. This user feedback provides an excellent basis for the sentiment classification experiments of this work.

The reviews that are submitted are typically provided by a patient or close relative who has experienced the healthcare system within a hospital. When submitting feedback, the user is asked to split their feedback into various fields, as opposed to submitting a single documents detailing all the comments of the user. During corpus compilation, each comment was extracted verbatim, so spelling mistakes remain in the developed corpus. All punctuation also remains in order to enable future experiments to be carried out on either the sentence or phrase level within each comment.

Corpus	D	W	$D_{avglength}$	V
Expressive				
Positive	1152	75052	65.15	6107
Negative	1108	76062	68.65	6791
Persuasive				
Positive	768	46642	60.73	4679
Negative	864	113632	131.52	7943

Table 1: Persuasive & expressive corpus statistics.

In developing the corpus, we leverage the fact that the data was separated into subfields, as opposed to one long review, where the all data is merged into a single document. We extracted comments which came under three categories in the NHS Patient Feedback dataset: *Likes*, *Dislikes* and *Advice*. The *Likes* were assumed to express positive sentiment and highlight elements of the health service that patients appreciated. Conversely, the documents given under the *Dislikes* header were assumed to convey a negative sentiment. These two subsets make up the *Expressive* subset of the compiled corpus. The *Advice* documents did not have an initial sentiment associated with them, so each comment was labelled by two independent annotators at the document level as being either a positive or negative comment. These *Advice* comments contributed to the *Persuasive* subcorpus. In compiling the persuasive document sets, we automatically discarded those comments that contained the term *"N/A "* or any of its derivative forms.

3 Method

The aim in this work was to examine the effect of training a supervised classifier on a corpus whose discourse function differs to that of the training set. We experimented with three standard supervised machine learning algorithms: standard Naïve Bayes (NB), multinomial Naïve Bayes (MN NB) and Support Vector Machines (SVM) classification. Each has proven to be effective in previous sentiment analysis studies (Pang et al. , 2002), so as this experiment is rooted in sentiment classification, these methods were also assumed to perform well in this cross-discourse setting.

For the cross-discourse sentiment classification

[1] http://nhs.uk

experiments, two variants of the Naïve Bayes algorithm are used. The difference between the standard NB and MN NB is the way in which the features for classification, the words, are modelled. In the standard NB learning method, a binary presence approached is taken in modelling the words of the training documents. This differs to the MN NB classifier, which takes into account term frequency when modelling the documents. Each has proven to be a high performing classifier across various sentiment analysis domains, but no distinction has been given as to which is the preferable method to use. Therefore in this paper, both were implemented.

In the literature, results from the use of SVMs in classification based experiments have outperformed other algorithms (Joachims, 1998; Pang et al. , 2002). For these cross-discourse experiments we use the Sequential Minimal Optimization training algorithm (Platt, 1998), in order to achieve the maximal hyperplane, and maximise the potential of the created classifier. Traditionally SVMs have performed well in text classification, but across discourse domains the results of such classification has not been examined.

Each document in the corpus was modelled as a bag of words. Features used within this representation were unigrams, bigrams and bigrams augmented with part-of-speech information. Due to this, and observing the results of preliminary experimentation that included rare features, it was decided to remove any feature that did not occur more than 5 times throughout the training set. A stopword list and stemmer were also used.

Each supervised classification technique was then trained using a random sample of 1,100 documents from both the positive and negative subsections of the expressive corpus. Following this we tested the classifiers on a set of 1,500 randomly selected persuasive documents, using 750 documents from each of the positive and negative subcorpora.

The results of cross-validation (Table 2) suggested that unigram features may outperform both bigram and part-of-speech augmented bigrams for all learning methods. In particular, the accuracy results produced by the NB algorithm surpassed the results of other classifiers in the tenfold cross-validation. This suggests that within a single discourse domain, presence based features are preferable to considering the frequency of a term when generating a machine learning model.

Features	NB	Multinomial NB	SVM
Unigrams	**79.65**	**78.14**	**76.11**
Bigrams	57.79	60.84	63.36
Bigrams + POS	74.25	75.71	72.83

Table 2: Average tenfold cross-validation accuracies on only the expressive corpus. Boldface: best performance for a given classifier.

4 Results

Table 3 shows the classification accuracies achieved in all experiments. For each classifier, with each feature set, if we take the most basic baseline for the two-class (positive/negative) problem to be the random baseline of 50% classification accuracy, then this is clearly exceeded. However if we take the results of the tenfold cross-validation as a baseline for each classifier in the experiments, then only the results given by the MN NB classifier with unigram and bigram features are able to surpass this.

The results given from the NB and the MN NB classifier imply that using frequency based features are preferable to using presence based features when performing cross-discourse sentiment classification. The MN NB is one of the few classifiers tested that exceeds the results of the cross-validated model. These results support experiments carried out for topic based classification using Bayesian classifiers by McCallum and Nigam (1998), but differs from sentiment classification results from Pang et al. (2002) that suggest that term-based models perform better than the frequency-based alternative. This also differs to the results that were returned during the cross-validation of the classifiers, where presence based features produced the greatest classification accuracy.

In our tests, the feature set which yielded the highest degree of classification accuracy across all classifiers is the unigram bag of words model. Tan et al. (2002) suggest that using bigrams enhances text classification, but as sentiment classification goes beyond this task, the assumption does not hold, as the results here show. The difference in discourse function could also contribute to bigrams yielding

		Positive			Negative		
	Accuracy	Precision	Recall	F_1	Precision	Recall	F_1
NB Uni	76.07	78.29	72.13	75.09	74.17	80.00	76.97
NB Bi	58.93	55.19	**94.93**	69.80	81.90	22.93	35.83
NB Bi + POS	65.00	71.84	49.33	58.50	61.42	80.67	69.74
MN NB Uni	**83.53**	**82.04**	85.87	**83.91**	**85.17**	81.20	**83.14**
MN NB Bi	57.00	63.78	32.40	42.97	54.69	**81.60**	65.49
MN NB Bi + POS	69.97	69.59	69.87	69.73	69.75	69.47	69.61
SVM Uni	69.00	68.43	70.53	69.47	69.60	67.47	68.52
SVM Bi	55.40	60.98	30.00	40.21	53.58	80.80	64.43
SVM Bi + POS	63.27	63.11	63.87	63.49	63.43	62.67	63.04

Table 3: Results of experimentation, with the expressive corpus as the training set, and the persuasive corpus as the test set. Boldface indicates the best performance for each metric.

the lowest accuracy results. Bigrams model quite specific language patterns, but as the expressive and persuasive language differs in structure and content, then the patterns learnt in one domain do not accurately map to another domain. Bigrams contribute the least to sentiment classification in this cross-discourse scenario, and only when they are augmented with part of speech information does the accuracy sufficiently pass the random baseline. However for good recall, using bigram based features produces excellent results, at the sacrifice of adequate precision, which suggests that bigram models overfit when they are used as features in such a learned model.

The SVM classifier with a variety of features does not perform as well as the multinomial Naïve Bayes classifier. Joachims (1998) suggests that for text categorization, the SVM algorithm regularly outperforms other classifiers, but unfortunately the outcome of our experiments do not correlate with these results. This suggests that SVMs struggle with text classification when the discourse function between the training and test domains differ.

5 Discussion

The results produced through training supervised machine learning methods on an expressive corpus, and testing on a corpus which contains documents with a persuasive discourse function indicate that cross-discourse sentiment classification is feasible.

The best performance occurred when the classifier took frequency based features into account, as opposed to solely presence based features. The reasoning for this could be attributed to the way that patients were asked to submit their feedback. Instead of asking a patient to submit a single comment on their experience with the health service, they were asked to submit three distinct comments on what they liked, disliked and any advice that they had. This gave the user the opportunity to separate their sentiments, and clearly communicate their thoughts.

It is of interest to note that the cross-discourse accuracy should surpass the cross-validation accuracy on the training set. This was not to be expected, due to the differences in discourse function, and therefore features used. However, where just the presence of a particular word may have made the difference in a single domain, across domains, taking into account the frequency of a word in the learned model is effective in correctly classifying a comment by its sentiment. Unigram features outperform both the bigram and bigrams augmented with part-of-speech features in our experiments. By using single tokens as features, each word is taken out of the context that its neighbours provide. In doing so the language contributing to the relative sentiment is generalised enough to form a robust model which can then be applied across discourse domains.

6 Related Work

A number of studies (Cambria at al. , 2011; Xia et al. , 2009) have used patient feedback as the domain for their sentiment classification experiments. However our work differs to these studies as we consider the effect that cross-discourse evaluation has on the classification outcome. Other work that has considered different discourse functions in sentiment analysis, have experimented on detecting arguments (Somasundaran et al. , 2007) and the stance of political debates (Thomas et al. , 2006).

Machine learning approaches to text classification have typically performed well when using a Support Vector Machine (Joachims, 1998) classifier or a Naïve Bayes (McCallum and Nigam, 1998) based classifier. Pang et al. (2002) applied these classifiers to the movie review domain, which produced good results. However the difference in domain, and singularity of discourse function differentiates the scope of this work from theirs.

7 Conclusion & Future Work

In this study we focused on the cross-discourse development of supervised machine learning algorithms in the clinical domain, that trained and tested across the expressive and persuasive discourse functions. We demonstrated that despite the differences in function of a corpus of patient feedback, the greatest classification accuracy was achieved when considering word frequency in the features of the learned model.

This study centred on the expressive and persuasive discourse functions, but it would be interesting to examine other such functions that convey a sentiment, such as argumentation. Another interesting avenue of investigation for this work would be to explore the lexical semantics of the different discourse functions, that could be used in sentiment classification, and factor this into the evaluation of the overall sentiment of persuasive documents within a corpus.

References

Douglas Biber. 1988. *Variation Across Speech and Writing.* Cambridge University Press.

John Blitzer, Mark Dredze 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*, pp. 440–447.

Erik Cambria, Amir Hussain and Chris Eckl. 2011. Bridging the Gap between Structured and Unstructured Health-Care Data through Semantics and Sentics. In *Proceedings of ACM WebSci*, Koblenz.

Andrea Esuli and Fabrizio Sebastiani. 2006. SentiWordNet: A Publicly Available Lexical Resource for Opinion Mining In *Proceedings of Language Resources and Evaluation (LREC)*, pp 417–422.

Thorsten Joachims. 1998. Text categorization with support vector machines: learning with many relevant features. In *Proceedings of ECML-98,* 10th *European Conference on Machine Learning*, pp. 137–142.

Andrew McCallum and Kamal Nigam. 1998. A Comparison of Event Models for Naive Bayes Text Classification. In *Proceedings of the AAAI/ICML-98 Workshop on Learning for Text Categorization*, pp. 41–48.

Bo Pang, Lillian Lee and Shivakumar Vaithyanathan. 2002. Thumbs up? Sentiment Classification using Machine Learning Techniques. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 79–87.

John Platt. Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines. In *Advances in Kernel Methods - Support Vector Learning*.

Swapna Somasudaran and Josef Ruppenhofer and Janyce Wiebe. 2007. Detecting Arguing and Sentiment in Meetings. In *Proceedings of the SIGdial Workshop on Discourse and Dialogue*, pp.26–34.

Chade-Meng Tan, Yuan-Fang Wang and Chan-Do Lee. 2002. The use of bigrams to enhance text categorization. In *Information Processing & Management*, 38(4) pp. 529–546.

Matt Thomas, Bo Pang and Lillian Lee. 2006. Get out the vote: Determining support or opposition from Congressional floor-debate transcripts. In *Proceeding of the 2006 Conference on Emperical Methods in Natural Language Processing (EMNLP)*, pp.327–335.

Lei Xia, Anna Lisa Gentile, James Munro and José Iria. 2009. Improving Patient Opinion Mining through Multi-step Classification. In *Proceedings of the 12th International Conference on Text, Speech and Dialogue (TSD'09)*, pp. 70–76.