Effectiveness Gain of Polarity Detection Through Topic Domains

Faiza Belbachir
University of Toulouse, IRIT UMR 5505 CNRS
118, route de Narbonne
F-31062 Toulouse cedex 9. France
Faiza.Belbachir@irit.fr

Malik M. S. Missen
The Islamia University of Bahawapur
Departement of Computer Science
and IT. Pakistan
Saad.missen@gmail.com

***Abstract*—Most of the work on polarity detection consists in finding out negative or positive words in a document using sentiment lexical resources. Indeed, some versions of such approaches have performed well but most of these approaches rely only on prior polarity of words and do not exploit the contextual polarity of words. Sentiment semantics of a term vary from one domain to another. For example, the word "unpredictable" conveys a positive feeling about a movie plot, but the same word conveys negative feeling in context of operating of a digital camera. In this work, we demonstrate this aspect of sentiment polarity. We use TREC Blog 2006 Data collection with topics of TREC Blog 2006 and 2007 for experimentation. The results of our experiments showed an improvement (95%) on polarity detection. The conclusion is that the context plays a role on the polarity of each word.**

Keywords-opinion; polarity; blogs; information retrieval; query categorization.

I. INTRODUCTION

Opinion retrieval aims at relating documents that are both relevant to the query (topic) and express opinions about it. It suffers from problems that are different from the ones that occur in classical information retrieval where the subject is identified only by keywords [14][15].

The opinion conveyed by a text can be expressed by very subtle and varied words, therefore it is often difficult to exactly determine it. The classification of sentiments (polarity) is a sub-task in opinion detection [23][27]. It consists in determining whether an opinion in a given document is positive or negative, which has been challenged at Text Retrieval Conference (TREC) Blog Track since 2006 [28]. The approaches explored by track participants can be devised in two types of approaches for opinion and polarity detection. Some of them are based on the lexicon of opinion words, others on machine learning [17][20].

The first type of approach uses a lexicon of opinion words. This lexicon can be general (such as SentiWordNet [21], General Inquirer [22], Subjective Lexicon [25]), built manually or generated automatically from the corpus (words that contain an opinion are taken directly from the corpus). Each word in the lexicon is associated with opinion and polarity scores. These scores are exploited by different approaches to compute the opinion (or polarity) score of a document. A simple method is to assign a score equal to the total number of words containing an opinion (or polarity) in the document [4][20].

The second type of approach is based on machine learning. This type of approach has two aspects: the level of the features (it is the characteristics of opinion word that determine whether a document contains opinions or not), and the type of classifier. The main features that are used are: single words, bi-grams, trigrams, part of speech and the main classifiers that are used in the polarity detection are: SVM, Naive Bayes, Logistic Regression [5][20]. Other works use a mixed approach (machine learning and lexicon) [13][14].

However, most of previous work do not take into account the context of words. The context can be defined by negation, word senses, syntactic role of words around the given word, intensifiers (or diminishers), or the domain of the topic. The prior polarity of a word is sometimes subject to changes under its context. The new polarity of the word defined by its context is called its contextual polarity. Let us take examples to illustrate what contextual polarity is:

- Negation: Polarity assigned to the term happy is positive, but if this term is preceded by negation word such as "not" or "never", its polarity changes and becomes negative.
- Word sense: the word "Car" has different meanings. For example it means "a motor vehicle with four wheels; usually propelled by an internal combustion engine" or "the compartment that is suspended from an airship and that carries personnel and the cargo".
- Intensifiers: "very bad" (intensifiers), "little problem" (diminishers).
- Domain of topic: the word "unpredictable" gives a positive feeling while writing a movie plot but the same word is negative about the features of a digital camera.

The above examples show that a word changes meaning (polarity) according to several characteristics (Negation, Word sense, Domain of topic). These characteristics are part of polarity context. We are interested in one part of the polarity context, it is the domains of the topic. Our basic assumption is that a word changes its polarity from one topic to another, e.g., "unpredictable". To investigate this question

 ISBN: 978-1-61208-293-6

we propose to categorize the topics into classes (domain), so that an opinion word has the same polarity for all topics of the same class. Then, we determine the polarity for each class.

In this paper, we show the impact of the context in the polarity detection by conducting experiments on data sets of various domains. We use TREC Blog (Text Retrieval Conference) 2006 Data collection with topics of TREC Blog 2006 and 2007 for experimentation purposes [19]. We use a machine learning system and simple features as number of positive words, number of negative words, number of neutral words, and the number of adjectives in a text to the polarity detection. We categorize the topics into six classes (Films, Person, Organization, Event, Product, Issue), and show that this categorization improves the opinion detection. The goal isn't to use sophisticated level of linguistic analysis but it is to show the impact of topic domain on polarity detection.

The remainder of this paper is organized as follows. In the Section 2, we present the related work. In Section 3, we describe the Text Retrieval Conference (TREC). Sections 4, 5 and 6 describe our experiments. We, then, conclude the paper and give some remarks about the related future work.

In this work, we have evaluated the effectiveness of using topic domains on sentiment detection using a standard data collection. It is found that using topical knowledge of topics helps increasing effectiveness of sentiment detection.

II. Related work

Few works exist that have proposed approaches to identify the contextual polarities in opinion expressions [7][9][12]. Yi, Nasukawa, Bunescu and Niblack [9] use a lexicon and manually developed high quality patterns to classify contextual polarity. Their approach shows good results with high precision (75-95%) over the set of expressions that they evaluate.

Popescu and Etzioni [7] use an unsupervised classification technique called relaxation labeling [10] to recognize the contextual polarity of words. They adopt a three-stage iterative approach to assign final polarities to words. They use features that represent conjunctions and dependency relations between polarity words.

Suzuki, Takamura and Okumura [12] use a bootstrapping approach to classify the polarity of tuples of adjectives and their target nouns in Japanese blogs. Negations (such as "only" and "not") were taken into account when identifying contextual polarities. The problem with the above approaches is their limitation to specific items of interest, such as products and product features, or to tuples of adjectives and nouns.

In contrast, the approach proposed by Wilson, Wiebe and Homan [11] classifies the contextual polarity of all instances of the words in a large lexicon of subjectivity clues that appear in the corpus. Included in the lexicon are not only adjectives, but nouns, verbs, adverbs, and even modals. They dealt with negations on both local and long-distance levels. Besides this, they also included clues from surrounding sentences. It was the first work to evaluate the effects of neutral instances on the performance of features for discriminating between positive and negative contextual polarity.

III. Text Retrieval Conference TREC

Text Retrieval Conference (TREC) was stated in year 1992 with the sponsor of U.S. Department of Defense and U.S. National Institute of standards and Technology (NIST). The objective of the TREC is to support and encourage IR by providing an infrastructure for evaluation of text retrieval methodologies. This infrastructure is composed by: a test data collection (Table I), a set of queries (Table II) and a set of relevance assessments (qrels) (Table III).

Table I
TREC Blog 2006 Collection Details [28]

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

Table II
Standard TREC Blog Topic Format

<top>
<num> Number: 851 </num>
<title> March of the Penguins </num>
<desc> Description:
Provide opinion of the film documentary "March of the Penguins".
</desc>
<narr> Narrative:
Relevant documents should include opinions concerning the film documentary "March of the Penguins".
Articles or comments about penguins outside
the context of this film documentary are not relevance.
</narr>
</top>

Many tracts are considered by TREC as blog Track, many tasks are defined in this Track for example: Opinion Finding Retrieval Task and Polarity Opinion Finding Retrieval Task. Several data collection with their relevance judgments(baseline) for different IR tasks were provided par TREC. For the blogs Track, TREC has released two data collections: Blog 2006 and Blog 2008. From 2006 to 2009,

 ISBN: 978-1-61208-293-6

Table III
TREC BLOG RELEVANCE JUDGEMENTS LABELS

Label	Caption	Description
-1	Not Judged	A label of -1 means that this document was not examined at all due to offensive URL or Header
0	Not Relevant	The post and its comments are not at all relevant to the topic
1	Relevant	The post or its comments contain some information about the topic but no opinion found about the topic concerned
2	Relevante, Negative Opinions	The post is relevant and contain a negative sentiment for the topic
3	Relevant, Mixed Positive and Negative Opinions	The post is relevant and contain both positive and negative opinions about the topic
4	Relevant, Positive Opinions	The post is relevant and explicitly positive about the topic

Table IV
JUDGMENT OF ANNOTATORS FOR DIFFERENT TOPICS

ANNOTATOR 1	**ANNOTATOR 2**						
	TV	**PE**	**OR**	**EV**	**PR**	**IS**	**TOT**
TV	12	0	0	3	0	1	**16**
PE	0	20	1	0	0	0	**21**
OR	0	0	14	3	0	2	**19**
EV	0	1	0	7	0	2	**10**
PR	0	0	1	0	13	3	**17**
IS	0	0	0	2	0	15	**17**
TOT	12	21	16	15	13	23	**100**

Table V
THE FINALE TOPIC CATEGORIZATION

CLASS	**TOPICS 2006**	**TOPICS 2007**	**TOT**
TV	9	3	**12**
PE	11	10	**21**
OR	9	8	**17**
EV	3	10	**13**
PR	5	10	**15**
IS	13	9	**22**

TREC has been providing 50 new topics each year. For our work, we choose to evaluate experimentation using TREC blog 2006 data collection with topics of year 2006 and 2007.

IV. CATEGORIZATION OF TOPICS

We propose to classify the topics of TREC blogs 2006 and TREC blogs 2007 into six classes: TV (TV), Person (PE), Organization (OR), Event (EV), Product (PR), Issue (IS). This categorization was built manually and inspired by [20] (Table IV).

Each topic of TREC blog 2006 and 2007 was marked by two people (PHD students) called annotators. In the instructions, annotators were asked:

- to read the descriptions, the title of each topic.
- to assign one class among the available classes.

We showed that there is small disagreement (Kappa = 0.77) between the annotators: for the topics of year 2007 "15 disagreements" and only one for 2006. To solve the disagreements of the two annotators, a third annotator was asked to classify these topics. Table V shows the results.

We conducted experiments on the polarity detection using this topic categorization. We worked only with relevant documents of these topics. We then analyzed the effects of this categorization. We performed experiments in two phases. In the first phase, we performed experiments of polarity detection without categorization of topics. In the second phase, we use the categorization of topics to detect polarity. The result of those experimentations was compared with the relevance judgment of TREC.

V. POLARITY DETECTION WITHOUT CATEGORIZATION OF TOPICS

We used a logistic regression model for our experiments. We chose some simple and common features of polarity detection (number of positive words, number of negative words, number of neutral words, and the number of adjectives), as already used in [1]. The experiments for the polarity detection without categorization of topics are devised in three different environments. All experiments and their parameters are explained below:

A. *First experiment*

The experiment was performed using the same features as those explained above. A cross-validations were performed for topics of 2007. The evaluation measures being used to report results are MAP (Mean Average Precision) and P@10 (Precision at 10 documents). More these measures are higher, more the detection of polarity is better. Table VI shows the results of polarity finding MAP and Precision. In this experiment, the data used in the learning phase are much larger than the data used for the testing, because of that the results are not significant. Therefore, before discussing other causes that could improve these results, we conduct another experiment using a small number of learning data for experiments without topics categorization.

Table VI
RESULTS OF THE FIRST EXPERIMENTATION

RUN	POS		NEG	
	MAP	P@10	MAP	P@10
EXPERIMENT 1	0.099	0.200	0.065	0.060

B. *Second experiment*

In this context, the learning data was reduced from 40 to 22 topics. 22 is the maximum number of topics in a group categorization (Table IV) and the choice of topics of the test was done in numerical order: the first test was done for the topics from 901 to 910, the second for the topics from 911 to 920, the third for topics from 921 to 930, the fourth for topics from 931 to 940 and the fifth for topics from 941 to 950.

Table VII
RESULTS OF THE SECOND EXPERIMENTATION

RUN	POS		NEG	
	MAP	P@10	MAP	P@10
EXPERIMENT 2	0.163	0.200	0.062	0.058

The problem that can arise is that the topics of the same class may be in the test and in the learning, which should be avoided. Therefore, we conduct another experiment using a third parameter.

C. *Third experiment*

For this experiment, we wondered about performance when an item of an unknown class has to be processed for polarity detection. To test robustness, we designed an experiment for which we train the classifier on all classes (e.g., Event, Product, TV, Person, Organization) but one (e.g., Issue which acts as the unknown class). For the testing phase, we submitted topics of "Issue" class to the classifier and measured performance. This process was repeated for all the 6 classes. Then, we averaged the results, which are showed in Table VIII. Notice that this intends to evaluate our classifier in the worst situation.

The results of this last experiment are even worse than other results. This leads to the conclusion that a model learned from a data of this topic is not suitable for data of another topic. Next, we present our experimentation with the categorization of topics, and compare the results.

Table VIII
EXPERIMENTS ON THE POLARITY WITHOUT CLASSIFICATION OF TOPICS

RUN	POS		NEG	
	MAP	P@10	MAP	P@10
EXPERIMENT 3	0.055	0.072	0.036	0.054

VI. DETECTION OF POLARITY WITH CATEGORIZATION OF TOPICS

In this section, we used the same features as those used for the detection of polarity without categorization, namely: number of positive words, number of negative words, number of neutral words, and the number of adjectives. A group of topics has been created for each class. We considered the topics of TREC 2006 and TREC 2007 classified in six classes (films, person, organization, event, product, issue). (N-1) cross validation was performed among topics in each group, using a Logistic Regression model [17], where N is the number of classes (6).

The comparison between "with categorization of topics" and "without categorization of topics" (that is, in the worst case) intends to show the benefit of topic categorization. The results are shown in Table IX, where MAP and P@10 are averaged across all topics.

Table IX
COMPARISON OF RESULTS FOR THE POLARITY

	POS		NEG	
	MAP	P@10	MAP	P@10
WITHOUT CATEGORIZATION	0.055	0.072	0.036	0.054
WITH CATEGORIZATION	0.109	0.146	0.068	0.068
% IMPROVEMENT	98.18	102.77	85.24	26.87

These results show that the classification of topics has improved the results for all experiments that have been made. We considered the last experiment (the third experiment in Section 4) as the baseline for comparisons because it represents the worst case situation (with a new class to process). A considerable improvement (98.18% Map) can be noted in the results. These results showed that the categorization of topics can improve the detection results of the polarity. It should be noted that the purpose of this work was not to improve the previous work to detect the polarity, but rather to analyze the effects of classification on the task of detecting opinions.

Figures 1 and 2 show an improved measurement of each MAP TREC topic 2007 for positive and negative polarities. These figures showed that the topics for which significant improvement (was validated through t-test (with $p < 0.05$)) was found in both polarities, are those belonging to classes "Event","Issue" or "Person"(902, 907, 908, 924, 938, etc.).

 ISBN: 978-1-61208-293-6

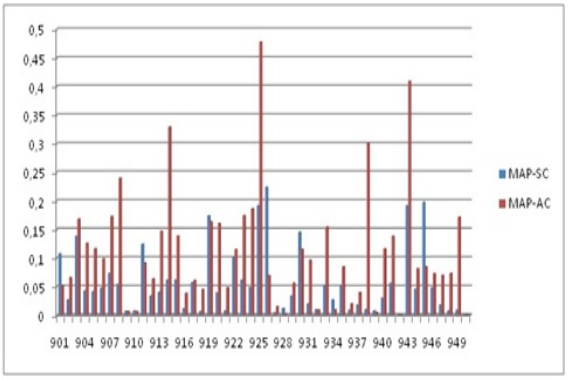

Figure 1. The result of Positive MAP in the various topics of TREC, using the two approaches: with categorization "MAP-AC" and without categorization "MAP-SC".

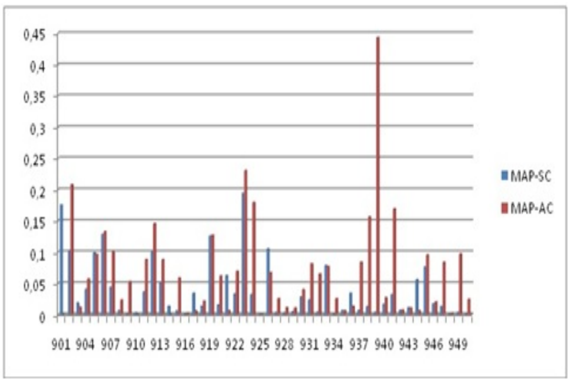

Figure 2. The result of Negative MAP in the various topics of TREC, using the two approaches: with categorization "MAP-AC" and without categorization "MAP-SC".

Tables X and XI show the improvement of few topics of this classes. The MAP of positive and negative words for the first approach ("without categorization") is very low compared to the second approach ("with categorization").

Table X
THE RESULT OF FEW TOPIC (TREC) FOR POSITIVE WORDS

TOPIC	MAP-SC	MAP-AC	Improvement %
902	0.027	0.066	144.444
907	0.073	0.172	134.690
908	0.0541	0.239	342.513
924	0.047	0.186	288.726

One reason why a significant improvement is obtained in these classes may be due to the number of topics in the training data sets. Topics number of class "Issue" and class "Person" are, respectively, 22 and 21.

Table XI
THE RESULT OF FEW TOPIC (TREC) FOR NEGATIVE WORDS

TOPIC	MAP-SC	MAP-AC	Improvement %
902	0.102	0.207	102.239
907	0.044	0.100	127.272
908	0.006	0.023	270.312
924	0.031	0.179	463.836

However, this justification does not hold for the class "Events" where we have 13 topics in total which is less than the class "Org" (17 topics) and the class "Prod (15)". One possible reason could be the classification itself of the topics. We observed that most conflicts encountered during the categorization of topics were to decide between the topics classified as an "Event" and the topics classified as "Issue". For example, it was difficult to decide whether the "Speech" of the president is an "Issue" or an "Event".

VII. CONCLUSION AND FUTURE WORK

Our work focuses on the detection of polarity in blogs. We assume that the context plays a role on the polarity of each word. One word changes meaning (polarity) when used in different subjects. We proposed two approaches. The first approach uses simple features to determine the polarity. The second approach introduces a categorization of topics and documents relevant to these topics. For each class we use the simple features and Logistic Regression classifier. A comparison of these two methods is made, the second method gives better results than the first with more than 95% improvement. The conclusion is that the domain context improves the result for the polarity detection.

In our work, the ranking of the topics was built manually; in the future, we propose to use categorization algorithms of machine learning (i.e., Support Vector Machine (SVM) [29]) and directory services (Yahoo, Dmoz, etc.) and use different features for each class to improve the polarity detection.

REFERENCES

[1] L. Hoang, S. Lee, G.Hong, J. Lee, and H. Riml, "A Hybrid Method for opinion finding task", (KUNLP at TREC Blog Track), 2008.

[2] C. Yejin, C. Claire, R. Ellen, and P. Siddharth, "Identifying sources of opinions with conditional random fields and extraction patterns", in Conference on Human Language Technology and Empirical Methods in Natural Language Processing, Vancouver, British Columbia, Canada, 2005, pages 355-362.

[3] D. Hannah, C. Macdonald, J. Peng, B. He and I. Ounis, University of Glasgow at TREC 2007, " Experiments in Blog and enterprise Tracks with Terrier", in TREC: Proceedings of the Text Retrieval Conference, 2007.

[4] K. Yang, N. Yu, and H. Zhang, WIDIT in TREC 2007 Blog Track, "Combining Lexicon-Based Methods to Detect Opinionated Blogs", in TREC: Proceedings of the Text Retrieval Conference, 2007.

[5] M. M. S. Missen and M. Boughanem, "Sentence-level opinion-topic association for opinion detection in blogs", ACIS-ICIS, 2009, pages 733-737.

[6] P. Kolari, A. Java, T. Finin, T. Oates, and A. Joshi, "Detecting spam blogs: A machine learning approach", in proceeding AAAI, 2006, pages 1351-1356.

[7] A. Popescu and O. Etzioni, "Extracting product features and opinions from reviews", in HLT '05: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing, Morristown, NJ, USA, 2005, pages 339-346.

[8] Y. Suzuki, H. Takamura, and M. Okumura, "Application of semi-supervised learning to evaluative expression classification", in Proceedings of CICLing-06, the 7th international conference on Computational Linguistics and Intelligent Text Processing, Mexico City, MX, 2006, pages 502-513.

[9] J. Yi, T. Nasukawa, R. Bunescu, and W. Niblack, "Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques", in Proceedings of the Third IEEE International Conference on Data Mining (ICDM-03), Washington, DC, USA, 2003, pages 427-434.

[10] R. Hummel and S. Zucker, "On the foundations of relaxation labeling processes", Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1987, pages 585-605.

[11] T. Wilson, J. Wiebe, and P. Homann "Recognizing contextual polarity in phrase-level sentiment analysis", in HLT '05: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing, Morristown, NJ, USA, 2005, pages 347-354.

[12] Y. Suzuki, H. Takamura, and M. Okumura, "Application of semi-supervised learning to evaluative expression classification", in Proceedings of CICLing-06, the 7th international conference on Computational Linguistics and Intelligent Text Processing, Mexico City, MX, 2006, pages 502-513.

[13] Y. Lee, S. Na, J. Kim, S. Nam, H. Jung, and J. Lee, "KLE at TREC 2008 Blog Track: Blog Post and Feed Retrieval", in TREC Proceedings of the Text Retrieval Conference, 2008.

[14] T. Huifeng, T.Songbo and C. Xueqi "A survey on sentiment detection of reviews", Journal Expert System with Application 36, 2009, volume Special Publication, pages 10760-10773.

[15] R. Santos, B. He, C. Macdonald and I. Ounis "Integrating proximity to subjective sentences for blog opinion retrieval", in ECIR, Toulouse, France, 2009, pages 325-336.

[16] S. Na, Y. Lee, S. Nam, and J. Lee, "Improving opinion retrieval based on query-specific sentiment lexicon", in ECIR 09, Proceedings of the 31th European Conference on IR Research on Advances in Information Retrieval, Berlin, Heidelberg, 2009, pages 734-738.

[17] C. Fautsch and J. Savoy, "UniNE at TREC 2008: Fact and Opinion Retrieval in the Blogosphere", in TREC Proceedings of the Text Retrieval Conference, 2008.

[18] C. Macdonald and S. Ounis, "Overview of the TREC 2007 Blog Track", in Proceedings of the TREC 2007.

[19] E. Voorhees and D. Harman, "TREC Experiment and Evaluation in Information Retrieval". Information Retrieval Journal, Springer, 2008, pages 473-475.

[20] G. Zhou, Joshi H., and C. Bayrak, "Topic categorization for relevancy and opinion detection". In TREC Proceedings of the Text Retrieval Conference, 2007.

[21] A. Esuli and F. Sebastiani, "Sentiwordnet: A publicly available lexical resource for opinion mining", in Proceedings of the 5th Conference on Language Resources and Evaluation (LREC-06), Genao, Italy, 2006, pages 417-422.

[22] Z. Zhang, Q. Ye, R. Law, and Y. Li, "Automatic detection of subjective sentences based on Chinese subjective patterns", in Computer and Information Science, Springer Berlin Heidelberg, pages 29-36.

[23] Y. Choi and C. Cardie, "Adapting a Polarity Lexicon using Integer Linear Programming for Domain-Specific Sentiment Classification", in Proceedings of Conference on Empirical Methods in Natural Language Processing, Singapore, 2009, pages 590-598.

[24] S. Kim and E. Hovy, "Identifying opinion holders for question answering in opinion texts", in Proceedings of AAAI Workshop on Question Answering in Restricted Domains, Pittsburgh, Pennsylvania 2005.

[25] O. Vechtomova, "Using Subjective Adjectives in Opinion Retrieval from Blogs". In Proceedings of Text Retrieval Conference, TREC 2007.

[26] C. Macdonald and I. Ounis, "The TREC Blogs06 collection: creating and analysing a blog test collection". In Proceedings of Text Retrieval Conference, TREC 2006.

[27] H. Yulan, L. Chenghua, and A. Harith, "Automatically extracting polarity-bearing topics for cross-domain sentiment classification", Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Oregon, Portland, 2011, pages 123-131.

[28] I. Ounis, M. Rijke, C. Macdonald, G. Mishne, and I. Soboroff, "Overview of the TREC-2006 Blog Track", In Proceedings of Text Retrieval Conference, TREC 2006.

[29] C. Corinna and V. Vladimir, "Support-Vector Networks", In Proceeding of Kluwer Academic Publishers, 1995, pages 273-297.

 ISBN: 978-1-61208-293-6