Title	Sentiment analysis of online media
Authors(s)	Salter-Townshend, Michael, Murphy, Thomas Brendan
Publication date	2012
Publication information	Salter-Townshend, Michael, and Thomas Brendan Murphy. “Sentiment Analysis of Online Media,” 2012.
Conference details	Paper presented at the DAGM-GfKl/IFCS 2011, Joint Conference of the German Classification Society (GfKl) and the German Association for Pattern Recognition (DAGM), August 31 to September 2, 2011 and at the IFCS 2011 Symposium of the International Federation of Classification Societies (IFCS), August 30, 2011, Frankfurt am Main, Germany
Item record/more information	http://hdl.handle.net/10197/3574

Sentiment Analysis of Online Media

Michael Salter-Townshend and Thomas Brendan Murphy

Abstract A joint model for annotation bias and document classification is presented in the context of media sentiment analysis. We consider an Irish online media data set comprising online news articles with user annotations of negative, positive or irrelevant impact on the Irish economy. The joint model combines a statistical model for user annotation bias and a Naive Bayes model for the document terms. An EM algorithm is used to estimate the annotation bias model, the unobserved biases in the user annotations, the classifier parameters and the sentiment of the articles. The joint modeling of both the user biases and the classifier is demonstrated to be superior to estimation of the bias followed by the estimation of the classifier parameters.

1 Introduction

Sentiment analysis involves extracting contextual information from documents [6]. Media sentiment has been shown to be of importance in economic contexts [10]. We examine a corpus of Irish news articles that have been annotated by a number of inexpert volunteers as having a sentiment which has positive, negative or irrelevant impact on the Irish economy. The aim of the analysis is to develop a classification method that can estimate the correct labelling of the articles in the corpus as well as the correct classification for other news articles. A core goal is to increase the accuracy of both the annotation based labelling and the classifier. Whilst the methods outlined herein are developed in the context of the media sentiment, they are readily applicable in any context where a classifier is trained on (potentially) biased and noisy annotations.

The media sentiment analysis involves a classification task where the sample labels are noisy and biased user annotations. Many existing classifiers do not take into account user (annotator) bias in reporting and a simple majority vote is used to select the true article type from the observed annotations; this majority vote labelling is particularly problematic in the presence of user bias. Some previous work has been proposed to help address the annotator bias issue. [9] applied the method of [3] to correct for annotator bias and estimate the true labeling before developing a classifier in an object recognition problem. Most recently [8] and [7] propose methods to address the problem of multiple imperfect annotations and classification. [8]

Michael Salter-Townshend and Thomas Brendan Murphy
School of Mathematical Sciences & Complex and Adaptive Systems Laboratory, University College Dublin, Dublin 4, Ireland.
{michael.salter-townshend,brendan.murphy}@ucd.ie

deals with the labelling of clinical reports and uses Bayesian models with Gaussian processes for classification and ordinal regression. [7] address the problem of training a classifier with multiple imperfect annotations by extending the model of [3] to learn a classifier at the same time as the annotator biases via maximum likelihood; this work is similar to the approach developed herein. Specifically, they train a logistic regression classifier and learn the sensitivity and specificity of the annotators in the context of binary labelling. The model that we present differs from that paper in that we explore a trinary labelling system (an arbitrary finite number of categories is possible) and train a Naive Bayes classifier. The contribution of our work is to demonstrate the method with another classifier, a greater number of potential labels and to report upon the comparative effectiveness of our approach in the context of the Irish online media sentiment analysis.

We validate our approach on a simulated dataset and calculate performance scores for both the decoupled estimator (learn the biases and then train the classifier) and the joint estimator model. We demonstrate the superiority of the joint estimator for various levels of bias and then apply it to the media dataset.

1.1 Sentiment Data

The Irish media dataset that we analyze is a subset of the data described in detail in [1, 2]. The dataset is comprised of 1024 articles collected from 3 online Irish news services (`rte.ie`, `irishtimes.com` and `independent.ie`), collected from July to October 2009. 31 volunteers have annotated an average of 834 of these articles as having either negative, positive or irrelevant impact on the Irish economy at time of press. There are 70873 word terms appearing in these articles. In order to reduce the impact of words that are too common (such as “at”, “the”, “and”, etc) we eliminate words that appear in more than 1000 articles. We also eliminate words that appeared in less than 30 articles. To further reduce the dimensionality of the data, we selected the top 300 most negative words (as indicated by a simple majority vote classifier), the top 300 most positive words and the top 300 most irrelevant words only.

[2] note that 45% of the articles do not have consensus annotations and that “there is some evidence that the learning process would be better off without them [articles with low consensus]”. The authors of that paper examined k-nearest neighbours (kNN) and support vector machine (SVM) classifiers also but settled on Naive Bayes following an assessment of the performance of the methods under cross-validation.

2 Model

Let $y_a^{(k)} = (y_{a1}^{(k)}, y_{a2}^{(k)}, \ldots, y_{aJ}^{(k)})$ be the annotation of article a by annotator k, where $y_{aj}^{(k)} = 1$ if article a is annotated as being of type j and $y_{aj}^{(k)} = 0$ otherwise. We model the annotator bias as per [3]. Error rates, or biases in reporting, are modelled via a matrix of conditional probabilities for each annotator, that is, the probability that annotator k records annotation j given a *true* (but unobserved) type i is denoted by $\pi_{ij}^{(k)}$. These probabilities sum to unity across j for each i and k. The observed annotations are thus a probabilistic (multinomial) function of these π matrices.

If we let the true type of article a be T_a, where $T_{ai} = 1$ if the article is of type i and $T_{ai} = 0$ otherwise. Then, the likelihood for the recorded annotations $y_a = (y_a^{(1)}, y_a^{(2)}, \ldots, y_a^{(K)})$ on article a given a true type T_a is given by

$$\mathscr{L}(\pi|y_a, T_a) \propto \prod_i^J \left\{ \prod_k^K \prod_j^J (\pi_{ij}^{(k)})^{y_{aj}^{(k)}} \right\}^{T_{ai}} \quad (1)$$

where J is three for our sentiment levels (negative, positive and irrelevant).

Hence, the complete-data likelihood of the full annotation dataset (including unobserved true types) across all A articles is

$$\mathscr{L}(\pi, p|y_1, y_2, \ldots, y_A, T_1, T_2, \ldots, T_A) \propto \prod_a^A \prod_i^J \left\{ p_i \prod_k^K \prod_j^J (\pi_{ij}^{(k)})^{y_{aj}^{(k)}} \right\}^{T_{ai}} \quad (2)$$

where p_i is the marginal probability of type i.

Another goal of the sentiment analysis described in [2] is to train a classifier to distinguish which word terms appear in which types of article. The trained classifier may then be used to automatically label un-annotated articles. Although word-term frequencies are available in the dataset, we model only the presence or absence of these features (word terms). Let $w_a = (w_{a1}, w_{a2}, \ldots, w_{aN})$ be a binary vector that indicates the presence and absence of words in document a. We employ a Bernoulli likelihood for term w_a given that the article is of type i, that is $T_{ai} = 1$. That is, we use a Naive Bayes classifier to learn the probability of an article type given the words that appear in the article. Although the Naive Bayes assumption is unlikely to hold exactly in practice, there is much evidence to suggest that it can yield excellent classification results [4, 5].

The product of Bernoullis likelihood for all N word terms w_a appearing in article a given T_a is then

$$\mathscr{L}(\theta|w_a, T_a) = \prod_i^J \left\{ \prod_n^N (\theta_{ni})^{w_{an}} (1-\theta_{ni})^{1-w_{an}} \right\}^{T_{ai}} \quad (3)$$

where θ_{ni} is the probability that word term w_n appears in an article of type i.

The full likelihood for the data is then a product of Equation (2) and a term in the form of Equation (3) for each article, yielding Equation (4),

$$\mathscr{L}(\theta, p, \pi|w, y, T) = \prod_a^A \prod_i^J \left\{ p_i \prod_k^K \prod_j^J (\pi_{ij}^{(k)})^{y_{aj}^{(k)}} \prod_n^N (\theta_{ni})^{w_{an}} (1-\theta_{ni})^{1-w_{an}} \right\}^{T_{ai}} . \quad (4)$$

3 EM Algorithm

Since T, p and π are unknown in Equation (2), we proceed as per [3]. We then extend the EM algorithm to yield a joint estimation that learns θ within the same EM iterations as it learns the values of missing data T, the marginal probabilities p and annotator bias matrices π. The algorithm proceeds as follows:

1. **for all articles** a**:**
2. initialize T using $\hat{T}_{ai} = \mathbf{E}[T_{ai}] = \sum_k y_{ai}^{(k)}/K$
3. initialize p using $\hat{p}_i = \sum_a T_{ai}/A$
4. estimate all π values via maximum likelihood expression

$$\hat{\pi}_{ij}^{(k)} = \frac{\sum_a \hat{T}_{ai} y_{aj}^{(k)}}{\sum_j \sum_a \hat{T}_{ai} y_{aj}^{(k)}}. \tag{5}$$

5. estimate all θ and p via maximum likelihood expressions

$$\hat{\theta}_{ni} = \frac{\sum_a w_{an} \hat{T}_{ai}}{\sum_a \hat{T}_{ai}} \quad \text{and} \quad \hat{p}_i = \frac{\sum_a \hat{T}_{ai}}{A}. \tag{6}$$

6. re-estimate T using

$$\hat{T}_{ai} = \mathbf{E}[T_{ai}] = \frac{p_i \prod_k^K \prod_j^J (\hat{\pi}_{ij}^{(k)})^{y_{aj}^{(k)}} \prod_n^N (\hat{\theta}_{ni})^{w_{an}} (1-\hat{\theta}_{ni})^{1-w_{an}}}{\sum_{i'} p_{i'} \prod_k^K \prod_j^J (\hat{\pi}_{i'j}^{(k)})^{y_{aj}^{(k)}} \prod_n^N (\hat{\theta}_{ni'})^{w_{an}} (1-\hat{\theta}_{ni'})^{1-w_{an}}}. \tag{7}$$

7. **repeat 4 to 6 until convergence**

with convergence assumed once the change in log-likelihood fell below 10^{-4}.

In contrast, the decoupled estimator of the above method estimates the biases π, document types T and marginal probabilities p first, as in [3]. The Naive Bayes parameters θ are then fitted using the final estimates of the missing data values from the first stage; the decoupled estimation approach is similar to that taken by [9].

4 Results

4.1 Simulated Data

To test and compare the algorithm described in Section 3 with the decoupled estimator, we simulated data two hundred times. For each run, we use the marginal probabilities $p = (0.3, 0.3, 0.4)$ of each of the three types of "article". True types for A "articles" are simulated directly with these marginal probabilities. We then construct K conditional probability matrices $\pi^{(k)}$ of size 3×3, one for each "annotator". The value of $\pi_{ij}^{(k)}$ gives the probability that annotator k annotates an article of type i with label j. Finally, we also simulate observed word terms w for each article using the conditional probabilities of words occurring in each type of article as given in θ.

Two hundred such simulated data sets were analysed and for each data set the biases were randomly sampled uniformly over the range 0.1 to 0.5 and split evenly between the two wrong types with the balance allocated to the correct type. This was done identically for all simulated annotators which is equivalent to having a single random annotator performing multiple annotations and the number of these annotators was sampled uniformly between 2 and 6. The words were assigned a type according to p and the word-type probabilities θ were 0.1 to appear in an article of opposite type and 0.8 to appear in an article of the same type.

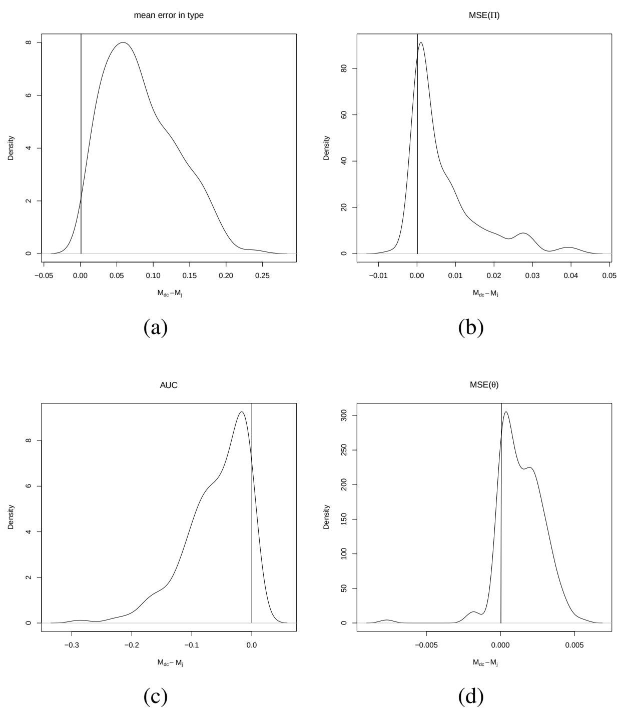


Fig. 1 Kernel density estimates of comparative performance measurements across multiple simulations. Two hundred runs of the simulated dataset analysis were performed and the difference in performance measure is computed for decoupled model (M_{dc}) and the joint model (M_j).1(a) shows the difference in mean error of type T. 1(b) shows the difference in mean squared error of bias π. 1(c) shows the AUC difference and 1(d) shows the mean squared error difference of word association θ.

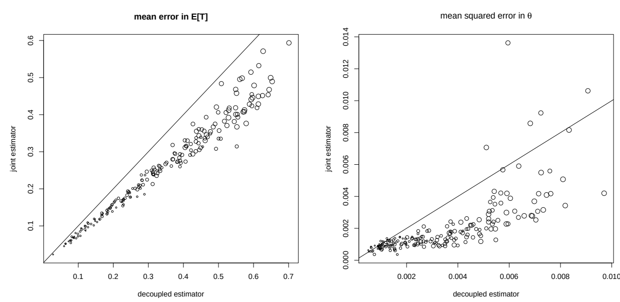
Both models are then evaluated on four performance metrics:

1. The mean error in expectation of type:

$$\sum_a (1 - \mathbf{E}[T_{ai}])/A \tag{8}$$

where the true value of article a is type i.
2. The mean squared error from the π matrix of bias probabilities.
3. The mean area under the ROC curve (AUC) for each of the 3 possible types.
4. The mean squared error from the θ matrix of word-type probabilities.

We subtracted the above four statistics under the joint estimation model M_j from the decoupled estimation model M_{dc} for repeated simulations. The mean paired difference between the above performance measures were $0.193, 0.009, -0.103$ and 0.009, respectively. All four were strongly statistically significantly different from zero under a t-test for the paired differences with p-values all less than 2.2×10^{-16}. Figure 1 shows kernel density estimates of these differences for the above statistics across the 200 simulation runs. Figure 2 indicates that the joint estimator's increase in performance is greater for higher biases. The size of the circles in the plot is proportional to the sampled bias and each circle represents a single run.



(a) (b)

Fig. 2 Comparison of performance across 200 iterations of simulated data. 2(a) shows the mean error in type T (as per Equation (8)) and 2(b) shows the mean squared error in word-to-type association θ. The size of the circles in the plot is proportional to the bias and each circle represents a single run. Lines with unit slope are added for reference.

4.2 Sentiment Data

We next present our results on the sentiment dataset. The interquartile range for the bias matrix off diagonal terms is $(0.110, 0.517)$, indicating a level of bias comparable to the simulated dataset. Table 1 shows the breakdown of classification with model for the media sentiment dataset. Figure 3 depicts tag clouds for word terms that have the strongest power for the negative, positive and irrelevant article types, under the joint estimation procedure. These tag clouds appear to show sensible word term associations to both positive and negative sentiment; for example, the names of the finance minister ("Brian", "Lenihan") and the new agency to deal with toxic debt ("NAMA") are included in the negative tag cloud and words like ("Germany",

Table 1 Cross-tabulation of article classification and model.

	Majority Vote	Decoupled Estimator	Joint Estimator
Negative	540	493	424
Positive	288	289	206
Irrelevant	196	242	394

"recovery") are included in the positive tag cloud. The tag cloud for the word terms for with the strongest predictive power for the irrelevant article types are given in Figure 3(c). Interestingly, most of the words in this tag cloud are non economic terms.

(a) (b) (c)

Fig. 3 Tag clouds for the top 100 word terms most strongly associated with 3(a) negative and 3(b) positive and 3(c) irrelevant articles. Most of the words appear to have an intuitively correct association with article type.

5 Discussion

We have demonstrated that the joint estimator makes use of the word term association with article type and thus outperforms the decoupled estimator for both bias estimation and classification. This boost in performance is related to the ratio of information in the features to the biases; if the annotators are all in agreement then the word term classifier will contribute little to the model. If there is bias in the annotations and the word terms are influenced by the article type then they will have a larger impact on the model and the joint estimation model will outperform the decoupled estimation model.

The joint estimator can achieve a target level of accuracy in article labelling using fewer biased annotators than the decoupled or majority vote labeling. This suggests that our method could be used to generate savings in the context of crowdsourcing with inexpert or otherwise biased annotators. There is a computational cost associated with the joint estimation; the time to perform 100 iterations for the decoupled and joint algorithms was approximately 16 and 50 seconds respectively. The joint algorithm does not seem to take more iterations to converge; for example, using the criterion that a change in log-likelihood of less than 10^{-3} required 38 and 36 iterations respectively. For a change of less than 10^{-2} they took 33 and 35.

The methodology outlined in the paper could be easily adapted to other model-based classifiers where samples are labeled using noisy annotations.

6 Acknowledgements

This work is supported by the Science Foundation Ireland under Grant No. 08/SRC/I1407: Clique: Graph & Network Analysis Cluster.

References

1. Brew, A., Greene, D., Cunningham, P.: The interaction between supervised learning and crowdsourcing. In: NIPS Workshop on Computational Social Science and the Wisdom of Crowds (2010)
2. Brew, A., Greene, D., Cunningham, P.: Using crowdsourcing and active learning to track sentiment in online media. In: H. Coelho, R. Studer, M. Wooldridge (eds.) ECAI 2010 - 19th European Conference on Artificial Intelligence, pp. 1–11. IOS Press (2010)
3. Dawid, A., Skene, A.: Maximum likelihood estimation of observer error-rates using the EM algorithm. Journal of the Royal Statistical Society. Series C (Applied Statistics) **28**(1), 20–28 (1979)
4. Domingos, P., Pazzani, M.: On the optimality of the simple Bayesian classifier under zero-one loss. Machine Learning **29**, 103–130 (1997)
5. Hand, D. J., Yu, K.: Idiot's Bayes — Not so stupid after all? International Statistical Review **69**(3), 385–398 (2001)
6. Pang, B., Lee, L.: Opinion mining and sentiment analysis Foundations and Trends in Information Retrieval **2**(1-2), 1–135 (2008)
7. Raykar, V., Yu, S., Zhao, L., Valadez, G., Florin, C., Bogoni, L., Moy, L.: Learning from crowds. Journal of Machine Learning Research **11**, 1297–1322 (2010)
8. Rogers, S., Girolami, M., Polajnar, T.: Semi-parametric analysis of multi-rater data. Statistics and Computing **20**, 317–334 (2010)
9. Smyth, P., Fayyad, U.M., Burl, M.C., Perona, P., Baldi, P.: Inferring ground truth from subjective labelling of venus images. In: G. Tesauro, D.S. Touretzky, T.K. Leen (eds.) Advances in Neural Information Processing Systems 7, pp. 1085–1092. MIT Press (1994)
10. Tetlock, P.C.: Giving content to investor sentiment: The role of media in the stock market. The Journal of Finance **62**(3), 1139–1168.