Effective and Efficient Polarity Estimation in Blogs based on Sentence-Level Evidence

Jose M. Chenlo
Centro de Investigación en Tecnoloxías da Información (CITIUS)
Universidade de Santiago de Compostela, Spain
josemanuel.gonzalez@usc.es

David E. Losada
Centro de Investigación en Tecnoloxías da Información (CITIUS)
Universidade de Santiago de Compostela, Spain
david.losada@usc.es

ABSTRACT

One of the core tasks in Opinion Mining consists of estimating the polarity of the opinionated documents found. In some scenarios (e.g. blogs), this estimation is severely affected by sentences that are off-topic or that simply do not express any opinion. In fact, the key sentiments in a blog post often appear in specific locations of the text. In this paper we propose several effective and robust polarity detection methods based on different sentence features. We show that we can successfully determine the polarity of documents guided by a sentence-level analysis that takes into account topicality and the location in the blog post of the subjective sentences. Our experimental results show that some of our proposed variants are both highly effective and computationally-lightweight.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms

Theory, Experimentation

Keywords

Opinion mining, blog retrieval, sentence retrieval, polarity estimation, efficiency

1. INTRODUCTION

The rise of social networks and blogs has led to new opportunities and challenges regarding how to effectively deal with the opinionated nature of these resources. Our research is focused on one of the most important web sentiment-oriented resources, the blogosphere. People frequently read blogs to determine the viewpoints of others. In order to build an effective information retrieval (IR) system that helps users to find out what other people think, we need to have an understanding regarding whether or not an opinion is present in a text and, if so, whether that opinion is positive or negative [14].

In recent years, several studies have been conducted to determine opinions in blog posts. Most of these address this challenge as a two-stage process that involves a topic retrieval stage (i.e. retrieve on-topic documents given a user query), and a re-ranking stage that takes into account opinion-based features [12]. In blogs, this second Opinion Mining (OM) stage usually involves two further subtasks: an opinion-finding task, where the main aim is to find opinionated posts related to the query, and a subsequent polarity task to identify the orientation of a blog post with respect to the topic (e.g. positive, negative or neutral).

We are concerned here with polarity estimation, a challenging area that is much more demanding than topicality estimation [14]. One problem in polarity estimation is that there may be conflicting opinions in a blog post. For example, a blog writer may summarise pros and cons of a particular argument before settling on an overall recommendation. This mixed set of opinions severely affects the quality of automatic methods designed to estimate the overall orientation of the post. This issue is illustrated particularly well in the following example (taken from a popular film reviews' blog[1]). Despite the start of the post being predominantly negative, with several negative comments being made, the overall recommendation is positive:

> *Gran Torino also includes a few easy outs built into the story ... And even without those easy outs, the storytelling's fairly obvious ... Gran Torino is a curdled mess, politically ... but considering that Gran Torino's heading towards the sunset of Eastwood's acting career, that's a good enough reason to watch it go by.*

Observe that the location of the subjective sentences may offer important clues when attempting to establish the polarity of the blog post. In the example above the last sentence is the one that expresses the overall view. Nevertheless, existing literature on blog polarity estimation has disregarded this valuable information. For instance, the most effective polarity systems participating in the TREC blog tracks [12, 7] do not incorporate any feature based on this flow of sentiments, but rather focus on a document-level estimation

CIKM'11, October 24–28, 2011, Glasgow, Scotland, UK.
Copyright 2011 ACM 978-1-4503-0717-8/11/10 ...$10.00.

[1] `http://blog.moviefone.com`

of polarity that combines relevance to the topic with some sort of global orientation of the sentiments in the document (e.g. counting positive/negative terms). We argue that this is a rather strong simplification and claim that more effective polarity estimation methods can be designed using a sentence-level approach.

In more restricted scenarios, such as corpus of movie or product reviews, some authors have found that the location of the sentiments could be valuable [13, 15, 1]. However, in blogs, the presence of noise (off-topic information or on-topic information that is non-opinionated) makes it difficult to locate the key polar sentences. In this respect, our thorough study of polarity gives useful insights on how sentiments are expressed in blog posts.

In the literature, it has been shown that the noise introduced by off-topic content in documents is a major issue that needs to be addressed to facilitate progress in OM [3, 19]. Therefore, we propose a refined analysis of the documents that takes into account not only the location of the sentiments, but also their relevance to the query. More specifically, we propose effective algorithms that consider two main factors when determining the key sentences for polarity estimation: the relatedness of the sentence and the query topic, and the location of a sentence in the post. We argue that this information, combined with evidence of polarity (i.e. positive/negative terms in the sentence), is extremely valuable when attempting to detect the overall orientation of a post.

The contributions of this paper are:

- We show that sentence-level methods are able to yield state-of-the-art performance in polarity estimation.
- We demonstrate that small subsets of subjective sentences obtained from certain locations of the post are good indicators of the overall polarity.
- We show that estimating polarity from these narrower parts of the document leads to substantial gains in terms of efficiency.

The rest of the paper is organized as follows. In section 2 we explain the blog subjectivity and polarity estimation methods, and the methodology followed to combine sentence-level information with document scores. Sections 3 and 4 report the experiments and analyze their outcomes. Section 5 presents related work. The paper ends with Section 6, where we present the conclusions and outline our ideas for future work.

2. METHOD

As we argued above, the noise introduced by off-topic content in documents may severely harm OM performance because documents might have query terms in a wrong context. Moreover, finding robust OM techniques that can be applied effectively across different underlying topic retrieval baselines is a real challenge. In the past TREC Blog tracks most polarity approaches did not give any added value over the topic retrieval baseline (meaning that the baseline, with no polarity-oriented capabilities, is not beaten by these approaches) [12]. Actually, it is interesting to note that only one participant in TREC 2008 had on average improved the polarity performance of the five topic retrieval baselines provided by the task. This illustrates how difficult it is to design effective polarity estimation methods.

To deal with this challenging problem, we define a general OM polarity approach that involves searching for on-topic polar sentences and location-aware estimation of the document polarity. Given an initial topic retrieval baseline, we work at sentence level to find positive and negative sentences related to the query. Next, our method builds a ranking of positive posts and a ranking of negative posts by aggregating relevance scores and sentence-level polarity information. Within this process we study different location-aware strategies to represent the overall view of the post. An alternative approach with no polarity capabilities is also studied. This final method, which is explained in subsection 2.4, simply promotes subjective documents (regardless of their polarity) and serves as a reference comparison.

The resources utilized in this research to estimate the opinions expressed in texts, and the methods to compute polarity and subjectivity scores based on these resources are explained below.

2.1 OpinionFinder

To estimate subjectivity and polarity we use OpinionFinder (OF) [22][2]. OF is a state of the art subjectivity classifier that works as follows. First, the text is processed using part-of-speech tagging, name entity recognition, tokenization, stemming, and sentence splitting. Next, a parsing module builds dependency parse trees where subjective expressions are identified using a dictionary-based method. This is powered by Naive Bayes classifiers that are trained using subjective and objective sentences. These sentences are automatically generated from a large corpus of unannotated data by two high-precision, rule-based classifiers.

Sentences are classified by OF as subjective or objective (or unknown if it cannot determine the nature of the sentence). Two classifiers are implemented: accuracy classifier and precision classifier. The first one yields the highest overall accuracy. It tags each sentence as either subjective or objective. The second classifier optimizes precision at the expense of recall. It classifies a sentence as subjective or objective only if it can do so with confidence. Furthermore, OpinionFinder marks various aspects of the subjectivity in the sentences, including the words that are estimated to express positive or negative sentiments or the confidence of the decisions made for the accuracy classifier (*diff*).

2.2 Finding On-Topic Polar Sentences

In order to have a precise representation of the mixed set of opinions in a blog post, we compute polarity at sentence level.

With the polar terms tagged by OF [23] we can naturally define the positive or negative polarity score of a sentence. To promote polar sentences that are on-topic, we run a sentence retrieval process to determine the relatedness between the query and each polar sentence. More specifically, we use the Lemur [20] implementation of tf-idf, with BM25-like weights[3].

The combination of relevance and polarity is done through

[2] `www.cs.pitt.edu/mpqa/opinionfinderrelease`

[3] We build a sentence-level index and apply the well-known BM25 suggested configuration ($k1 = 1.2, b = 0.75$), which has proved to be very robust in many retrieval experiments [18].

linear interpolation:

$$pol(S,Q) = \beta\cdot rel_{norm}(S,Q) + (1-\beta)\cdot pol(S) \qquad (1)$$

where $rel_{norm}(S,Q)$ is the Lemur's tf-idf score after a query-based normalization into [0,1] and $pol(S)$ represents the number of positive (resp. negative) terms tagged in the sentence S divided by the total number of terms in S[4]. $\beta \in [0,1]$ is a free parameter.

2.3 Document Polarity Score

Our objective in this paper is to apply sentence-level features combined with location information to improve blog polarity estimation. To this end we score sentences using eq. 1, but only take into account those subjective sentences that have at least one term tagged as positive/negative (i.e. sentences with $pol(S)$ equal to 0 are discarded). We refer to the these sentences as *polar* sentences. To aggregate the individual sentence polarity scores in a document-level polarity measure we work with the following alternatives to define a document polarity score ($pol_S(D,Q)$):

- *PolMeanAll:* The mean of *pol* scores computed across all polar sentences in the document. This measure is a natural choice to estimate the overall polarity of a document.

- *PolMeanBestN:* The mean of *pol* scores from the n sentences with the highest *pol* scores (sentences with the highest aggregated score of topicality and polarity). Focusing on the on-topic sentences with high polarity (e.g. the most controversial contents of the post) we expect to detect properly the polarity of a document.

- *PolMeanFirstN* and *PolMeanLastN:* The mean of *pol* scores from the first/last n polar sentences in the document. As argued above, the position of the sentence in the post may be an important clue when attempting to understand the polarity of the document. Therefore, we study whether the subsets consisting of the first/last polar sentences are good indicators of the overall view in a post. Observe that these strategies are more sophisticated than simply splitting the document into parts. In fact, the polar sentences selected by *PolMeanFirstN* and *PolMeanLastN* depend on the flow of sentiments of the documents, which is specific to each post. For instance, a post whose last part is objective might have its last polar sentence in the middle of the post.

Finally, we combine relevance and polarity evidence as follows:

$$pol(D,Q) = \gamma\cdot rel_{norm}(D,Q) + (1-\gamma)\cdot pol_S(D,Q) \qquad (2)$$

where rel_{norm} is the document's relevance score (obtained from the initial topic retrieval baseline) after a query-based normalization in [0,1], $pol_S(D,Q)$ is one of the aggregation alternatives sketched above and $\gamma \in [0,1]$ is a free parameter. Note that some aggregation techniques have an extra parameter: the number of sentences (n). By studying the behavior of this parameter we might discover valuable patterns about the way in which bloggers express their views.

2.4 Document-Level Subjectivity Estimation

As an alternative approach focused only on subjectivity, we use the proportion of subjective sentences in each retrieved post and the accumulated confidence about their subjectivity (OF's confidence [17, 21]) as subjectivity indicators. This approach has been adopted successfully in other studies [5]:

$$subj(D) = sumdiff\cdot \frac{\#subj}{\#sent} \qquad (3)$$

where $\#subj$ and $\#sent$ are the number of subjective sentences and the number of sentences in a document, respectively. *sumdiff* is the sum of the confidence values of the subjective sentences in the document.

Next, we combine the relevance and subjectivity scores to promote subjective documents that are on-topic. The function used to combine the topic and subjective score of a document is simply:

$$subj(D,Q) = \alpha\cdot rel_{norm}(D,Q) + (1-\alpha)\cdot subj_{norm}(D,Q) \quad (4)$$

where rel_{norm} is the normalized relevance score (obtained from the initial topic retrieval baseline) and $subj_{norm}$ is a query-based normalization of eq. 3. $\alpha \in [0,1]$ is a free parameter. This method, with no polarity capabilities, serves as a reference comparison for polarity-oriented approaches.

3. EXPERIMENTAL DESIGN

In our evaluation we aim at answering the following research questions:

1. Are sentence-level methods effective for polarity estimation?

2. Is sentence location valuable in blog polarity estimation?

3. How does performance change as we focus the estimation of polarity on narrower parts of the documents? What are the implications in terms of efficiency?

To answer the first question, we apply the sentence-level methods defined in subsection 2.3 to a set of different baselines provided by TREC. This helps to assess the impact of our polarity techniques across a range of different topic-relevance baselines. We also compare the performance of our strategies against the best polarity methods participating in TREC and against the subjectivity method sketched in subsection 2.4. We address the second question by analyzing whether we can achieve competitive performance taking only a few sentences from each post (the initial/last polar sentences). We answer the third question by studying the relationship between the number of sentences used and the performance achieved by our location-aware methods. This study takes into account both effectiveness and efficiency.

The remainder of this section describes the experimental setup used to support these investigations.

3.1 Collection and Topics

Our experiments are based on the BLOGS06 collection [8], which is one of the most renowned blog test collections with relevance, subjectivity and polarity assessments. Some statistics of the collection are reported in Table 1.

[4] For positive document retrieval $pol(S)$ is the percentage of positive terms in the sentence, and for negative document retrieval $pol(S)$ is the ratio of negative terms in the sentence.

Number of Unique Blogs	100,649
RSS	62%
Atom	38%
First Feed Crawl	06/12/2005
Last Feed Crawl	21/02/2006
Number of Feeds Fetches	753,681
Number of Permalinks	3,215,171
Number of Homepages	324,880
Total Compressed Size	25GB
Total Uncompressed Size	148GB
Feeds (Uncompressed)	38.6GB
Permalinks (Uncompressed)	88.8GB
Homepages (Uncompressed)	20.8GB

Table 1: The main statistics of the BLOGS06 collection. This collection was utilized in the TREC 2006, TREC 2007 and TREC 2008 blog tracks

We worked with the TREC 2006, TREC 2007 and TREC 2008 blog track's benchmarks. All of them utilize the BLOGS06 document collection as the reference collection for the retrieval experiments. Every year a new set of topics was provided and new judgments were made according to the documents retrieved by the participants. Details about these topics are reported in Table 2.

Blog Track	Topics(#)
2006	851-900 (50)
2007	901-950 (50)
2008	1001-1050 (50)

Table 2: Topics provided in the TREC Blog tracks

Each TREC topic contains three different fields (i.e. title, description and narrative) but we only utilized the title field, which is short and the best representation of real user web queries, as reflected in the official TREC Blog track literature [11, 9, 12]. Documents and topics were preprocessed with Krovetz stemmer and 733 English stopwords were removed.

For the assessment, the content of a blog post was defined as the content of the post itself and the contents of all comments to the post (i.e. permalink document). Documents were judged in two different levels by TREC assessors:

- relevance level: A post can be relevant, not relevant or not judged.

- opinion level: If the post or its comments are not only on target, but also contain an explicit expression of opinion or sentiment about the target, showing some personal attitude of the writer(s), then the document is tagged as positive, negative or mixed (if the opinion expressed is ambiguous, mixed, or unclear). Note that a post tagged as positive (negative) can still contain some negative (positive) opinions provided that the overall document expresses clearly a positive (negative) view with respect to the topic. For instance, the BLOGS06 document below was assessed as positive for the topic *'MacBook Pro'*. Observe that in spite of the presence of conflicting opinions the document was not tagged as mixed because the overall sentiment seems to be positive.

> *[...]the MacBook Pro doesn't come with a modem [...] If you're a business traveller then you WILL be in a situation where the only way to phone home is on an actual phone. You can always add a modem to the MacBook Pro, but that's another expense and another thing to carry. And that's fine, really. Since most people won't need the modem, take it out and gain back the space.*

Blog web pages have noisy information within their internal structure (e.g. links and advertisements). To remove such noise we built a preprocessing unit which can identify the main permalink components (i.e. title, post and comments) and discard the rest of the documents' content. This unit uses a common HTML parser [6] to process the structure of permalinked documents and a set of heuristics to find the core components. The main idea is to detect pieces of text in different HTML blocks and then classify them according to their positional information and size. This type of heuristic has been used successfully in other contexts [16]. We only used the information from title and post, because the comments could be misleading. Deciding how to use comments to effectively guide the estimation of polarity of the document is an interesting challenge that is out of the scope of this paper.

3.2 Retrieval Baselines

In TREC 2008, to allow the study of the performance of a specific opinion-finding technique across a range of different topic-relevance baseline systems, a set of five topic-relevance baselines was provided. These standard baselines use a variety of different retrieval approaches, and have varying retrieval effectiveness[5]. Participants were encouraged to apply their opinion-finding techniques on as many standard baselines as possible. This aims at drawing a better understanding of the most effective and stable opinion-finding techniques, by observing their performance on common standard topic-relevance baselines. Here we adopt this evaluation methodology and apply our methods on these five topic-relevance baselines to assess the robustness of our techniques.

3.3 Polarity Task

The polarity task was introduced in TREC 2007 as a natural extension to the opinion-finding task. This task was initially defined as a classification problem where the systems had to identify the real polarity of a blog post (i.e. positive, negative or mixed). To draw a better simulation of a real user search scenario the task was redefined in TREC 2008 as a typical adhoc retrieval task where the systems had to return a ranking of positive opinionated documents and a ranking of negative opinionated documents. We follow this

[5] The baselines were selected by TREC from the runs submitted to the initial ad-hoc retrieval task in the TREC blog track. These baselines were introduced in 2008 but runs were provided not only for the TREC 2008 topics but also for earlier topics.

Label	Training	Testing
TREC 2007	2006	2007
TREC 2008	2006, 2007	2008

Table 3: Training and testing configurations.

	Int.	Step	Desc.	Form.
α	[0..1]	0.1	Doc. Subjectivity (Comb.)	eq. 4
β	[0..1]	0.1	Sentence Polarity (Comb.)	eq. 1
γ	[0..1]	0.1	Doc. Polarity (Comb.)	eq. 2
n	[1..10]	1	Number of Sentences	eq. 2

Table 4: Parameters to train: the interval, the step used to train, a description and the formula affected by the parameters.

definition of the polarity task. The measures applied to evaluate performance are mean average precision (MAP), and precision at 10 documents (P@10).

3.4 Training and Testing

We built two realistic and chronologically organized query datasets with the topics provided by TREC. Details about these configurations are shown in Table 3. The training topics were used to set all the parameters of our methods. In Table 4 we report some details about the parameters and their characteristics.

Two different training-testing processes focused on maximizing MAP were ran: one for positive polarity retrieval and another for negative polarity retrieval.

4. EXPERIMENTAL RESULTS

In this section, we analyze the results obtained with the datasets described above and address our three research questions. Subsection 4.1, which is related to our first question, reports the results of our polarity estimation methods. Subsection 4.2 addresses the second question, by investigating how location-aware methods work in comparison with other approaches that do not take into account location-based information. Finally, in subsection 4.3 we cover the third research question by studying the impact of the number of sentences used by our polarity approaches and the performance obtained.

4.1 Polarity Task

Tables 5 and 6 show the results of the polarity retrieval approaches. Each run is evaluated in terms of its ability to rank positive (resp. negative) opinionated permalinks higher up in the ranking. In order to have an overall performance for each method, we compute the mean of the positive and negative MAPs and P10s of each run (denoted Mix MAP and Mix P10 respectively)[6]. The best value in each column for each baseline is underlined. Statistical significance was estimated using the paired t-test at the 95% level. The symbols △ and ▽ indicate a significant improvement or decrease over the corresponding baseline. To specifically measure the benefits of our polarity methods we also compare their performance against the results obtained from the subjectivity method (eq. 4, results reported in the rows labeled as Subj). The symbols ▲ and ▼ indicate a significant improvement or decrease over the subjectivity method.

Sentence-level polarity methods. The technique that shows the best performance across all different baselines is *PolMeanBestN*. In TREC 2007, *PolMeanBestN* is the best method in 17 out of 30 cases, showing usually significant improvements in performance with respect to the baseline and with respect to the subjectivity method. *PolMeanAllN* performs the best in 6 cases and *PolMeanLastN* is the best approach in 4 cases. Although *PolMeanFirstN* was never the best option, their results are close to the best ones in most scenarios. We will go back to this issue in subsection 4.2. Observe also that, on average (mix column), some methods yield to a statistically significant decrease in performance for one of the baselines in TREC 2007 (baseline5) but *PolMeanBestN* does not. In TREC 2008, the relative merits of the methods remain the same.

Subjectivity Method. Not surprisingly, subjectivity information alone is not useful in polarity scenarios (the subjectivity method hardly shows any significant improvement in performance with respect to the baseline).

Positive vs Negative rankings. Another observation is that the performance of negative document rankings is quite poor. It is interesting to note that TREC systems (see Table 8) show similar trends. We argue that this is due to the difficulty to retrieve negative posts. As a matter of fact, these collections have many more positive documents than negative ones. The difference is greatest in TREC 2007, where the number of positive documents is 2960 and the number of negative documents is 1844. In TREC 2008, the difference between the number of positive and negative documents is not so marked (3338 against 2789).

Comparison against TREC Systems. To put things in perspective, we report in Table 8 how our methods compare with those proposed by teams participating in TREC [12]. Here, we show the mean of the relative improvements over the five standard baselines. Observe that this polarity task is quite challenging: most TREC polarity systems failed to retrieve more positive or negative documents than the baselines[7]. The methods proposed in this paper perform as well as the best TREC polarity approach (KLE, Pohang University of Science and Technology) [7], showing better performance for some configurations. Observe that our methods and these TREC systems were evaluated under the same testing conditions. Going back to our first research question, these results show that sentence-level methods are an effective strategy for polarity estimation, performing comparably to state-of-the-art TREC systems.

Parameters Trained. Table 7 reports the parameter values trained for each method. Observe that, although the methods proposed have up to three parameters, their optimal values are quite stable across collections.

The subjectivity approach gets a high value of α (0.9). This parameter controls the relative weight of relevance over subjectivity in eq. 3. The value of this parameter indicates

[6] Do not confuse with mixed polarity documents, which refer to documents with mixed opinions.

[7] We can only report the 2008 results because the polarity task was not defined as a ranking process until TREC 2008. Therefore, there are not official results for systems with earlier topics.

	Negative		Positive		Mix	
	MAP	P@10	MAP	P@10	MAP	P@10
baseline1	.0569	.0620	.1779	.2640	.1174	.1630
+Subj	.0603	.0920△	.1599	.2540	.1101	.1730
+PolMeanAll	.0737	.0980△	.1673	.2680	.1205	.1830
+PolMeanBestN	.0818▲	.1240△ ▲	.1819▲	.2880	.1318▲	.2060△ ▲
+PolMeanFirstN	.0742	.0960△	.1668	.2660	.1205	.1810
+PolMeanLastN	.0731	.0980△	.1718	.2640	.1224▲	.1810
baseline2	.0657	.0640	.1590	.2260	.1124	.1520
+Subj	.0656	.0800	.1582	.2260	.1119	.1530
+PolMeanAll	.0719△ ▲	.0740	.1673△ ▲	.2420	.1196△ ▲	.1580
+PolMeanBestN	.0723	.0960	.1624△ ▲	.2320	.1174▲	.1640
+PolMeanFirstN	.0715△ ▲	.0840	.1624△ ▲	.2300	.1170△ ▲	.1570
+PolMeanLastN	.0715△ ▲	.0760	.1655△ ▲	.2360	.1185△ ▲	.1560
baseline3	.0787	.0940	.1919	.2660	.1353	.1800
+Subj	.0792	.0940	.1927△	.2640	.1360△	.1790
+PolMeanAll	.0842△ ▲	.1000	.1956△ ▲	.2780	.1399△ ▲	.1890
+PolMeanBestN	.0843	.1080	.1933△	.2720	.1388	.1900▲
+PolMeanFirstN	.0837△ ▲	.1020	.1933△	.2720▼	.1385△ ▲	.1870
+PolMeanLastN	.0839△ ▲	.1000	.1948△ ▲	.2740△	.1394△ ▲	.1870
baseline4	.0872	.0780	.2176	.2760	.1524	.1770
+Subj	.0878	.0760	.2171	.2740	.1524	.1750
+PolMeanAll	.0912	.0860	.2235△ ▲	.2780	.1574	.1820
+PolMeanBestN	.0899	.1120△ ▲	.2208△ ▲	.2820	.1554	.1970△ ▲
+PolMeanFirstN	.0896	.0860	.2212△ ▲	.2760	.1554	.1810
+PolMeanLastN	.0915	.0840	.2235△ ▲	.2900▲	.1575	.1870
baseline5	.0931	.0960	.2239	.2860	.1585	.1910
+Subj	.0926	.1120△	.2093▽	.2600	.1510▽	.1860
+PolMeanAll	.0843	.1080	.1922▽▼	.2600	.1382▽▼	.1840
+PolMeanBestN	.0785▼	.1100	.2273▲	.2880	.1529	.1990
+PolMeanFirstN	.0818▽▼	.1080	.2181▲	.2820	.1500▽	.1950
+PolMeanLastN	.0834	.1100	.2032▽▼	.2700	.1433▽▼	.1900

Table 5: Polarity Retrieval Results in TREC 2007.

	Negative		Positive		Mix	
	MAP	P@10	MAP	P@10	MAP	P@10
baseline1	.1175	.1700	.1364	.1860	.1270	.1780
+Subj	.1148	.1580	.1379	.1760	.1264	.1670
+PolMeanAll	.1223	.1860	.1477	.2300△ ▲	.1350	.2080▲
+PolMeanBestN	.1280	.1920	.1498	.2200▲	.1389	.2060▲
+PolMeanFirstN	.1315	.2100▲	.1489△ ▲	.2360△ ▲	.1402▲	.2230△ ▲
+PolMeanLastN	.1212	.1920	.1453△	.2200▲	.1332	.2060▲
baseline2	.0865	.1420	.0952	.1400	.0908	.1410
+Subj	.0865	.1380	.0934▽	.1360	.0900▽	.1370
+PolMeanAll	.1026△ ▲	.1480	.1000△ ▲	.1440	.1013△ ▲	.1460
+PolMeanBestN	.0981△ ▲	.1700	.1019△ ▲	.1520	.1005△ ▲	.1610▲
+PolMeanFirstN	.1049△ ▲	.1500	.0975△ ▲	.1420	.1012△ ▲	.1460
+PolMeanLastN	.1000△ ▲	.1460	.0980△ ▲	.1400	.0990△ ▲	.1430
baseline3	.1266	.1520	.1376	.1680	.1321	.1600
+Subj	.1275△	.1540	.1378	.1680	.1326△	.1610
+PolMeanAll	.1333△ ▲	.1700△	.1398△ ▲	.1660	.1366△ ▲	.1680
+PolMeanBestN	.1358△	.1900△ ▲	.1410△ ▲	.1760	.1384△ ▲	.1830△ ▲
+PolMeanFirstN	.1325△ ▲	.1640	.1386△ ▲	.1680	.1356△ ▲	.1660
+PolMeanLastN	.1317△ ▲	.1660	.1386△	.1680	.1352△ ▲	.1670
baseline4	.1288	.1600	.1532	.1980	.1410	.1790
+Subj	.1294	.1640	.1529	.1880▽	.1412	.1760
+PolMeanAll	.1388	.1660	.1576	.2060	.1482△ ▲	.1860
+PolMeanBestN	.1333	.1820	.1559	.1940	.1446	.1880
+PolMeanFirstN	.1423△ ▲	.1900△	.1555△ ▲	.1980	.1489△ ▲	.1940▲
+PolMeanLastN	.1380	.1820	.1552	.2020▲	.1466△ ▲	.1920▲
baseline5	.1085	.1680	.1229	.1780	.1157	.1730
+Subj	.1087	.1620	.1232	.1800	.1160	.1710
+PolMeanAll	.0971	.1640	.1301	.1860	.1136	.1750
+PolMeanBestN	.0988	.1760	.1204	.1980	.1096	.1870
+PolMeanFirstN	.1051	.1780	.1270	.1940	.1160	.1860
+PolMeanLastN	.0991	.1740	.1357	.2000	.1174	.1870

Table 6: Polarity Retrieval Results in TREC 2008.

	TREC 2007		TREC 2008	
	Negative	Positive	Negative	Positive
Subj	$\alpha = 0.9$	$\alpha = 0.9$	$\alpha = 0.9$	$\alpha = 0.9$
PolMeanAll	$\beta = 0.6, \gamma = 0.8$	$\beta = 0.4, \gamma = 0.9$	$\beta = 0.6, \gamma = 0.8$	$\beta = 0.3, \gamma = 0.7$
PolMeanBestN	$\beta = 0.5, \gamma = 0.6, n = 1$	$\beta = 0.1, \gamma = 0.8, n = 1$	$\beta = 0.6, \gamma = 0.6, n = 3$	$\beta = 0.2, \gamma = 0.5, n = 1$
PolMeanFirstN	$\beta = 0.6, \gamma = 0.8, n = 6$	$\beta = 0.2, \gamma = 0.9, n = 5$	$\beta = 0.5, \gamma = 0.8, n = 3$	$\beta = 0.2, \gamma = 0.9, n = 9$
PolMeanLastN	$\beta = 0.6, \gamma = 0.8, n = 10$	$\beta = 0.3, \gamma = 0.9, n = 1$	$\beta = 0.5, \gamma = 0.8, n = 3$	$\beta = 0.2, \gamma = 0.8, n = 1$

Table 7: Parameters trained.

that the relevance component is much more important than the subjectivity component. This seems to indicate that the subjectivity approach is extremely sensitive to off-topic material.

Regarding β, we observe different trends in positive and negative polarity rankings. Positive rankings have lower values of β (the value of this parameter is around 0.2 for positive document retrieval and around 0.5 for negative document retrieval). The β parameter controls the trade-off between relevance and polarity at sentence level (see eq. 1). This means that in positive rankings the polarity evidence is more important than content-match evidence. This might be due to a more reliable estimation of polarity for positive sentences (i.e. OF might be more reliable for positive polarity estimation) or it might be due to the presence of more noisy text (off-topic sentiments) in negative documents. This will be subject to further research in our future work.

Another important trend found affects the number of sentences used by *PolMeanFirstN* and *PolMeanLastN* (i.e. the parameter n). In general, *PolMeanFirstN* takes more sentences to estimate polarity than *PolMeanLastN*. This fact seems to indicate that bloggers briefly summarise their views in the last part of the post. In contrast, if we want to have a reliable summary of the overall opinion obtained from the initial part of the post we need to take a longer subset of sentences. From Table 7 it is also interesting to observe that the number of sentences used by *PolMeanBestN* was 1 in most of the cases. This indicates that we can use the highest *pol*-sentence as the best guidance to understand the overall sentiment of a blog. This finding can be also useful, for example, to build polarity-biased snippets in a blog retrieval scenario.

4.2 Number of Polar sentences needed to achieve state-of-the-art performance

The results reported above suggest that the best way to estimate the overall polarity of a post is to take the highest-*pol* sentences as a representation of the sentiments of the author. However, the methods based on sentences taken from specific document locations work often quite well. Figure 1 depicts the evolution of performance of *PolMeanFirstN* and *PolMeanLastN* against the number of polar sentences taken. For each point in the plot, a ▼ indicates a significant decrease in performance over *PolMeanBestN*, while a ● indicates a non-significant difference in performance with respect to *PolMeanBestN*. With few polar sentences the performance is not statistically different to the performance achieved by the best method. Interestingly, the number of sentences needed to achieve similar performance with respect to the best method differs between *PolMeanFirstN* and *PolMeanLastN*. With *PolMeanLastN*, the last two polar sentences are enough to have a level of effectiveness that is not statistically different to *PolMeanBestN* (for both MAP and P@10). With *PolMeanFirstN*, the initial four polar sentences seem to be a good choice to estimate polarity (with fewer sentences we obtain statistically significant decreases for some measure in some of the collections).

We have therefore successfully addressed our second research question: the use of location information is valuable in blog post polarity estimation, because the first four or last two polar sentences of a blog are good indicators of the overall sentiment.

4.3 Effectiveness vs Efficiency

In the previous section we have compared the performance of the best blog polarity estimation method (*PolMeanBestN*) against the location-aware methods. The reader might wonder why we should bother with these location-based methods if we can achieve state-of-the-art performance with *PolMeanBestN*. In this respect, we argue that there are important implications in terms of efficiency. *PolMeanBestN*, *PolMeanAllN* and *Subj* need to classify all sentences in the post to compute the polarity score of a document. In contrast, the location-based methods just need to classify a small set of sentences.

In the literature, many authors have expressed their concerns about efficiency when using tools such as OF for opinion-finding [5, 4]. We argue that by reducing the amount of data we can substantially decrease the computational cost associated to polarity estimation. To further explore this issue, we took a random sample of 100 documents from the BLOGS06 text collection that had a mean of 6.5 polar sentences according to OF[8]. For each document we created new files based on the first four or the last two polar sentences (appropriate configurations for the *PolMeanFirstN* and *PolMeanLastN* methods, respectively, as discussed in subsection 4.2). For example, for the first four polar sentences of a document, we built a new file that contains the text of all sentences until the fourth polar sentence is found. Finally, we applied OF on each file and on the original document and recorded the average time needed to process each file (preprocessing, tagging and classification). In Table 9 we report the results of this experiment. The use of location-aware methods to estimate the polarity of a blog has a very positive impact in terms of efficiency. *PolMeanLastN* and *PolMeanFirstN* reduce substantially the computation time

[8] The mean of polar sentences per document in the collection is 6.45. The standard deviation is 27.03. This high deviation is likely because of the presence of smap documents, which tends to be large.

	Negative		Positive		Mix	
	MAP	△ MAP	MAP	△ MAP	MAP	△ MAP
KLE	**.1180**	**3.51%**	**.1370**	**6.08%**	**.1274**	**4.86%**
UoGtr	.1103	-2.76%	.1226	-4.62%	.1165	-3.77%
UWaterlooEng	.0987	-12.33%	.1252	-1.69%	.1119	-6.70%
UIC_IR_Group	.0568	-49.60%	**.1313**	**2.12%**	.0941	-22.10%
UTD_SLP_Lab	.0799	-29.23%	.1068	-17.51%	.0934	-22.96%
fub	.0569	-50.18%	.0521	-59.81%	.0545	-55.26%
tno	.0260	-77.02%	.0312	-75.93%	.0286	-76.42%
UniNE	.0584	-48.49%	.0775	-39.41%	.0680	-43.68%
PolMeanAll	**.1189**	**4.80%**	**.1350**	**4.75%**	**.1269**	**4.92%**
PolMeanBestN	**.1190**	**4.89%**	**.1338**	**3.86%**	**.1264**	**4.37%**
PolMeanFirstN	**.1234**	**9.18%**	**.1335**	**3.45%**	**.1284**	**6.07%**
PolMeanLastN	**.1180**	**4.08%**	**.1346**	**4.39%**	**.1263**	**4.36%**

Table 8: Comparison against TREC systems using all 5 of the standard baselines and TREC 2008 topics. TREC results are reported in the first set of rows (top 8 rows). The performance of the polarity methods proposed in this paper is reported in the second set of rows (bottom 4 rows). Positive improvements with respect to baselines are bolded.

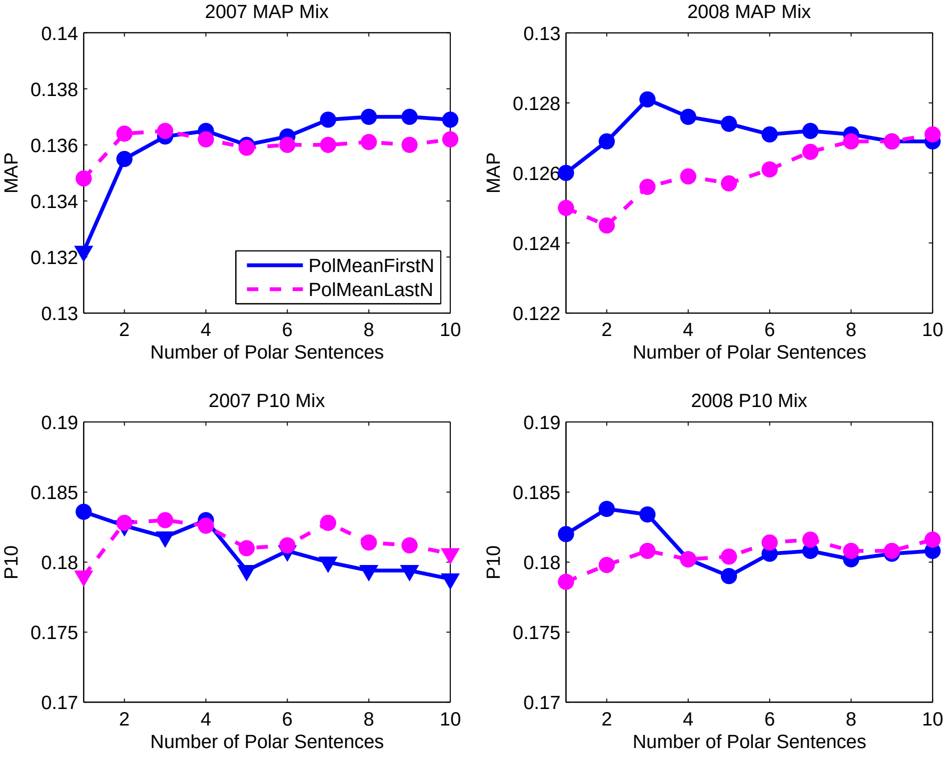

Figure 1: Performance of polarity methods against the number of sentences utilized. A ▼ indicates a significant decrease in performance over the PolMeanBestN method, while a ● indicates a non significant difference in performance with respect to the PolMeanBestN method.

	Avg. Time(s)	$\triangle$ %
Complete Document	2.6	
Last Two Polar Sentences	1.26	−51.5%
First Four Polar Sentences	1.81	−30.4%

Table 9: Average time taken to classify complete documents vs the time taken to classify narrower subsets containing the first/last polar sentences.

with respect to the full document approach (time required is reduced by 51.5% and 30.4%, respectively).

Observe that this classification, which is required to compute $pol(S)$, can be done offline (at indexing time) but, still, there are also benefits on-line. With *PolMeanBestN* it is necessary to process all sentences at query time (to compute $pol_S(D, Q)$ in eq. 2) while *PolMeanFirstN* and *PolMeanLastN* only need to score a small set of sentences. Observe also that the best TREC polarity system (KLE) also treats the full document to find opinionated terms [7]. We argue that location-aware methods are more convenient because they get to similar levels of effectiveness with little computational effort. Furthermore, our findings are potentially applicable not only to learning approaches such as those based on OF but also to other methods that currently process whole documents.

This study answers our third research question: we can substantially improve the efficiency of the OM processes by focusing on small sets of sentences (initial/last polar sentences), and this reduction in the representation of the post does not affect the effectiveness of the polarity estimations.

5. RELATED WORK

Many opinion detection approaches have been proposed in the literature. Among the most successful studies in this subject are those focused on finding document contents that are both opinionated and on-topic [19, 3]. To meet this aim, some papers consider term positional information to find opinionated information related to the query. Santos et al. [19] applied a novel OM approach that takes into account the proximity of query terms to subjective sentences in a document. Gerani et al. [3] proposed a proximity-based opinion propagation method to calculate the opinion density at the position of each query term in a document. These two studies led to improvements over state of the art baselines for blog opinion retrieval. Although we focus on polarity estimation (rather than opinion finding), we also need to filter out off-topic material. In this respect, we worked here with simple sentence retrieval methods and showed that they are consistent to estimate which polar sentences are on-topic.

Pang and Lee [13] considered the impact of the location of the opinionated sentences on the accuracy of two state-of-the art polarity classifiers of film reviews. They built polarity classifiers based on sentences from different parts of a document (e.g. first sentences, last sentences), however these classifiers were not able to overcome local-unigram state-of-the-art systems. Nevertheless, the results obtained showed that the last sentences of a document might be a good indicator of the overall polarity of the review.

Pang et al. [15] considered the impact of term positions in polarity classifiers and argued that the position of a word in the text might make a difference (e.g. movie reviews normally conclude by summarising the author's overall view). Each word was tagged according to whether it appeared in the first quarter, last quarter, or middle half of the document and this information was incorporated in a state-of-the-art unigram classifier. The results did not differ greatly from those obtained using unigrams alone, but the authors argue that the study of more refined notions of positions could be useful in polarity retrieval scenarios.

Beineke et al. [1] proposed several sentiment summarisation approaches based on the analysis of data from a popular film reviews website[9]. This study revealed that the first and the last sentences of the reviews are more important for summarising opinions. To show the importance of the sentence locations, an automatic classifier was built based on two kind of sentence-location features: location within paragraph (i.e. opening, ending, interior or complete paragraph) and location within document (as the fraction of the document that has been completed until the sentence appears). These features were utilized in film reviews to predict whether a particular span of text should be chosen as a summary sentence. The authors found that the use of location-based features alone were insufficient to create proper summaries, being the best results achieved by a classifier that incorporated both term frequencies in the sentences and positional information of the sentences.

In [10], Mao and Lebanon predict the global sentiment of a document by analyzing the sentiment flow at sentence-level. The results indicate that the classification performance is better than a bag of words approach.

In our paper we revisited these issues and studied whether this location information is also a good guidance in blogs (where, typically, we have much more noise than in movie reviews). The techniques proposed here are more sophisticated than the methods applied in film or product reviews, which simply take into account the first or the last part of a document. We analyzed here the first/last sentences *that are subjective with respect to the query*. This allowed us to successfully incorporate location-based information into a large-scale multi-topic scenario, such as the blogosphere. Furthermore, we conducted experiments to understand the implications of location-aware methods in efficiency.

6. CONCLUSIONS AND FUTURE WORK

In this paper we have investigated the impact of sentence-level information in a challenging problem: polarity estimation in blogs. In particular, we have deeply studied different ways to aggregate sentence-level evidence into a document polarity measure. We have also assessed the impact of sentence locations in polarity estimation and evaluated the performance of these techniques in terms of effectiveness and efficiency.

From the results obtained, we found that location-aware polarity methods yield state-of-the-art performance, which is robust across different topic-relevance baselines. We were also able to detect some patterns related to the way in which people write in blogs. More specifically, the overall polarity of posts relies on a few specific sentences (taken from the beginning, from the end, or from the set of high polarity sentences related to the query). This result could be also valuable for creating polarity-biased snippets. We have also

[9] `www.rottentomatoes.com`

demonstrated that we can improve efficiency with no impact on effectiveness.

Most of the methods proposed in this paper are based on simple combinations of polarity and topicality. We are aware that there might be better and more formal ways to approach this combination of evidence (e.g. subjectivity and relevance might be combined using formal methods to learn query-independent weights [2]). This will be explored in the near future. Another problem relates to the number of free parameters to train. Although the optimal parameter values seem to be stable across collections, we plan to study alternative ways to introduce location information in our models. Related to this, we are also interested in studying more refined ways of representation of the sentiment flow of the documents. We also expect to explore the benefits of the use of sentence location for creating opinion-biased summaries.

Acknowledgments

The authors thank the financial support given by Ministerio de Ciencia e Innovación through research project ref. TIN2010-18552-C03-03. We would also like to thank David Elsweiler, Álvaro Barreiro and Leif Azzopardi for their useful comments and suggestions.

7. REFERENCES

[1] P. Beineke, T. Hastie, C. Manning, and S. Vaithyanathan. Exploring sentiment summarization. In *Proceedings of the AAAI Spring Symposium on Exploring Attitude and Affect in Text Theories and Applications*, pages 12–15, 2004.

[2] N. Craswell, S. E. Robertson, H. Zaragoza, and M. J. Taylor. Relevance weighting for query independent evidence. In *SIGIR*, pages 416–423, 2005.

[3] S. Gerani, M. J. Carman, and F. Crestani. Proximity-based opinion retrieval. In *Proc. 33rd Annual International ACM SIGIR Conference*, pages 403–410, 2010.

[4] B. He, C. Macdonald, J. He, and I. Ounis. An effective statistical approach to blog post opinion retrieval. In *CIKM*, pages 1063–1072, 2008.

[5] B. He, C. Macdonald, and I. Ounis. Ranking opinionated blog posts using opinionfinder. In *SIGIR*, pages 727–728, 2008.

[6] M. Jericho. Jericho HTML parser. `http://jericho.htmlparser.net/docs/index.html`, 2009.

[7] Y. Lee, S.-H. Na, J. Kim, S.-H. Nam, H.-Y. Jung, and J.-H. Lee. KLE at TREC 2008 blog track: Blog post and feed retrieval. In *Proc. TREC 2008, the 17th Text Retrieval Conference*, Gaithersburg, United States, 2008.

[8] C. Macdonald and I. Ounis. The TREC Blogs 2006 collection: Creating and analysing a blog test collection. Technical Report TR-2006-224, Department of Computing Science, University of Glasgow, 2006.

[9] C. Macdonald, I. Ounis, and I. Soboroff. Overview of the TREC 2007 blog track. In *Proc. TREC 2007, the 16th Text Retrieval Conference*, Gaithersburg, United States, 2007.

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

[11] I. Ounis, C. Macdonald, M. de Rijke, G. Mishne, and I. Soboroff. Overview of the TREC 2006 blog track. In *Proc. TREC 2006, the 15th Text Retrieval Conference*, 2006.

[12] I. Ounis, C. Macdonald, and I. Soboroff. Overview of the TREC 2008 blog track. In *Proc. TREC 2008, the 17th Text Retrieval Conference*, Gaithersburg, United States, 2008.

[13] B. Pang and L. Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Pr. of the Association for Computational Linguistics*, pages 271–278, 2004.

[14] B. Pang and L. Lee. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135, 2007.

[15] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up? sentiment classification using machine learning techniques. In *Pr. of the Conference on Empirical Methods in Natural Language Processing*, 2002.

[16] J. Parapar and A. Barreiro. An effective and efficient web news extraction technique for an operational newsir system. In *XIII Conferencia de la Asociación Española para la Inteligencia Artificial CAEPIA - TTIA 2007*, pages 319–328, Salamanca, Spain, November 2007. Actas Vol II.

[17] E. Riloff and J. Wiebe. Learning extraction patterns for subjective expressions. In *Proc. of Conference of Empirical Methods in Natural Language Processing*, pages Pages 105–112, 2003.

[18] S. Robertson. How okapi came to TREC. *E.M. Voorhees and D.K. Harman (eds.), TREC: Experiments and Evaluation in Information Retrieval*, pages 287–299, 2005.

[19] R. L. T. Santos, B. He, C. Macdonald, and I. Ounis. Integrating proximity to subjective sentences for blog opinion retrieval. In *Proc. 31st European Conference on Information Retrieval , ECIR 2009*, pages 325–336, 2009.

[20] The Lemur Project. `http://www.lemurproject.org/`, 2009.

[21] J. Wiebe and E. Riloff. Creating subjective and objective sentence classifiers from unannotated texts. In *Proc. of 6th International Conference on Intelligent Text Processing and Computational Linguistics*, 2005.

[22] T. Wilson, P. Hoffmann, S. Somasundaran, J. Kessler, J. Wiebe, Y. Choi, C. Cardie, E. Riloff, and S. Patwardhan. Opinionfinder: A system for subjectivity analysis. In *HLT/EMNLP*, 2005.

[23] T. Wilson, J. Wiebe, and P. Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proc.of Human Language Technologies Conference/Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP 2005)*, 2005.