Sentiment classification of blog posts using topical extracts

Zhixin Zhou **Xiuzhen Zhang** **Phil Vines**

School of Computer Science and Information Technology
RMIT University,
GPO Box 2476 Melbourne VIC 3001 Australia,
Email: {zhixin.zhou, xiuzhen.zhang, phil.vines}@rmit.edu.au

Abstract

Unlike news stories and product reviews which usually have a strong focus on a single topic, blog posts are often unstructured, and opinions expressed in blog posts do not necessarily correspond to a specific topic. This can lead to unsatisfactory performance of sentiment classification. In this paper we report our pilot study on addressing topic drift in blogs. We examine this phenomenon by manual inspection and extablish a ground truth. Our annotations have shown that topic drift is indeed very common, with all documents sampled showing a considerable degree of drift, averaging over 80%. The topical sentences are extracted from each post to produce an extract data set. We propose to address the topical drift problem by classifying the blog posts using the sentence-level polarities of topical extracts. We propose and evaluate two models for aggregating the sentence polarities by comparing their performance to that of a popular word-based model. Our preliminary results suggest that topical extracts can provide a concise but more accurate representation of the sentiment polarity of the blog posts. More importantly, sentence-level polarities are potentially a more reliable evidence than word distributions with regard to document polarity prediction.

Keywords: sentence polarity, sentiment classification, topical extract

1 Introduction

The task of opinion mining, or sentiment classification, can be defined as *"to find opinions targeting topic X"* (Ounis et al. 2006). The blogosphere has attracted a lot of interest in the opinion mining community as blog posts provide a great source of web users' opinions both in terms of the size of the collection and the coverage of topics. However, unlike product reviews and movie reviews, the blog posts are less structured; and unlike news stories, the blog posts are not subject to an editorial process. The author of a blog post can talk about anything in any style of writing.

Blog authors often start with commenting on another related entity before expressing his/her opinion on the topic of his/her true interest. It is particularly common in the political comments. The following example was extracted from a document on the topic of *State of the Union Address by President George W Bush.*

"*Though surprised today to find out that Hamas had a unexpectedly large VICTORY in the Palestinian elections, I am not shocked, and such a development should not have been unexpected if you follow politics, and the actions of George Walker Bush since he was selected to be our president back in the year 2000.*"

Before attacking the actions of the Bush administration the author commented on another political event, the latter was used merely for contrasting. It is also worth noting that in this example, opposite opinions on two different entities are expressed in one single sentence.

It is also common for a post to have a few lines of description related to the author's recent activity, but then followed by some reflections on his/her past experience. This is typical in movie and product reviews, where the author first briefly mentions the movie or product, and may then drift away to write about the emotions or memories aroused by the movie or product. The example below from a document on the topic of *March of the Penguins* typifies this approach.

"*This is a must-see documentary about the mating ritual of Antarctic penguins... I sniffled and nodded – I know what that feels like... We received news yesterday that the donor's ultrasound exam was normal, so we are good to go. Am feeling detached and ambivalent again...*"

The author of this post watched the movie *The March of the Penguins*, and the birth-giving scenes in the movie caused her to reflect upon her own pregnancy in the past. In fact, the majority of this post in about her own experience, and not the movie.

Topic drift may lead to serious problems in sentiment classification. Previous studies on sentiment analysis of blogs have mostly focused on word-based models using the whole document (Ounis et al. 2006, Macdonald et al. 2007). However, the drifting portion of the document would change the word distribution and potentially compromise the accuracy of the classification models. To address this problem, an intuitive solution would be to extract the topical snippets first and perform classification on the topical extracts instead of the original documents. Both Pang (Pang & Lee 2004) and Lloret (Lloret et al. 2010) have done similar studies, but neither has observed consistently better results on the extracts (note that both works extracted the sentences that are topical and subjective). It has been shown that the length of the extract affects the classification accuracy (Pang & Lee 2004), which suggests that poor-quality extracts might be the reason of the counter-intuitive observation. More recent work by McDonald integrates

Table 1: An outline of the two topics we used in the collection

Topic	#[1]	Description
869	83	Find opinions worldwide to the cartoons depicting the Muslim prophet-Muhammad printed in a Danish newspaper.
903	80	Find documents stating opinions about Apple CEO Steve Jobs.

[1] The number of posts we kept in the collection for the topic

sentence-level information into the learning process. While his main focus is on a joint-structured model that classifies the documents as well as the sentences at once, a cascaded model is also described which predicts the sentence labels first and then pass the labels to the document classification model. In his work however, all the sentences in the document are used.

To our best knowledge, no existing work has applied sentiment classification on manual extracts. Because of this, it is not able to tell whether the problems result from quality of the extracts, or the method employed. In order to isolate these effects we have manually extracted relevant sentences. In our study we annotated randomly sampled documents from two different topics used in Blog06 and Blog07 tasks in TREC. The descriptions of these two topics are shown in Table 1. We adopted the document-level labels annoated by NIST judges, and adapted their annotation scheme to suit our case. We labeled each sentence in the documents both by topicality and sentiment orientation.

We propose two approaches to classify the document polarity with topical extracts. Both approaches are based on sentence-level polarities. We contrast their performance with that of a word-based classifier. Our preliminary experiments have shown that one of our approaches has better performance than that of the word-based classifier using the original posts on both topics. We also note that the word-based approach performs better on the extract than the original post on one topic, but has a lower average accuracy on the other. This suggests that sentiment classification on the topical extracts may indeed be more accurate than on the original posts, but models based on the word-distributions may not work well with the extracts. Pang extracted topical and subjective sentences with the minimum-cut algorithm and Lloret used the SUMMA toolkit (Saggion 2008).

We describe our manual annotation process in Section 3, and introduce the classification models in Section 4. The implementation of the classification experiments are detailed in Section 5. After that, we conclude our work and provide an outlook into the future work in Section 6.

2 Related Work

Sentiment classification, or opinion mining, can be done at different levels of natural language structure. Typically it is done at the document level (Turney 2002, Dave et al. 2003, Pang & Lee 2004, Das & Chen 2007). Turney (2002) extracted potentially opinion-bearing phrases from the documents with a POS tagger and calculated the sentiment orientation score for each phrase. The averaged score of the phrases in a document was used to classify the document as either positive or negative. The voting model proposed in this paper is similar to their work in that we used accumulated length of sentences of different polarities, instead of phrase scores, as the indicators of document polarities.

The most similar work to ours is McDonald et.al's cascaded model briefly described in (McDonald et al. 2007) , where sentence polarities are used to influence the document level predictions. However, we use only the topical extracts, while they use the whole document; the sentence polarities are used jointly with other features in their approach, while our approaches are based solely on the sentence polarities. We also note that their main focus is a structured model that jointly predicts the document polarity and the sentence polarities at once.

Another piece of work by Pang et al (Pang & Lee 2004) proposed a more fine-grained model which first classifies the sentences by topicality and subjectivity so that only the topical and opinionated portion of the document is used for classification. Their approaches achieved a classification accuracy comparable to that on the original documents, and suggested that the extracts were not only more concise, but also probably "cleaner" representations of the intended polarity. Our study also compared the performance on extracts and original documents with a bag-of-words model, but we extended our scope to more topics. While Pang's work was based on movie review data, we tested our approaches on a data set on political events and another data set relating to opinions towards a public figure.

Another piece of relevant work that contrasts classification performance on extracts and full documents is described in Lloret et.al's (Lloret et al. 2010) paper on the problem of rating-inference, which aims to assign numeric rating numbers to the documents. This is different from our work as the documents in our study are associated with either positive or negative label.

The voting model and the logistic regression model proposed in this work employ sentence sentiment polarities to predict the document sentiment polarity. While manual annotations are used in this pilot study, automatic sentence-level classification techniques are to be integrated into the proposed models in our future work. Existing sentence-level studies include subjectivity studies (Hatzivassiloglou & Wiebe 2000), which only identifies whether a sentence is subjective; and polarity classification (Hu & Liu 2004, Kim & Hovy 2004, Nasukawa & Yi 2003, Popescu & Etzioni 2005, Khan et al. 2011), which predicts the sentiment orientation (positive, negative, etc).

Overall, our work focuses on predicting document polarities with regard to a user-given topic. And instead of making use of an opinionated lexicon (Hu & Liu 2004, Kim & Hovy 2004), our approach is based on a higher-level linguistic construct and employs sentence polarities.

There are several datasets (Wilson et al. 2005, Pang & Lee 2004, Ku et al. 2007) labelled by sentiment orientation. Wilson and Wiebe (Wilson et al. 2005) conducted an annotation experiment at the word- and phrase-level, and produced the MPQA dataset. The annotation scheme they adopted is rather fine-grained and captures the nested structures of the private states and speech events. The articles in this dataset are from the world press and are thus all news stories. Pang (Pang & Lee 2004) built a sentence-level dataset to test their classification model. All the sentences are extracted from movie reviews and are labelled as either subjective or objective. Ku (Ku et al. 2007) annotated news documents from the NTCIR collection over 32 topics, and assigned *positive*, *negative* or *neutral* to each sentence. However, neither of the two data sets suffices for our study. First, they are restricted to certain domains, in which sentiment tends to be expressed in a more constrained way compared to a collection with unrestricted topic domains. Second, the sentences in these collections were not labelled by topicality. Therefore, we annotated our own data sets as part of this work.

3 Addressing Topic Drift in Blogs

To address the topic drift in blogs, we need first to examine how common this phenomenon is. This requires the ability to judge whether or not each sentence in the document is relevant to the given query. While there are existing technologies to accurately classify documents by topicality, it is much harder to classify by topicality at the sentence level. Relevant documents usually contain certain query terms about the topic, but a relevant sentence may have no topical term at all. For example,

- *"Ok, here's why this is stupid. 1) The menu ... you can also get the information in the store, by asking a McDonalds ... 2) ...* ", extracted from BLOG06-20051206-033-0009894627
- *"... I love french fries and when I eat them, especially Mickey Ds, I feel all warm and fuzzy ... Anyhow, french fries are my comfort food ...* "

In the first example, the first sentence *"Ok, here's why this is stupid."* is relevant, but looking at that sentence alone, it is impossible to tell what *"this"* really is. One would have to read the text after this sentence to link the comment to the brand *McDonald*. Similarly, in the second example, a reader couldn't tell whether the *"french fries"* is from *McDonald* unless the context of the sentence wherein the phrase appeared has been examined.

We propose a simple annotation scheme to study the topic drift phenomenon. The scheme can be described as a set of rules, as explained below. We adopt the topic descriptions provided in the TREC Blog 06 data set to judge whether a sentence is topical. The descriptions for the two topics we used are shown in Table 1, and all the following examples are extracted from posts of topic 869, which is about *"the cartoons depicting the Muslim prophetMuhammad printed in a Danish newspaper"*.

3.1 Sentence Annotation Scheme

3.1.1 Topicality

A sentence is either *topical* or *non-topical*.

- A sentence that directly mentions the query topic is considered *topical*.

  A topical example:

  They claim he's racist, largely because he reports stories like these: about appeasement in Europe:The Danish newspaper Jyllands-Posten recently published cartoons of Mohammed, and Danish Muslims went crazy, rioting and threatening the newspaper.

  A non-topical example:

  Little Green Footballs has interesting posts (but the comments sections isn't usually very edifying, just so you know), but a lot of people don't like Charles Johnson, the host of the site.

- If it can be implied from the context that the sentence is about the topic, the sentence should also be considered *topical*. Both sentences before and after the target sentence should be taken into consideration.

  "The UN is Appeasing Muslims - Again.", *followed by "UN Concerned over Prophet Cartoons by Ole Damkjaer..."*

  This sentence itself does not directly mention the cartoons, but the word *"appeasing"* actually refers to the UN's response to the cartoon incident, and is thus topical.

- A sentence that addresses more than one topic, including the target topic, is still labelled as *topical*.

  "Fjordman, the Norwegian blogger (how sorely his invaluable reports from Scandinavia will be missed when he quits blogging next week) that , the , has over the 12 cartoons [see them] depicting the prophet Muhammad which were published in the Danish newspaper Jyllands-Posten last September."

  This sentence is *topical*, although it also mentioned comments on Fjordman's quit from blogging.

- Sentences from quoted content are processed in the same way as the sentences written by the author.

  "In a letter to the 56 member countries of the Organization of the Islamic Conference (OIC), she states: "I understand your concerns and would like to emphasize that I regret any statement or act that could express a lack of respect for the religion of others". ... the 56 Islamic governments have asked Louise Arbour to raise the matter with the Danish government "to help contain this encroachment on Islam, so the situation won"t get out of control."

  This sentence is topical because "a lack of respect for the religion of others" refers to the cartoon incident.

3.1.2 Sentiment Orientation

A topical sentence can have one of the four polarities in our system: *neutral, negative, mixed, positive.* The label *unknown* is automatically assigned to non-topical sentences, since such sentences are discarded in the classification models we proposed. The definition of the other four labels are the same as defined in the TREC Blog track.

- The polarity of the opinion is evaluated with regard to the query topic. If the sentence is about multiple topics, only the relevant opinion towards the target topic is considered when labeling. Opinions expressed in non-topical sentences are not labeled, and the sentiment polarity of all non-topical sentences is defaulted to *unknown*.

  This is because our task is to find and classify opinions targeted at a specific topic, therefore the irrelevant sentences are not useful.

  "Running a newspaper is a tough business these days." This sentence is opinionated, but not on the topic. We label it as *unknown*.

  "While one cartoon was particularly offensive because it showed the prophet as wearing a turban with a bomb attached to it, a great deal of the

anger had to do with the mere depiction of the prophet."

This sentence is right on topic, and explicitly showed negative opinion towards the cartoons: *"offensive"*, thus labeled as *negative.*

- We do not identify the opinion holders. Whether or not the opinion was expressed by the author is not examined in the annotation process. In fact, many posts quote statements from newspapers and other media but not showing any opinion of the author himself/herself. Ideally, only the author's opinions should be considered, as most users seem to be more interested in the author's opinions. Nonetheless, our sentence-level annotation rules must be consistent with the document-level annotations, which we adopted from NIST annotators at the TREC conference. In their annotation scheme, the holder of the opinion was not taken into consideration. To be consistent, we followed their scheme.

  "In a letter to the 56 member countries of the Organization of the Islamic Conference (OIC), she states: "I understand your concerns and would like to emphasize that I regret any statement or act that could express a lack of respect for the religion of others". ... the 56 Islamic governments have asked Louise Arbour to raise the matter with the Danish government "to help contain this encroachment on Islam, so the situation won"t get out of control."

  The quoted content bears negative opinion toward the cartoons, though not expressed by the author. We label this sentence as *negative.*

- When a sentence bears opinions towards a statement which is related to the topic, we must take the opinion expressed in that statement into consideration.

  "The Islamic governments have expressed satisfaction with the reply from Louise Arbour."

  Since Louise Arbour holds negative opinions towards the cartoons, the Islamic governments holds negative opinions as well. Thus the sentence is labeled as *negative.*

3.1.3 The Labeling Procedure

Two topics from the TREC Blog Track 2006 and 2007 tasks were used in the annotation. NIST annotators have already labelled the documents by topicality as well as sentiment polarity. We adopted the document-level annotations made by NIST assessors. For our study, we only used the posts labelled as 2 or 4, as people are generally more interested in these opinions rather than mixed and neutral. In order to reach more reliable conclusions in the classification experiments, we chose the topics which have a relatively large number of posts with a reasonable length. We used 300-words as the threshold for selecting posts, and after filtering out the short posts, we chose only the topics which have at least 50 posts in both the negative and the positive classes. When building the corpus for each topic, we sampled as many posts as possible, while keeping the number of posts the same in each class. The details of the topics are listed in Table 1, and the steps we followed to build the collection are explained below.

Table 2: Topic drift on the topics we used in the collection

Topic	Doc	Neu	Neg	Mix	Pos	Drifted
869	14323	0.23%	7.19%	0.06%	6.16%	86.36%
903	12429	6.56%	3.23%	0.14%	6.68%	83.40%

[1] Measured by # of characters and averaged over all documents in the topic

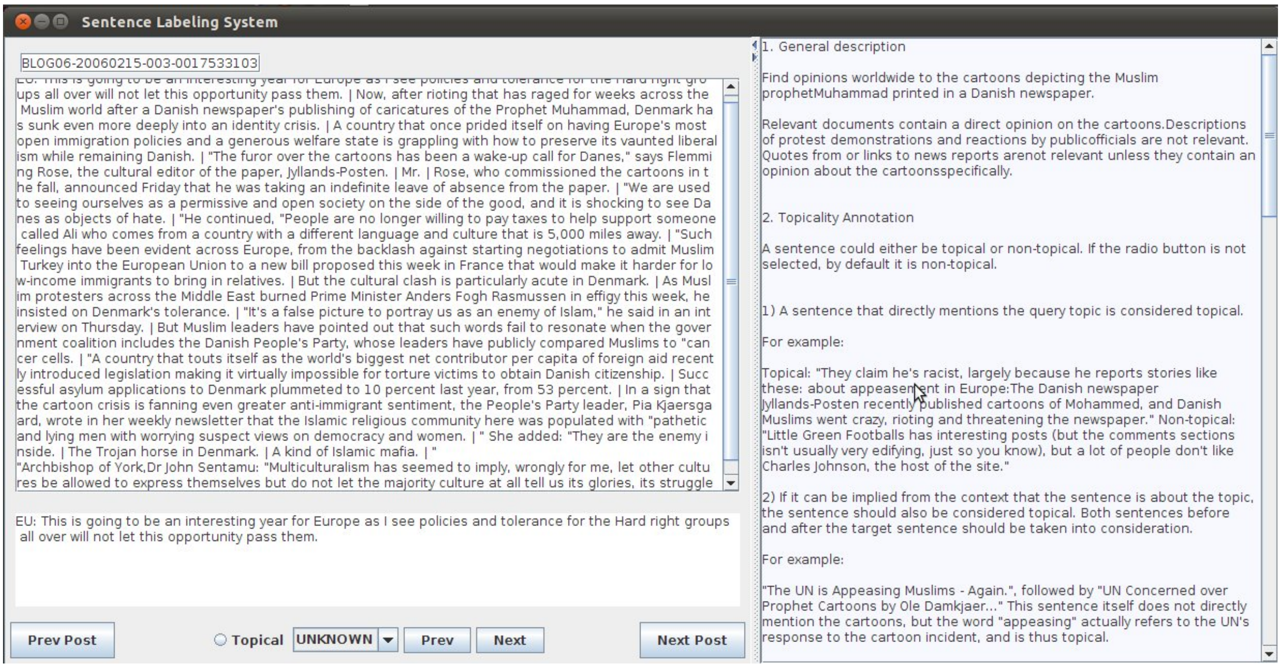
1. We extract the html section from the Blog06 collection, and run a preprocessing program to keep only the original post and the replies of web users. We save each post in a single document.
2. For each post, we split the text into sentences. We used a regex expression to perform this task. Note however, there is some noise in the collection due to sloppy punctuation and therefore the sentences were not perfectly segmented. There are cases where multiple sentences are grouped as one, and cases where a single sentence is mistakenly split into two. Nonetheless, this does not significantly affect the accuracy of our classification models, as such cases are rare, and none of our models is based directly on the number of sentences. More explanation follows in Section 4, when the classification models are introduced.
3. By default, each sentence was labelled as irrelevant, and the sentiment polarity is labelled as *Unknown.*
4. The sentences and the posts were then uploaded to a database. The labels that an annotator applied to the sentence were also kept in a database.

Our annotators were able to view the full post when labelling each sentence, so that the context was taken into consideration. The sentences were shown according to the sequence they appeared in the blog post, but the annotator could always move backwards to change the labels made previously. As the task itself was rather subjective, we did not provide a detailed annotation guide. Instead, we showed the annotation rules on the side panel of the labeling interface, as indicated by Figure 1, and allowed the annotator to relate to their background knowledge when making the judgement.

3.2 Results

With our manually annotated corpus we were able to show some statistics on the phenomenon of topic drift. Table 2 shows some statistical details about the two topics. The drifting sentences constitutes as much as 86.36% and 83.40% on topic 869 and 903 accordingly. The amount of the four types of sentences has been normalized by the document length, and are shown in averaged percentages over all documents in each topic. It is also noteworthy that all documents in our collection have drifting portions with regard to the query.

The distribution of the sentences by sentiment polarity are shown in boxplots. Figure 2 shows the distribution of the sentences in the negative documents of topic 869, Figure 3 shows the distribution of the sentences in the positive documents of topic 869, and Figure 4 shows the distribution of the sentences in the negative documents of topic 903, and Figure 5 shows the distribution of the sentences in the positive documents of topic 903. It can be seen from the four graphs that sentences labeled as bearing *mixed* opinions are quite rare on both topics. *Neutral* sentences, on the other hand, have very different distributions on the two topics in the collection. This is mainly



BLOG06-20060215-003-0017533103

EU: This is going to be an interesting year for Europe as I see policies and tolerance for the Hard right gro
ups all over will not let this opportunity pass them. | Now, after rioting that has raged for weeks across the
Muslim world after a Danish newspaper's publishing of caricatures of the Prophet Muhammad, Denmark ha
s sunk even more deeply into an identity crisis. | A country that once prided itself on having Europe's most
open immigration policies and a generous welfare state is grappling with how to preserve its vaunted liberal
ism while remaining Danish. | "The furor over the cartoons has been a wake-up call for Danes," says Flemmi
ng Rose, the cultural editor of the paper, Jyllands-Posten. | Mr. | Rose, who commissioned the cartoons in t
he fall, announced Friday that he was taking an indefinite leave of absence from the paper. | "We are used
to seeing ourselves as a permissive and open society on the side of the good, and it is shocking to see Da
nes as objects of hate. | "He continued, "People are no longer willing to pay taxes to help support someone
called Ali who comes from a country with a different language and culture that is 5,000 miles away. | "Such
feelings have been evident across Europe, from the backlash against starting negotiations to admit Muslim
Turkey into the European Union to a new bill proposed this week in France that would make it harder for lo
w-income immigrants to bring in relatives. | But the cultural clash is particularly acute in Denmark. | As Musl
im protesters across the Middle East burned Prime Minister Anders Fogh Rasmussen in effigy this week, he
insisted on Denmark's tolerance. | "It's a false picture to portray us as an enemy of Islam," he said in an int
erview on Thursday. | But Muslim leaders have pointed out that such words fail to resonate when the gover
nment coalition includes the Danish People's Party, whose leaders have publicly compared Muslims to "can
cer cells. | "A country that touts itself as the world's biggest net contributor per capita of foreign aid recent
ly introduced legislation making it virtually impossible for torture victims to obtain Danish citizenship. | Succ
essful asylum applications to Denmark plummeted to 10 percent last year, from 53 percent. | In a sign that
the cartoon crisis is fanning even greater anti-immigrant sentiment, the People's Party leader, Pia Kjaersga
ard, wrote in her weekly newsletter that the Islamic religious community here was populated with "pathetic
and lying men with worrying suspect views on democracy and women. | " She added: "They are the enemy i
nside. | The Trojan horse in Denmark. | A kind of Islamic mafia. | "
"Archbishop of York,Dr John Sentamu: "Multiculturalism has seemed to imply, wrongly for me, let other cultu
res be allowed to express themselves but do not let the majority culture at all tell us its glories, its struggle

EU: This is going to be an interesting year for Europe as I see policies and tolerance for the Hard right groups all over will not let this opportunity pass them.

1. General description

Find opinions worldwide to the cartoons depicting the Muslim prophetMuhammad printed in a Danish newspaper.

Relevant documents contain a direct opinion on the cartoons.Descriptions of protest demonstrations and reactions by publicofficials are not relevant. Quotes from or links to news reports arenot relevant unless they contain an opinion about the cartoonsspecifically.

2. Topicality Annotation

A sentence could either be topical or non-topical. If the radio button is not selected, by default it is non-topical.

1) A sentence that directly mentions the query topic is considered topical.

For example:

Topical: "They claim he's racist, largely because he reports stories like these: about appeasement in Europe:The Danish newspaper Jyllands-Posten recently published cartoons of Mohammed, and Danish Muslims went crazy, rioting and threatening the newspaper." Non-topical: "Little Green Footballs has interesting posts (but the comments sections isn't usually very edifying, just so you know), but a lot of people don't like Charles Johnson, the host of the site."

2) If it can be implied from the context that the sentence is about the topic, the sentence should also be considered topical. Both sentences before and after the target sentence should be taken into consideration.

For example:

"The UN is Appeasing Muslims - Again.", followed by "UN Concerned over Prophet Cartoons by Ole Damkjaer..." This sentence itself does not directly mention the cartoons, but the word "appeasing" actually refers to the UN's response to the cartoon incident, and is thus topical.

Figure 1: The Labeling Interface

due to a rather special topic description: "*Find opinions worldwide to the cartoons depicting the Muslim prophetMuhammad printed in a Danish newspaper.*" As the description specifically asks for opinionated content, the annotator ignored most of the neutral content. The amounts of positive and negative content have a rather strong correlation with the document polarities on both topics, and particularly so with topic 869.

In our data set every sentence has a sequence number that identifies its position in the document. With this information we are able to show the locations where subjective sentences appeared in the documents. We normalized the position by dividing the sequence number by the total number of sentences in each document, and plotted Figure 6 and 7 for topic 869 and 903 respectively. In the documents on both topics the first 20% of the text is a highly probably region where subjective sentences appear. The distribution of subjective sentences in other portions of the text varies significantly in the two topics.

4 Predicting Document Polarities

To tackle the problem of topic drift in the blogs we propose to do sentiment classification on the topical extracts. We validate the effectiveness of our approach by contrasting the performance on the full posts with that on the extracts. Our baseline approach uses the Naive Bayes classifier, which is a traditional classifer commonly used in text mining, and treats the blog post as a whole in the classification process. Reasonable performance of this model has been reported in existing literature (Pang et al. 2002). Also, when comparing our proposed approaches with this baseline system our main objective is not to improve the document-level classification models, but rather, to compare the performance of classification on full posts with that on topical snippets. For this purpose, this model is sufficient to validate our hypothesis.

Unlike probabilistic classifiers such as Naive Bayes, humans judge document polarities not by examining the word distributions, but by aggregating the polarities of sub-documental natural language structures. Motivated by this, we explore models that simulate this process. In this study we treat the sentence as the basic opinion-bearing structure, and aim to predict the document polarity by aggregating the sentence polarities. Existing work (Hu & Liu 2004, Kim & Hovy 2004, Nasukawa & Yi 2003, Popescu & Etzioni 2005, Khan et al. 2011) have shown that automatic sentence-level prediction can be done at a reasonable accuracy. Combined with our approach, it will then be possible to predict the document polarity with a reasonable performance. Such techniques are subject to further study and are not the focus of this paper. In our work we directly use the groud truth - human-annotated sentence polarities, and propose two models: a voting model and a logistic regression model.

Although our techniques could be applied to multi-class problems, in this paper we limit our scope to a binary-class problem, i.e. we are only interested in classifying documents bearing either positive or negative opinions. This decision is made primarily due to the lack of documents in other classes.

4.1 The Word-based Approach

We adopted a simple Naive Bayes model similar to the approach reported in Pang's work (Pang et al. 2002). In this model, all of the words, aside from the stop words, from the post are included as features. Each document is represented as a vector consisting of the tf-idf values of the symbols[1], and ended with the the class label.

$$\overrightarrow{D} = \{tf \cdot idf_{w_1}, tf \cdot idf_{w_2}, ..., tf \cdot idf_{w_i}, class\} \quad (1)$$

where $tf \cdot idf$ is defined as,

$$tf \cdot idf = tf(t,d) \times \log \frac{|D|}{|\{d : t \in d\}|} \quad (2)$$

[1] Here a symbol refers to any alphabetic token in the post, without any filtering process by means of a dictionary. As such, incorrectly spelled words, as well as commonly used abbreviations on the web (e.g. AFAIK, WTF, IMHO), are all included.

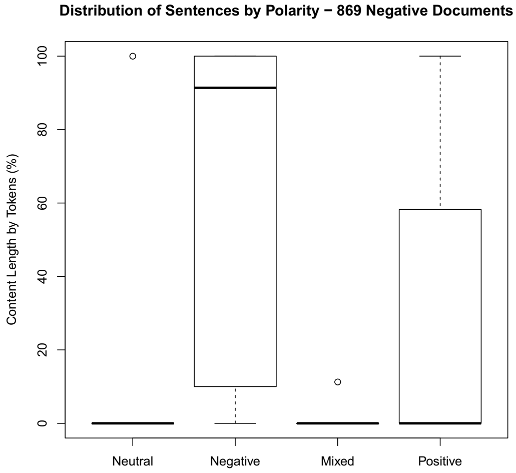


Figure 2: Negative Documents in Topic 869

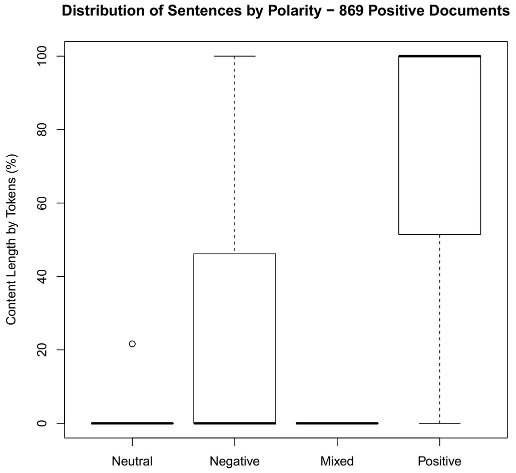


Figure 3: Positive Documents in Topic 869

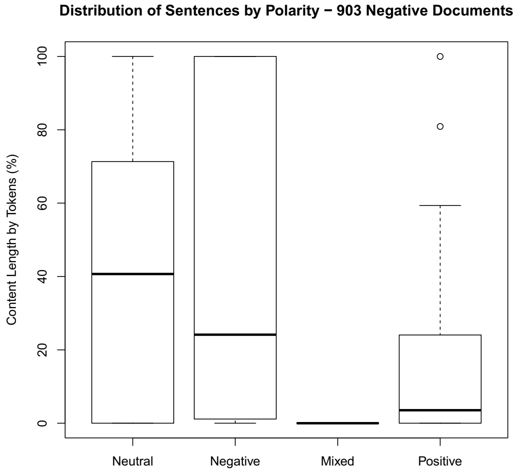


Figure 4: Negative Documents in Topic 903

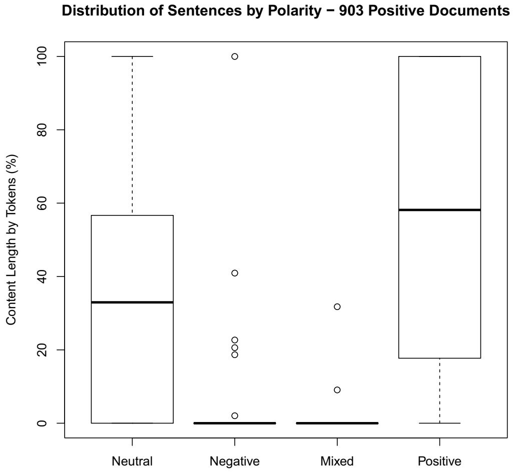


Figure 5: Positive Documents in Topic 903

Here, $tf(t,d)$ is the frequency of the term in the document, and $\log \frac{|D|}{|\{d:t\in d\}|}$ is the number of documents containing the term inverted by the total number of documents in the collection.

In the training process, the prior probabilities $p(t_i|C)$ of each term appearing in the documents of each class are calculated. Then based on the Bayes Theorem, we have

$$p(C|t_1,...,t_i) = \frac{1}{Z}p(C)\prod_{i=1}^{n} p(t_i|C) \qquad (3)$$

where Z is a scaling factor based on the vector $\{t_1,...,t_i\}$, and $p(C)$ is the prior probablitiy of the classes. In our experiment we adopted the implementation in the Weka data mining toolkit[2].

4.2 The Voting Model

Intuitively, a document with a larger portion of positive content is more likely to be positive than negative. Motivated by this, we propose a voting model that classifies the documents by comparing the amount of positive and negative contents in a document. The rule is simple, the majority class in the topical sentences is assigned to the document. The polarity $P(D)$ of a document D is given by,

$$P(D) = \begin{cases} positive & if\ |D_{positive}| >= |D_{negative}| \\ negative & if\ |D_{positive}| < |D_{negative}| \end{cases} \qquad (4)$$

In measuring the amount of the contents $|D_{positive}|$ and $|D_{negative}|$, either the number of tokens or the number of characters can be used. Experiment results are shown in Section 6.

Intuitively this model is not perfect, due to the subtlety of human language. It has been reported in previous work by Pang, et.al (Pang et al. 2002) that a specific problem by the name of thwarted exceptions exists, where the author sets up a deliberate contrast to his/her earlier discussion. In this case, our voting model would fail, as despite of the fact that the document may have more positive content, it should be classified as negative. Measures to tackle this problem is subject to further study. Surprisingly however, in our experiments we observed a significantly better performance than other approaches on one of the topics we used, and a slightly inferior performance on

[2] http://www.cs.waikato.ac.nz/ml/weka/

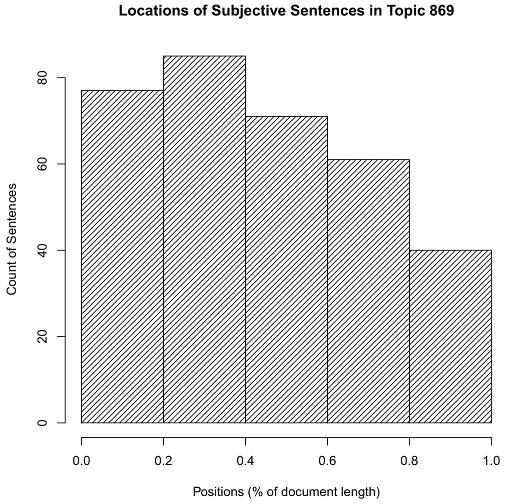


Figure 6: Positions of Subjective Sentences in Topic 869

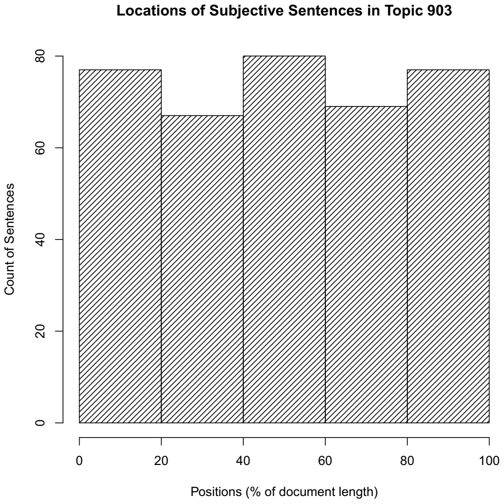


Figure 7: Positions of Subjective Sentences in Topic 903

the other topic.

4.3 The Logistic Regression Model

The logistic regression model describes the relationship between one or more independent variables and a binary response variable, expressed as a probability, that has only two values, which matches our problem. This model predicts the value of the response variable with the logistic function,

$$f(z) = \frac{e^z}{1 + e^z} \tag{5}$$

where $f(z)$ represents the probability of an outcome, and z is the input, whose value is determined by the features. We use the accumulated lengths of sentences with each different sentiment polarity as features, and assume a linear relationship between the features,

$$z(f_1, ..., f_i) = \sum_{i=0}^{n} x_i f_i \tag{6}$$

where x_0 is the intercept of the linear function, and f_0 arbitrarily set to 1. In our experiments, only the accumulated lengths of different types of sentences were used as features, and the predicted class label is given in the form of a probability value calculated with $f(z)$.

When we use all four types of the sentences, the linear function is given in the follwoing form,

$$z(f_i) = x_0 + x_{neu} f_{neu} + x_{neg} f_{neg} + x_{mix} f_{mix} + x_{pos} f_{pos} \tag{7}$$

where x_0 is the intercept of the function, x_{neu} is the accumulated count of the tokens in neutral sentences, x_{neg} is the accumulated count of the tokens in negative sentences, x_{mix} is the accumulated count of the tokens in mixed sentences, x_{pos} is the accumulated count of the tokens in positve sentences.

5 The Classification Experiment

We conducted our experiments on the annotated collection introduced in Section 3. All the documents we used have been annotated by NIST assessors as topical, and bearing either positive or negative opinion. We use their sentiment polarity labels as the golden standard for document polarities, and our sentiment-level annotations as the gold standard for sentence polarities. In this experiment, we only focus on a binary class problem, i.e. only documents labeled as either positive or negative were used, and those labeled as neutral or mixed were discarded. When constructing the corpus we have also intentionally kept an even class distribution.

The objective of this set of experiments is to validate our hypothesis that we can achieve better classification performance on the extracted topical extracts than on the original full posts. The steps below were followed to carry out the experiments,

1. We first generated the topical snippet collection by extracting the topical sentences, whose topicality labels are *Relevant*, from the full posts.

2. Stop words and non-alphabetic symbols were then removed from both collections. Words that appear less than 3 times in the documents of each class (either positive or negative) were also removed for robustness, to be consistent with usual data mining practice.

3. When classifying with the Naive Bayes classifier, each blog post document were transformed into a vector consisting of features which are computed from statistical information of each term. In our experiments we used the tf-idf values as features.

4. Five-fold cross validation was applied to evaluate the performance of the all the approaches. Note that the voting model we proposed is unsupervised. However, by applying the cross validation process we are thus able to compare the robustness of this approach. We did not use ten-fold cross validation mainly because we had only a limited number of documents in the collection.

The experiment results are shown in Figure 8 and Figure 9. With the word-based approaches *full_NB* and *summ_NB*, significant improvement in the classification accuracy has been observed on the extract data set of topic 869. On topic 903 however, the average accuracy dropped dramatically. This is probably caused by the much smaller feature set (254 words) for the extract data set compared to the full post data set, which has 4067 features. Note though, this

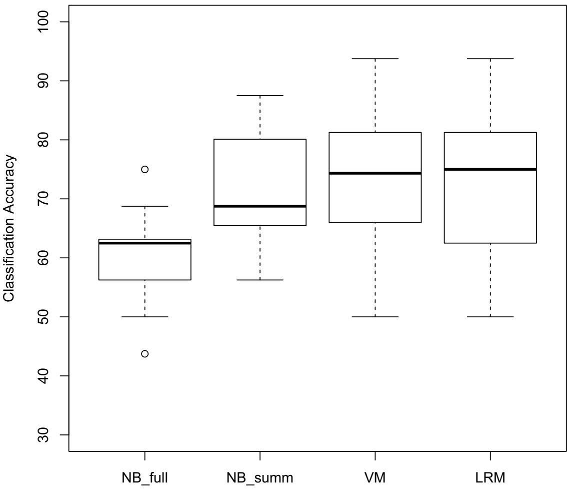


Figure 8: Classification on Topic 869

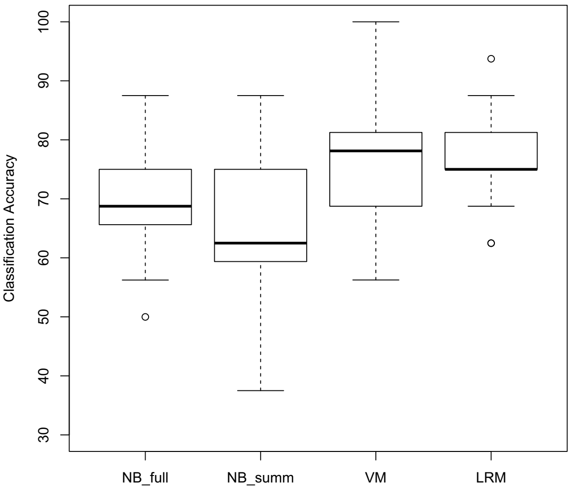


Figure 9: Classification on Topic 903

drop in accuracy is not statistically significant, as is shown in Figure 4. Considering the size of the topical snippets is only 13.64% of the full post on 869 and 14.37% on 903, this system may still be favored in a commercial setting. However, we cannot arrive at the conclusion that the system based on topical snippets has higher efficiency, as the computation cost of snippet generation is subject to further study, and a larger data set with more topics is needed to evaluate how the system works in different domains.

For the voting model *summ_VM*, we used the accumulated count of the tokens to evaluate the amount of the positive and negative content. As mentioned in Section 3.1, some noise were introduced when splitting the document into sentences, where either a group of sentences were treated as one, or one sentence was splitted into several. With our evaluation measure, only the former case (a group of sentences were treated as one) will affect the classification process, in which case the amount of the content with the labeled polarity will be inflated. Such cases are quite rare in the collection. From Figure 8 and Figure 9 we can see that the voting model outperformed the word-based model on both topics. However, as is shown in Table 5, its improvement over the word-based approach is not statistically significant on topic 869.

For the logistic regression model introduced in section 4.3, we used the accumulated count of tokens in each type of the sentences as features. The results shown in Figure 8 and Figure 9 were based on only two types of sentences, namely, positive and negative. The performance of this approach is consistently better than the performance of the approach that uses the full post. Although preliminary, this result suggests that the topical extracts are not only highly abridged in size, but may also provide a more accurate presentation of the opinions shown in the posts towards the specific topic.

We have also experimented with all four types of the sentences (*neutral, negative, mixed and positive*) with the logistic regression model. Note that this experiment was only applicable on topic 903, as documents with sentences labeled as *mixed* are too rare on topic 869 to carry out five-fold cross-valiation. Interestingly, we observed an accuracy of 75.72% with four features but 76.88% with two features (using only positive and negative sentences). This suggests that sentences labeled as *mixed* or *neutral* may not be useful in identifying the document polarity. This is also reflected in the coefficients x_i learnt in the five fold cross validation process, as is shown in Table 3. Strong correlation between positive sentences and positive documents is observed, and negative sentences correlates to negative documents; whereas the correlation between the neutral sentences and the negative documents is very weak, and no reliable correlation between mixed sentences and the document polarity is observed.

Table 3: Analysis of the linear function learnt

Fold	x_0	x_{neu}	x_{neg}	x_{mix}	x_{pos}
1	0.4302	-0.0029	-0.0435	0.0399	0.0106
2	0.8287	-0.0065	-0.0444	0.2237	0.0135
3	0.4220	-0.0034	-0.0365	0.0364	0.0097
4	1.0508	-0.0066	-0.0812	-0.0463	0.0171
5	0.6475	-0.0004	-0.0400	0.0407	0.0056

[1] The logistic regression model was trained with R [a]

[a] http://www.r-project.org/

Table 4: Statistical Significance Test (extract vs. full post)

Topic	NB	VM	LRM
869	**0.00076**	**0.00083**	**0.00001**
903	0.20470	0.08988	**0.03371**

[1] The statistical significance test is done with Wilcoxon's signed rank test.

6 Conclusions and Future Work

In this paper we have studied the problem of topic drift in blogs. By manual inspection of the blog posts we observed a high level of topic drift, and hypothe-

Table 5: Statistical Significance Test (the three approaches)

Topic	NB vs. VM	VM vs. LRM	LRM vs. NB
869	0.57300	0.56690	0.79240
903	**0.02797**	0.80480	**0.00470**

[1] The statistical significance test is done with Wilcoxon's signed rank test.

sised that classification on only the relevant portion would be more efficient both in terms of accuracy and efficiency. To validate this hypothesis, we have annotated a small collection of posts from two topics at the sentence level, thus able to produce manual topical extracts as ground truth. We should note that the snippets for both topics are much smaller in size compared to the full posts (13.64% for 869, and 14.37% for 903).

We have proposed two cascaded models that build upon the sentiment polarities of the topical sentences to predict the document polarity. On both topics, the logistic regression model has achieved higher accuracies than what the word-based approach achieved with the original blog posts. The voting model had a similar performance but its improvement over the word-based approach is not statistically significant. Our preliminary experiments have confirmed our hypothesis that classification on the topical extracts may result in better accuracy, but also suggests that approaches based on word-distribution may be less robust than those based on sentence-level polarities when using the extracts.

Our main contributions in this work are first the sentence-level annotations, which could be used for further analysis in the research community, and sencond the new approaches to classify the document polarity with sentiment polarities. Our future work will focus on automating the sentence classification process and expanding the data set, so that the effectiveness of our models can be tested in a more robust context.

References

Das, S. R. & Chen, M. Y. (2007), 'Yahoo! for amazon: Sentiment extraction from small talk on the web', *Management Science* **53**(9), 13751388.

Dave, K., Lawrence, S. & Pennock, D. M. (2003), Mining the peanut gallery: opinion extraction and semantic classification of product reviews, *in* 'Proceedings of the 12th international conference on World Wide Web', p. 519528.

Hatzivassiloglou, V. & Wiebe, J. M. (2000), Effects of adjective orientation and gradability on sentence subjectivity, *in* 'Proceedings of the 18th conference on Computational linguistics-Volume 1', p. 299305.

Hu, M. & Liu, B. (2004), Mining and summarizing customer reviews, *in* 'Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining', p. 168177.

Khan, A., Baharudin, B. & Khan, K. (2011), 'Sentiment classification using sentence-level lexical based semantic orientation of online reviews'.

Kim, S. M. & Hovy, E. (2004), Determining the sentiment of opinions, *in* 'Proceedings of the 20th international conference on Computational Linguistics', p. 1367es.

Ku, L. W., Lo, Y. S. & Chen, H. H. (2007), Test collection selection and gold standard generation for a multiply-annotated opinion corpus, *in* 'Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions', p. 8992.

Lloret, E., Saggion, H. & Palomar, M. (2010), Experiments on summary-based opinion classification, *in* 'Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text', p. 107115.

Macdonald, C., Ounis, I. & Soboroff, I. (2007), 'Overview of the TREC 2007 blog track', *Proc. TREC-2007 (Notebook)* p. 3143.

McDonald, R., Hannan, K., Neylon, T., Wells, M. & Reynar, J. (2007), Structured models for fine-to-coarse sentiment analysis, *in* 'ANNUAL MEETING-ASSOCIATION FOR COMPUTATIONAL LINGUISTICS', Vol. 45, p. 432.

Nasukawa, T. & Yi, J. (2003), Sentiment analysis: Capturing favorability using natural language processing, *in* 'Proceedings of the 2nd international conference on Knowledge capture', p. 7077.

Ounis, I., De Rijke, M., Macdonald, C., Mishne, G. & Soboroff, I. (2006), Overview of the TREC-2006 blog track, *in* 'Proceedings of TREC', Vol. 6.

Pang, B. & Lee, L. (2004), A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts, *in* 'Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics', p. 271.

Pang, B., Lee, L. & Vaithyanathan, S. (2002), Thumbs up?: sentiment classification using machine learning techniques, *in* 'Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10', p. 7986.

Popescu, A. M. & Etzioni, O. (2005), Extracting product features and opinions from reviews, *in* 'Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing', p. 339346.

Saggion, H. (2008), 'SUMMA: a robust and adaptable summarization tool', *Traitement Automatique des Languages* **49**, 103125.

Turney, P. (2002), Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews, *in* 'Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL'02)'.

Wilson, T., Wiebe, J. & Hoffmann, P. (2005), Recognizing contextual polarity in phrase-level sentiment analysis, *in* 'Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing', p. 347354.