More than Bag-of-Words: Sentence-based Document Representation for Sentiment Analysis

Georgios Paltoglou
Faculty of Science and Technology
University of Wolverhampton
Wulfruna Street, WV1 1LY, UK
g.paltoglou@wlv.ac.uk

Mike Thelwall
Faculty of Science and Technology
University of Wolverhampton
Wulfruna Street, WV1 1LY, UK
m.thelwall@wlv.ac.uk

Abstract

Most sentiment analysis approaches rely on machine-learning techniques, using a bag-of-words (BoW) document representation as their basis. In this paper, we examine whether a more fine-grained representation of documents as sequences of emotionally-annotated sentences can increase document classification accuracy. Experiments conducted on a sentence and document level annotated corpus show that the proposed solution, combined with BoW features, offers an increase in classification accuracy.

1 Introduction

Sentiment analysis is concerned with automatically extracting sentiment-related information from text. A typical problem is to determine whether a text is positive, negative or neutral overall. Most of the proposed solutions are based on supervised machine-learning approaches, with some notable exceptions (Turney, 2002; Lin and He, 2009), although unsupervised, lexicon-based solutions have also been used, especially in non review-based corpora (Thelwall et al., 2010).

This paper deals with the problem of detecting the overall polarity of a document. A common theme with a significant number of proposed solutions is the bag-of-words (BoW) document representation, according to which a document is represented as a binary or frequency-based feature vector of the tokens it contains, regardless of their position in the text. Nonetheless, significant semantic information is lost when all positional information is discarded. Consider, the following extract of a movie review (taken from Pang (2008)):

> This film should be brilliant. It sounds like a great plot, . . . a good performance. However, it cant hold up.

Most of bag-of-words machine learning or lexicon-based solutions would be expected to classify the extract as *positive* because of the significant number of positive words that it contains. However, a human reader studying the review, recognizes the *change of polarity* that occurs in the last sentence, a change that is hinted at by the first sentence ("should be brilliant") but is only fully realized at the end. In fact, this phenomenon of "thwarted expectations" is particularly common in reviews and has been observed by both Pang et. al (2002) and Turney (2002) who noted that "the whole is not necessarily the sum of the parts".

In this work we propose a solution to the aforementioned problem by building a meta-classifier which models each document as a sequence of emotionally annotated sentences. The advantage of this modeling is that it implicitly captures word position in the whole document in a semantically and structurally meaningful way, while at the same time drastically reducing the feature space for the final classification. Additionally, the proposed solution is conceptually simple, intuitive and can be used in addition to standard BoW features.

2 Prior Work

The commercial potential of sentiment analysis has resulted in a significant amount of research and Pang (2008) provides an overview. In this section, we limit our presentation to the work that is most relevant to our approach.

McDonald et al. (2007) used structured models for classifying a document at different levels of granularity. The approach has the advantage that it allows for classifications at different levels to influence the classification outcome of other levels. However, at training time, it requires labeled data at all levels of analysis, which is a significant practical drawback. Täckström and McDonald (2011) attempt to elevate the aforementioned requirement, focusing on sentence-level sentiment

Proceedings of Recent Advances in Natural Language Processing, pages 546–552,
Hissar, Bulgaria, 7-13 September 2013.

analysis. Their results showed that this approach significantly reduced sentence classification errors over simpler baselines.

Although relevant to our approach, the focus of this paper is different. First, the overall purpose of our approach is to aid document-level classification. Second, the algorithm presented here utilizes sentence-level classification in order to train a document meta-classifier and explicitly retains the position and the polarity of each sentence.

Mao and Lebanon (2006) use isotonic Conditional Random Fields, in order to capture the *flow* of emotion in documents. They focus on sentence-level sentiment analysis, where the context of each sentence plays a vital role in predicting the sentiment of the sentence itself. They also present some results for predicting global sentiment, but convert the sentence-based flow to a smooth length-normalized flow for the whole document in order to compare documents of different length and use a *k*-nearest neighbor classifier using L_p distances as a measure of document similarity.

Our work can be seen as an extension of their solution, where the fine-grained sentiment analysis is given as input to the meta-classifier in order to predict the overall polarity of the document. Nonetheless, in our modeling we retain the structural coherence of the original document by representing it as a discrete-valued feature vector of the sentiment of its sentences instead of converting it to a real-valued continuous function.

3 Sentence-based document representation

The algorithm proposed in this paper is simple in its inception, intuitive and can be used in addition to standard or extended (Mishne, 2005) document representations. Although the approach isn't limited to sentiment classification and can be applied to other classification tasks, the fact that phenomena such as "thwarted expectation" occur mainly in this context, makes the approach particularly suitable for sentiment analysis.

3.1 Sentence classification

At the first level classification, the algorithm needs to estimate the affective content of the sentences contained in a document. The affective content of each sentence is characterized in two dimensions; subjectivity and polarity. The former estimation will aid in removing sentences which contain no or little emotional content and thus don't contribute to the overall polarity of the document and the latter estimation will be used in the final document representation as a surrogate for each sentence. Therefore, for each sentence we need to estimate its subjectivity and polarity, that is, build a *subjectivity* and a *polarity detector*.

Polarity detector: Given a set of positive and negative documents, the algorithm initially trains a standard unigram-based polarity classifier. In our experiments we tested Naive Bayes and Maximum Entropy classifiers, but focus on the former since both classifiers perform similarly, due to space constraints. The classifier utilizes the labels of the training documents as positive and negative instances. The trained classifier will be used at the second-level classification in order to estimate the polarity of individual sentences.

Subjectivity detector: As above, in this stage the algorithm trains a unigram-based subjectivity classifier, that will be used at a later stage for filtering out the sentences that don't contribute to the overall polarity of the document. Training such a classifier is less straight-forward than training the polarity classifier, because of the potential lack of appropriate training data. We propose two solutions to this problem. The first one is based on using a static, external *subjectivity* corpus. The second partly elevates the need for a full subjectivity corpus, by requiring only a set of objective documents, which are usually easier to come by (e.g. wikipedia). In the this case, we can use the training documents as subjective instances and the objective documents as objective instances[1]. We present results with both approaches in section 5.

3.2 Document classification

Having built the unigram-based subjectivity and polarity classifiers in the first stage of the process, the sentence of each training document is classified in terms of its subjectivity and polarity. The former estimation is used in order to remove objective sentences which do not contribute to the overall polarity of the document and also helps in "normalizing" documents to a common length.

More specifically, the sentences are ranked in reference to their probability of being subjective and only the top M are retained, where M is a predetermined parameter. In section 5 we present

[1]During n-fold cross-validation, we utilize only the documents in the training folds as subjective instances.

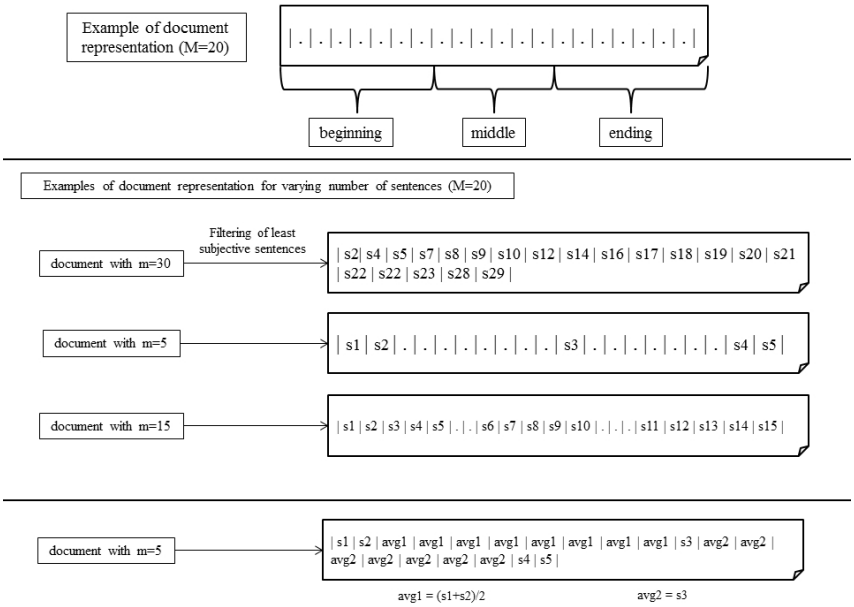

Figure 1: Examples of document representation.

results with various threshold values, but experiments show that a value for M in the $[15, 25]$ interval performs best. A natural question is how does the algorithm deal with documents which have less than M sentences. We provide the answer to this question subsequently, after we explain how the remaining sentences are ordered and utilized in producing the final document representation.

Having removed the least subjective sentences, the remaining are ordered in reference to their relative position in the original document, that is, sentences that precede others are placed before them (see first example in middle section of Figure 1). Using the polarity classifier built on the first stage of the algorithm, we estimate the polarity of each sentence and use this information in order to represent the document as a sequence of emotionally annotated sentences. Alternatively, we can use the probability of polarity of the sentences (e.g., $Pr(+1|sentence)$) in order to represent a document. In fact, the latter representation retains more information than the simple polarity, for example distinguishing between a "barely" positive and a "highly" positive sentence. Although the polarity of both sentences would be the same (i.e., +1) retaining information about the probability provides the document-level classifier with additional information. This decision contrasts with the way that sentence-based sentiment analysis is utilized by Mao and Lebanon (2006) and the experiments presented in section 5 indicate that it typically results in increased accuracy.

The modeling serves two purposes: first of all, by retaining only the more subjective sentences, we remove all sentences which do not contribute to the final polarity of the document. Secondly, by ordering the remaining sentences by their relative original position, we maintain positional information about the emotional content of the most subjective sentences in the document and thus may be able to extract useful positional patterns.

3.3 Dealing with small documents

One of the main problems with the aforementioned approach is the document "normalization" issue, that is, how to represent documents as an equal number of sentences. The retaining of only the most M subjective sentences solves the problem for longer documents and provides a predefined feature vector space, but the problem of effectively representing smaller documents remains.

In order to deal with the problem of small documents, we propose the following solution. Initially, we assume that each document can be represented on an abstract level as having a "beginning" section, a "middle" section and a "ending" section. Depending on the value of M each section is required to be populated by a specific number of sentences. If M is a multiple of 3, then each section will have an equal number of sentences ($M/3$). In the other cases, initially all sec-

tions are attributed the maximum equal number of sentences and the remaining sentences are attributed as follows: if $Mmod3 = 1$, then the extra sentence is added to the middle section and if $Mmod3 = 2$ then each one of the extra sentences are added to the beginning and last sections. For example if $M = 15$, then the distribution of sentences is $\{5, 5, 5\}$, if $M = 16$ then the distribution is $\{5, 6, 5\}$ and if $M = 17$ then $\{6, 5, 6\}$. Clearly, the decision of representing a document as three sections is ad-hoc, and some prior evidence suggests that a 4-way split is better for sentiment analysis (Pang et al., 2002), but we believe it is more in accordance with the intuitive interpretation of documents (Kress, 1999). See top section of Figure 1 for an example with $M = 20$.

Having determined the number of sentences that should be allocated to each section, the next logical step is distributing the existing document sentences to each[2]. We adopt the same process as above, using the number of sentences in the document m instead of M. Therefore, if for example a document has 7 sentences, then their distribution would be $\{2, 3, 2\}$. The placement of the sentences for the beginning and ending sections begin at the first and last position respectively while for the middle section, they are placed around the center. The middle section of Figure 1 provides examples for different m values.

Two final issues subsequently need to be resolved. The first one refers to the filling of the *empty* positions and the second refers to the distribution of sentences on the middle section when m and M differ in terms of their parity (odd vs. even). For the first issue we propose two solutions; the first one fills the empty positions with zeros and the second one fills them with the average of the proceeding polarities or probabilities (e.g., the average of $s1$ and $s2$ in the first example, see lower section of Figure 1). For the second problem, we propose two possible solutions; a "forward weighting" approach where the sentences in the middle section are placed one position toward the beginning of the document and the "backward weighting" approach in which the reverse happens. For example in the middle section of Figure 1 the former approach is used.

[2]Recall that this process is only adopted for documents with less than M sentences.

3.4 Training and testing

To summarize the whole process, during training the algorithm is given a set of positive and negative documents, and initially trains a unigram-based polarity classifier using the labels of the documents. A subjectivity classifier is also built either using a separate *subjectivity* corpus or alternatively, utilizing the documents in the training set as *subjective* instances and only a separate set of *objective* documents as objective instances. Using those classifiers, every sentence in the original training documents are classified in terms of subjectivity and polarity. The sentences are ranked in terms of their probability of being subjective and only the top M are retained, where M is a predefined threshold. Next, the sentences are ordered in reference to their position at the document and their polarity or probability of polarity is used to represent the document and train the second-level, sentence-based classifier.

During testing time, the unigram based classifiers that were built from the training corpus are utilized in order to classify all the sentences in the testing documents in terms of their subjectivity and polarity. As described previously, only the M most subjective sentences are kept and they are reordered in reference to their position in the original document. The learnt sentence-based classifier is applied and a final polarity prediction is made.

4 Experimental Setup

For our experiments, we used a corpus of customer reviews containing reviews of books, DVDs, electronics, music and videogames, split by polarity (henceforth referred to as the *consumer reviews* dataset). The dataset was introduced by Täckström and McDonald (2011) and is freely available[3]. It comprises 97 positive, 98 neutral and 99 negative reviews, annotated by two human assessors both at the document and at the sentence level. Overall inter-annotator agreement is 86% and Cohen's κ value is 0.79. More information about the dataset can be found at Täckström and McDonald (2011). The existence of a set of *neutral* documents and the fact that the corpus is also annotated at the sentence level make it very appropriate for our purposes. Alternatively, we could have utilized the corpus presented by McDonald

[3]The dataset can be obtained from `http://www.sics.se/people/oscar/datasets`.

Training Dataset	Naive Bayes	MaxEnt
Subjectivity corpus (whole documents)	60.75%	59.04%
Subjectivity corpus (filtered documents)	64.87%	62.00%
Consumer reviews (whole documents)	59.81%	63.12%
Consumer reviews (filtered documents)	62.69%	67.73%

Table 1: Subjectivity detection accuracy on the consumer reviews dataset. Result for the last two rows are based on 10-fold cross validation.

et al. (2007), but due to licensing issues, it is currently not publicly available.

For building the subjectivity classifier we use two different approaches. First, we utilize the objective documents of the corpus as objective instances and the training documents as subjective. Two parameterizations are tested: in the first case, we train the classifier on the whole documents and in the second we train the classifier only on the objective/subjective sentences for each category respectively. This way, we'll be able to test whether using much less noisy training data significantly aids the effectiveness of the classifier. In the second approach, we use a static, external corpus to train the subjectivity classifier. In this paper, we use the *subjectivity* corpus by Pang et al. (2002). The corpus is larger than the current dataset, but is only partly relevant to it, as it was built primarily for movie reviews while the consumer dataset that we are utilizing contains reviews from multiple domains.

As baselines, we use the standard unigram representation with presence-based features, with and without length normalization. The first-stage unigram based sentence classifiers are built using the MALLET toolkit (McCallum, 2002). For the final document classification, either using unigram or sentence-based features, we use the SVM implementation from Chang and Lin (2001). Experiments are based on 10-fold cross-validation.

5 Results

5.1 Sentence classification

We begin the analysis of the results by reporting the effectiveness of the subjectivity unigram classifiers in Table 1.

Approach	Accuracy (setting 1)	Accuracy (setting 2)
Baselines		
Unigrams	69.81%	69.81%
Unigrams (N.)	71.76%	71.76%
S-based (M=5)		
Standard	65.55%	64.55%
+ Unigrams	74.39%	72.81%
+ Unigrams (N.)	**75.39%**	72.81%
S-based (M=10)		
Standard	69.21%	69.71%
+ Unigrams	**74.86%**	**76.42%**
+ Unigrams (N.)	73.76%	72.31%
S-based (M=20)		
Standard	69.55%	65.10%
+ Unigrams	75.39%	74.42%
+ Unigrams (N.)	**77.42%**	**76.42%**
S-based (M=30)		
Standard	68.63%	65.60%
+ Unigrams	74.92%	74.92%
+ Unigrams (N.)	**76.92%**	**76.42%**
S-based (M=max)		
Standard	67.13%	67.13%
+ Unigrams	74.92%	74.9%
+ Unigrams (N.)	**76.42%**	**76.42%**

Table 2: 10-fold cross-validation accuracy. *S-based* denotes the sentence-based approach. For the $M = max$ setting we use the number of sentences of the longest document.

The results overall indicate that subjectivity detection on the specific dataset is particularly difficult. More specifically, training either a Naive Bayes or Maximum Entropy classifier on the subjectivity corpus and evaluating in on the consumer corpus, testing either on the whole documents (i.e., "whole documents") or only the objective and subjective sentences (i.e., "filtered documents") results in an accuracy of 64.87% at best. The results are slightly better using an subjectivity classifier trained on the same dataset. In this case, training and testing on only the objective and subjective sentences results in an accuracy of 67.73% at best, while using the whole documents produces an accuracy of 63.12% at best. It will be interesting therefore to see how the sentence-based document classification is affected by the subjectivity detection accuracy.

5.2 Document classification

Due to the number of variations of different document representations that can be explored (e.g., values of parameter M) and space constraints in this section we will present results with the optimal settings that we've discovered for those parameters[4]. Therefore in this section, all presented results are based on backward balancing where the documents are represented as a sequence of probabilities $Pr(+1|sentence)$ and the empty features for small documents are set to 0.

Table 2 presents results for various values of M, with and without additional unigram-based features. The *Standard* approach is based on using only sentences while the *+Unigrams* additionally adds unigram tokens as features. Lastly, we denote full document length-normalization with "*(N.)*". The results on the 2^{nd} column of Table 2 (i.e., setting 1) are based on using the *objective* documents of the dataset for training the objectivity classifier while the results in the 3^{rd} column are based on using the *subjectivity* corpus (i.e., setting 2).

The first rows of the tables present the unigram-based classification accuracy. As already stated, the proposed algorithm can be used in combination with other approaches, so we've opted to utilize this simple approach as a baseline in order to demonstrate its applicability. The baseline results indicate that the specific dataset offers particular challenges, with the standard unigram approach with a length-normalized document vector obtaining an accuracy of 71.76%, much lower that the typical 88% typically reported for other datasets (Pang et al., 2002). Using the sentence-based document representation of documents, initially doesn't provide any significant advantage, maintaining the accuracy effectiveness roughly at the same levels for most values of M. Especially when utilizing the external subjectivity corpus, the effectiveness seems to drop by approximately 6% (Table 2, 3^{rd} column, *Standard* approach for $M = 20$ and $M = 30$).

Nonetheless, using the sentence-based document representation in combination with standard presence-based unigram features always results in an increase in classification accuracy, especially for values of M in the $[20, 30]$ range, reaching an accuracy of 76% in most cases, a rough increase of 6% and 77.42% at best, with $M = 20$. The results overall indicate that the algorithm is quite robust to the value of parameter M. The algorithm retains the high level of effectiveness even when M is set to the number of sentences of the longest document, that is, no sentences are removed and the approach presented in section 3.3 for small documents is applied to the rest of the documents.

The observed differences between using the external *subjectivity* corpus and the objective documents of the dataset aren't as pronounced as expected. Although the observed accuracy for a low value of M in this case is decreased, overall the accuracy levels for higher M values remain stable, typically higher than 76%. The results indicate the potential robustness of the algorithm in reference to the effectiveness of the subjectivity classifier and demonstrate that a static external subjectivity corpus can provide comparable performance.

Limitations: In addition to the experiments presented here, some experiments were also conducted on the MovieReview dataset (Pang et al., 2002) and initial results showed smaller improvements in accuracy. This fact may indicate that the proposed method is more suited for datasets with only limited training data or when unigram features alone attain reduced accuracy.

6 Conclusion

In this paper, we presented a simple and intuitive method of document representation that both implicitly retains word position in documents and explicitly trains a document classifier on the sequence of sentence-based opinions expressed in the document. The proposed algorithm aims to overcome some of the drawbacks of the standard bag-of-words representation, by offering a structurally and semantically meaningful way of effectively representing documents for sentiment analysis.

An obvious extension of the proposed algorithm is the utilization of sequential models, such as CRFs (Lafferty et al., 2001) and structurally-based features (Täckström and McDonald, 2011) in order to increase the effectiveness of the sentence polarity detection, as it was shown that increased sentence classification accuracy typically resulted in increased document classification accuracy.

[4]Detailed results with different parameter values are available from the authors upon request.

References

Chih-Chung Chang and Chih-Jen Lin, 2001. *LIBSVM: a library for support*

vector machines. Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.

Nancy Kress. 1999. *Beginnings, Middles and Ends (The elements of fiction writing)*. Writer's Digest Books.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Chenghua Lin and Yulan He. 2009. Joint sentiment/topic model for sentiment analysis. In *Proceeding of the 18th ACM conference on Information and knowledge management*, CIKM '09, pages 375–384, New York, NY, USA. ACM.

Y. Mao and G. Lebanon. 2006. Sequential models for sentiment prediction. *ICML Workshop on Learning in Structured Output Spaces*.

Andrew Kachites McCallum. 2002. Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu.

Ryan McDonald, Kerry Hannan, Tyler Neylon, Mike Wells, and Jeff Reynar. 2007. Structured models for fine-to-coarse sentiment analysis. In *Proceedings of the Association for Computational Linguistics (ACL)*, pages 432–439, Prague, Czech Republic, June. Association for Computational Linguistics.

G. Mishne. 2005. Experiments with mood classification in blog posts. In *1st Workshop on Stylistic Analysis Of Text For Information Access*.

B. Pang and L. Lee. 2008. *Opinion Mining and Sentiment Analysis*. Now Publishers Inc.

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)*.

O. Täckström and R. McDonald. 2011. Discovering fine-grained sentiment with latent variable structured prediction models. In *Proceedings of the 33rd European Conference on Information Retrieval (ECIR 2011), Dublin, Ireland*.

Mike Thelwall, Kevan Buckley, Georgios Paltoglou, Di Cai, and Arvid Kappas. 2010. Sentiment strength detection in short informal text. *J. Am. Soc. Inf. Sci. Technol.*, 61:2544–2558, December.

Peter D. Turney. 2002. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In *ACL*, pages 417–424.