Sentiment Classification via a Response Recalibration Framework

Phillip Smith & Mark Lee
School of Computer Science,
University of Birmingham,
Birmingham,
UK

Abstract

Probabilistic learning models have the ability to be calibrated to improve the performance of tasks such as sentiment classification. In this paper, we introduce a framework for sentiment classification that enables classifier recalibration given the presence of related, context-bearing documents. We investigate the use of probabilistic thresholding and document similarity based recalibration methods to yield classifier improvements. We demonstrate the performance of our proposed recalibration methods on a dataset of online clinical reviews from the patient feedback domain that have adjoining management responses that yield sentiment bearing information. Experimental results show the proposed recalibration methods outperform uncalibrated supervised machine learning models trained for sentiment analysis, and yield significant improvements over a robust baseline.

1 Introduction

Probabilistic classifiers are a typically used method for the classification of documents by the sentiment they convey (Pang et al., 2002). Given an unlabelled document, a trained probabilistic model is able to determine an appropriate labelling in relation to a given confidence for the proposed labelling. In the two-class sentiment classification problem a labelling confidence that is greater than 0.5 will lead to a particular sentiment being attached to the input document. However, it is questionable whether a classifier confidence output of 0.51 is sufficiently suitable for the application of any given label. This low confidence poses a problem for sentiment classification, and can often lead to documents being labelled incorrectly, to the detriment of a sentiment analysis system.

A low classifier confidence in sentiment analysis may be produced due to inherent linguistic difficulties that plague systems developed for natural language processing. For example, documents where a sentiment is conveyed implicitly, ambiguously, or in a sarcastic manner can cause problems for machine learning approaches to sentiment classification. Methods have been proposed to deal with such facets of language in sentiment analysis. These tend to focus on hand-crafted lexicons (Balahur et al., 2011) or intra-document contextual cues (Greene and Resnik, 2009) to disambiguate the polarity of a document. We propose a method that takes into consideration related documents in the classification process, and duly adjusts classification output using a *sentiment recalibration framework*.

Our proposed method takes into account external but relevant documents during the sentiment recalibration process. We use these documents to make adjustments to classifier outputs, in an adjustment and correction phase. To our knowledge, this is the first work in sentiment classification to attempt the recalibration of a sentiment classifier given relevant documents. We attribute this ability to the dataset used for our experiments.

The remainder of the paper is structured as follows: in Section 2 we describe the data and annotation experiments devised to observe the relevance of a response to a comment. Section 3 then outlines the motivation for sentiment recalibration, and section 4 details our proposed methodology for the construction of a calibration framework for sentiment classification. Section 5 gives the baseline for evaluation. Section 6 details the results of experimentation with the framework, and discusses the implications. Section 7 describes related work and we conclude and give direction for future work in Section 8.

2 Data and Annotation

The monologic nature of current datasets for evaluating sentiment classifiers, while valuable to the development of the field, are not applicable to our proposed recalibration framework. Most relevant to our work is the forum post data set (Murakami and Raymond, 2010). However, this is too general for the purposes we are examining due to deviation in discourse topic. Therefore we have developed a dataset for sentiment classification with the related documents that are required for the response recalibration framework. We use patient feedback data provided by the National Health Service (NHS). This has been used before (Smith and Lee, 2014), however author responses were not a feature examined in this work. In this dataset, each feedback item consists of a patient's comment and a response from the NHS. Unlike other online reviews used to investigate the potency of sentiment classification algorithms, this dataset does not contain a user ranking or score to accompany their comment. An annotation phase is therefore required in order to use the documents as an evaluation dataset for our algorithms.

We annotate a subset of 4,059 comments and their related responses for their sentiments expressed and responded to, at the document level. The comments contained 254,611, of which 10,325 were unique. Responses contained 403,315 words, of which 9,115 were unique. Despite a larger average document size, the response vocabulary was smaller that the comment vocabulary. This indicates that the responses given were constrained in nature. An initial pass of the data highlighted that reviews were not merely binary, but often weighed up mixed sentiments before giving a conclusion. Due to this observation, we initially annotate the data with a five-class annotation scheme. This includes, neutral, mixed-positive and mixed-negative categories. The mixed categories denote that varying sentiments are present in the document, but one sentiment is more salient than the other.

Results of this annotation are presented in table 1. Given the annotations, we calculate inter-category agreement using Cohen's kappa coefficient. Between all categories $\kappa = 0.4294$, and observing positive and negative only $\kappa = 0.761$, a good level of agreement. This agreement is indicative of the level to which the sentiment expressed in a comment is mirrored and acknowledged in a related response. Due to this result, we proceed with our experiments using only the positive and negative data.

$S_{Comment}$ \ $S_{Response}$	-2	-1	0	+1	+2
-2	3	139	6	5	12
-1	8	2,022	92	33	117
0	0	153	25	102	44
+1	4	251	83	671	187
+2	1	68	1	15	17

Table 1: Comment-response sentiment label confusion matrix.

3 Sentiment Classification Recalibration

This work examines the potential for the outcome of sentiment classifiers to be recalibrated given the presence of a related document. In this work, the response to a sentiment-bearing comment is the related document under consideration. Typical approaches to recalibration may rely on the Platt scaling or binning methods. Platt scaling trains a logistic regression model on the output of an SVM classifier, enabling the production of posterior classification probabilities (Platt and others, 1999). Binning is another calibration method that is particularly effective for classification (Zadrozny and Elkan, 2001). Such recalibration methods focus on statistical methods of recalibrating classifier output. However, when dealing with related natural language documents, we can use inferences from the content of the related text to guide the recalibration process. We therefore propose the use of the response to recalibrate the labelling of the initial comment. This takes a response directed at a comment, and uses the outcome of its classification as a starting point for recalibration. We discuss in further detail the recalibration protocols in the following section.

4 Method

The notion of applying supervised learning methods to classify related documents (Taskar et al., 2001; Jensen et al., 2004) and the post-processing of classification (Benferhat et al., 2014) has been examined in the literature. Based on this, we propose three approaches to leverage the acknowledged sentiment in a response in the recalibration process. The first two methods observe the outcomes of probabilistic classifiers appointed with

the role of classifying comment and response sentiment individually. A probabilistic classifier will output a confidence associated with a particular label. Given a confidence greater than .5 a label will be applied. However, a confidence of .51 is probably of no more use than a random guess in the two-class sentiment classification task. Given a complete iteration of the confidence thresholds at .01 confidence intervals, we examine the classifier performance. At each interval, given each instance, where the response label differs to the comment's we assign this label to the review. The third method judges the level of lexical similarity between the comment and response before making a judgement regarding recalibration.

4.1 Probabilistic Threshold Calibration

In classification, the probability of labelling a document with a certain category is just as important as the labelling itself. A classifier may not be overly confident with its initial labelling, and so an external but relevant source of information may help guide and recalibrate the outcome of the initial calibration. Recalibration methods attempt to determine at what threshold the labelling would be most effective. These are typically guided by a line of best fit related to a posterior probability (Zadrozny and Elkan, 2001). We propose a recalibration framework to examine the effects of recalibrating the threshold. The first approach is to iterate over the probabilities at intervals of 0.01, and recalibrate the labelling if the confidence of given labelling is below the threshold. The recalibration is given through the labelling of the response relative to the original comment instance. In the protocol experiments, the confidence of the response classifier is not observed, only the labelling. The label is then commuted to the comment.

The second set of experiments observing probability thresholds imposes the constraint that the response classifier must yield a more confident classification outcome than that of the comment classifier. Both may exhibit the same sentiment, but in order to overcome any confusion due to ambiguous or implicit expressions we commute the response labelling if and only if the confidence output by the response classifier is higher than that of the comment classifier.

4.2 Document Similarity

Classifier confidence is just one potential method of determining cases for instance relabelling where sentiment classifiers may yield incorrect classifications. Another method that deserves consideration is determining the level of lexical similarity between the comment and response. The assumption is made that is that if a response is replying to the content of the original comment, there will be elements of language reuse in the response. Then, the greater the similarity, the more likely the relative document sentiments are homogeneous. We implement the Greedy String Tiling algorithm (Wise, 1993) as a measure of document similarity. The algorithm outputs a score between [0,1], given the level of similarity. As with the previous experiments with relabelling given a classifier confidence, we take the same approach here. Our experiment iterates over varying thresholds, with a 0.01 interval at each step. However, we do not make any adjustments for classifier confidence, only taking the binary labelling as the primary label.

4.3 Baseline

Comments in our dataset receive responses that both acknowledge and concisely respond to the content of the original message. We identify features of these responses that are useful to the sentiment classification process. We employ a rule-based system based on these observations to test the hypothesis that given the presence of these features, the sentiment of the response mirrors that of the original comment. Using a small set of regular expressions for frequent word stems we achieve a recall of .9004. Given the categorisation of the terms we then classify the sentiment of the response and compare the labelling to the gold-standard labelling of the comment. This yields an accuracy of 0.6634. We also cross-validate the three classifiers on the dataset to form another baseline, results of which are shown in Table 2.

	Acc.	Prec.	Recall	F_1
Comment				
NB +1	0. 692	0.502	0.765	0.606
NB -1		0.862	0.659	0.747
MNB +1	0.871	0.784	0.805	0.794
MNB -1		0.911	0.9	0.906
SMO +1	0.856	0.771	0.759	0.765
SMO -1		0.893	0.899	0.896

Table 2: Baseline for sentiment classification (+1 = positive -1 = negative)

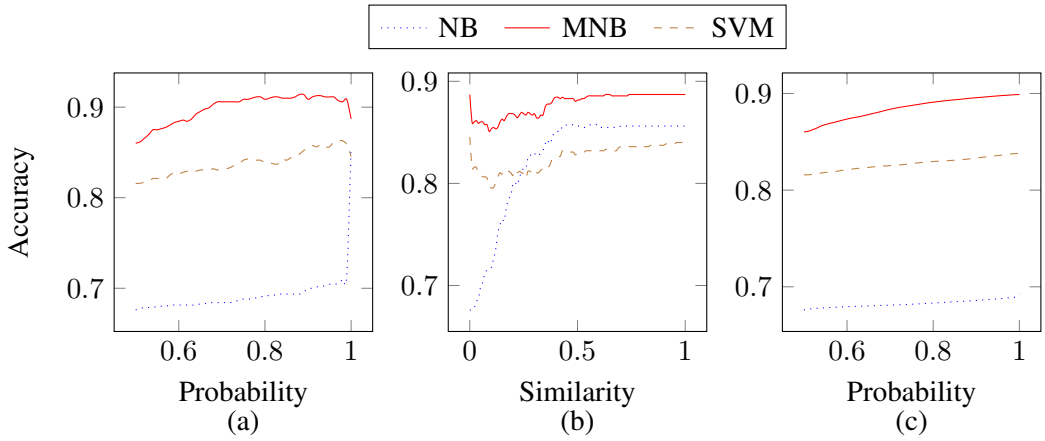

Figure 1: Accuracy comparison graphs for our recalibration methods: (a) confidence threshold (b) similarity threshold (c) confidence threshold where $Pr(R) > Pr(A)$.

5 Results and discussion

Results of the probabilistic relabelling experiment highlight an improvement in classifier accuracy as the probability threshold increases, contrary to expectation. We must look to both Figures 1 and 2 in order to understand this result. One would expect that a classifier outputting a low confidence of classification would yield a higher accuracy given the recalibration process. Results highlight the relative over-confidence of all classifiers when predicting the class of an instance as the number of candidates eligible for relabelling is relatively low. In particular, if we observe the candidates returned for the NB this classifier does not exhibit a great variance in confidence, with the majority of labellings being ≥ 0.99. The hubristic nature of the NB labelling confidence is not beneficial where results are unable to be recalibrated. Given the total relabelling scenario for the NB classifier, whereby all labels from the responses are commuted to annotate the comment, there is a significant increase in classification accuracy of 0.15. In the case of the other two classifiers, such a scenario leads to a decrease in performance. This indicates the poor quality of model initially produced by the NB learner. This also shows the relative strength in model building qualities of the MNB and SVM learners.

The SVM outperforms the NB, but falls short of MNB performance. Figure 2 indicates that potentially poor relabelling choices contribute to this. The success ratio drops dramatically as SVM confidence tends towards 1. This trait is similarly present in both the NB and MNB, also. This is

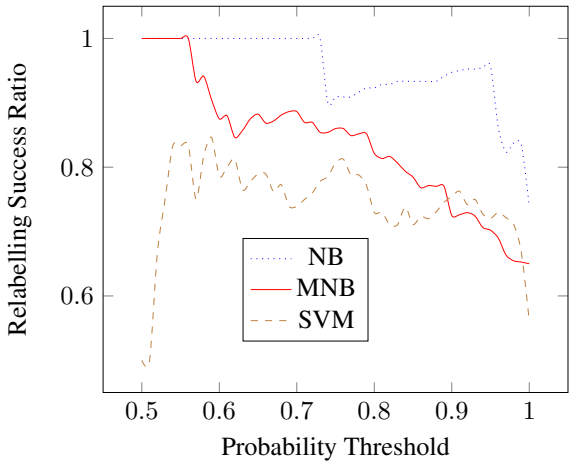

Figure 2: Relabelling success rate given varying classifier confidence thresholds.

to be expected however, and suggests that a highly confident initial judgement by the comment sentiment classifier should not be altered.

Contrasting the probabilistic thresholding results with the similarity threshold relabelling experiments we notice that classifier confidence is a substantially better calibration method. Given the similarity thresholding experiments, decreases in classification accuracy are shown for MNB and SVM. However for NB, significant gains in classifier accuracy are made. This is in contention with the results of the probabilistic thresholding experiments, where gains for NB classification were minimal. We can attribute this to the relaxed string matching method of the implemented similarity measure.

Precision and recall results (although not shown due to space constraints) highlight the dominance

of the MNB for in the recalibration framework. Recall increases over all iterations ≤ 0.99, which then yields a drop when full relabelling is applied. Precision peaks at a threshold of 0.78 and decreases following this. The SVM follows suit, however performance degrades almost 0.06 when full relabelling is applied. Precision sees a drop for the negative class, although the NB exhibits reconciling traits. The SVM suffers from a drop in precision of 0.05. The positive class however shows signs of strengthening as the similarity threshold increases. The MNB remains the dominant classifier throughout precision comparison. The recall follows an inverse pattern. As similarity threshold increases for in our negative class experiments, recall gradually increases for MNB and SVM, however makes gains of 0.22 for the NB. The positive recall drops however for all classifiers.

The strong response classification experiments impose a constraint that labelling of the comment can only be commuted from the response classification if and only if the response classifier confidence is higher than that of the comment classification of a given instance. The constraint appears to have a stabilising quality. Comparing graphs (a) and (c) in Figure 1, we see a substantially smoother gradient to the curve of the strong response classification curves in comparison to the general threshold commutation experiments. We do not see the drops in performance for the MNB and SVM classifiers, much to the benefit of the overall classification, but similarly, we do not see the steep climb in classifier accuracy demonstrated by the NB. The precision and recall rates achieved by strong response classification only mimics that of the general probabilistic threshold experiments. Closer comparison of the two shows marginal differences.

Results indicate that there is no requirement for the confidence of the response classifier output to be higher than that yielded for the corresponding instance classified by the model trained on the comment data. Comparing Figure 1 with the comment baseline given in Table 2, accuracy results from the classifiers in the experiments marginally succeed the baseline for the MNB classifier, but for the SVM and NB, accuracy is detrimentally effected. We can conclude that in this case of response relabelling the constraint is too strict.

6 Related Work

Work has observed the useful nature of relationships between documents hen classifying stocks based on the contents of related posts on social networks (Si et al., 2014) and classifying sentiment in posts on online forums based on user relationships, or user stances in online debates (Murakami and Raymond, 2010).

Work on bagging in sentiment classification is somewhat related to our work (Dai et al., 2011; Nguyen et al., 2013). Bagging trains a number of models on a similar set of training data. During classification, each model then classifies the given instance, and a voting protocol labels the instance with the majority label suggested. Our framework however does not train multiple classifiers, although the framework could be extended to incorporate this. Instead, a related document is used to guide and recalibrate the outcome of the initial classification. Our method does not suffer from the issue of low classifier trustworthiness, as we have shown the results of response only classification to be reliable in our baselines. The need for further methods such as stacking is therefore eliminated.

The use of management response in online reviews has been examined to empirically determine the effectiveness in improving a firm's reputation (Proserpio and Zervas, 2014). Analysis has shown moderate improvements where a management response was given. This work did not computationally evaluate the content of reviews, however.

7 Conclusion

We have examined the role of sentiment recalibration in the domain of patient feedback. The proposed classification recalibration method considered acknowledged sentiment in a comment response in order to recalibrate classifier output. Our framework examined three methods for recalibration, two probabilistic and one similarity based. We found that all classifiers exhibited improvements in classification performance when subject to recalibration over varying probability thresholds. Results suggest that the MNB classifier is most suited to the recalibration methods, and yields the best performance, with a 4.2% increase in classification accuracy over our baseline. Our proposed method is suitable where a dataset contains a number of related documents. As the wealth of data for sentiment classification

increases, we would like to examine and evaluate our method on additional datasets.

References

Alexandra Balahur, M. Jesús Hermida, and Andrès Montoyo. 2011. Detecting implicit expressions of sentiment in text based on commonsense knowledge. In *Proceedings of the Second Workshop on Computational Approaches to Subjectivity and Sentiment Analysis*, pages 53–60. ACL.

Salem Benferhat, Karim Tabia, Mouaad Kezih, and Mahmoud Taibi. 2014. Post-processing a classifier's predictions: Strategies and empirical evaluation. In *ECAI*, pages 965–966.

Lin Dai, Hechun Chen, and Xuemei Li. 2011. Improving sentiment classification using feature highlighting and feature bagging. In *ICDMW*, pages 61–66. IEEE.

Stephan Greene and Philip Resnik. 2009. More than words: Syntactic packaging and implicit sentiment. In *ACL-HLT*, pages 503–511.

David Jensen, Jennifer Neville, and Brian Gallagher. 2004. Why collective inference improves relational classification. In *ICKDDM*, pages 593–598.

Akiko Murakami and Rudy Raymond. 2010. Support or oppose?: Classifying positions in online debates from reply activities and opinion expressions. In *COLING*, pages 869–875.

Quoc Dai Nguyen, Quoc Dat Nguyen, and Bao Son Pham. 2013. A two-stage classifier for sentiment analysis. In *IJCNLP*, pages 897–901.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? sentiment classification using machine learning techniques. In *EMNLP*.

John Platt et al. 1999. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74.

Davide Proserpio and Georgios Zervas. 2014. Online reputation management: Estimating the impact of management responses on consumer reviews. Technical Report 2521190, Boston U. School of Management.

Jianfeng Si, Arjun Mukherjee, Bing Liu, Jialin Sinno Pan, Qing Li, and Huayi Li. 2014. Exploiting social relations and sentiment for stock prediction. In *EMNLP*, pages 1139–1145.

Phillip Smith and Mark Lee. 2014. Acknowledging discourse function for sentiment analysis. In *CICLING*, volume 8404 of *LNCS*, pages 45–52.

Benjamin Taskar, Eran Segal, and Daphne Koller. 2001. Probabilistic classification and clustering in relational data. In *IJCAI*, volume 17, pages 870–878.

Michael J Wise. 1993. String similarity via greedy string tiling and running karp-rabin matching. *Online Preprint, Dec*, 119.

Bianca Zadrozny and Charles Elkan. 2001. Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. In *ICML*, volume 1, pages 609–616.