UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Departamento de Electrónica e Computación

Ph.D. Thesis

EXPLOITING MULTIPLE SOURCES OF EVIDENCE FOR OPINION SEARCH IN THE WEB

Author:

José Manuel González Chenlo

May 2014

Dr. **David E. Losada Carril**, Profesor Titular de Universidad del Área de Ciencias de la Computación de la Universidad de Santiago de Compostela

HACE CONSTAR:

Que la memoria titulada **EXPLOITING MULTIPLE SOURCES OF EVIDENCE FOR OPINION SEARCH IN THE WEB** ha sido realizada por **D. José Manuel González Chenlo** bajo mi dirección en el Departamento de Electrónica e Computación de la Universidade de Santiago de Compostela, y constituye la Tesis que presenta para optar al grado de Doctor por la Universidade de Santiago de Compostela (Programa de doutoramento en Investigación en tecnoloxías da información del Departamento de Electrónica e Computación).

Mayo 2014

Firmado: Dr. David E. Losada Carril
Director de la Tesis

A mi madre, mi padre y mis abuelos,
a mis tíos y mi primo,
a Marián

Attitude is a little thing that makes a big difference.

Sir Winston Leonard Spencer-Churchill

Success is the ability to go from one failure to another with no loss of enthusiasm.

Sir Winston Leonard Spencer-Churchill

Do what you can, with what you have, where you are.

Theodore Roosevelt

I have nothing to offer but blood, toil, tears and sweat.

Sir Winston Leonard Spencer-Churchill

Acknowledgments

If I had to choose a soundtrack for this thesis I would pick Joe Cocker's version of *With a little help from my friends*. This work would not have been possible without the support and encouragement of many research fellows. First of all, I would like to thank my advisor Dr. David E Losada for his great support and guidance, being an exceptional supervisor and a better friend through these years. The most important gift he gave me was a strong commitment with excellence; and this lesson will endure the rest of my life.

This PhD thesis, and all the related papers were financially supported by Ministerio de Ciencia e Innovación (under projects TIN2012-33867 and TIN2010-18552-C03-03), by Xunta de Galicia (under project PGIDT07SIN005206PR) and by the Galician network on Natural Language Processing & Information Retrieval (2006-2008/2012-2014). Additional funds for research activities related to this thesis were supplied by the Universidade de Santiago de Compostela (USC), "Centro de Investigación de Tecnoloxías da Información" (CiTIUS) and the Intelligent Systems Group (GSI). I also acknowledge the organisers of the ACM Conference on Information and Knowledge Management (CIKM) for giving me a grant to travel to CIKM 2011 conference.

I thank Dr. Jordi Atserias and Dr. Roi Blanco for supervising my work during my research internship at Yahoo! Labs Barcelona. They showed me how a large company as Yahoo! works every day and also taught me a different way of working. They were more than good supervisors for a young researcher, and become the best friends that a visitor could have. Thanks to them and the rest of Yahoo! staff (special mention to Estefanía and Natalia), I felt like one more in Barcelona and in the lab. Additionally, I would also like to thank the rest of members of the Semantic Search Research Group and its leader Peter Mika for their kind support.

I would like to thank all my colleagues from my lab at the Departamento de Electrónica e Computación and Centro de Investigación en Tecnoloxías da Información (CiTIUS) at the Universidade de Santiago de Compostela, with whom I enjoyed so much and celebrated some parties. Particularly, I would like to mention José Carlos (JC) and Ronald, two IR@GSI folks, with whom I shared very special moments of my life. At the same way, thanks to Miguel, Cris and Fabi, for their support through the first steps of my journey. Also special thanks to the people from the IRLab at the University of A Coruña. Thank you Álvaro for your wise comments and inputs. Thank you Javier for being the mirror and the reference of what a researcher should be. Also thanks to Edu, Isma and Xose, so many special moments with them that can be summarised in one sentence: *If you have it you don't need it and if you need it you don't have it*. Thanks to Alexander Hogenboom for working with us to expand the horizons of our research. Thanks to David Elsweiler and Leif Azzopardi for their feedback about our papers.

On the other hand, I want to acknowledge all my friends in general (some of them research buddies), with whom I have spent many good (and bad) moments. Among all of them, I would like to name here Alberto, Angel, Jose María, Fran, César & Coruña's people, Juan Angel, Bea, Noe, Enrique, Victor, Adrián, Fernando, Roi, Arturo, and Juan. Special thanks to the IT guys that work every day to facilitate research and experimentation: Jorge, Fernando, Diego and Óscar from the University of Santiago de Compostela and Diego and Fabián from Yahoo! Labs Barcelona.

I would also like to acknowledge my girlfriend Marian, the most important person in my life. I would not be here without her support and love. She rescued me in the middle of a heavy storm and now she has become the anchor of my life.

All I have done so far has been the result of tons of work and it would be impossible without the help of my family. I would like to acknowledge my parents, my aunt and uncle. If I am here writing these words, specially from where we belong, it's all because of their effort and sacrifice during their life for giving me the opportunities that they did not have. I have nothing but love and respect for them.

Thanks!!!!!!!!!!

Contents

List of Figures

List of Tables

Abstract

In this thesis we study Opinion Mining and Sentiment Analysis and propose a fine-grained analysis of the opinions conveyed in texts. Concretely, the aim of this research is to gain an understanding on how to combine different types of evidence to effectively determine on-topic opinions in texts. To meet this aim, we consider content-match evidence, obtained at document and passage level, as well as different structural aspects of the text.

Current Opinion Mining technology is not mature yet. As a matter of fact, people often use regular search engines, which lack evolved opinion search capabilities, to find opinions about their interests. This means that the effort of detecting what are the key relevant opinions relies on the user. The lack of widely accepted Opinion Mining technology is due to the limitations of current models, which are simplistic and perform poorly. In this thesis we study a specific set of factors that are indicative of subjectivity and relevance and we try to understand how to effectively combine them to detect opinionated documents, to extract relevant opinions and to estimate their polarity. We propose innovative methods and models able to incorporate different types of evidence and it is our intention to contribute in different areas, including those related to i) search for opinionated documents, ii) detection of subjectivity at document and passage level, and iii) estimation of polarity. An important concern that guides this research is efficiency. Some types of evidence, such as discourse structure, have only been tested with small collections from narrow domains (e.g., movie reviews). We demonstrate here that evolved linguistic features –based on discourse analysis– can potentially lead to a better understanding of how subjectivity flows in texts. And we show that this type of features can be efficiently injected into general-purpose opinion retrieval solutions that operate at large scale.

Introduction

Information retrieval (IR) is a computer science branch that deals with finding material of an unstructured nature (usually textual documents) that satisfies an information need from within large collections [MRS08]. An IR system represents, stores, organises, and gives access to massive volumes of information items [BYRN08]. IR tools are designed to effectively and efficiently retrieve information from a given source. It is assumed that the information exists in the source and that a well-formed query will retrieve it. Web search engines (e.g., Yahoo!, Google, Bing) are prototypical examples of IR systems. These systems support different types of information needs (usually expressed as queries) and retrieve a ranked list of documents related to the user query.

IR technology is currently present in personal (e.g., desktop/email/mobile search) and professional (e.g., enterprise search) user environments. Original IR systems were mostly oriented to text. With the rise of the Web, IR technology had to evolve to adapt to this new scenario and to a wide range of information formats, such as text, images, audio or video. Nowadays, document (or text) retrieval is still popular but other tasks have emerged in the shadow of the Web.

Text retrieval is about searching for textual documents –or spans of text within documents– that satisfy an information need. Documents can be stored in a single machine or distributed among several computers (distributed search) and the information need is often expressed by users as a sequence of terms (textual query). The cognitive process to translate an information need into a sequence of terms is difficult and many users fail to clearly express their needs. Many queries are ambiguous or vague. Moreover, users are reluctant to write more than a couple of query terms [SMHM99]. This is a major issue because the query is the driving source of evidence to estimate relevance.

Many advanced search tasks, e.g., in the area of Opinion Mining (OM) –also known as Sentiment Analysis (SA)–, need to go beyond a ranked list of relevant documents. Social networks, blogs and other websites have rapidly emerged to become leading sources of opinions in the Web. Every day, more and more people make their opinions available on the Internet [PL07]. These repositories of opinions have become one of the most effective ways to influence people's decisions. According to a recent study, comments and recommender systems influence online shopping behaviours; and there is a positive relation between comments and recommendations and shopping experience, shopping satisfaction and shopping intention [Che12].

But the consumption of services and products is not the only motivation of the people. Political information is another important factor. In a survey on American adults about social media and politics, Rainie et. al. [RSS+12] concluded that 66% of social media users have employed social media platforms, such as blogs, Twitter or Facebook, to, e.g., post their thoughts about civic and political issues. According to this study:

- 35% of social media users have used social media to encourage people to vote.
- 34% of social media users have used social media to post their own thoughts or comments on political and social issues.
- 33% of social media users have used social media to re-post content related to political or social issues that was originally posted by someone else.
- 31% of social media users have used social media to encourage other people to take action on a political or social issue that is important to them.

Companies are aware of the power of social media and try to monitor their reputation over social networks or blogs, e.g., to infer what people think about their products and to get early warnings about reputation issues. However, web retrieval technologies need to make progress on how to support these opinion seeking tasks in ways that minimise the users' effort.

Given a query, an effective opinion retrieval method needs to: i) search for documents related to the query topic, ii) determine what opinions are conveyed in those documents, and iii) organise those opinions in a comprehensive way to support the user's decision. In order to deal with these subtasks, we have to consider opinions as a core element within the retrieval process [PL07].

Many research efforts have been made to introduce OM capabilities into IR systems. Developing a complete opinion search application involves attacking each of the following problems [PL07]:

1. Determining whether the user is in fact looking for subjective material.
2. Determining which documents or portions of documents contain opinionated material.
3. Identifying the type of sentiments expressed within the opinionated text.
4. Presenting the sentiment information in some reasonable summary fashion.

These challenges have been studied with different sources of data, e.g., film reviews, blogs, or tweets [PL07, OMdR+06, SMM+12, OMLS11]. Most of the efforts focused on the last three problems. The first challenge, determining if the user is looking for opinions, has been shown to be simpler [PL07]. Queries demanding opinions tend to contain indicator terms like *review* or *opinions*. Alternatively, the software application could provide a way for the user to indicate that the system should retrieve opinions. To address the other three problems, the most popular approach is to apply 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 [OMS08]. The first stage usually involves ad-hoc search, where classic Information Retrieval (IR) models work reasonably well. The second and third stages (finding on-topic opinions and organising them in a proper way) are difficult tasks with many unresolved issues (e.g. irony, off-topic opinions or contrasting opinions). Finally, the opinionated extracts need to be presented to the user. Different summarisation approaches (e.g., tag clouds, multi-document summaries, opinion-oriented charts) have been proposed in the literature [PL07]. A simple method to present results is to return a ranking of documents in decreasing order of their estimated subjectivity (or positivity/negativity) with respect to the query.

In spite of the research efforts made in OM, current technology is still not mature. There are many popular engines for topic-oriented web searches (e.g., Google, Yahoo!, Bing) but, the availability of general-purpose opinion search engines is limited. In fact, people often use regular search engines, which lack evolved opinion search capabilities, to find opinions about their interests. This means that the effort of detecting what are the key relevant opinions relies on the user. The lack of widely accepted OM technology is due to the limitations of current models, which are rather rough and perform poorly.

Many opinion mining algorithms are based on global methods that compute document-level statistics to estimate subjectivity. For instance, supervised learning techniques have been applied to construct text classifiers that discriminate subjective material from non-subjective material. This is usually based on document-level features, e.g., frequencies of positive and negative terms provided by external opinion-based lexicons. We argue that the location of key sentiments within documents is a difficult passage-level task that cannot be merely solved with matching or count-based techniques alone. On-topic opinions are often scattered through documents and appear only in certain locations. There is a lack of research on effective and general-purpose technology able to extract opinions from these narrow bits of information.

In this thesis we go beyond current research and we try to make a more fine-grained analysis of the opinions conveyed in texts. Concretely, the aim of this research is to gain an understanding on how to combine different types of evidence to effectively determine on-topic opinions in texts. To meet this aim, we consider content-match evidence, obtained at document and passage level, as well as different structural aspects of the text.

Limits to the research are noted from the start. Our work is focused on relatively lengthy pieces of text (e.g., blog posts or news). We do not consider microblog sites such as Twitter. In our experiments we focus on texts written in English. Most of the techniques proposed here can be adapted and expanded to other languages, but this adaptation and the subsequent analysis is out of the scope of this thesis. It is our intention to focus on large-scale and multi-topic opinion retrieval. In this context, there exists a wide variability of opinionated texts and there is a need of computationally lightweight methods. Therefore, our research is oriented towards general and efficient techniques able to work in a wide range of conditions. A main reference task in this thesis consists of searching for documents –or document parts– that convey opinions about a given query topic. And we are interested in the study of the subjectivity and polarity of these documents. This work does not explore the aggregation of opinions across different texts. This is an interesting summarisation challenge but it is out of the scope of this thesis. Finally, we do not study other aspects related to the authority of opinions, such as opinion credibility or opinion spam detection.

One intended outcome is to identify a specific set of factors that are indicative of subjectivity and relevance and, therefore, could act as a valuable guidance to detect opinionated documents, to extract relevant opinions and to estimate their polarity. A second intended outcome of the study is the proposition of innovative methods and models able to combine different types of evidence –obtained at document and passage level– to determine on-topic

opinions in texts. It is our intention to contribute in different areas, including those related to i) search for opinionated documents, ii) detection of subjectivity at document and passage level, and iii) estimation of polarity. Another important concern that guides this research is efficiency. Some types of evidence, such as discourse structure, have only been tested with small collections from narrow domains (e.g., movie reviews). It is our firm intention to demonstrate that evolved linguistic features –based on discourse analysis– can potentially lead to a better understanding of how subjectivity flows in texts. And we will show that this type of features can be efficiently injected into general-purpose opinion retrieval solutions that operate at large scale. A common goal across the thesis is to propose tangible solutions to mine opinions from massive volumes of multi-topic contents (e.g., the Web).

This thesis consists of five further chapters. In Chapter 1 we situate our study in related literature and we set the research methodology. We review the state-of-the art in Opinion Mining and Sentiment Analysis and study the historical context, current practice, and the role of OM in current IR systems. We also introduce the research methodology as well as the evaluation methods and the set of resources used in this study. Chapters 2, 3 and 4 are the core of this thesis and analyse the three dimensions needed when accounting for on-topic opinions: relevance, subjectivity and polarity.

In Chapter 2 we focus on relevance. First, we introduce sentence-level features to improve the estimation of document relevance. We consider features based on the overlapping between the sentences and the query topics. We also analyse state-of-the-art IR methods using specific social media repositories, such as blogs.

In Chapter 3 we focus on Opinion Finding. We study some structural aspects of documents –e.g., document's parts– as well as the importance of discourse analysis to detect opinions. Specifically, we propose several query expansion techniques to enhance opinion retrieval based on specific parts of social media documents (comments). We also explore the importance of positional information and passage retrieval to detect opinionated blog posts. Finally, we study structural and discourse features to improve the classification of subjective sentences in news articles.

Chapter 4 explores the use of sentence and document level evidence to estimate polarity. First, we define effective and efficient models to detect what are the key on-topic opinions in a document. We define, implement and evaluate different approaches to estimate polarity at sentence level. Then we analyse several aggregation techniques to compute the overall sen-

timent score at document level. Finally we study the role of rhetorical structure information, both at sentence and document level, to enhance polarity estimation.

Last but not least, Chapter 5 contains the conclusions of this thesis and suggests future lines of work.

Chapter 1

Background

This chapter introduces some aspects of the context and current theory in Information Retrieval and other related fields such as Opinion Mining. We pay special attention to some of the problems that arise when we deal with opinionated texts. Ad-hoc search and opinion finding are two main components of many opinion mining tools and, therefore, we summarise here the main advances in these areas.

IR systems have to support at least three different processes [Cro93]: i) representing the documents of a given collection, ii) representing a user's query, and iii) comparing both representations for relevance estimation purposes. IR retrieval systems often utilise inverted indexes to represent documents. An inverted index is a structure that stores a mapping from content, such as words or numbers (i.e., the vocabulary), to their locations within a document or within a set of documents (posting list). The purpose of an inverted index is to allow fast full text search. This comes at a cost of increased processing time when a document is added to the index. The process of adding documents to the index is called indexing. This indexing process usually involves straightforward preprocessing methods such as down-casing the text and splitting it into tokens. Other preprocessing techniques are stemming and stopword removal. Stemming refers to a crude heuristic process that chops off the ends of words in the hope of reducing inflectional forms –and sometimes derivationally related forms of a word– to a common base. This process often removes derivational affixes [MRS08]. The most common algorithm for English, and one that has repeatedly been shown to work effectively, is Porter's algorithm [Por80, MRS08]. Another typical preprocessing stage is stopword removal. Some extremely common words are of little value in helping to select documents that match a user

need. These words are called *stopwords* and can be entirely excluded from the vocabulary of the inverted index [MRS08]. Traditionally, stopwords are taken from a fixed list that includes prepositions, adverbs and other common words. Another strategy for determining a stop list consists of sorting the terms by collection frequency (the total number of times each term appears in the document collection), and then including the most frequent terms into the stop list.

In text retrieval, documents and queries are sequences of terms. To facilitate matching, query words are converted into query terms by following the same preprocessing approach applied for documents (e.g., stemming, stopword removal). Observe that query processing and matching needs to be carried out after the query is entered to the system (online). On the other hand, document preprocessing and indexing can be done without user interaction (offline). However, both processes have to be efficient and scalable to deal with large amounts of data.

Translating an information need into a sequence of terms is difficult and many users fail to clearly express their needs (e.g., a query might be ambiguous). Moreover, web users are reluctant to write more than a couple of query terms [SMHM99]. To tackle this problem, the initial user's query can be modified to produce a better query. Since early days of IR research, Relevance Feeedback (RF) and Query Expansion techniques have been considered as an efficient, effective and natural way to reformulate queries [Roc71]. RF methods use the information provided by relevant documents (obtained from an initial retrieval) to construct a new query that, hopefully, is more precise than the original query [RL03]. However, RF is not always feasible because of the lack of relevance judgements. Pseudo Relevance Feedback (PRF) strategies [CH79, LAD96] do not need explicit relevance judgements and work under the assumption that some of the top documents retrieved by the search system are relevant to the original query. This permits to improve the original query with no user interaction.

Retrieval engines are based on the principle of ranking documents according to their relevance with respect to the query. This means that the IR system needs to incorporate some ranking function able to determine which documents to retrieve (and in which order) according to their estimated relevance to a given user query. This task is commonly defined as *ad-hoc search* [BYRN08]. IR models have rapidly evolved since the early boolean formulations [LF73]. Boolean models are based on boolean representations of queries and set-based representations of documents. These models have severe limitations. For instance, they do not provide a graded relevance score for each document. Therefore, it is not possible to build

a ranking of relevant documents. The vector space model [SWY75], is a popular alternative that represents documents and queries as vectors of terms. Relevance is estimated using some measure of closeness between these vectors (e.g., the cosine of the angle between the query vector and every document's vector). One of the problems of the vector space model is its heuristic nature[1]. Over the years, new models based on strong probabilistic principles have been designed. BM25 [RWJ+94] is one of the strongest and powerful probabilistic models used in the Information Retrieval field. BM25 has shown its merits in many tasks [AMWZ09]. Another important family of retrieval functions are those obtained from Language Models (LMs) [PC98]. In the Language Modelling approach to Information Retrieval, one considers the query as a textual sample and computes the probability of a query as being "generated" by a probabilistic model based on a document. The standard Language Modelling approach in Information Retrieval is the so-called query likelihood, which is based on: i) estimating a statistical Language Model for each document d, and ii) computing the probability of generating the query q for each of the document models.

Besides document-level matching, the use of short fragments of documents, called passages, has been deeply studied. Passage retrieval can benefit retrieval processes in many different ways. For instance, passage ranking provides convenient units of text to return to the user, can avoid the difficulties of comparing documents of different length, and enables identification of short blocks of relevant material amongst otherwise irrelevant text [KZ01].

Another important way to improve textual retrieval systems is to apply Computational Linguistics. For instance, Natural Language Processing (NLP) can be used to increase the quality of the original user's query by incorporating synonyms, correcting misspelled words, detecting name entities, or resolving acronyms and ambiguous terms. Applying Discourse analysis in IR is also an intriguing avenue of research. Discourse analysis is concerned with how meaning is built up in the larger communicative process. Such an analysis can be applied on different levels of abstraction, i.e., within a sentence, within a paragraph, or –typically– within a document or conversation. The premise is that each part of a text has a specific role in conveying the message of a piece of natural language text. Rhetorical Structure Theory (RST) [MT88] is one of the leading discourse theories. The theory can be used to split texts into segments that are rhetorically related to one another. Each segment may in turn be split as well, thus yielding a hierarchical rhetorical structure. Within this structure, text segments can be either nuclei or satellites, with nuclei being assumed to be more significant than satellites

[1] The weights associated to each dimension are often computed from heuristic weighting schemes, such as *tf-idf*.

with respect to understanding and interpreting a text. Many types of relations between text segments exist; the main paper on RST defines 23 types of relations [MT88]. A satellite may for instance be an elaboration on what is explained in a nucleus. It can also form a contrast with respect to matters presented in a nucleus. This linguistically advanced method is promising for IR. For instance, in [LLL12], Lioma et. al. designed a LM that takes into account linguistic information to estimate the relevance of a document to a query in web search. Their experiments showed that features based on discourse analysis lead to important gains in performance over state-of-the-art retrieval methods.

Many of the models and techniques discussed so far are not parameter-free. For example, $k1$ and b are well known parameters in BM25. Likewise, Language Models for IR have different smoothing parameters. These parameters often need to be tuned to adapt the models to specific settings. But parameter tuning is far from trivial. Another issue is how to combine different ranking models. Many models have been proposed in the literature, and it is natural to investigate how to combine them to create more effective retrieval functions. This is, however, not straightforward either. In general, all those methods that apply Machine Learning (ML) technology to solve the problem of ranking are called Learning to Rank methods [Liu11]. These methods have the capability of combining a large number of features to learn how to rank documents according to a training dataset. It is easy to incorporate any new progress on a retrieval model by including the output of the model as one dimension of the features. Such a capability is necessary in real search engines because the complex information needs of web users cannot be merely solved with simple retrieval functions.

1.1 Ad-hoc Search in Opinion Mining

Search for on-topic documents is commonly a first step in many OM systems. It has been shown that the overall effectiveness of opinion search engines is highly influenced by the quality of the initial *ad-hoc* search [MHOS08]. This initial retrieval process is a relatively standard *ad-hoc* search and, therefore, the goal is to retrieve as many relevant documents as possible. However, opinion-rich web resources such as blogs or news have specific characteristics that need to be taken into account. For instance, the presence of noise or adversarial content, and the high proportion of off-topic material within the documents are two additional difficulties that often arise [SMM+12].

Most opinion datasets crawled from the web include the raw content of the web pages. For instance, blog posts usually include noisy data such as links to other blog posts, information related to the blog (but not to the post), and many advertisements. This data might severely harm retrieval performance because documents that have query terms in a wrong context can be retrieved. It is necessary to apply an effective preprocessing strategy to identify the key elements of the permalink (title, post and comments) and discard noisy pieces of information [PLCB10a]. The most successful approach to remove such noisy content is based on textual patterns. This family of methods has been widely applied to discard noisy pieces of text in Web documents [CKP07, Eve08, VdSP+06, PLCB10a]. In the case of the blogosphere, one of the most recognised algorithms is *DiffPost* [NNLL09]. This method discards irrelevant content from blog posts by assuming that posts from the same blog follow the same HTML template. Hence, by performing a *diff* process it is possible to detect the genuine content of each blog post.

Another related issue is the presence of spam. Web spam pages contain information automatically generated to gather audience or to act as a link farm to increase the authority of other sites [CD11]. IR systems are severely affected by this type of noise. For example, in a blog retrieval setting, MacDonald and his colleagues concluded that almost 10% of all retrieved documents are spam [MOS09]. Many researchers approached this problem with supervised learning technology e.g., using state-of-the-art machine learning classifiers (Support Vector Machines) to filter out spam documents [KFJ06, KJF+06, LSC+07, RK13].

Text cleaning and spam detection are important, but relevance estimation is the key strategic stage in building a robust retrieval system of opinionated content. Once the noisy content from Web documents is removed, the estimation of relevance can be regarded as a standard retrieval problem, in which state-of-the-art methods such as LMs and BM25 are expected to work well. However, some characteristics of social media can lead to a poor estimation of relevance. For instance, informal writing style, poor description of the information needs (weak queries), and dispersion of on-topic passages result in a poor overlapping between documents and queries. Several methods have been proposed to mitigate this problem. Some authors experimented with passage retrieval techniques as a way to penalise off-topic parts of the documents that are not related to the query [LhNK+08]. The use of the position and distribution of high-scored sentences to learn a probabilistic model that can distinguish relevance-flow patterns for relevant documents was studied in the context of news retrieval by Seo and Jeon

[SJ09]. However, the impact of these features for retrieving opinionated documents remains unknown.

Another promising technique to improve retrieval performance in opinion repositories is query expansion [SMM+12]. Expansion can be done from external resources (e.g., Wikipedia, Freebase) [WdMDR09] or from document contents of the target collection [LhNK+08]. The use of external resources such as Wikipedia was studied in [JYZ08]. Essentially, the original query was expanded with synonyms of concepts identified in the query. Weerkamp and de Rijke [WdMDR09] used the Wikipedia and news resources that are temporally aligned with the collection of documents as sources of data for expanding the original query. One of the most effective approaches that gets expansion terms from document contents was proposed by Lee et al. [LhNK+08], who employed a passage retrieval process to drive the selection of terms needed to expand the original query.

1.2 Opinion Finding

The second stage in typical OM systems consists of searching for opinions related to the query. These opinions are sought within the estimated relevant documents (obtained from the initial topic retrieval stage). Depending on the nature of the methods applied, research studies in this area can be roughly categorised into two different classes: Lexicon-based methods and classification methods.

1.2.1 Lexicon-based Methods

Lexicon-based approaches work from lists of terms with known semantic orientation (*opinionated lexicon* or *sentiment dictionary*) [SMM+12]. Some studies proposed ad-hoc methods to extract collection-dependent opinionated lexicons [AAB+08, HMHO08]. These lexicons are used to process documents following different techniques (e.g., considering the lexicon terms as features for a Machine Learning method). One of the most interesting studies was done by Gerani et al. [GCC09], who investigated methods to mine the lexicon from training data through feature selection. Their approach tries to determine which are the most discriminative terms for subjectivity purposes. Other authors adapted and refined general-purpose lexicons in order to increase effectiveness in specific scenarios [LhNK+08].

Sentiment lexicons are also important to guide query expansion toward opinionated content. For instance, Huang and Croft [HC09] proposed a query expansion technique based on

Relevance Models that expands the original query with terms provided by both the document collection and external opinionated lexicons. This approach showed satisfactory results.

The information provided by the document structure has been largely ignored in OM. This is unfortunate because, the comments associated to textual entries or posts supply valuable information. This has been demonstrated in, e.g., ad-hoc retrieval [WdR08, Mis07], summarisation [HSL07] and snippet generation [PLCB10b].

Some researchers considered term positional information when searching for opinionated spans related to the query [SHMO09, GCC10]. For instance, Santos et al. [SHMO09] proposed a novel OM approach that takes into account the proximity of query terms to subjective sentences. Gerani et al. [GCC10] designed 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.

1.2.2 Classification Methods

Classification approaches build opinion classifiers based on training data. The classifier learnt is used to determine opinions in a test collection. The first attempt to employ Machine Learning in OM was done by Pang et al. [PLV02]. They worked with movie reviews and analysed the impact of term positions on polarity classification. The experimental results showed 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 appears in the first quarter, last quarter, or middle half of the document. 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 argued that the study of more refined notions of positions could be useful for polarity estimation.

Pang and Lee [PL04] 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). These classifiers were not able to overcome baseline unigram alternatives. Nevertheless, the results obtained showed that the last sentences of a document might be a good indicator of the overall polarity of the review.

Beineke et al. [BHMV04] proposed several sentiment summarisation approaches based on the analysis of data from a popular film reviews website[2]. This study revealed that the first

[2] www.rottentomatoes.com

and the last sentences of the reviews are more important for summarising opinions. In order to analyse the impact of sentence locations, an automatic classifier was built based on two types 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 utilised 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 –within sentences– and positional information of the sentences.

Gerani et al. [GCC09] combined relevance and opinion for building a ranked list of opinionated blog posts. They demonstrated that learning can be suitable for both generating opinion scores for individual documents and detecting opinion-based ranking functions.

In [ML06], Mao and Lebanon predicted the global sentiment of a document by analysing the sentiment flow at sentence-level. Rather than using bag of words classifiers, they modelled the sequential flow of sentiment throughout the document using a conditional model. They defined the concept of local sentiment as the sentiment associated with a particular part of the text and they assumed that the global sentiment of a document is a function of the local sentiments. The experimental results indicated that the flow-based approach is better than a bag of words approach. Estimating the local sentiment was a key step in predicting the global sentiment and Mao and Lebanon also demonstrated the usefulness of the approach in selecting subjective sentences. Their experiments were done with a movie review dataset.

In [ZNSS11], Zirn et al. presented an automatic framework for fine-grained sentiment analysis at sub-sentence level. They estimated polarity of product reviews by jointly combining several sentiment lexicons, neighbourhood information and discourse relations. Their experiments demonstrated that the use of structural features improves the accuracy of polarity predictions.

Finally, Somasundaran et al. [SNWG09] showed the importance of general discourse analysis in polarity classification of multi-party meetings. Related to this, Heerschop et al. [HGH$^+$11] worked with film reviews and used rhetorical features to determine the importance of every piece of text. By dividing the text into important and less important parts, depending on their rhetorical role according to a sentence-level analysis, they were able to outperform a document level approach based on lexicons. One of the main issues that the authors found was the processing time required for identifying/classifying discourse structure in natural language

text. This problem seems to have prevented the application of these methods in large-scale scenarios.

1.2.3 OpinionFinder

OpinionFinder (OF) [WHS+05][3] is a key reference in the field to estimate subjectivity at passage level. It is an effective sentence-level subjectivity classifier that can be used to extract subjective extracts from documents. Some studies have combined OF with query-dependent evidence. For instance, OF was used in [HMO08, SHMO09] to search for opinions related to a query topic.

OpinionFinder 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, objective, or unknown (if it cannot determine the nature of the sentence). Two classifiers are implemented: an accuracy classifier and a 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. OpinionFinder also 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 [RW03, WR05]. In Figure 1.1 we show an example of a sentence tagged by OF. This sentence was tagged as unknown by the precision classifier (*autoclass*1) and as subjective by the accuracy classifier (*autoclass*2) with a confidence value of 32.8 (*diff* score). OF also tagged the presence of the negative terms *very* and *sorry* in the sentence.

OpinionFinder is a well-known method that has been used in many experimentations as a reference baseline. It is also common to combine the information provided by OpinionFinder with other methods to build more advanced OM techniques.

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

```
<MPQASENT autoclass1="unknown" autoclass2="subj" diff="32.8">
      I am <MPQAPOL autoclass="negative">very</MPQAPOL>
        <MPQAPOL autoclass="negative">sorry</MPQAPOL>
            to learn that Henry has been sick.
</MPQASENT>
```

Figure 1.1: Example of a sentence tagged by OpinionFinder. OF marks two terms as negative and, overall, the sentence is classified as subjective by the accuracy classifier and as unknown by the precision classifier.

1.3 Final Remarks

In this section we have briefly reviewed the current state-of-the-art in OM related areas. Most of the studies in the literature either work with a very focused dataset (e.g. movie reviews) or apply topic retrieval as an initial stage. We have also remarked that the nature of opinion repositories in the Web introduces challenging issues. Researchers try to search for relevant opinions by applying passage-level, query expansion and NLP techniques. A common trend among successful proposals is the analysis of narrow parts of the documents to infer the overall sentiment.

Lexicon and classification approaches have shown their merits in different empirical studies. The most successful lexicon methods incorporate some sort of distance measure between the subjective extracts and the relevant parts of the documents. This information seems to be essential to detect on-topic opinions. Nevertheless, most of these approaches do not take advantage of other aspects of the text, such as discourse structure or positional evidence.

Classification is advantageous to reveal collection-dependent features, e.g. to discover terms that may play a specific –domain-dependent– role in terms of opinion. Another important characteristic of classification methods is their ability to handle and combine a wide range of features. For instance, combining classical features (e.g., n-grams) with more advanced linguistic and positional features. However, most of the supervised learning studies were oriented to small-scale classification problems (e.g., small film reviews datasets). Hence, the role of these models and features in large-scale problems remains unknown.

Number of Unique Blogs	100649
RSS	62%
Atom	38%
First Feed Crawl	06/12/2005
Last Feed Crawl	21/02/2006
Number of Feeds Fetches	753681
Number of Permalinks	3215171
Number of Homepages	324880
Total Compressed Size	25GB
Total Uncompressed Size	148GB
Feeds (Uncompressed)	38.6GB
Permalinks (Uncompressed)	88.8GB
Homepages (Uncompressed)	20.8GB

Table 1.1: Main statistics of the BLOGS06 collection. This collection was utilised in the TREC 2006, TREC 2007 and TREC 2008 blog tracks.

1.4 Evaluation Methodology: Collections and Metrics

In this section we review some standard benchmarks that have been constructed over the years to facilitate experimental research in OM.

1.4.1 TREC Blog Track

The Text REtrieval Conference (TREC)[4], co-sponsored by the National Institute of Standards and Technology (NIST) and the U.S. Department of Defense, was initiated in 1992. Its purpose is to support research within the Information Retrieval community by providing the infrastructure necessary for large-scale evaluation of text retrieval methodologies. The TREC blog track has been one of the most renowned evaluation challenges for large-scale OM [SMM+12]. In this thesis we focus on the TREC 2006, TREC 2007 and TREC 2008 blog track's benchmarks [OMdR+06, MOS07, OMS08]. In these benchmarks opinion seeking was the main aim of the competition[5]. The BLOGS06 TREC collection [MO06] was the reference collection for these tracks. Some statistics of the collection are reported in Table 1.1.

The construction of the BLOGS06 corpus lasted four months and went through the following stages:

[4] http://trec.nist.gov/

[5] TREC blog track was also organised until 2010, but opinion seeking was no longer supported since 2008.

- **Selection of suitable blogs to crawl**: The blogs included in the collection were predetermined before the fetching phase. In total, 100649 blogs were selected for the BLOGS06 collection. These came from several sources, including general interest blogs, such as news, sport, politics (US & UK) or health. Spam blogs were also manually included to study their impact on retrieval algorithms.
- **Fetching the appropriate content from the Web**: The content of the collection was fetched over an eleven week period. Fetching the content from the blogs over this period was broken down into two tasks: regularly fetching the feeds and homepages of each blog; and fetching newly found permalinks (i.e. blog posts) that were extracted from the feeds.
- **Organising the collection into a reusable form**: The collection was organised in a day-by-day format, one directory for each day of the collection. For each day, the feeds, homepages, and permalink documents were placed in separately named files. Each feed, homepage, and permalink document were given unique identifiers. A DOCNO uniquely identifies one permalink document. From the DOCNO, it can be determined what day the permalink URL was first discovered, what file number the document is stored in, and the offset within the file.
- **Assessment**: NIST organised the assessment procedures for the opinion retrieval tasks. The judgement of a document for a topic was only made by one assessor, meaning that no assessor disagreement studies can be made. More details about the assessment process will be reported later in this chapter.

Every year a new set of topics was provided and new judgements were made according to the documents retrieved by the participants. Since these judgements are produced by human assessors, it is very difficult to judge every document in the collection. Instead, the assessors are asked to judge a subset of the collection that is formed by taking the top p documents from each participant's ranking (usually $50 \leq p \leq 100$). This set of documents is referred to as the *pool* of documents, and this assessment procedure is known as *pooling* [MRS08]. This method provides a reliable benchmark [Zob98] in which it is possible to assess the performance of different systems, while minimising the effort needed to create the gold standard. Details about the TREC blog track topics are reported in Table 1.2. Each TREC topic contains three different fields (title, description and narrative). In Figure 1.2 we present an example of a TREC topic.

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

Table 1.2: Topics provided in the TREC Blog tracks across different years.

```
<top>
  <num> Number: 1004 </num>
  <title> Starbucks </title>

  <desc> Description:
    What do people think about the Starbucks chain
    of coffee shops?
  </desc>

  <narr> Narrative:
    Any opinion of Starbucks and their products and
    services is relevant.
    Opinions of Starbucks' business practices, their
    ubiquity, etc are also relevant.
  </narr>
</top>
```

Figure 1.2: Example of a TREC Blog Track topic.

During the three years of the TREC Blog Track different tasks were proposed [OMdR+06, MOS07, OMS08]. In this thesis we are concerned with three of them:

– **Topic Retrieval Task (TR).** This task was officially introduced as a sub-task of the competition in 2008 [6]. This is an ad-hoc search task in blogs: *"Find blog posts related to X"*, being *X* the TREC topic.

– **Opinion Finding Task (OF).** This is the only task that ran during the three years. It is about finding opinions related to a specific topic expressed by a TREC topic. The task can be summarised as: *"Find blog posts that express some opinion about X"*, being *X* the TREC topic.

[6] Ad-hoc search was an important part of the competition since the beginning of the track but it was only introduced as an official subtask in 2008.

Task	2006	2007	2008
Topic Retrieval Task (TR)			*x*
Opinion Finding Task (OF)	*x*	*x*	*x*
Polarity Task (PL)		*x*	*x*
Overview paper	[OMdR$^+$06]	[MOS07]	[OMS08]

Table 1.3: Tasks proposed in the TREC Blog track from 2006 to 2008. The last row of the table cites the overview paper that summarises the overall experience of TREC participants in each year of the track.

– **Polarity Task (PL).** The polarity task was introduced in 2007 as a natural extension to the opinion finding task [MOS07]. Initially, this task was seen as a classification task, in which the final orientation of a given post should be determined. In 2008, the task was redefined as a re-ranking task, in which participants had to build a ranking of positive and a ranking of negative blog posts related to a TREC topic. This task can be seen as: *"Find negative (resp. positive) blog posts related to X"*, being *X* a TREC topic.

These tasks are summarised in Table 1.3. After the first two years of the competition, it was observed that most participants in the OF and PL tasks approached the problem as a two-stage process [OMdR$^+$06]. The first stage being topic retrieval (i.e. retrieve on-topic documents given a user query), and the second stage being a re-ranking phase that takes into account opinion-based features. This poses some problems for properly comparing opinion search algorithms. The effectiveness of applying a particular opinion finding technique strongly depends on the initial topic retrieval stage. Therefore, it was not possible to compare different pure OM methods because they were applied to different topic retrieval baselines. To address this issue the Topic Retrieval Task was created in 2008. From the runs submitted by TREC participants to this task, five different runs were selected to provide a standard benchmark to test the effectiveness of opinion detection techniques. 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. These baselines have been commonly used as a reference in many empirical validations to measure the quality of opinion finding and polarity estimation.

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

– relevance level: A post can be relevant, not relevant or not judged with respect to the topic.

– 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 presented in Figure 1.3 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.

Some statistics about the number of documents judged for the tasks are reported in Table 1.4.

Blog Track	Rel.	Subj.	Pos.	Neg.	Mix.
TREC 2006	19891	11530	4159	3707	3664
TREC 2007	12187	7000	2960	1844	2196
TREC 2008	11735	8797	3338	2789	2670

Table 1.4: Number of relevant, subjective, positive, negative and mixed documents in TREC 2006, 2007 and 2008.

1.4.2 Movie Review Benchmarks

Movie Review sites such as *imdb*[7] or *rottentomatoes*[8] are film review repositories that can be easily crawled to build a benchmark. One of the main advantages of these datasets is that opinion judgements can be easily derived from the information provided by the authors of the reviews.

[7] `http://www.imdb.com/`
[8] `http://www.rottentomatoes.com/`

```
 [...]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.
```

Figure 1.3: Example of a blog post judged as positive for the topic *'MacBook Pro'*.

For instance, the overall recommendation of a review is explicitly encoded by user ratings. Depending on the interpretation of the user ratings we can build different types of datasets:

- Opinion benchmarks: Film reviews with ratings up/below a certain threshold are considered as subjective examples. The rest of examples can be considered as neutral.
- Polarity benchmarks: A threshold is defined to split reviews into positive and negative examples. For instance, with a five-star system, a review with more than three stars can be considered positive and a review with less than three stars can be considered negative. One of the most popular polarity datasets in the literature was created by Pang and Lee [PL04]. The testbed is a collection of 1000 positive and 1000 negative movie reviews, which have been extracted from movie review websites[9].
- Graded benchmarks: Collection of documents whose labels represent the full rating scale. This leads to benchmarks to test OM techniques able to detect different grades of polarity. One of the most famous benchmark in this area is the collection compiled in [PL05] [9].

Other datasets are composed of sentences extracted from different sources. For instance, in [PL04], Pang and Lee mined the Web to create a large, automatically-labeled sentence corpus. To gather subjective sentences (or phrases), they collected 5000 film review snippets from *rottentomatoes*. To obtain objective sentences, they took 5000 sentences from plot summaries available from *imdb*.

[9] http://www.cs.cornell.edu/people/pabo/movie-review-data

Movie review collections (or similar datasets, e.g. restaurant/hotel reviews) are popular in experimental studies that apply ML or NLP for Opinion Mining or Sentiment Analysis [PL04, PL05, PL04, HGH+11, BHMV04].

1.4.3 NTCIR-7 English MOAT Research Collection

The NTCIR Workshop[10] is a series of evaluation workshops designed to enhance research in Information Access (IA) technologies including Information Retrieval, Question Answering, Text Summarisation and Information Extraction. This workshop provides large-scale test collections reusable for experiments and common evaluation infrastructures allowing cross-system comparisons. In this thesis we are concerned with the multilingual Opinion Analysis Task (MOAT). The reference collection for this task was the NTCIR-7 (English) MOAT Formal Research Collection, provided by the 7th NTCIR Workshop (2007/2008). The collection contains 14 topics[11], and pre-segmented documents that were assessed as relevant to the topics. The collection also provides annotated data at sentence level. An example of document tagged by NTCIR assessors is presented in Figure 1.4. This information includes both relevance and subjectivity labels, as well as the identification of the opinion holders. In Figure 1.5 we present an example of sentence tagged for the topic *"I would like to know about the background and details of the incident that happened with then Nepalese Royal Family"*. The opinion judgements were made by three different assessors. The English version of these collections contains news from different sources:

- **Mainichi Daily News.** English articles published in Japan in the years of 1998-2001.
- **Korea Times.** English news articles published in Korea in the years of 1998-2001.
- **Hong Kong Standard.** English news articles published in Hong Kong, China PRC in the years of 1998-1999.
- **Straits Times.** English news articles published in Singapore in the years of 1998-2001.

Some statistics about the ground-truth are reported in Table 1.5.
In the NTCIR-7 MOAT, five different subtasks were proposed:

[10] `http://research.nii.ac.jp/ntcir/index-en.html`
[11] Textual representations of user needs. The information provided include title and narrative statements.

Number of News	80
Number of Sentences	3584
Number of topics	14
# of Relevant sentences	878
# of Opinionated sentences	887
# of Positive sentences	179
# of Negative sentences	417
# of Neutral sentences	291

Table 1.5: Statistics from NTCIR-7 (English) MOAT Formal Research Collection.

– Opinion subtask: Systems have to classify each sentence of the collection as subjective or objective.

– Topic relevance subtask: Systems have to classify each sentence as either relevant or non-relevant to the topic.

– Polarity subtask: Systems have to determine the polarity orientation of the sentences with respect to the query.

– Holders & targets task: Participants have to detect the opinion holders and targets within each opinionated sentence.

Despite being a small dataset, NTCIR-7 MOAT research collection is valuable in the context of this thesis. It has subjectivity and polarity judgements at sentence level and, therefore, it allows us to evaluate OM passage-level techniques. This collection is a good complement to the TREC Blog Track, which does not provide a sentence-level ground truth. The next three collections (MPQA, FSD and PL) also supply subjectivity and polarity judgements at sentence level.

1.4.4 Multi-Perspective Question Answering dataset (MPQA)

The Multi-Perspective Question Answering initiative[12] provides a wide range of resources to facilitate research in SA. These resources include annotated corpus, subjectivity lexicon or subjectivity sense annotations. The Multi-Perspective Question Answering dataset (MPQA) is one of the key components of this initiative. It contains news articles manually annotated using

[12] http://mpqa.cs.pitt.edu/

```
<DOC>
  <DOCNO>
    KT2001_06313
  </DOCNO>
  <LANG>
    EN
  </LANG>
  <TEXT>
    <STNO>0001</STNO>
    KoreaTimes : Impact of Attacks on US Beyond Tomorrow
    <SEN_REL type="NA"></SEN_REL>
    <SEN_OP type="NO"></SEN_OP>

    ...

    <STNO>0020</STNO>
    <SUBSEN_ATTITUDE type="SUP">
      <HOLDER HolderNum=-2 HolderText="POST_AUTHOR"></HOLDER>
      <TARGET TargetNum=48 TargetText="selling of stocks for
      cash and other safe investments are likely temporary">
      </TARGET>
      Even<MARKTARGET TargetNum=48> the selling of stocks for
      cash and  other safe investments are likely temporary
      </MARKTARGET>.
    </SUBSEN_ATTITUDE>
    <SEN_REL type="YES"></SEN_REL>
    <SEN_OP type="YES"></SEN_OP>
    <SEN_ATTITUDE type="POS"></SEN_ATTITUDE>

    ...

    <STNO>0028</STNO>
    <SUBSEN_ATTITUDE type="SUP">
      <HOLDER HolderNum=-2 HolderText="POST_AUTHOR"></HOLDER>
      <TARGET TargetNum=55 TargetText=" Americans will resume
        normal economic activities"></TARGET>
      Within a matter of weeks, <MARKTARGET TargetNum=55>
      Americans will resume normal economic activities</MARKTARGET>,
      trusting that government will respond to these attacks
      appropriately.
    </SUBSEN_ATTITUDE>
    <SEN_REL type="YES"></SEN_REL>
    <SEN_OP type="YES"></SEN_OP>
    <SEN_ATTITUDE type="POS"></SEN_ATTITUDE>

    ...

</TEXT>
</DOC>
```

Figure 1.4: Example of News article tagged by NTCIR-7 assessors.

an annotation scheme for opinions and other private states (i.e., beliefs, emotions, sentiments or speculations).

```
<STNO>0022</STNO>

<SUBSEN_ATTITUDE type="SUP">

<HOLDER HolderNum=1 HolderText="Mr Manohar Bikram Thakuri,
 a construction supervisor">
</HOLDER>
<TARGET TargetNum=1 TargetText="I love him more
 than I love myself." ">
 </TARGET>

In Singapore, the depth of feeling in the Nepalese community
here was best encapsulated by
<MARKHOLDER HolderNum=1>
Mr Manohar Bikram Thakuri, a construction supervisor
</MARKHOLDER>
who said tearfully of the slain monarch:
"<MARKTARGET TargetNum=1>I love him more than I love myself."
</MARKTARGET>

</SUBSEN_ATTITUDE>

<SEN_REL type="YES"></SEN_REL>
<SEN_OP type="YES"></SEN_OP>
<SEN_ATTITUDE type="POS"></SEN_ATTITUDE>
```

Figure 1.5: Example of a news article tagged by NTCIR-7 assessors. This sentence has been tagged as relevant, subjective and positive for the topic *"I would like to know about the background and details of the incident that happened with then Nepalese Royal Family"*. The holder of the opinion has been tagged as *Mr Manohar Bikram Thakur, a construction supervisor*, and the opinion is that *"I love him more than I love myself"*.

We followed existing practice [RWW03, RW03] that applies annotations patterns to label sentences as subjective or objective[13]. The same patterns can be easily extended for assigning positive and negative labels. Once we applied these patterns to the sentence collection we obtained 7333 subjective sentences —over a total of 15802– and 4881 polar sentences (1626 positive and 3255 negative). An example of a MPQA sentence can be found in Figure 1.6.

[13]For instance, a sentence that contains a phrase labelled as highly subjective is regarded as a subjective sentence.

```
The 450ft long vessel Nisha, carrying 26,000 tonnes metric
tons of raw sugar, was stopped on Friday morning 21 December
amid fears it could be transporting noxious, hazardous or
dangerous substances.
```

Figure 1.6: Example of a negative sentence found in the MPQA collection.

1.4.5 Finegrained Sentiment Dataset

There are several freely available data sets annotated with sentiment at various levels of granularity, but most of them lack neutral documents. To fill this gap, Täckström and McDonald created the Finegrained Sentiment Dataset (FSD) collection [TM11]. The Finegrained Sentiment Dataset (FSD) collection [TM11] contains 294 product reviews from various online sources. The reviews are approximately balanced with respect to domain (books, DVDs, electronics, music, and videogames) and overall review sentiment (positive, negative, and neutral). Two annotators assigned sentiment labels to sentences. The identified sentence-level sentiment is often aligned with the sentiment of the associated reviews, but reviews from all categories contain a substantial fraction of neutral sentences, as well as both positive and negative sentences. The FSD collection includes a total of 2243 polar sentences: 923 positive sentences and 1320 negative sentences. An example of tagged review can be found in Figure 1.7.

1.4.6 Pang & Lee subjectivity dataset

The Pang & Lee subjectivity dataset (PL) is an automatically labelled sentence corpus [PL04]. To gather subjective sentences (or phrases), 5000 review snippets were crawled from a popular film reviews site[14] (e.g., "bold, imaginative, and impossible to resist"). Sentences estimated as objective were obtained from plot summaries of the Internet Movie Database[15]. Examples of subjective and objective sentences can be found in Figure 1.8 and Figure 1.9, respectively.

1.4.7 Evaluation Measures

We consider several measures to evaluate performance. These measures can be categorised as either measures for ranked retrieval sets or measures for unranked retrieval sets.

[14] www.rottentomatoes.com
[15] www.imdb.com

```
nr        Kevin Vanhoozer is the editor of this book which is
        a compilation of essays about each book of the Old
        Testament (OT).
nr        Each essay is written by a different person who
        sticks somewhat to a similar pattern.
...

pos        * Each chapter looks to see how we can relate the
        OT books to the New Testament, specifically to
        Jesus. * The book is highly readable.
nr        Each author writes in a manner I could describe as
        pastoral.
pos        The goal was to teach the reader, not impress them
        or bore them.
pos        * The book is well researched.
pos        Most authors examine all angles of interpretation
        through the ages and give a decent bibliography at
        the end of each chapter.

...

pos        You could return to it for information over and over
        again.
pos        I think it should be read by any serious Bible student
        as well as any casual Bible reader.
pos        I give it 4 out of 5 stars.
pos        Good work.
nr       -Don-
```

Figure 1.7: Example of a positive book review in FSD. It contains polarity tags for each sentence in the document (first column). For instance, the sentence *"Good work"* was tagged as positive.

Evaluation of unranked retrieval sets

These measures are suitable for retrieval problems in which the output is an unordered set of objects (e.g., documents or sentences)[16]. Most measures for these problems are based on the number of false positives, false negatives, true positives, and true negatives produced by the system (see Table 1.6).

[16]Throughout this thesis, depending on the evaluation task, a retrieval object can be either a document or a sentence.

```
a haunting , rich film...[fraser] and caine blend beautifully
with their sweet-and-sour mix of acting.

the story is familiar from its many predecessors ;
like them , it eventually culminates in the not-exactly
stunning insight that crime doesn't pay.

the first fatal attraction was vile enough.do we
really need the tiger beat version ?
```

Figure 1.8: Example of subjective sentences extracted from `www.rottentomatoes.com`.

```
the movie begins in the past where a young boy named sam
 attempts to save celebi from a hunter.

she, among others excentricities, talks to a small rock,
gertrude, like if she was alive.

renata is a young high-class girl and ulises is a poor guy.
```

Figure 1.9: Example of objective sentences extracted from `www.imdb.com`.

	Predicted Positive	Predicted Negative
Real Positive	t_p	f_n
Real Negative	f_p	t_n

Table 1.6: Confusion Matrix that reports the number of true positives (t_p), false negatives (f_n), false positives (f_p) and true negatives (t_n).

Two common measures for assessing the effectiveness of non-ranked retrieval outputs are precision and recall [MRS08]. Precision (P), is the fraction of retrieved objects that are relevant and is computed as follows:

$$Precision = \frac{t_p}{t_p + f_p} \tag{1.1}$$

Recall (R), is the fraction of relevant objects that are retrieved by the system:

$$Recall = \frac{t_p}{t_p + f_n} \tag{1.2}$$

An alternative to these measures is to judge a retrieval system by its accuracy. Accuracy is the fraction of the system's decisions that are correct [MRS08]:

$$Accuracy = \frac{t_p + t_n}{t_p + f_p + t_n + f_n} \tag{1.3}$$

However, accuracy is not commonly used in retrieval experiments. Most retrieval scenarios are highly imbalanced (i.e., the number of relevant objects is very small when compared with the number of non-relevant objects) and applying accuracy is problematic. For instance, given a dataset in which 90% of the objects are non relevant, a trivial system that marks all documents as non relevant is 90% accurate.

Having separate figures of performance, one can optimise precision or recall depending on the circumstances. For instance, web users would like every web result on the first page to be relevant (high precision) but have not the slightest interest in looking at every document in the collection that is relevant. In contrast, professional searchers such as paralegals and intelligence analysts are very concerned with having high recall, and will tolerate fairly low precision results in order to see every relevant object [MRS08]. However, the two quantities clearly trade off against one another: you can always get maximum recall by retrieving all objects but this likely harms precision. Precision usually decreases as the number of objects retrieved is increased. In general, we want to get some amount of recall while tolerating only a certain percentage of false positives. A single measure that trades off precision versus recall is the F measure, which is the weighted harmonic mean of precision and recall. In this thesis we work with a variation of this measure, the so called $F1$ score [Rij79], which is computed as follows:

$$F1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \tag{1.4}$$

We will employ this metric in situations in which the output of the system is a set (e.g., the output of a binary classifier).

Evaluation of ranked retrieval sets

We consider three different measures to assess the performance of a given ranking of objects: Precision at 10 objects retrieved (P@10), Mean Average Precision (MAP) and Robust Index (RI). P@10 is the proportion of the top 10 retrieved objects that are relevant, i.e.:

$$P@10 = \frac{\#(\text{relevant objects retrieved in the top 10})}{10} \tag{1.5}$$

Given a set of queries, their respective P@10 values are averaged out to get a single P@10 figure. MAP (Mean Average Precision) provides a single-figure measure of quality across recall levels. For a single information need, average precision is the average of the precision values obtained for the set of top k objects existing after each relevant object is retrieved. This value is then averaged over queries [MRS08], i.e.:

$$MAP = \frac{1}{|Q|} \cdot \sum_{j=1}^{|Q|} \frac{1}{m_j} \cdot \sum_{k=1}^{m_j} Precision(R_{jk}) \tag{1.6}$$

where, given the set of relevant objects for a query $q_i \in Q$, R_{jk} is the set of ranked retrieval results from the top result until you get to object o_k , m_j is the number of relevant objects for query q_j , and

$$Precision(R_{jk}) = \begin{cases} \frac{\#(\text{relevant objects retrieved in } R_{jk})}{R_{jk}} & ,\text{when } o_k \text{ is relevant} \\ 0 & ,\text{otherwise} \end{cases} \tag{1.7}$$

Finally, the Robustness Index (RI) of a ranked retrieval set with respect to a baseline ranking was formulated by Sakai et al. in [SMK05]:

$$RI(Q) = \frac{n_+ - n_-}{|Q|} \tag{1.8}$$

where Q is the set of queries over the RI has to be calculated, n_+ is the number of improved queries, n_- the number of degraded queries and $|Q|$ is the total number of queries in q. RI is a measure commonly used in PRF evaluation to compare the robustness of the improvements achieved by systems over a reference baseline.

Across this section we resorted to relevance as the reference notion to obtain the ground truth. However, depending on the task, the ground truth will be composed of the appropriate target objects. For instance, in a system that has to retrieve a ranking of positive blog posts, P@10 will be the proportion of positive blog posts retrieved in the top 10.

Statistical Significance

In IR evaluation, the performance values are averaged out across queries to get a single performance score for a given algorithm. However, this score might be not representative of the overall behaviour of an algorithm. For instance, a system might work very well with a small amount of queries, showing higher performance than another method that improves the

performance of most of the queries with less effectiveness. Although the first method has a higher overall performance, we cannot say that it is better than the second one. This is because the improvements come from a small set of queries and this small sample might not be representative.

To determine if two systems have a significant different behaviour in terms of performance we use statistical significance tests [CMS09]. Most IR evaluation experiments are based on *paired* observations. Two sets of observations are paired if each observation in one set has a special correspondence or connection with exactly one observation in the other set. This is often the case in IR experimentation: when comparing IR algorithms, each query has usually an individual performance figure for each method being compared.

Every significance test is based on a *null hypothesis*. In IR, this hypothesis is that there is no difference in effectiveness between two retrieval algorithms. On the other hand, the *alternative hypothesis* is that there is a difference. The tests are based on assuming the *null hypothesis* true and try to refute it. Most statistical significance tests follow the next steps when comparing two ranking algorithms (A and B):

1. Compute a performance measure for every query in both rankings (e.g., P@10).

2. Compute a *test statistic* based on a comparison of the effectiveness measures for each query (e.g., the difference: $P@10A - P@10B$).

3. The test statistic is used to compute a *P-value*, which is the probability that observations as extreme as the current data would occur if the null hypothesis were true (i.e., assuming that both algorithms perform the same what is the probability that we find this particular difference of performance?).

4. The null hypothesis is rejected if the *P-value* is $\leqslant \alpha$, being α the *significance level*. α values are small, (e.g., 0.05 and 0.01) to reduce the chance of errors.

In our ranking experiments, statistical significance was estimated using the two-tailed t-test. This test assumes that the performance differences follow a normal distribution. The t-test has been shown to be a robust significance test for information retrieval, obtaining similar results than other non-parametric tests [SAC07]. The null hypothesis is that the mean of the distribution of differences is zero ($\mu_A = \mu_B$). The alternative hypothesis is that the mean of

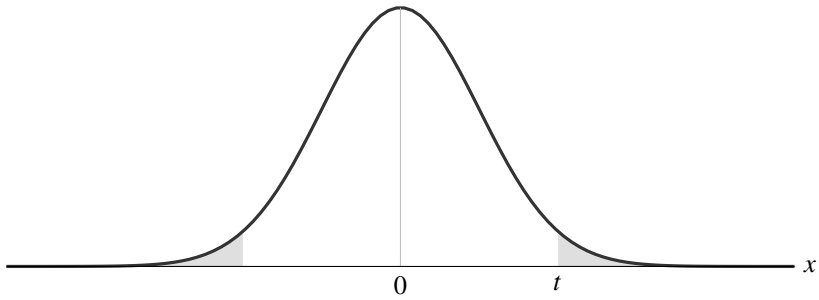

Figure 1.10: Normal distribution for the mean difference of performance. Where t is the test statistic value associated to the data.

the distribution of differences is not equal to zero ($\mu_A \neq \mu_B$)[17]. In Figure 1.10 we sketch an example of two-tailed t-test distribution. The test statistic associated to this paired t-test is:

$$t = \frac{\overline{B-A}}{\sigma_{B-A}} \cdot \sqrt{N} \tag{1.9}$$

were $\overline{B-A}$ is the mean of the differences in performance of both systems, σ_{B-A} is the standard deviation of the differences, and N is the number of queries used in the experiments. If we want to reject the null hypothesis, we have to ensure that the probability mass of the two tails is less that α. This type of tests permit to assess whether a given difference in performance between two systems is significant from a statistical point of view (and not simply due to sampling randomness).

In our classification experiments, we measured statistical significance with a paired, two-sided micro sign test [YL99]. This test compares two systems based on their binary decisions and applies the Binomial distribution to compute the p-values under the null hypothesis of equal performance.

[17] Observe that in a one-tailed t-test the alternative hypothesis would be $\mu_A > \mu_B$.

CHAPTER 2

AD-HOC SEARCH

In this chapter we investigate topic retrieval in blogs. First, we study how to adapt classic IR retrieval models to this scenario. Then, we propose a novel topic retrieval method based on the distribution of salient sentences.

2.1 Information Retrieval and Blogs

In this section we describe the main design issues that need to be considered in order to build a highly effective IR system in blogs. The characteristics of blogs affect the following aspects of the IR system:

– *Retrieval Unit:* Many studies on retrieval in the blogosphere chose the blog permalink[1] as the retrieval unit. The permalink document contains complete information about the blog entry: title, post, comments, and many noisy components that is necessary to deal with. Other studies opted for fine-grained retrieval units. Recently, Lee et al. have successfully applied retrieval at passage level [LhNK+08]. In their study, passage scores were aggregated to build a permalink retrieval system. In general, the selection of the retrieval unit depends on both the application domain and the user's demands.

– *Document preprocessing:* Blog pages are noisy. Most permalink documents contain off-topic data, such as links to other blog posts, information related to the blog (but not

[1] The permanent link to a specific page within a blog or post that remains unchanged. Permalinks are useful for bookmarking or tagging a specific blog post for future reference.

with the post) and many advertisements. These noisy elements might severely harm retrieval performance by promoting documents that have query terms in a wrong context. Effective preprocessing to extract the key components of the permalink (title, post and comments) is required to design a good search system.

– *Topic Retrieval Method:* State of the art search models (e.g., BM25 or Language Models) have performed well to support information retrieval in blogs, being extremely difficult to beat [OMS08].

2.1.1 Retrieval Unit and Document preprocessing

We opted for permalink documents as retrieval units. This is common practice in the TREC blog track and simplifies the evaluation because the track demands constructing a ranking of permalinks (ordered decreasingly by presumed relevance).

To remove noisy elements we built a preprocessing unit that extracts the main permalink components (title, post and comments) and discards the rest of the documents' content. This unit uses a HTML parser[2] to process the structure of permalinks 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 positional information and size. This type of heuristics has been also employed by others researchers in the past [PB07].

2.1.2 Topic Retrieval Method

We adopted BM25 as the reference retrieval model [RWJ+94, RZ09]. BM25 is one of the strongest and powerful methods used since 1994 in the information retrieval field [AMWZ09]. We used the Lemur's implementation of BM25 matching function[3]:

$$w = \log\left(\frac{N-n+0.5}{n+0.5}\right) \tag{2.1}$$

$$BM25(D,Q) = \sum_{t \in Q} w \cdot \frac{(K_1+1)tf_{t,D}}{K_1\left((1-b)+b\times(L_D/L_{ave})\right)+tf_{t,D}} \frac{(K_3+1)tf_{t,Q}}{K_3+tf_{t,Q}} \tag{2.2}$$

where N is the total number of documents in the collection, n is number of documents that contain the term t, $tf_{t,D}$ is the frequency of t in document D, $tf_{t,Q}$ is the frequency of t in

[2] `http://jericho.htmlparser.net/docs/index.html`.
[3] `http://www.lemurproject.org/`.

	MAP		P@10	
	default $b = 0.75$ $K1 = 1.2$	trained $b = 0.3$ $K1 = 0.7$	default $b = 0.75$ $K1 = 1.2$	trained $b = 0.3$ $K1 = 0.7$
TREC 2007	.3489	**.4017▲**	.6080	**.6440**
△%		(+15.13%)		(+5.92%)
TREC 2008	.3237	**.3812▲**	.6340	**.6640**
△%		(+17.76%)		(+4.73%)

Table 2.1: BM25 results in TREC 2007 (topics 901-950) and TREC 2008 (topics 1001-1050) topic retrieval task. The BM25 parameters were trained with TREC 2006 (topics 851-900). Statistical significance was estimated using the t-test at the 95% level. The symbols ▲ and ▼ indicate a significant improvement or decrease over the original BM25 configuration.

query Q, L_D and L_{ave} are the length of document D and the average document length in the whole collection.

BM25 has three parameters: $K1$, which controls term frequency; b, which is a length normalisation factor; and $K3$, which is related to query term frequency. Parameter tuning was done with the TREC 2006 topics, and the TREC 2007 and TREC 2008 topics were used for testing. We fixed $K3$ to 0 (we work with short queries and, therefore, the effect of $K3$ is negligible[4]) and experimented with $K1$ and b values from 0 to 2 and from 0 to 1, respectively (steps of 0.1).

The measures applied to evaluate retrieval performance were mean average precision (MAP) and Precision at 10 documents (P@10). Statistical significance was estimated using the t-test at the 95% level. The symbols ▲ and ▼ indicate a significant improvement or decrease over the default BM25 configuration ($K1 = 1.2$ and $b = 0.75$).

Performance results are reported in Table 2.1, Figure 2.1 and Figure 2.2. The parameter setting learnt in the training collection performed better than the default BM25 setting. This is somehow surprising because the default BM25 setting has proved to be very robust in many document retrieval experimentations [Rob05]. Our optimal setting fixed $K1$ to 0.7 (instead of 1.2, which is the default value). Still, we found that performance is not very sensitive to $K1$ (the default $K1$ setting led to quasi-optimal performance). The main difference between the default BM25 configuration and our configuration lies in the b parameter. The default value for b is 0.75 but we observed a significant improvement in performance when b was smaller

[4] As a matter of fact, $K3$ is barely used today [RZ09].

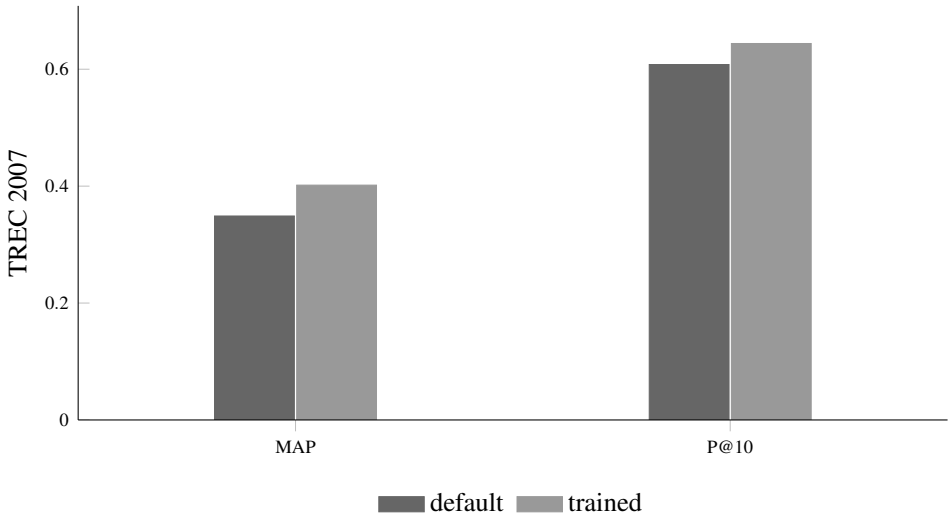

Figure 2.1: BM25 performance (MAP and P@10) in TREC 2007 obtained by the default and by a trained parameter setting.

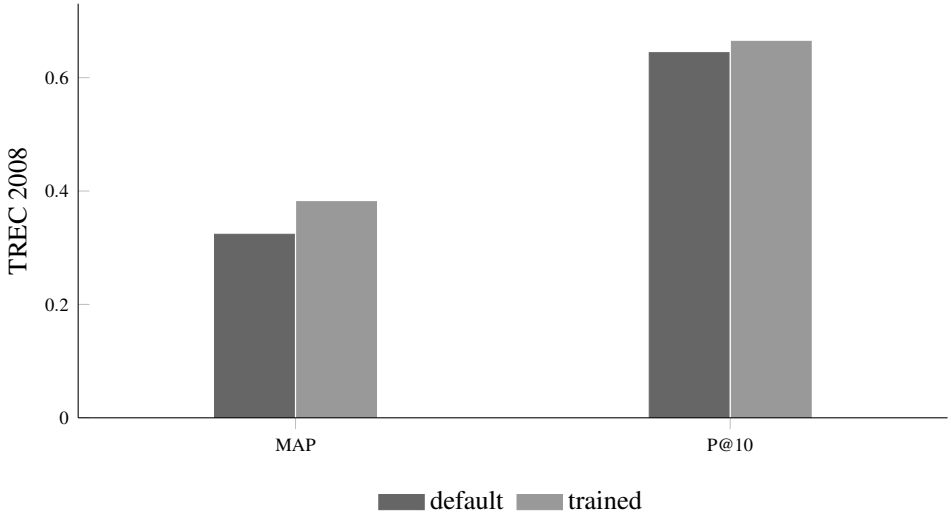

Figure 2.2: BM25 performance (MAP and P@10) in TREC 2008 obtained by the default and by a trained parameter setting.

(optimal $b = 0.3$). We hypothesise that this is related to the nature of the documents. b is a length normalisation parameter. High b values increase the penalty for long documents. In the BLOGS06 collection the distribution of lengths might be more uniform than the distribution of lengths in standard text collections (the variability of lengths in blog entries might be smaller than the variability of lengths in generic web or adhoc collections).

With proper preprocessing and parameter setting we have obtained a strong baseline that is competitive when compared to TREC 2007 and 2008 blog retrieval systems [MOS07, OMS08]. The optimal BM25 parameters ($b = 0.3$ and $K1 = 0.7$) were fixed for the subsequent experiments.

2.2 Combining Document and Sentence Scores for Blog Topic Retrieval

In this section we describe a new method to support blog IR based on the distribution of salient sentences. A high sentence score is associated to a good matching with the query. By promoting documents with high score sentences, we try to identify documents that contain terms related to the query in the right context. We hope that the flow of the presumed relevant sentences help us to determine relevance-flow patterns.

We consider several document features such as the ratio of high-scored sentences in a document (peaks), the median number of unique terms matched by the document's sentences, the variance of sentence scores in a document, and the maximum score of the document's sentences. These features provide valuable information about the way in which a document matches a query.

2.2.1 Sentence Scores

The sentence retrieval module is composed of two components: a preprocessing component and a weighting component. Given the collection of blogs, we split the documents into sentences. We used a Java port of the Carnergie Mellon University link grammar parser[5]. This software is included in the MorphAdorner project[6], developed by Northwestern University of Chicago. The link grammar parser is a natural language parser based on link grammar theory. Given a sentence, the system assigns a syntactic structure consisting of a set of labelled

[5] `http://www.link.cs.cmu.edu/link/`.
[6] `http://morphadorner.northwestern.edu/`.

links connecting pairs of words. The parser also produces a "constituent" representation of a sentence (including, e.g., noun phrases or verb phrases). For efficiency reasons, the sentences were stored in a sentence-level inverted index. At query time, we only had to retrieve all sentences that matched at least one query term.

Sentence Retrieval (SR) models are based on matching the query and every sentence. A vector-space approach, the tfisf model [AWB03], is a simple but very effective SR method. We adopted tfisf because it is parameter-free and performs at least as well as tuned SR methods based on Language Models or BM25 [Los10]. The tfisf matching function is:

$$tfisf(S,Q) = \sum_{t \in Q} \log(tf_{t,Q} + 1) \log(tf_{t,S} + 1) \log\left(\frac{n+1}{0.5 + sf_t}\right) \tag{2.3}$$

where $tf_{t,Q}$ and $tf_{t,S}$ are the number of occurrences of the term t in the query Q and sentence S, respectively; sf_t is the number of sentences where t appears, and n is the total number of sentences in the collection.

2.2.2 Combining Document and Sentence Scores

Given the query-document similarity score (BM25) and the sentence's scores (tfisf), it is necessary to define a combination method that assigns an overall score to every document. This score will be used to re-rank the top-ranked documents retrieved by $BM25\left(D_{T_q}\right)$[7].

First, we applied the following normalisation, which scales the scores into [0,1]:

$$BM25_{norm}(D,Q) = \frac{BM25(D,Q)}{max_{D_i \in C} BM25(D_i,Q)} \tag{2.4}$$

$$tfisf_{norm}(S,Q) = \frac{tfisf(S,Q)}{max_{S_i \in D_{T_q}} tfisf(S_i,Q)} \tag{2.5}$$

where C is the collection of documents and S_i are the sentences belonging to the top retrieved documents.

Sentence-level features are potentially indicative of the pattern of matching between relevant documents and queries. The features used in our study are the following[8]:

– *Ratio of peaks:* the ratio of peaks in a document, being a peak each sentence of the document that has a normalised score greater than 0.5. We calculate the ratio of peaks

[7] For each query, we re-ranked the first 1000 documents from the initial BM25 ranking.

[8] Variance and Median are calculated only for sentences that match at least one query term.

as $(\#peaks/\#sentences)$. This measure is promising to favour documents with several sentences highly relevant to the query with respect to documents that contain many query terms scattered through the text. We chose this threshold because the number of sentences considered as a peak should be low (as a matter of fact many documents have no peaks). We hypothesise that this threshold will be useful to detect the most salient sentences. The same threshold has been used in other studies to estimate what sentences are salient [SJ09].

– *Variance:* the variance of non-zero sentence scores in the document. With this feature we can model how query matching varies across the document. It is interesting to detect these matching trends in relevant documents and to study how they differ from the non-relevant documents' trends.

– *MedianU:* the median of the number of unique query terms matched by sentences in the document. We hypothesise that documents with high median score are potentially relevant because they contain many sentences on topic. In contrast, documents with low median scores have sentences with poor matching with the query.

– *Max:* the maximum score of the document's sentences. This measure could be used to promote documents with a highly relevant sentence. This relates to passage-based retrieval methods that estimate relevance using the highest scoring passage.

The function used to combine the document and sentence-level score features is simply:

$$sim(D,Q) = \alpha \cdot BM25_{norm}(D,Q) + \beta \cdot SF_{norm}(D,Q) \tag{2.6}$$

where $SF_{norm}(D,Q)$ is one of the four scores explained above, and α and β are free parameters trained by linear regression [Cro00].

We only considered the individual incorporation of these features in our model. Combining multiple features and applying more formal ways to combine them –taking into account dependencies– will be studied in the near future.

2.2.3 Experiments

From the TREC blog track datasets, we built two realistic and chronologically organised test beds (see Table 2.2).

Label	Train	Test
TREC 2007	2006	2007
TREC 2008	2006, 2007	2008

Table 2.2: Train and test configurations.

Table 2.3, Figure 2.3 and Figure 2.4 report the results of our approach against the high performing BM25 configuration described in Section 2.1.2 ($b = 0.3$ and $K1 = 0.7$). The first column refers to the baseline (BM25) and the rest of the columns refer to our model with different sentence-level features: *ratio of peaks*, *medianU*, *variance* and *max*. The best value in each row is bolded.

Two features yielded to improvements in performance over the baseline: *RatioPeaks* and *MedianU*. *RatioPeaks* is the feature that performs the best, significantly outperforming the baseline in terms of MAP and also consistently improving P@10. *MedianU* consistently improves MAP but is slightly less competitive in terms of P@10. Moreover, *MedianU*'s improvements are not statistically significant.

The weights obtained at training time help to understand the combination model. For instance, *RatioPeaks* gets a negative weight (β) for the feature score, meaning that we promote documents with few *peaks*. We only re-rank top-retrieved documents and, therefore, the sentence-level features are applied to documents with high query-document similarity score. By promoting top-ranked documents with few peaks we are selecting documents that have the query-document score highly concentrated in a few sentences. On the other hand, documents with many peaks have the score distributed over many places. Our results suggest that query topics discussed in a few concentrated locations are indicative of relevance.

MedianU is assigned a positive weight, meaning that we promote documents with high *medianU*. We are giving a lower weight to documents that contain many sentences with poor overlapping with the query (few query terms), and hence, vaguely related to the query. For example, consider two documents, D_1 and D_2, and a query that matches with two D_1 sentences and six D_2 sentences. If the number of unique query terms matched by D_1 sentences and D_2 sentences are $\{3,4\}$ and $\{1,1,1,2,1,1\}$, respectively[9] then: D_2 has more matching sentences than D_1, but D_1 sentences are more highly related to the query. Hence, D_1 is likely more relevant than D_2 (*medianU* values are 3.5 and 1, respectively).

[9]Each document is represented by the number of unique terms matched by their sentences. Note that we only consider sentences that match at least one query term.

		$\alpha \cdot BM25_{norm} + \beta \cdot SF_{norm}$			
	BM25 $b = .3$ $K1 = .7$	RatioPeaks	MedianU	Var	Max
		TREC 2007			
(α, β)		(2.0751,-0.3484)	(2.0916,0.0642)	(1.9952,1.8750)	(1.9701,0.1716)
MAP	.4017	**.4100▲**	.4060	.3987	.3965
$\triangle\%$		(+2.07%)	(+1.07%)	(-0.75%)	(-1.31%)
P@10	.6440	**.6880▲**	.6360	.6140	.6000
$\triangle\%$		(+6.83%)	(-1.26%)	(-4.89%)	(-7.33%)
		TREC 2008			
(α, β)		(1.5944,-0.2436)	(1.5345,0.1042)	(1.4578,5.0085)	(1.3233,0.4103)
MAP	.3812	.3863▲	**.3922**	.3567	.3665
$\triangle\%$		(+1.34%)	(+2.89%)	(-6.87%)	(-4.01%)
P@10	.6640	**.6900**	.6780	.5460	.5740
$\triangle\%$		(+3.92%)	(+2.11%)	(-21.61%)	(-15.68%)

Table 2.3: Topic retrieval results in TREC 2007 and TREC 2008. The symbols ▲ and ▼ indicate a significant improvement or decrease over the baseline. The table reports the α and β values learnt in the training process.

The analysis of *RatioPeaks* and *MedianU* suggests that we should prefer documents with a few focalised (and high-quality) sentences on topic rather than documents with many (low-quality) sentences poorly related to the query. The other features (*var* and *max*) did not give any added value. In some document retrieval studies [SJ09], the variance of sentence scores was extremely effective to estimate relevance. Our study with the blog collection reveals otherwise. This might be due to a higher variance of scores in adhoc collections when compared to the more focused collection of blog posts.

2.3 Conclusions

In this chapter we have described a novel way to incorporate sentence-level features into blog topic retrieval baselines. We worked with an effective BM25 parameter configuration (which outperforms the default BM25 configuration) and we re-ranked an initial retrieval baseline in a way that incorporates new document features based on sentence scores.

We have tested the effectiveness of our approach by combining four sentence-level features and evaluating them against two different datasets. We found two features (ratio of peaks and median of unique terms) that offer a good performance and yield to combined models that outperform state-of-the-art models for blog topic retrieval.

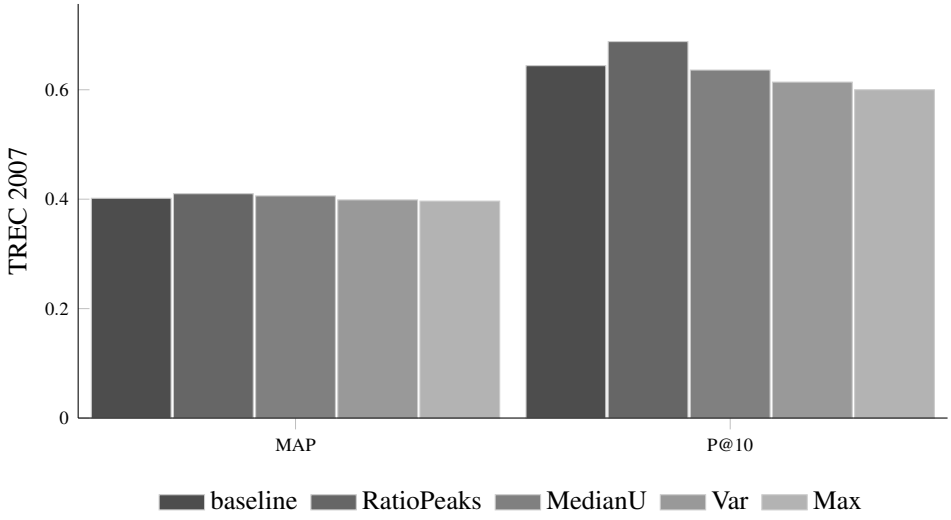

Figure 2.3: Topic retrieval performance in TREC 2007 (MAP and P@10).

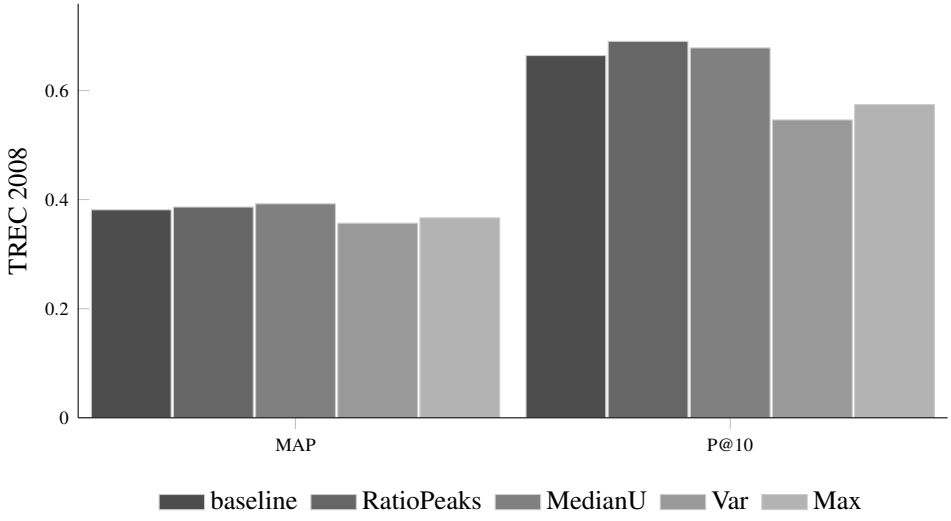

Figure 2.4: Topic retrieval performance in TREC 2008 (MAP and P@10).

Chapter 3

Opinion Finding

In this chapter we concentrate on Opinion Finding within textual documents. First of all, we propose novel query expansion techniques based on structural aspects of documents. Next, we explore the importance of positional information and passage retrieval to detect opinionated blog posts. Finally, we study structural and discourse features to improve the classification of subjective sentences from news articles and product reviews. This chapter, therefore, focuses on subjectivity detection –both at document and sentence level– and employs different types of features to determine opinions (regardless of their polarity).

3.1 Query Expansion

A common way to understand people's opinions about a given topic is to read the comments associated to documents where the topic is discussed (e.g., the comments attached to a blog post or the comments associated to a news story). In the comments' section, people tend to express opinions related to the topic of the document. We argue that these comments are more densely populated by opinions than other parts of the text. Therefore, terms from comments are likely opinionated and on-topic, and a simple feedback technique that takes advantage of these specific words is promising to improve opinion finding. To the best of our knowledge, this is the first attempt to apply query expansion powered by the document's comments.

3.1.1 Relevance Models

Relevance Models (RM) have emerged as one of the most effective and efficient Pseudo Relevance Feedback (PRF) methods. Relevance Models explicitly introduced the concept of relevance in the Language Modelling (LM) framework [LC01]. The original query is considered as a short sample of words obtained from a relevance model R; and relevant documents are larger samples of text from the same model. From the words already seen (original query), the relevance model is estimated. If more words from R are needed then the words with the highest estimated probability are chosen. The terms in the vocabulary are therefore sorted according to these estimated probabilities. Two estimations were originally presented in [LC01]: RM1 and RM2. RM1 is defined as:

$$P(w|R) \propto \sum_{d \in C} P(d) \cdot P(w|d) \cdot \prod_{i=1}^{n} P(q_i|d) \tag{3.1}$$

Usually, the prior $P(d)$ is assumed to be uniform. $\prod_{i=1}^{n} P(q_i|d)$ is the query likelihood given the document model, which is traditionally computed using Dirichlet smoothing. $P(w|d)$ accounts for the importance of the word w within the document d. The process follows four steps:

1. Initially, the documents in the collection (C) are ranked using a standard LM retrieval model (e.g., query likelihood with Dirichlet smoothing [ZL04]).

2. The top r documents from the initial retrieval are taken for the estimation of the relevance model. In the following, this pseudo relevant set will be referred to as RS.

3. The relevance model's probabilities, $P(w|R)$, are calculated using the estimate presented in Eq. 3.1, using RS instead of C.

4. The expanded query is built with the top e terms with the highest estimated $P(w|R)$.

RM3 [AjAC$^+$04] is a later extension of RM that performs better than RM1. RM3 interpolates the terms selected by RM1 with a LM computed from the original query:

$$P(w|q') = (1-\lambda) \cdot P(w|q) + \lambda \cdot P(w|R) \tag{3.2}$$

The expanded query is used to get a second ranking of documents using negative cross entropy.

3.1.2 Our Proposal: $RM3_C$

We propose an alternative RM3 estimation to promote terms that appear in the comments of online documents:

$$P(w|R) \propto \sum_{d \in RS} P(d) \cdot P(w|d_{comm}) \cdot \prod_{i=1}^{n} P(q_i|d) \tag{3.3}$$

where w is any word appearing in the set of comments associated to documents in RS and $P(w|d_{comm})$ is computed as the probability of w in the set of comments associated to document d. In this way, comments act as proxies of the documents in terms of opinion. The estimation of query likelihood remains at document level because the effect of topic relevance is better encoded using the whole document. The query likelihood factor acts as document-query similarity measure and, therefore, promotes terms from highly relevant documents; whereas $P(w|d_{comm})$ biases the computation towards terms from comments. $P(w|d_{comm})$ and $P(q_i|d)$ are estimated using Dirichlet smoothing:

$$P(w|d_{comm}) = \frac{tf_{w,d_{comm}} + \mu \cdot P(w|C_{comm})}{|d_{comm}| + \mu} \tag{3.4}$$

$$P(q_i|d) = \frac{tf_{q_i,d} + \mu \cdot P(q_i|C)}{|d| + \mu} \tag{3.5}$$

where $tf_{q_i,d}$ is the number of times that the query term q_i occurs in document d, and $tf_{w,d_{comm}}$ is the number of times that the word w appears in the document (d_{comm}) that is constructed by concatenating all comments associated to d. $|d|$ and $|d_{comm}|$ are the number of words in d and d_{comm}, respectively. $P(q_i|C)$ is the probability of q_i in the collection of documents C and $P(w|C_{comm})$ is the probability of w in the collection of comments. μ is an smoothing parameter.

Overall, this leads to a novel expansion approach that combines the strength of the RM formalism with the inherent structure of some types of documents (e.g., blog posts and comments in the blogosphere). Whether or not this comments-biased expansion performs better than a standard expansion is a question that we try to answer in the next subsection.

3.1.3 Experiments

The 100 topics provided by TREC 2006 and TREC 2007 blog tracks were used for training (optimising MAP) and, then, we used the 50 TREC 2008 topics as the test query set.

Documents were pre-processed and segmented into posts and comments following the heuristic method proposed in [PLCB10a]. This method relies on the DOM structure of webpages and the nodes' attribute values to extract the desired blog parts. Every webpage can be represented as a tree, known as the DOM (Document Object Model) and each content fragment in the page is represented by a node or a set of nodes. Each node may have several attributes, which describe properties of the node. In [PLCB10a], Parapar et al. defined several rules able to identify common patterns in popular blog generator software (e.g., Blogger, Wordpress). The method is based on the position of the nodes within the DOM tree and on the values of their attributes. Different parts of a blog post, including the comments, can be effectively identified with these patterns.

We removed 733 common words from documents and queries. As argued in Chapter 1, it is standard practice to use the topic-retrieval baselines provided by TREC 2008 as initial input for the opinion retrieval stage. We followed this evaluation design and applied the proposed RM estimation to re-rank the baselines. The parameters trained were the following: the smoothing parameter of Dirichlet, μ ($\mu \in \{10, 100, 1000, 2000, 3000, 4000, 5000, 6000\}$), the number of documents in the pseudo relevant set $r = |RS|$, ($r \in \{5, 10, 25, 50, 75, 100\}$), the number of terms selected for expansion e ($e \in \{5, 10, 25, 50, 75, 100\}$) and the interpolation weight λ ($\lambda \in \{0, .1, .2, .3, .4, .5, .6, .7, .8, .9, 1\}$). The parameters were tuned (independently for each baseline) for both the classical *RM*3 estimated from whole documents (post and comments) and for our proposal (labelled as $RM3_C$). We also tested other alternatives (e.g., comments alone, *RM*3 with post alone), but these expansion methods did not outperform the *RM*3 method with whole documents. The optimisation followed an exhaustive exploration process (grid search). The experiments were run under the Indri retrieval platform[1]. In order to apply our RM estimation under this framework, *RM*3 was implemented in the Indri's query language as follows:

$$\#weight\ (\lambda\ \#combine(\ q_1\ \cdots\ q_n)(1-\lambda)\ \#weight(\ P(t_1|R)\cdot t_1\ \cdots\ P(t_e|R)\cdot t_e)) \tag{3.6}$$

where $q_1 \cdots q_n$ are the original query terms, $t_1 \cdots t_e$ are the e terms with highest probability according to Equation 3.3, and λ is a free parameter to control the trade-off between the original query and the expanded terms. We selected Dirichlet [ZL04] as the smoothing technique for our experiments.

[1] `http://www.lemurproject.org/indri.php`

	orig.	*RM*3			*RM*3$_C$		
Baseline	MAP	MAP		RI	MAP		RI
baseline1	.3239	.3750▲	(+16%)	.60	.3653▲	(+13%)	.56
baseline2	.2639	.3117▲	(+18%)	.36	.3244▲	(+23%)	.52
baseline3	.3564	.3739	(+5%)	.08	.3753	(+5%)	.12
baseline4	.3822	.3652	(−4%)	-.04	.3688	(−4%)	-.08
baseline5	.2988	.3383▲	(+13%)	.44	.3385▲	(+13%)	.48
average	.3251	.3528▲	(+8%)	.29	.3545▲	(+9%)	.32
	orig.	*RM*3			*RM*3$_C$		
	P@10	P@10		RI	P@10		RI
baseline1	.5800	.6140	(+6%)	.18	.6360	(+10%)	.20
baseline2	.5500	.5560	(+1%)	.04	.6340▲ △	(+15%)	.18
baseline3	.5540	.5800	(+5%)	-.02	.6460▲ △	(+17%)	.30
baseline4	.6160	.6140	(−0%)	-.04	.6560	(+6%)	.18
baseline5	.5300	.5940	(+12%)	.18	.6660▲ △	(+26%)	.54
average	.5660	.5916	(+5%)	.07	.6476▲ △	(+14%)	.28

Table 3.1: Opinion finding results in TREC 2008. The symbols ▲(▼) and △(▽) indicate a significant improvement (decrease) over the original baselines and the *RM*3 method, respectively.

The measures adopted to evaluate opinion retrieval effectiveness were Mean Average Precision (MAP), Precision at 10 (P@10), and the Reliability of Improvement (RI) [SMK05], which is a commonly used robustness measure for PRF methods. The gold-standard was obtained from the documents that were assessed as subjective with respect to the query topic.

Table 3.1 and Figures 3.1 and 3.2 report the experimental results. Each run was evaluated in terms of its ability to retrieve subjective documents higher up in the ranking. The best value for each baseline and performance measure is underlined. Statistical significance was estimated using the t-test at the 95% level. The symbols ▲ and ▼ indicate a significant improvement or decrease over the original baselines and the symbols △ and ▽ indicate a significant improvement (resp. decrease) with respect to the standard *RM*3 method.

*RM*3 and *RM*3$_C$ generally outperform the baselines. This is not surprising because the baselines are topic retrieval runs with no opinion finding capabilities. Baseline 4 is the only case where expansion does not work. But the decrease in performance is not statistically significant. *RM*3$_C$ performs better than *RM*3. In terms of *MAP*, *RM*3 is able to achieve improvements that are similar to those found with *RM*3$_C$. However, in terms of *P*@10, *RM*3$_C$ usually shows significant improvements with respect to the baselines and with respect to *RM*3. Furthermore, *RM*3$_C$ shows higher values of *RI*. This indicates that the improvements obtained

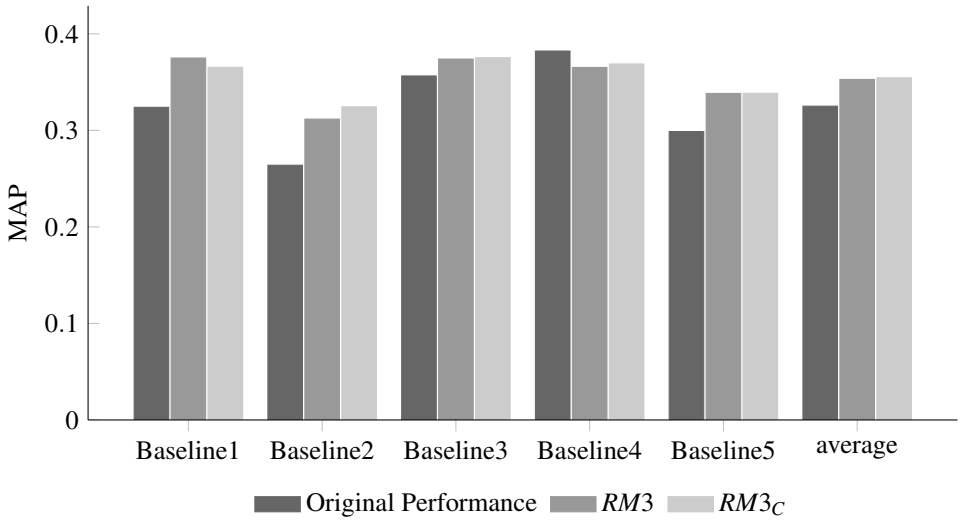

Figure 3.1: Opinion finding performance (MAP) in TREC 2008.

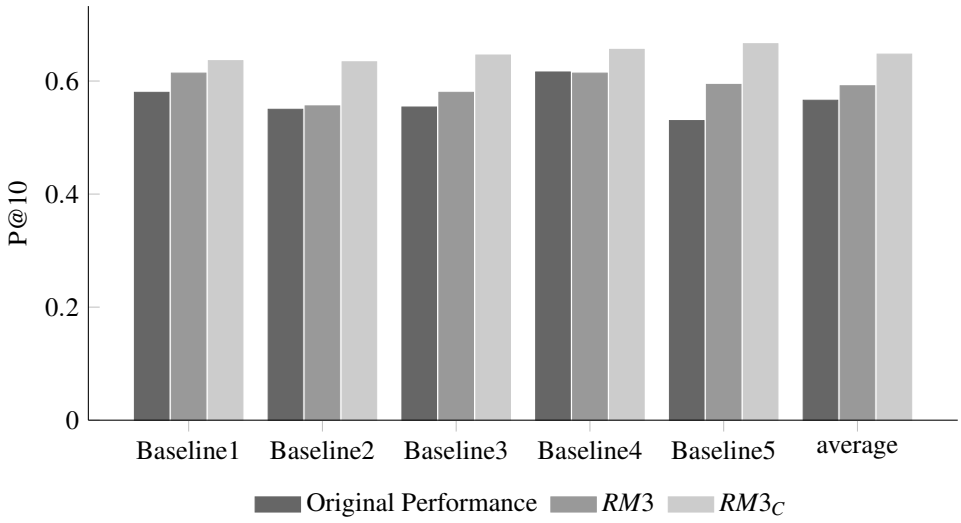

Figure 3.2: Opinion finding performance (P@10) in TREC 2008.

using queries expanded with terms from comments are more consistent than those obtained with terms from whole documents. These results highlight the importance of comments to

TREC Run	orig. MAP	*TREC run*+$RM3_C$ MAP		RI	orig. P@10	*TREC run*+$RM3_C$ P@10		RI
uicop1bl1r	.3614	.3524	(−2%)	-.18	.6020	.6264	(+4%)	.14
B1PsgOpinAZN	.3565	.3558	(−2%)	.10	.6204	.6512▲	(+5%)	.30
uogOP1PrintL	.3412	.3510	(+3%)	.10	.5964	.6320▲	(+6%)	.25
NOpMM107	.3273	.3532▲	(+8%)	.38	.5744	.6432▲	(+12%)	.37
UWnb1Op	.3215	.3538▲	(+10%)	.33	.6068	.6500	(+7%)	.25
FIUBL1DFR	.2938	.3520▲	(+20%)	.61	.4804	.6392▲	(+33%)	.76
UniNEopLRb1	.2118	.2121	(+0%)	.18	.6156	.6464	(+5%)	.29
uams08b1pr	.1378	.3347▲	(+43%)	.93	.1284	.6100▲	(+375%)	1.0

Table 3.2: Average opinion finding performance of different TREC 2008 opinion finding systems against the results achieved by $RM3_C$ on top of those systems. The symbols ▲(▼) indicate a significant (resp. decrease) improvement over the TREC systems. TREC systems that were able to outperform the original 5 topic-retrieval baselines are in bold.

enhance precision without harming recall (MAP is roughly the same with either expansion method). This suggests that subjective words estimated from comments lead to a more accurate query-dependent opinion vocabulary. Furthermore, the independence of $RM3_C$ of any external lexicon is convenient because there is a lack of good opinion resources in many domains and languages.

$RM3_C$ simply re-ranks documents based on a comments-oriented query expansion method that works from an initial ranked set of documents. This brings us the opportunity to apply our methods on top of effective opinion finding methods. To test this ability, we considered the systems proposed by teams participating in the last TREC blog opinion retrieval task (TREC2008) [OMS08]. This task was quite challenging: half of TREC systems failed to retrieve more subjective documents than the baselines [OMS08]. In Table 3.2 and Figures 3.3 and 3.4 we report the mean performance (averaged out over the five baselines) of the TREC systems against the mean performance achieved by applying $RM3_C$ on top of those systems' runs[2]. The systems in bold were the only ones able to show improvements with respect to the original five retrieval baselines (in terms of MAP).

$RM3_C$ is often able to improve the performance of the TREC systems, showing usually significant improvements in terms of $P@10$, as well as good RI scores. This demonstrates that our method is a good complement to strong opinion finding algorithms. Table 3.2 also shows that our expansion approach is robust: $RM3_C$ outperforms all types of opinion retrieval systems regardless of their original performance. Observe also that the average $P@10$ of

[2]We thank TREC for providing us with these runs.

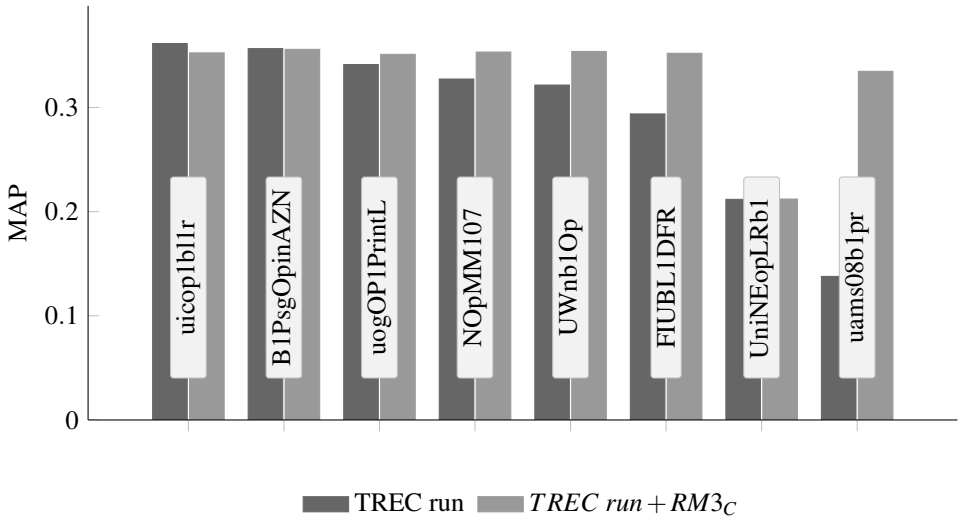

Figure 3.3: Average MAP performance of different TREC 2008 systems against the results achieved by $RM3_C$ on top of these systems.

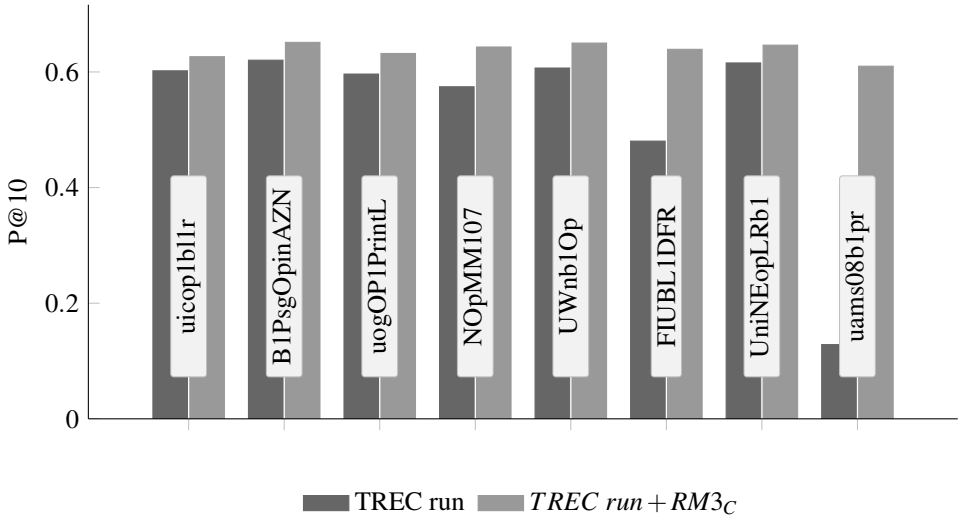

Figure 3.4: Average P@10 performance of different TREC 2008 systems against the results achieved by $RM3_C$ on top of these systems.

our method in Table 3.1 (.6476) is clearly higher than the $P@10$ obtained by any TREC participant.

3.2 Search for opinionated documents

In the previous section we presented a robust query expansion strategy for opinion finding. The method can work from topic relevance baselines or from rankings already biased towards opinionated documents. However, the quality of the original ranking is crucial. In fact, we can see in Table 3.2 that the best P@10 is achieved by expanding $B1PsgOpinAZN$, which is the method with the highest original P@10 performance. Therefore, it is essential to provide the expansion methods with document rankings that contain many opinionated documents.

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 Opinion Finding [SHMO09, GCC10]. We propose a passage-level analysis of texts that takes into account the location of the sentiments and their relevance to the query. More specifically, we propose effective algorithms that consider two main factors when determining the key subjective sentences: the relatedness of the sentence and the query topic, and the location of a sentence in the text. We argue that this information, combined with other evidence of subjectivity (i.e., positive/negative terms in the sentence), is extremely valuable when attempting to detect opinionated documents. This leads to a general opinion finding method that searches for on-topic subjective sentences and estimates subjectivity in a location-aware way.

The method works as follows: given an initial baseline, we work at sentence level to find opinionated sentences related to the query. Next, we build a ranking of subjective documents by aggregating relevance scores and sentence-level subjectivity information. Within this process we study different location-aware strategies to represent the document. To estimate the subjectivity of the sentences we employ OpinionFinder (OF).

3.2.1 Subjectivity at Sentence Level

With the subjective terms tagged by OF [WWH05] we can naturally estimate the subjectivity of a sentence. An example of a sentence tagged by OF was presented in Figure 1.1. To promote subjective sentences that are on-topic, we run a sentence retrieval process to determine the relatedness between the query and each subjective sentence. More specifically, we use the

Lemur[3] implementation of tf-idf, with BM25-like weights[4]. This is a variation of the original tf-idf formula:

$$\textit{tf-idf}(S,Q) = \sum_{t \in S \cap Q} \frac{K_1 \cdot tf_{t,S}}{tf_{t,S} + K_1 \cdot ((1-b) + b \cdot \frac{|S|}{l_c})} \cdot \frac{N}{n_t} \cdot tf_{t,Q} \tag{3.7}$$

where $tf_{t,S}$ is the frequency of the term t in the sentence S, $tf_{t,Q}$ is the frequency of the term t in the query Q, $|S|$ and l_c are the length of the sentence and the average length of sentences in the collection, respectively. N is the number of documents in the collection, and n_t is the number of documents that contain the term t. K_1 and b are free parameters.

The combination of relevance and subjectivity is done through linear interpolation:

$$subj(S,Q) = \beta \cdot rel_{norm}(S,Q) + (1-\beta) \cdot subj(S) \tag{3.8}$$

where $rel_{norm}(S,Q)$ is the Lemur's tf-idf score after a query-based normalization into [0,1]:

$$rel_{norm}(S,Q) = \frac{\textit{tf-idf}(S,Q)}{\textit{tf-idf}(S_{max_q},Q)} \tag{3.9}$$

where S_{max_q} is the sentence that has the highest *tf-idf* score for query Q. $subj(S)$ represents the number of positive and negative terms tagged in the sentence S divided by the total number of terms in S (i.e., the ratio of subjective terms in the sentence). $\beta \in [0,1]$ is a free parameter.

Equation 3.8 allows us to combine query-independent evidence, provided by OF, with a query-dependent measure of similarity with respect to the query topic. By doing so, we expect to promote subjective sentences that are related to the query.

3.2.2 Subjectivity at Document Level

We apply a sentence-level analysis that takes into account location information to search for subjective documents. To this end, we score sentences using eq. 3.8, but we only consider those subjective sentences that have at least one term tagged as subjective (sentences with $subj(S)$ equal to 0 are discarded). To aggregate the individual sentence scores we considered the following alternatives of defining a document subjectivity score ($subj_S(D,Q)$):

– *SubjMeanAll:* The mean of $subj$ scores ($subj(S,Q)$) computed across all subjective sentences in the document. This measure is a natural choice to estimate the overall subjectivity of a document.

[3] `http://www.lemurproject.org/`.

[4] We built a sentence-level index and applied the well-known BM25 suggested configuration ($k1 = 1.2, b = 0.75$), which has proved to be robust in many retrieval experiments [Rob05].

– *SubjMeanBestN:* The mean of scores from the n sentences with the highest $subj$ scores (sentences with the highest aggregated score of topicality and subjectivity). Focusing on the on-topic sentences with high subjectivity (e.g., the most controversial contents of the post) we expect to detect properly those documents that are really subjective with respect to the query topic.

– *SubjMeanFirstN* and *SubjMeanLastN:* The mean of $subj$ scores ($subj(S,Q)$) from the first/last n subjective sentences in the document. The position of the sentence in the post may be an important clue when attempting to understand the subjectivity of the document. Therefore, we study whether the subsets consisting of the first/last subjective sentences are good indicators of the subjectivity of a post. Observe that these strategies are more sophisticated than simply splitting the document into parts. The subjective sentences selected by *SubjMeanFirstN* and *SubjMeanLastN* depend on the flow of sentiments, which is specific to each post. For instance, a blog post whose last part is objective might have its last last subjective sentence in the middle of the text.

Finally, we combine relevance and subjectivity evidence as follows:

$$subj_{SEN}(D,Q) = \gamma \cdot rel_{norm}(D,Q) + (1-\gamma) \cdot subj_S(D,Q) \tag{3.10}$$

where rel_{norm} is the document's relevance score (obtained from the initial topic retrieval baseline) after a query-based normalization in [0,1], $subj_S(D,Q)$ 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 opinion holders express their views.

3.2.3 A baseline approach: $subj_{DOC}(D)$

Finally, as an alternative approach focused only on subjectivity (i.e., topic-independent), we consider the proportion of subjective sentences in each retrieved document and the accummulated confidence about their subjectivity (OF's confidence [RW03, WR05]). This approach has been adopted successfully in other studies [HMO08]:

$$subj(D) = sumdiff \cdot \frac{\#subj}{\#sent} \tag{3.11}$$

	Int.	Step	Desc.	Form.
α	[0..1]	0.1	Doc. Subjectivity (Comb.)	eq. 3.12
β	[0..1]	0.1	Sentence Subjectivity (Comb.)	eq. 3.8
γ	[0..1]	0.1	Doc. Subjectivity (Comb.)	eq. 3.10
n	[1..10]	1	Number of Sentences	eq. 3.10

Table 3.3: Parameters to train: the interval, the step used to train, a description and the formula affected by the parameters.

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 relevance and subjectivity scores to promote subjective documents that are on-topic:

$$subj_{DOC}(D,Q) = \alpha \cdot rel_{norm}(D,Q) + (1-\alpha) \cdot subj_{norm}(D,Q) \tag{3.12}$$

where rel_{norm} is the normalized relevance score (obtained from the initial baseline) and $subj_{norm}$ is a query-based normalization of eq. 3.11. $\alpha \in [0,1]$ is a free parameter. This method serves as a reference comparison for $subj_{SEN}$. Observe that $subj_{DOC}$ combines relevance and subjectivity at document level. However, it does not take into account sentence-level evidence such as the location and the relevance of the sentences of the document.

3.2.4 Experiments

We experimented with the same datasets as in Section 2.2.3. Documents and topics were preprocessed with Krovetz stemmer and 733 English stopwords were removed. We only used the information from title and post in this experiment. Our method is focused on extracting key relevant subjective sentences from the flow of text of the documents. Incorporating comments into this sentence analysis process is potentially misleading because of the way in which comments are written. Deciding how to use comments to effectively guide the estimation of subjectivity of the document is an interesting challenge that is studied in other parts of the present thesis (e.g., for query expansion purposes).

The training topics were used to set all the parameters of our methods. In Table 3.3 we report some details about the parameters and their characteristics. The train process was focused on maximising MAP.

The results of our experiments are reported in Table 3.4, Figure 3.5 and Figure 3.6. In general, it is interesting to observe that the original retrieval baselines, which do not have subjectivity capabilities, are difficult to beat. In fact, $subj_{DOC}(D,Q)$ is inferior to them in most situations. On the other hand, our sentence-based methods work reasonably well, showing significant improvements with respect to both the original baselines and $subj_{DOC}(D,Q)$ in most cases. *SubjMeanBestN* is the best performing method. It consistently outperforms 4 out of the 5 original baselines and it is clearly superior to $subj_{DOC}(D,Q)$. The results also provide some evidence indicating that location information might be helpful to estimate subjectivity. Our location-based methods (*subjMeanFirstN*, *subjMeanLastN*) are able to outperform 3 different baselines. Still, this experimental evidence is not strong enough and the role of positional evidence for OM will be revisited in the next chapters of this thesis.

In Table 3.5 we report the value of the parameters trained. The methods proposed have up to three parameters, but their optimal values are quite stable across collections. The $subj_{DOC}(D,Q)$ method gets a high value of α (0.9). This parameter controls the relative weight of relevance and subjectivity (eq. 3.12). The value of this parameter indicates that the relevance component is much more important than the subjectivity component. This seems to indicate that $subj_{DOC}(D,Q)$ is extremely sensitive to off-topic material. Similarly, the γ parameter controls the trade-off between relevance and subjectivity at document level in our sentence-based methods (eq. 3.10). This parameter is quite stable across collections and has a lower value than the one obtained for $subj_{DOC}(D,Q)$ (γ gets values between 0.6 and 0.8). This might be due to a more reliable estimation of subjectivity obtained from the methods proposed in this work. Regarding β, we observe different trends depending on the aggregation strategy. *SubjMeanBestN* has high values of β (the value of this parameter is around 0.5 for *SubjMeanBestN* and around 0.2 for other methods). β controls the trade-off between relevance and subjectivity at sentence level (eq. 3.8). This means that in *SubjMeanBestN* content-match evidence is more important than content-match in the other models.

From Table 3.5 it is also interesting to observe that the number of sentences used by *SubjMeanBestN* was 1 in all cases. This indicates that the sentence with the highest $Subj(S,Q)$ score (aggregated score of relevance and subjectivity) is the best guidance to detect opinionated blog posts. For instance, in Figure 3.8 we show a blog post in which the author is expressing his opinion about the topic presented in Figure 3.7 (Starbucks coffee shops). This post has been judged as relevant and subjective for this topic by TREC assessors. The subjective sentence selected by *SubjMeanBestN* ($n = 1$) is bolded in Fig 3.8. This sentence is

highly indicative of the opinionated nature of this post and, hence, a good clue to determine the opinionated nature of the blog post. The effectiveness of *SubjMeanBestN* can be also fruitful, for example, to build opinion-biased snippets of blog posts.

	2007		2008	
	MAP	P@10	MAP	P@10
Baseline1	.2766	.4580	.3239	.5800
+ $subj_{DOC}$	.2743	.5580▲	.3291	.6240
+SubjMeanAll	.2747	.4660▽	.3335	.5800
+SubjMeanBestN	.3151▲ △	.5500▲	.3485	.6460
+SubjMeanFirstN	.2726	.4560▽	.3325	.5780
+SubjMeanLastN	.2744	.4520▽	.3318	.5780
Baseline2	.3034	.5320	.2640	.5500
+ $subj_{DOC}$	.3041	.5400	.2629	.5440
+SubjMeanAll	.3104▲ △	.5400	.2809▲ △	.5440
+SubjMeanBestN	.3314▲ △	.5580	.2896▲ △	.5920
+SubjMeanFirstN	.3107▲ △	.5480	.2803▲ △	.5540
+SubjMeanLastN	.3101▲ △	.5420	.2787▲ △	.5540
Baseline3	.3488	.5760	.3565	.5540
+ $subj_{DOC}$	.3515▲	.5780	.3584▲	.5580
+SubjMeanAll	.3553▲ △	.5840	.3651▲ △	.5780▲
+SubjMeanBestN	.3641▲ △	.5860	.3742▲ △	.5920
+SubjMeanFirstN	.3552▲ △	.5820	.3653▲ △	.5820
+SubjMeanLastN	.3553▲ △	.5840	.3633▲ △	.5600
Baseline4	.3784	.5340	.3822	.6160
+ $subj_{DOC}$	.3817▲	.5420	.3843▲	.6200
+SubjMeanAll	.3871▲ △	.5640▲ △	.3906▲	.6240
+SubjMeanBestN	.3976▲ △	.5760▲	.3971▲ △	.6580▲
+SubjMeanFirstN	.3872▲ △	.5580▲	.3905▲	.6340
+SubjMeanLastN	.3866▲ △	.5620▲	.3890▲	.6260
Baseline5	.3815	.5640	.2988	.5300
+ $subj_{DOC}$	.3725	.5680	.2998	.5400
+SubjMeanAll	.3537	.5360	.2680▼▽	.5120
+SubjMeanBestN	.3570▼▽	.5840	.2823▽	.5640
+SubjMeanFirstN	.3516▼▽	.5340	.2682▼▽	.5120
+SubjMeanLastN	.3520▼▽	.5300	.2747▼▽	.5280

Table 3.4: Opinion Finding in the TREC blog track. The symbols ▲(▼) and △(▽) indicate a significant improvement (decrease) over the original baselines and the $subj_{DOC}$ method, respectively.

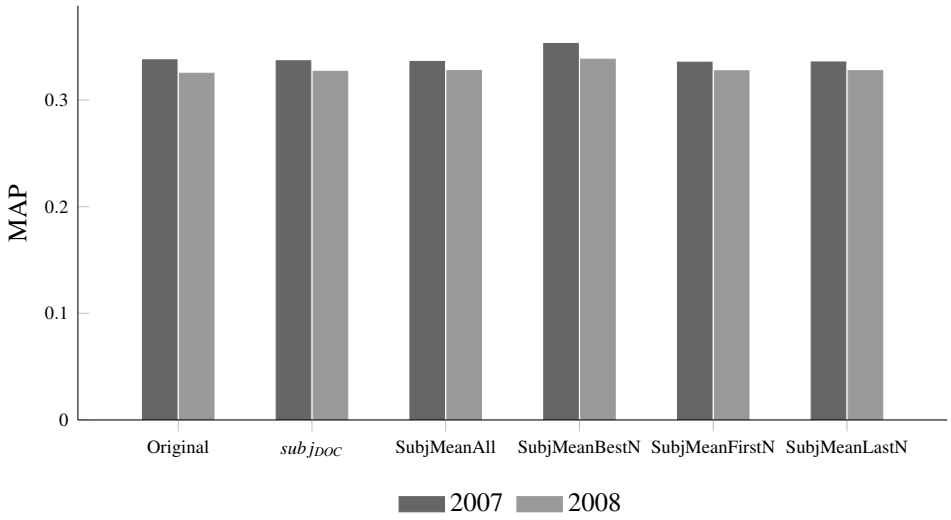

Figure 3.5: Average MAP performance obtained by our opinion finding methods for TREC 2007 and 2008 topics.

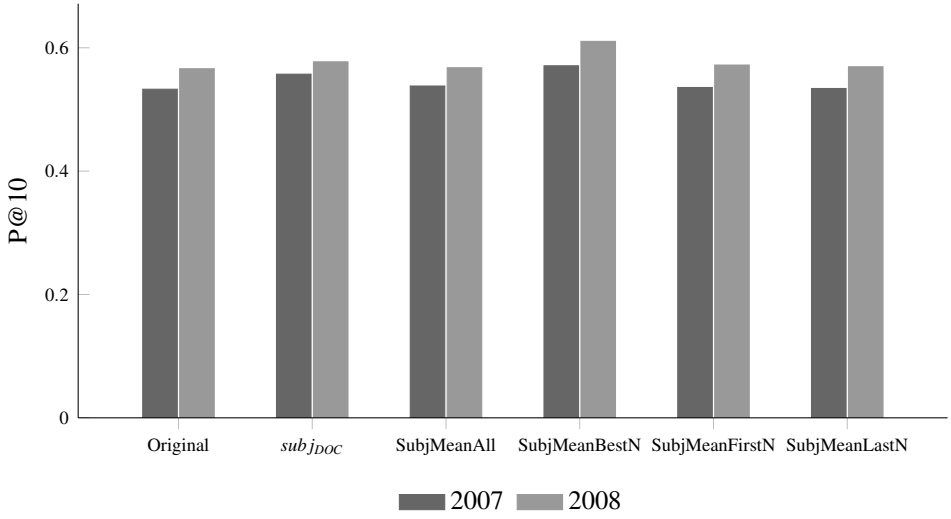

Figure 3.6: Average P@10 performance obtained by our opinion finding methods for TREC 2007 and 2008 topics.

	TREC 2007	**TREC 2008**
$Subj_{DOC}$	$\alpha = 0.9$	$\alpha = 0.9$
SubjMeanAll	$\gamma = 0.7, \beta = 0.1$	$\gamma = 0.6, \beta = 0.2$
SubjMeanBestN	$\gamma = 0.8, \beta = 0.5, n = 1$	$\gamma = 0.7, \beta = 0.4, n = 1$
SubjMeanFirstN	$\gamma = 0.7, \beta = 0.1, n = 6$	$\gamma = 0.6, \beta = 0.2, n = 6$
SubjMeanLastN	$\gamma = 0.7, \beta = 0.1, n = 6$	$\gamma = 0.7, \beta = 0.3, n = 5$

Table 3.5: Parameters trained for $Subj_{DOC}$, $SubjMeanAll$, $SubjMeanBestN$, $SubjMeanFirstN$ and $SubjMeanLastN$.

```
<top>
<num> Number: 1004 </num>
<title> Starbucks </title>

<desc> Description:
What do people think about the Starbucks chain of coffee
 shops?
</desc>

<narr> Narrative:
Any opinion of Starbucks and their products and services is
relevant.Opinions of Starbucks' business practices, their
 ubiquity, etc are also relevant.
</narr>

</top>
```

Figure 3.7: TREC Blog Track topic 1004: Starbucks

3.2.5 Query expansion and SubjMeanBestN

We applied the query expansion technique proposed in Section 3.1 ($RM3_C$) on top of the best subjective approach proposed here ($SubjMeanBestN$). In Table 3.6 we report the performance achieved by this expansion. In Figure 3.9 we depict the average MAP and P@10 performance over the 5 different $SubjMeanBestN$ baselines. The symbols ▲(▼) indicate a significant (resp. decrease) improvement over the original $SubjMeanBestN$ method.

In general, $RM3_C$ seems to be less effective when applied to $SubjMeanBestN$. Still, the benefits obtained in P@10 are consistent across baselines and query topics. However, we can see that $RM3_C$ shows modest levels of improvements in terms of MAP. These modest levels of improvement might have something to do with the nature of $SubjMeanBestN$. Observe that we did not include the comments to compute the score of subjectivity in $SubjMeanBestN$. Hence, documents are promoted by only taking into account the text in the blog post. This

Grande, no, er, Venti, Frappa-Mocha-Cappa Crappa

Just give me a fucking cup of coffee without making me learn a new goddamn language to get it. I've just now succumbed to the lure of Starbuck's, having formerly sneered derisively at all the fools who paid upwards of $3 for a simple cup of coffee. Little did I know the glories of the Venti Iced Caffe Mocha. First, you have to appreciate that I'm a chocolaholic of the highest order. If there were an Order of the Chocolate Empire, I'd be addressed as Lady Bronwen Bittersweet, Keeper of the Cocoa, Duchess of the Dutch Process. Second, I love me some iced coffee. I get a large iced coffee from Dunkin Donuts every morning - what can I say? I'm a creature of habit. Third, any beverage that routinely comes served with Whipped Cream is high on my preferred beverage list. So, I've become enamoured, nay, enchanted by the Venti Iced Caffe Mocha.
It takes me a few minutes to compose my Starbuck's order though, as the grammar and word-order rules to their lexicon are so damned confusing.
I mean, do you say Venti before iced when ordering? Is it an Iced Venti Caffe Mocha or a Venti Iced Caffe Mocha? Could you even say Caffe Mocha, Venti, Iced, as though you were Bond, James Bond? And this is an easy one, not including the placement of half-caf, double-shot, macchiato, or any of the other adjectives one can add. You'd think being able to speak three languages would enable me to easily order a cup of coffee in my own damn country. Don't you think you should be able to pay less for your coffee if you master the language of Starbuckland and manage to successfully order your drink without stuttering? I think a linguistically-gifted discount should be made available. Who's with me?

posted by Bronwen @ 10:32 PM

Figure 3.8: Example of a subjective blog post for the topic *"Starbucks"*. The bolded sentence is the key subjective sentence according to $SubjMeanBestN$ (n=1).

could populate the top of the ranking with document without comments (or with comments without subjective content). $RM3_C$ is driven by the comments of top ranked documents, and having a ranking in which top positions are populated by documents without comments might severely harm the overall performance of this method. Selective query expansion techniques might be appropriate to mitigate the risk of decrease in performance. For instance, by applying $RM3_C$ only on the top of selected rankings. We will investigate this in the near future.

3.3 Classification of subjective vs non-subjective sentences

So far we have been concerned with detecting opinionated documents. Essentially, given a document, we analysed its sentences as a way to estimate the subjective nature of the document. But the quality of sentence-level estimations was only indirectly analysed because the TREC Blog track lacks judgements associated to sentences or passages. In this section, instead, we explore and explicitly evaluate the more fine-grained task of detecting opinions at sentence level. To meet this aim we utilise the NTCIR07 English MOAT research collection,

Run	orig. MAP	$+RM3_C$ MAP	RI	orig. P@10	$+RM3_C$ P@10	RI
			2007			
baseline1+*SubjMeanBestN*	.3152	.3218	.08	.5500	.5880	.16
baseline2+*SubjMeanBestN*	.3314	.3253	.00	.5580	.5700	.08
baseline3+*SubjMeanBestN*	.3641	.3100▼	-.36	.5860	.5340	-.10
baseline4+*SubjMeanBestN*	.3976	.3503▼	-.24	.6240	.6680	.16
baseline5+*SubjMeanBestN*	.3570	.3789▲	.4	.5840	.6220	.10
average	.3506	.3373	-.10	.5728	.5876	.25
			2008			
baseline1+*SubjMeanBestN*	.3485	.3629	.28	.6460	.6400	.02
baseline2+*SubjMeanBestN*	.2896	.3241▲	.40	.5920	.6260	.16
baseline3+*SubjMeanBestN*	.3742	.3783	.20	.5920	.6560▲	.22
baseline4+*SubjMeanBestN*	.3971	.3790	-.20	.6580	.6680	.00
baseline5+*SubjMeanBestN*	.2823	.3380▲	.60	.5640	.6640▲	.20
average	.3383	.3565	.22	.6104	.6508	.27

Table 3.6: Opinion finding performance of the 5 different *SubjMeanBestN* baseline runs against the results achieved by $RM3_C$ on top of these baselines. The symbols ▲(▼) indicate a significant (resp. decrease) improvement over the TREC systems.

dataset	# subjective sentences	# objective sentences	#unique unigrams	#unique bigrams
MOAT	887	2697	2218	2812
MPQA	7333	8469	6463	9203
PL	5000	5000	4948	9103

Table 3.7: Test collections for experimentation in subjectivity classification at sentence level. The tables include the number of unique unigrams and bigrams after pre-processing. We did not apply stemming and we did not remove common words. We only removed terms that appeared in less than four sentences.

the Multi-Perspective Question Answering dataset (MPQA) and the Pang & Lee subjectivity dataset (PL). All these collections contain opinion judgements at sentence level (see Section 1). The main statistics of these collections are reported in Table 3.7.

With the raise of the Web, the large amount of information available on-line becomes a major issue. To mitigate this problem, many efforts have been made to automatically classify textual documents. There are different types of classification tasks. The most classical ones try to classify documents according to their topics (e.g., sports, economy). However, with the

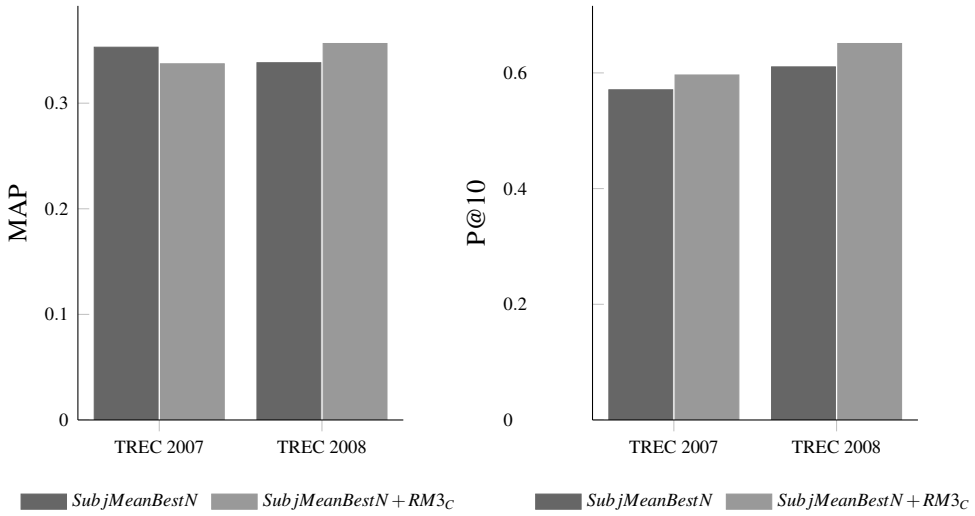

Figure 3.9: MAP and P@10 performance obtained by $SubjMeanBestN$ and $SubjMeanBestN + RM3_C$ in TREC 2007 and TREC 2008.

advent of the social web, opinions have become a key component in many on-line repositories [PL07]. These new opinion-oriented resources demand advanced classification processes able to skim off the opinionated texts to reveal the subjective parts.

Lexicon-based techniques are often applied to detect opinionated sentences. However, extracting opinions from text is challenging and poses many problems that cannot be merely solved with lexicon-based approaches. For example, the sentence *"What the author was thinking when he wrote this story?"* does not contain a single opinionated word, but it implicitly expresses a negative opinion. These difficulties are caused by the subjectivity of a text span being not so much conveyed by the sentiment-carrying words that people use, but rather by the way in which these words are used. We argue that the study of sentence positional information and intra-sentence discourse structure can help to tackle this issue. For instance, people tend to summarise their viewpoints at the end of the text. Moreover, the rhetorical roles of text segments can effectively guide the opinion detection process.

In this section we combine bag of words features, such as unigrams or bigrams, with features computed from sentiment lexicons and with more advanced positional and rhetorical features. To the best of our knowledge, this is the first attempt to combine rhetorical, content-based and positional features for a fine-grained (i.e., sentence-level) estimation of subjectivity.

As argued in the background chapter, other studies have explored the role of rhetorical features in Opinion Mining (OM) but previous efforts are mostly based on coarse-grained tasks (e.g., categorising the overall orientation of a movie review).

3.3.1 Sentence Features

We focus on a two-class (subjective vs. non-subjective) classification of sentences and take into account the following traditional and advanced features to build our classifiers:

- ***Vocabulary features.*** These are binary content-based features based on the appearance of unigrams and bigrams in the sentence[5]. These features are important to detect specific domain-dependent opinionated words. The discriminative power of intra-collection words in terms of opinion has been shown in several studies [GCC09, PLV02].

- ***Positional Features.*** As we claimed in Section 1.2.1, positional evidence might benefit the subjective classification process. In fact, the location-based methods (*subjMeanFirstN*, *subjMeanLastN*) were reasonably effective. To exploit the same intuition under this experimental setting, we encoded two positional features: the absolute position of the sentences within the document (e.g., 2 for the second sentence in the document) and its relative position (the absolute position normalised by the number of sentences in a document).

- ***Part-of-speech features.*** The part-of-speech (POS) of each word can be valuable for analysing sentiments [Tur02]. For each POS tag –e.g., JJ for adjectives– we defined one sentence feature: the count of occurrences of the tag in the sentence. Text was processed by the Stanford Log-linear Part-Of-Speech Tagger[6], which assigns Treebank POS tags [MSM93] (see Table 3.8).

- ***Syntactic Patterns features.*** Apart from sentiment words, many other language compositions express or imply sentiments and opinions [Liu12]. For instance, Turney [Tur02] defined five syntactic patterns to extract opinions from reviews. These patterns have become reference rules for discovering opinions [Liu12]. Turney's patterns are sequences of POS tags (see Table 3.9) and we encoded them as binary features (representing the appearance of every pattern in the sentence).

[5] Unigrams and brigrams with less than 4 occurrences in the collection were removed.
[6] `http://nlp.stanford.edu/downloads/tagger.shtml`

CC	Coordinating conjunction	PRP$	pronoun, possessive
CD	Cardinal Number	RB	Adverb
DT	Determiner	RBR	Adverb, comparative
EX	Existential *there*	RBS	Adverb, superlative
FW	Foreign word	RP	Particle
IN	Preposition or subordinating conjunction	SYM	Symbol
JJ	Adjective	TO	*to*
JJR	Adjective, comparative	UH	Interjection
JJS	Adjective, superlative	VB	Verb, base form
LS	List item marker	VBD	Verb, past tense
MD	Modal	VBG	Verb, gerund or present participle
NN	Noun, singular or mass	VBN	Verb, past participle
NNP	Proper noun, singular	VBP	Verb, non-3rd person singular present
NNPS	Proper noun, plural	VBZ	Verb, 3rd person singular present
NNS	Noun, plural	WDT	Wh-determiner
PDT	Predeterminer	WP	Wh-pronoum
POS	Possessive ending	WP$	Possessive wh-pronoum
PRP	Pronoun, personal	WRB	Wh-adverb

Table 3.8: Penn Treebank Part-Of-Speech (POS) tags.

First Word	Second Word	Third Word
JJ	NN or NNS	anything
RB,RBR, or RBS	JJ	not NN nor NNS
JJ	JJ	not NN nor NNS
NN or NNS	JJ	not NN nor NNS
RB,RBR, or RBS	VB,VBD,VBN, or VBG	Anything

Table 3.9: Patterns of POS tags defined by Turney [Tur02] for extracting opinions.

– ***Sentiment Lexicon Features.*** These features are based on counting the opinionated terms that appear in the sentence. The sentiment lexicon was obtained from OpinionFinder. We included the number and proportion of opinionated terms in a sentence as features for our classifiers. We also included the number and proportion of interrogations and exclamations in the sentence. These features have been widely applied in other fine-grained OM scenarios, such as sentiment detection in tweets [AXV+11].

– ***Rhetorical Features.*** Rhetorical Structure Theory (RST) [MT88] is one of the leading discourse theories. This theory explains how texts can be split into segments that are rhetorically related to one another. Within this structure, text segments can be either nuclei or satellites, with nuclei being assumed to be more significant than satellites with respect to understanding and interpreting a text. Many types of relations between text segments exist; the main paper on RST defines 23 types of relations [MT88]. A satellite may for instance be an elaboration, an explanation or an evaluation on what is explained in a nucleus. We used SPADE (Sentence-level PArsing of DiscoursE) [SM03], which creates RST trees for individual sentences, and we included binary features associated to the appearance of every type of RST relation in a given sentence (see Table 3.10). Observe that we make an intra-sentence RST analysis. The study of inter-sentence RST analysis is an interesting challenge that is out of the scope of this thesis. The rhetorical roles of text segments can effectively guide the opinion detection process. For example, the sentence *"Nevertheless it is undeniable that economic disparity is an important factor in this ethnic conflict"* contains an attribution relationship between the nucleus of the sentence (*"that economic disparity is an important factor in this ethnic conflict"*) and its satellite (*"Nevertheless it is undeniable"*). The presence of this relation helps to understand that the writer is expressing his/her point of view (satellite) about the statement presented in the nucleus. This type of rhetorical clue is potentially valuable to detect opinions.

– ***Sentiment RST features.*** These features are counts of the opinionated terms that occur in the nucleus and in every type of satellite. In this way, we individually represented the subjectivity of the nucleus, the subjectivity of an attribution satellite, the subjectivity of a contrast satellite, and so forth. Again, the representation is sparse because every sentence only contains one (top-level) satellite type. The opinionated terms were also obtained from the OF [WWH05] sentiment lexicon. We included absolute and relative counts (by normalising by the length of the discourse unit), and the number and proportion of exclamations and interrogations in the nucleus and satellites.

– ***Length Features.*** These features encode the length of the sentence, the length of the nucleus and the length of the satellite of the sentences (all of them computed as the total number of words). This information could be valuable to detect subjective sentences. For instance, a short length might be an indication of objectivity.

Relation	Description
attribution	Clauses containing reporting verbs or cognitive predicates related to reported messages presented in nuclei.
background	Information helping a reader to sufficiently comprehend matters presented in nuclei.
cause	An event leading to a result presented in the nucleus.
comparison	Clauses presenting matters which are examined along with matters presented in nuclei in order to establish similarities and dissimilarities.
condition	Hypothetical, future, or otherwise unrealized situations, the realization of which influences the realization of nucleus matters.
contrast	Situations juxtaposed to situations in nuclei, where juxtaposed situations are considered as the same in many respects, yet differing in a few respects, and compared with respect to one or more differences.
elaboration	Rhetorical elements containing additional detail about matters presented in nuclei.
enablement	Rhetorical elements containing information increasing a reader's potential ability of performing actions presented in nuclei.
evaluation	An evaluative comment about the situation presented in the associated nucleus.
explanation	Justifications or reasons for situations presented in nuclei.
joint	No specific relation is assumed to hold with the matters presented in the associated nucleus.
temporal	Clauses describing events with a specific ordering in time with respect to events described in nuclei.

Table 3.10: RST relation types taken into account.

Table 3.11 summarises the considered sentence features. We employed these feature-based representations to build *linear* classifiers (Support Vector Machines or Logistic Regression). Such classifiers base their decision rule on a weighted combination of the feature values, thus bringing the advantage of easily interpretable weights that are assigned to input features in the learning process. This can be seen as an extension of the work presented in Section 3.2. For instance, the linear combination defined in equation 3.8 (sentence score of subjectivity and relevance) can naturally be learned by these classifiers.

3.3.2 Experiments

In our experiments we used *liblinear* [FCH$^+$08], which is a highly effective library for large-scale linear classification. This library handles Support Vector Machines (SVMs) classification and Logistic Regression classification with different regularisation and loss functions.

Set	Feature
Vocabulary	Unigrams and bigrams (binary)
Length	Length of the sentence Length of the nucleus Length of the satellite Length of the document that contains the sentence
Positional	Absolute position of the sentence in the document Relative position of the sentence in the document
POS	Number of occurrences of every POS tags (one feature for each POS tag, see Table 3.8)
Syntactic Patterns	The presence of a POS syntactic pattern (one binary feature for each pattern defined in Table 3.9)
Sentiment Lexicon	Number and proportion of subjective terms in the sentence Number and proportion of exclamations and interrogations in the sentence
RST	Contains a satellite (binary) Contains specific satellite types (binary)
Sentiment RST	Number and proportion of subjective terms in the nucleus Number and proportion of subjective terms in satellites Number and proportion of exclamations and interrogations in the nucleus Number and proportion of exclamations and interrogations in satellites

Table 3.11: Sentence features for subjectivity classification. The features related to satellites are defined for each specific type of rhetorical relation mentioned in Table 3.10.

These types of classifiers have performed very well in many learning problems. We extensively tested all the classifiers supported by *liblinear* against the training collections and selected the classifiers that performed the best.

We randomly split the dataset into a training and test set, consisting of 75% and 25% of the sentences, respectively[7]. With the training set, we applied 5-fold cross validation to set all the parameters of the classifiers and also to select the best performing classifier[8].

In most collections, the two-class categorisation problem is unbalanced: fewer subjective sentences than objective sentences (see Table 3.7). When dealing with unbalanced problems, discriminative algorithms such as SVMs, which maximise classification accuracy, result in trivial classifiers that completely ignore the minority class [Nal04]. Some of the typical methods to deal with this problem include oversampling the minority class (by repeating minority examples), under-sampling the majority class (by removing some examples from the major-

[7] We repeated this process 10 times and we averaged out the performance achieved to obtain a reliable estimation of effectiveness.

[8] Usually, the best classifier was a Logistic Regression classifier.

ity class), or adjusting the misclassification costs. Oversampling the minority class results in considerable computational costs during training because it significantly increases the size of the training collection. Under-sampling the majority class is not an option for our problem because we have a small number of positive examples and we would need to remove most of the negative examples in order to have sets of positive examples and negative examples that are comparable in size. This massive removal of negative examples would result in much information being missed. We therefore opted for adjusting the misclassification costs to penalise the error of classifying a positive example as negative (i.e., subjective sentence classified as a non-subjective). The training process was designed to maximise the $F1$ score computed with respect to the subjective class. Next, we used the test set to evaluate the best performing classifier against unseen data.

In Table 3.12, Table 3.13 and Table 3.14 (and Figure 3.10, Figure 3.11 and Figure 3.12) we report the subjectivity classification performance achieved on MOAT, MPQA and PL, respectively. Vocabulary-based classifiers (unigrams only, or unigrams combined with bigrams) were regarded as baselines and we incorporated various combinations of features into the baseline classifiers: Length, Position, POS tags, POS syntactic Patterns, Sentiment Lexicon, RST, and Sentiment RST (see Table 3.11). Additionally, we ran experiments with all features included (All).

The results reveal the following trends. Length features do not contribute to discriminate between objective and subjective sentences. Syntactic Patterns seem to be slightly beneficial when used on top of unigram representations. But these linguistic cues do not help in combination with bigrams. This indicates that bigrams are already capturing some structural aspects of subjective sentences.

POS features are valuable: in MPQA and PL they led to statistical significant improvements over the baselines and, in MOAT, performance remained roughly the same. This confirms the usefulness of counting POS labels to detect subjective content [Tur02].

Positional features seem to work particularly well for discovering subjective content. Where available[9], positional information helped to improve recall of subjective sentences. However, its ability to classify objective sentences seems to be limited. This might indicate a tendency of using subjective sentences in specific parts of the document, e.g., in the end of the document as a conclusion.

[9]Observe that we do not have positional information in the PL collection.

	Subjective			Objective			microavg	micro
Features	Prec.	Rec.	F_1	Prec.	Rec.	F_1	F_1	sign test
Unigrams	.5207	.4295	.4707	.8185	.8667	.8419	.7565	
+ Length	.4933	.4907	.4920	.8287	.8301	.8294	.7446	$\ll$
+ Position	.5046	.5111	.5078	.8345	.8309	.8327	.7503	$\sim$
+ POS	.5282	.4362	.4778	.8205	.8687	.8439	.7597	$\sim$
+ Synt. Patterns	.5387	.4286	.4774	.8198	**.8763**	**.8471**	.7635	$\gg$
+ Sentiment Lexicon	**.5401**	.4566	.4949	.8259	.8690	.8469	**.7650**	$\gg$
+ RST	.5305	.4242	.4714	.8182	.8735	.8449	.7602	$\sim$
+ Sentiment RST	.5316	.4814	.5053	.8306	.8570	.8436	.7623	$>$
+ All	.5021	**.5926**	**.5436**	**.8538**	.8019	.8270	.7492	$\sim$
Uni and bigrams	.5802	.3577	.4426	.8083	.9128	.8574	.7728	
+ Length	.5041	.4371	.4682	.8184	.8551	.8363	.7497	$\ll$
+ Position	.5332	.4632	.4957	.8268	.8633	.8447	.7625	$\ll$
+ POS	.5864	.3639	.4491	.8099	**.9135**	.8586	.7750	$\sim$
+ Synt. Patterns	.5747	.3546	.4386	.8074	.9116	.8563	.7712	$\sim$
+ Sentiment Lexicon	**.5895**	.3941	.4724	.8163	.9075	**.8595**	**.7781**	$>$
+ RST	.5587	.3794	.4519	.8113	.8990	.8529	.7680	$<$
+ Sentiment RST	.5698	.4326	.4918	.8231	.8899	.8552	.7746	$\sim$
+ All	.5009	**.6011**	**.5464**	**.8558**	.7982	.8260	.7485	$\ll$

Table 3.12: Subjectivity classification results for the MOAT collection, in terms of precision, recall, and F_1 scores for subjective and objective sentences. For each vocabulary representation (i.e., unigrams, or unigrams and bigrams), the best performance for each metric is bolded. The symbol $\gg$ (resp. $\ll$) indicates a significant improvement (resp. decrease) with respect to the vocabulary-based baselines, with $p \leq .01$. The symbol $>$ (resp. $<$) indicates a significant improvement (resp. decrease) with respect to the vocabulary-based baseline with $0.1 < p \leq .05$. $\sim$ indicates that the difference was not statistically significant ($p > .05$).

Binary RST-based features did not work well. Apparently, the presence of particular rhetorical relations per se does not convey much more information than unigrams and bigrams do.

The best performing combination was the one that included the Sentiment Lexicon features. It was the only feature set able to statistically improve the baselines in all situations across the different test sets. Combining unigrams or bigrams with a sentiment lexicon is a way to account for both general purpose opinion expressions and domain-specific opinion expressions. This led to robust subjectivity classifiers.

Sentiment RST features, which weight opinionated terms within the RST spans of the sentences, led to modest improvements over the baselines. These improvements were inferior

	Subjective			Objective			microavg	micro
Features	Prec.	Rec.	F_1	Prec.	Rec.	F_1	F_1	sign test
Unigrams	.7172	.7222	.7197	.7597	.7552	.7574	.7399	
+ Length	.6683	.7007	.6841	.7315	.7010	.7159	.7009	$\ll$
+ Position	**.7311**	.7564	.7435	.7841	**.7608**	**.7723**	.7588	$\gg$
+ POS	.7217	.7248	.7232	.7625	.7597	.7611	.7436	$\gg$
+ Synt. Patterns	.7170	.7268	.7219	.7623	.7533	.7578	.7410	$\sim$
+ Sentiment Lexicon	.7250	.7383	.7316	.7714	.7592	.7653	.7495	$\gg$
+ RST	.7150	.7272	.721	.7620	.7507	.7563	.7399	$\sim$
+ Sentiment RST	.7236	.7370	.7302	.7702	.7579	.7640	.7483	$\gg$
+ All	.7111	**.8116**	**.7580**	**.8156**	.7164	.7628	**.7604**	$\gg$
Uni and bigrams	.7226	.7069	.7147	.7526	.7666	.7595	.7390	
+ Length	.6756	.7126	.6936	.7407	.7058	.7228	.7089	$\ll$
+ Position	.7248	.7551	.7396	.7816	.7534	.7672	.7542	$\gg$
+ POS	.7234	.7135	.7184	.7565	.7655	.7610	.7414	$>$
+ Synt. Patterns	.7191	.7126	.7158	.7548	.7607	.7577	.7384	$\sim$
+ Sentiment Lexicon	.7323	.7212	.7267	.7634	**.7733**	.7683	.7492	$\gg$
+ RST	.7188	.7123	.7155	.7545	.7604	.7574	.7382	$\sim$
+ Sentiment RST	.7250	.7254	.7252	.7638	.7634	.7636	.7458	$\gg$
+ All	**.7357**	**.7830**	**.7586**	**.8025**	.7581	**.7797**	**.7696**	$\gg$

Table 3.13: Subjectivity classification results for the MPQA collection, in terms of precision, recall, and F_1 scores for subjective and objective sentences. For each vocabulary representation (i.e., unigrams, or unigrams and bigrams), the best performance for each metric is bolded. The symbol $\gg$ (resp. $\ll$) indicates a significant improvement (resp. decrease) with respect to the vocabulary-based baselines, with $p \leq .01$. The symbol $>$ (resp. $<$) indicates a significant improvement (resp. decrease) with respect to the vocabulary-based baseline with $0.1 < p \leq .05$. $\sim$ indicates that the difference was not statistically significant ($p > .05$).

to those found with Sentiment Lexicon features. This suggests that Sentiment RST features are not more discriminative than pure lexicon-based features for subjectivity classification.

Finally, when combining all features into a single classifier we obtained a good classifier in terms of recall of subjective sentences but recall of objective sentences tended to fall. This led to classification performance that was sometimes worse than the baseline's performance (e.g., in MOAT, all features combined led to performance decreases that were statistically significant).

	Subjective			Objective			microavg	micro
Features	Prec.	Rec.	F_1	Prec.	Rec.	F_1	F_1	sign test
Unigrams	.8939	.8910	.8924	.8916	.8944	.8930	.8927	
+ Length	.8614	.8940	.8774	.8901	.8565	.8730	.8752	$\ll$
+ Position	–	–	–	–	–	–	–	
+ POS	**.9007**	**.9008**	**.9007**	**.9010**	**.9009**	**.9009**	**.9008**	$\gg$
+ Synt. Patterns	.8969	.8955	.8962	.8959	.8973	.8966	.8964	$\gg$
+ Sentiment Lexicon	.8926	.8995	.8960	.8989	.8920	.8954	.8958	$\gg$
+ RST	.8934	.8910	.8922	.8915	.8939	.8927	.8924	$\sim$
+ Sentiment RST	.8903	.9004	.8953	.8995	.8892	.8943	.8948	$\sim$
+ All	.8965	**.9008**	.8986	.9006	.8963	.8984	.8986	$\gg$
Uni and bigrams	.9043	.8942	.8992	.8956	.9055	.9005	.8999	
+ Length	.8829	.8811	.8820	.8816	.8834	.8825	.8822	$\ll$
+ Position	–	–	–	–	–	–	–	
+ POS	**.9099**	.8945	.9021	.8965	**.9116**	**.9040**	**.9031**	$>$
+ Synt. Patterns	.9047	.8930	.8988	.8946	.9062	.9004	.8996	$\sim$
+ Sentiment Lexicon	.9016	.8964	.899	.8973	.9024	.8998	.8994	$\sim$
+ RST	.9054	.8888	.8970	.8910	.9073	.8991	.8980	$\ll$
+ Sentiment RST	.9034	.8916	.8975	.8932	.9049	.8990	.8982	$\sim$
+ All	.9063	**.8986**	**.9024**	**.8997**	.9073	.9035	.9030	$>$

Table 3.14: Subjectivity classification results for the PL collection, in terms of precision, recall, and F_1 scores for subjective and objective sentences. For each vocabulary representation (i.e., unigrams, or unigrams and bigrams), the best performance for each metric is bolded. The symbol $\gg$ (resp. $\ll$) indicates a significant improvement (resp. decrease) with respect to the vocabulary-based baselines, with $p \leq .01$. The symbol $>$ (resp. $<$) indicates a significant improvement (resp. decrease) with respect to the vocabulary-based baseline with $0.1 < p \leq .05$. $\sim$ indicates that the difference was not statistically significant ($p > .05$).

3.3.3 Feature Weights

After obtaining a linear SVM model, the weights (w_i) of the separating hyperplane can be used to assess the relevance of each feature [CL08]. The larger $|w_i|$ is, the more important the i_{th} is in the decision function of the SVM. Only linear SVM models have this indication, which naturally facilitates the analysis of the classifiers. This useful property has been used to gain knowledge of data and, for instance, to do feature selection [CL08, MFM02]. A proper and direct comparison of the weights can only be done if all features are scaled into the same range. We focus our analysis on the *unigrams & bigrams + All* classifier obtained from the MOAT dataset after scaling the features into [0,1]. Table 3.15 presents the top 50 features ranked by decreasing absolute weight ($|w_i|$). A positive weight ($w_i > 0$) means that

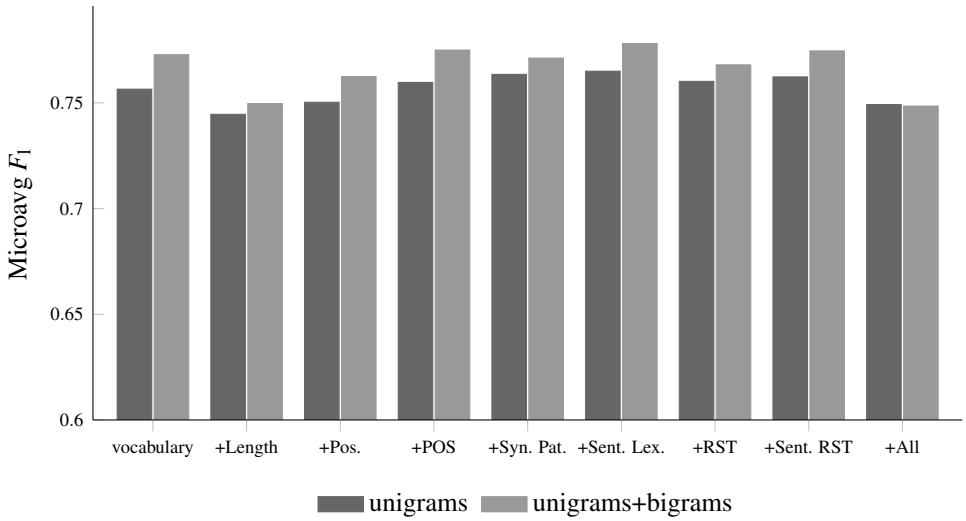

Figure 3.10: Microavg F_1 performance obtained by different opinion classifiers in the MOAT collection.

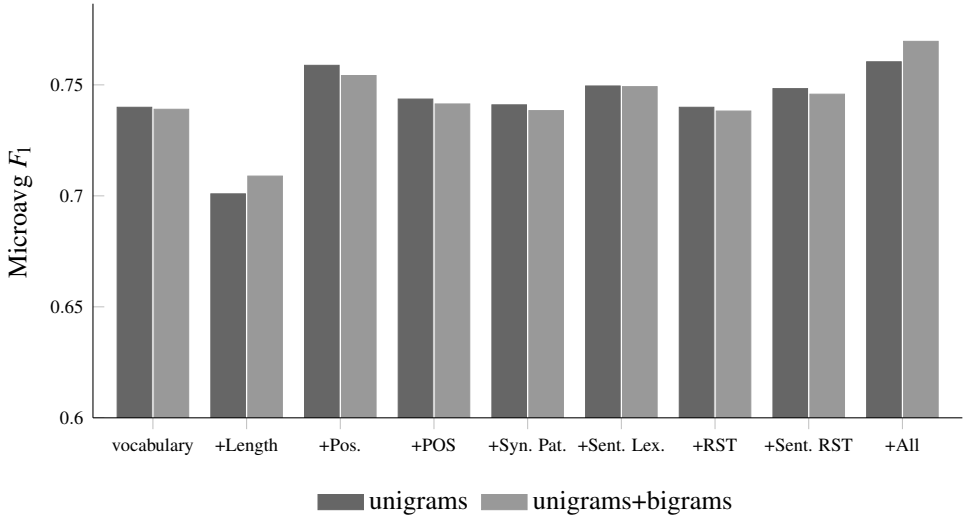

Figure 3.11: Microavg F_1 performance obtained by different opinion classifiers in the MPQA collection.

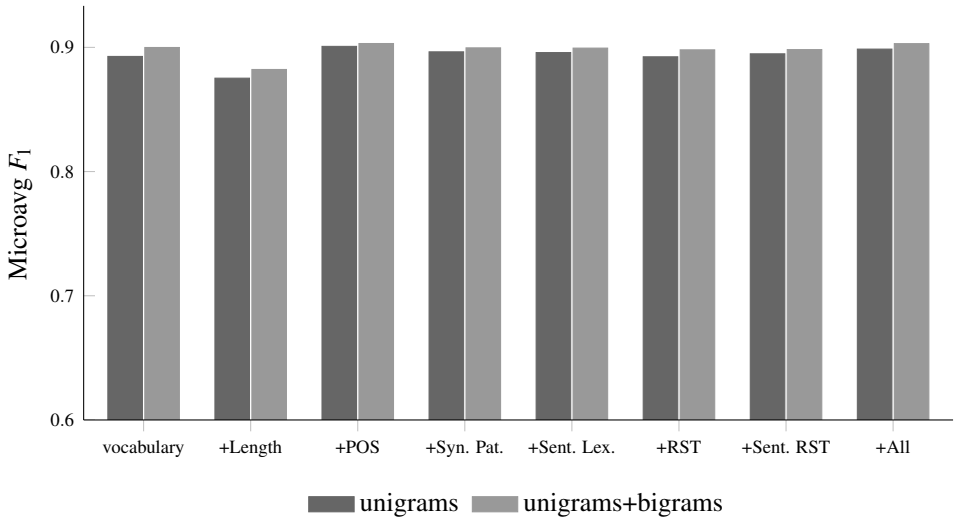

Figure 3.12: Microavg F_1 performance obtained by different opinion classifiers in the PL collection.

high values of the feature are indicative of the membership of the sentence into the subjective class. On the other hand, a negative weight ($w_i < 0$) means that high values of the feature are indicative of the membership of the sentence into the objective class. We can see that the most discriminative features are the number of subjective words in the sentence and the number of subjective words in the nucleus. This demonstrates the importance of the sentiment lexicon in the classification process. The subjective terms in the nucleus are more important than the subjective terms in satellites –not ranked among the top 50 features. This highlights the importance of the nucleus in subjectivity classification. The number of exclamations and interrogations has a high negative w_i score. This means that exclamations/interrogations are indicative of objectivity. This is intringuing, because exclamations or interrogations have been associated to subjective behaviour [AXV+11]. This outcome might be related to the nature of the documents (news articles). For instance, a journalist may write in a way that tries to attract people's attention (e.g., open questions).

Among the top 25 non-vocabulary features (Table 3.16), there are several POS features that help to detect subjective sentences. This demonstrates the usefulness of these features to detect opinions. For instance, the number of a comparative adjectives (JJR, see Table 3.8) is associated to subjectivity. Another important feature was the position of the sentence in

rank	w_i	feature	feature set	rank	w_i	feature	feature set
1	3.3334	#subj terms	Sent. Lex.	26	1.4520	leadership	vocab.
2	2.3455	#subj terms nuc.	Sent. RST	27	-1.4481	the economy	vocab.
3	-2.1189	#sent doc.	Length	28	-1.4430	charge	vocab.
4	-2.1123	#exc.int. nuc.	Sent. RST	29	1.4313	US economy	vocab.
5	1.9680	actions	vocab.	30	1.4274	maintain	vocab.
6	-1.9035	#POS(NNP)	POS	31	1.4262	prepared	vocab.
7	1.8896	they are	vocab.	32	1.4233	countries to	vocab.
8	1.8859	fear	vocab.	33	-1.4107	market	vocab.
9	-1.8012	#POS(CD)	POS	34	1.4054	policy	vocab.
10	-1.7807	key	vocab.	35	1.4031	stop	vocab.
11	1.7614	finally	vocab.	36	-1.4011	Russia	vocab.
12	1.7343	something	vocab.	37	-1.4010	million	vocab.
13	-1.6934	will have	vocab.	38	-1.4002	Although	vocab.
14	-1.6533	weather	vocab.	39	-1.3915	officials	vocab.
15	1.6528	way	vocab.	40	-1.3861	closer	vocab.
16	-1.6402	#POS(VBN)	POS	41	1.3795	terrorists	vocab.
17	1.6329	is that	vocab.	42	1.3730	service	vocab.
18	1.6252	Still	vocab.	43	-1.3637	accept	vocab.
19	-1.6085	financial	vocab.	44	1.3599	set to	vocab.
20	1.5446	should	vocab.	45	-1.3535	with a	vocab.
21	1.5323	capitalist	vocab.	46	1.3494	expressed	vocab.
22	-1.5277	interests	vocab.	47	1.3485	objections	vocab.
23	1.5171	The economic	vocab.	48	-1.3262	to use	vocab.
24	-1.4622	case	vocab.	49	1.3235	like	vocab.
25	1.4557	on this	vocab.	50	1.3208	Marzuki	vocab.

Table 3.15: List of the 50 features with the highest $|w_i|$ in the best(scaled) classifier. The features are ranked by decreasing $|w_i|$.

the document. The sentence position feature represents the order of a concrete sentence in its document (e.g., the third sentence of a document has a score of 3). A high w_i weight makes that the final sentences have more chance of being labelled as subjective. This suggests that writers tend to summarise their overall viewpoints at the end of the document. Finally, some satellites seem to be important for the classifier. For instance, *evaluation* and *attribution* relationships are highly indicative of opinionated sentences (weights 0.9420 and 1.0283, respectively) and *attribution* satellites occur when the writer (e.g., journalist) discusses others' opinions (e.g., *According to the new CEO, the future of the company is brilliant.*). On the other hand, *background* relationships are indicative of objective sentences (weight -0.7317).

In fact, *background* statements are ofen used to help a reader to sufficiently comprehend matters presented in nuclei but do not necessary convey opinions.

rank	w_i	feature	feature set
1	3.3334	#subj terms	Sent. Lex.
2	2.3455	#subj terms nuc.	Sent. RST
3	-2.1189	#sent doc	Length
4	-2.1123	#exc.int. nuc.	Sent. RST
5	-1.9035	#POS(NNP)	POS
6	-1.8012	#POS(CD)	POS
7	-1.6402	#POS(VBN)	POS
8	1.2974	sent. position	Position
9	1.2344	#POS(NNPS)	POS
10	-1.2036	#POS(VBD)	POS
11	1.0293	has satellite (Attribution)	RST
12	-0.9953	Position norm.	Position
13	0.9420	has satellite (Evaluation)	RST
14	0.8772	#suj terms sat. (Cause)	Sent. RST
15	0.8363	#POS(JJR)	POS
16	-0.8008	#exc.int. nuc. norm	Sent. RST
17	0.7605	#POS(POS)	POS
18	-0.7586	#POS(NNS)	POS
19	-0.7340	#POS(VB)	POS
20	-0.7326	#POS(CC)	POS
21	-0.7317	has satellite (Background)	RST
22	0.6980	#subj terms sat. norm (Cause)	Sent. RST
23	0.6043	length satellite	Length
24	-0.6000	#of POS(RP)	POS
25	0.5725	#subj terms sat. norm (Joint)	Sent. RST

Table 3.16: List of the top 25 non-vocabulary features with the highest $|w_i|$ in the $unigrams\&bigrams + All(scaled)$ classifier. The features are ranked by decreasing $|w_i|$.

3.4 Conclusions

In this chapter we studied the impact of different methods on searching for opinionated text. First, we considered the role of comments as a source of opinionated terms for query expansion in blogs. To this aim, we used RMs, a state of the art PRF approach, to expand the original factual query with terms provided by the comments associated to blogs posts. The proposed method significantly outperforms the classical $RM3$ estimation for an opinion finding task.

We provided experimental evidence showing that the comments are very useful to move the query towards opinionated words. This novel expansion approach is particularly consistent as a high precision mechanism. These results highlight the importance of comments to enhance precision without harming recall (MAP is roughly the same with either expansion methods). This suggests that subjective words estimated from comments lead to a more accurate query-dependent opinion vocabulary. Furthermore, the independence of our method of any external lexicon is important because, in many domains and languages, there is a lack of good opinion resources. In our study we applied an homogeneous treatment to all types of queries. However, in some cases this could be harming. In the future, we would like to study methods to dynamically adapt our expansion techniques depending on the quality of the initial query [SMK05].

We have also demonstrated that passage-level methods can help to estimate subjectivity in blogs. We showed that positional evidence can be a valuable clue in subjectivity classification. This opens a new line of research: how to incorporate new forms of location-aware features in opinion retrieval processes. Location features in combination with passage-level features are a good guidance for extracting opinions related to a query topic. Our experimental results indicate that the sentence with the highest aggregated score of relevance and subjectivity provides the best representation for subjectivity estimation blogs. This finding could be applied in the future to create opinion-biased summaries. We approached the combination of evidence –subjectivity location and topicality– in an adhoc way. We are aware that there might be better and more formal ways of combining evidence (e.g., subjectivity and relevance might be combined using formal methods to learn query-independent weights [CRZT05]). 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 into our models. Related to this, we are also interested in studying more refined ways of representing the sentiment flow of the documents.

In the last part of this chapter we proposed several classification methods to estimate whether or not a sentence is opinionated. We have studied classical features such as n-grams or lexicon-based counts in combination with more advanced features such as location and rhetorical features. Among all features tested, length and binary RST-based features did not give any added value. Positional features worked well as a recall-oriented mechanism for detecting subjective sentences and POS features were valuable for subjectivity classification.

Nevertheless, a score based on counting sentiment terms from a general-purpose vocabulary, i.e., Sentiment Lexicon, was at least as effective as accounting for POS labels. Syntactic patterns were only beneficial when the baseline classifier handled unigrams representations. With bigrams, syntactic patterns did not give any added value. Lexicon-based features consistently give high performance in sentence subjectivity classification, being the best approach (in combination with n-grams). Finally, Sentiment RST features were slightly inferior to pure lexicon-based features. Overall, classifying sentences based on sentiment lexicon scores and unigrams/bigrams is an effective and safe choice for subjectivity classification.

CHAPTER 4

POLARITY ESTIMATION

In this chapter we are concerned with polarity estimation, a challenging area that is more demanding than subjectivity estimation. We explore the role of positional information and passage retrieval techniques for analysing key evaluative statements and for determining the overall orientation of the text. We also study how beneficial structural and discourse features are for sentiment classification in documents and sentences.

4.1 Polarity Estimation at Document level

In this section we explore polarity estimation at document level. As argued above, polarity estimation (i.e., determining the overall orientation of a document) is a challenging task that is more difficult than subjectivity classification. For instance, there may be conflicting opinions in a given evaluative document (e.g., 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 final orientation of the document. This issue is illustrated in the blog posts presented in Figure 4.1 and Figure 4.2. In Figure 4.1, despite the start of the post being predominantly negative, with several negative comments being made, the overall recommendation seems to be positive. Similarly, in Figure 4.2 we observe a mixed set of opinions across the document, which is finished with an overall positive recommendation.

The location of subjective sentences may offer important clues when attempting to establish the polarity of the text. In the previous examples, the last sentences are the ones that express the overall view. Existing literature on blog polarity estimation has disregarded this

```
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.
```

Figure 4.1: Example of a Gran Torino's review taken from a popular film reviews' blog: `http://blog.moviefone.com`. Observe that the last sentence is the one that represents the overall recommendation of the writer about the film.

```
Last night, I watched the events in its entirely and I have to say
I'm pretty excited by what I saw. Let's get a few objections out
of the way first...
...
But I thing that the Apple announcement yesterday once again gave
us a peak into the future.
```

Figure 4.2: Example of blog post (by Joel Burslem) about Apple's iphone 4S event.

valuable information. For instance, the most effective polarity systems participating in the TREC blog tracks [OMS08, LhNK+08] did not incorporate any feature based on this flow of sentiments. Instead, they applied a document-level estimation 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 we claim that more effective polarity estimation methods can be designed using a sentence-level approach.

Off-topic content in documents may be harming. Documents might have query terms in a wrong context and it is challenging to design robust polarity detection techniques that can be applied effectively across different underlying topic retrieval baselines. For instance, in 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, was not inferior to most of the algorithms) [OMS08]. Actually, only one TREC 2008 participant had on average improved the polarity performance of the five topic retrieval baselines provided by the task. This illustrates the difficulty of designing effective polarity estimation methods and encourages us to define more evolved polarity estimation models.

In Section 3.2 of this thesis we defined some subjectivity estimation methods that integrate sentence-level subjectivity and topicality. Some of the proposed variants performed well, but the methods that take sentence position into account were not very effective. We hypothesise that the position of the opinions is more important in polarity classification. The position of a sentence can determine its influence on the overall recommendation of the document. For instance, the last subjective sentences of the document might represent the final recommendation of the writer. To test this hypothesis, we extend the methods proposed in the Section 3.2 for polarity estimation.

4.1.1 Polarity Estimation at Sentence Level

In order to have a precise representation of the mixed set of opinions in a document, we compute polarity at sentence level. From the polar terms tagged by OF we can define the positive or negative polarity score of a sentence. To promote polar sentences that are on-topic, we combine relevance and polarity scores as follows:

$$pol(S,Q) = \beta \cdot rel_{norm}(S,Q) + (1-\beta) \cdot pol(S) \tag{4.1}$$

where $rel_{norm}(S,Q)$ is the sentence-query similarity measure (with Lemur's tf-idf weights) 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[1]. $\beta \in [0,1]$ is a free parameter.

4.1.2 Document Polarity Score

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 ($pol_S(D,Q)$)[2]:

– *PolMeanAll:* The mean of $pol(S,Q)$ scores computed across all polar sentences in the document. This measure is a natural choice to estimate the overall polarity of a document.

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

[2] Observe that these alternatives are the polarity-oriented variants of the subjectivity aggregation methods proposed in Section 3.2.

– *PolMeanBestN:* The mean of $pol(S,Q)$ scores from the n sentences with the highest $pol(S,Q)$ 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 $pol(S,Q)$ 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:

$$pol(D,Q) = \gamma \cdot rel_{norm}(D,Q) + (1-\gamma) \cdot pol_S(D,Q) \tag{4.2}$$

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. Again, some aggregation techniques have an extra parameter: the number of sentences (n). By studying the behaviour of this parameter we might discover valuable patterns about the way in which bloggers express their views.

4.1.3 Experiments

The experimental setting is the same presented in Section 3.2, but we now evaluate the ability of the algorithms to search for positive or negative blog posts. Tables 4.1 and 4.2 show the results of the polarity estimation approaches against the two reference collections. 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 metric for each method, we compute the mean of the MAPs (resp. P@10s) from the two rankings (positive and negative). This is denoted as Mix MAP and Mix P@10 respectively[3]. The best value in each column

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

for each baseline is underlined. Statistical significance was estimated using the paired t-test at the 95% level. The symbols $\triangle$ and $\triangledown$ 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 proposed in section 3.2 (eq. 3.12, $subj_{DOC}(D)$). This method, with no polarity capabilities, serves as a reference comparison for polarity-oriented approaches. The symbols ▲ and ▼ indicate a significant improvement or decrease over the subjectivity method. We also report the average MAP and P@10 scores (computed across the five baselines) of our methods for positive and negative rankings (Figure 4.3 and Figure 4.4).

The technique that looks superior across all cases 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 4.1.4. 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. Not surprisingly, subjectivity information alone is not useful in polarity estimation (the subjectivity method hardly shows any significant improvement in performance with respect to the baseline).

Another observation is that the performance of negative document rankings is quite poor. It is interesting to note that TREC systems (see Table 4.3) 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 larger 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).

To put things in perspective, we report in Table 4.3 how our methods compare with those proposed by teams participating in TREC [OMS08]. 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

	Negative		Positive		Mix	
	MAP	P@10	MAP	P@10	MAP	P@10
baseline1	.0569	.0620	.1779	.2640	.1174	.1630
+$subj_{DOC}$	.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
+$subj_{DOC}$	.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
+$subj_{DOC}$	.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
+$subj_{DOC}$	.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
+$subj_{DOC}$	.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 4.1: Polarity Retrieval Results in TREC 2007. 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. The symbols ▲ and ▼ indicate a significant improvement or decrease over the subjectivity method.

	Negative		Positive		Mix	
	MAP	P@10	MAP	P@10	MAP	P@10
baseline1	.1175	.1700	.1364	.1860	.1270	.1780
+$subj_{DOC}$	.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
+$subj_{DOC}$	.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
+$subj_{DOC}$	.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
+$subj_{DOC}$	.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
+$subj_{DOC}$	.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 4.2: Polarity Retrieval Results in TREC 2008. 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. The symbols ▲ and ▼ indicate a significant improvement or decrease over the subjectivity method

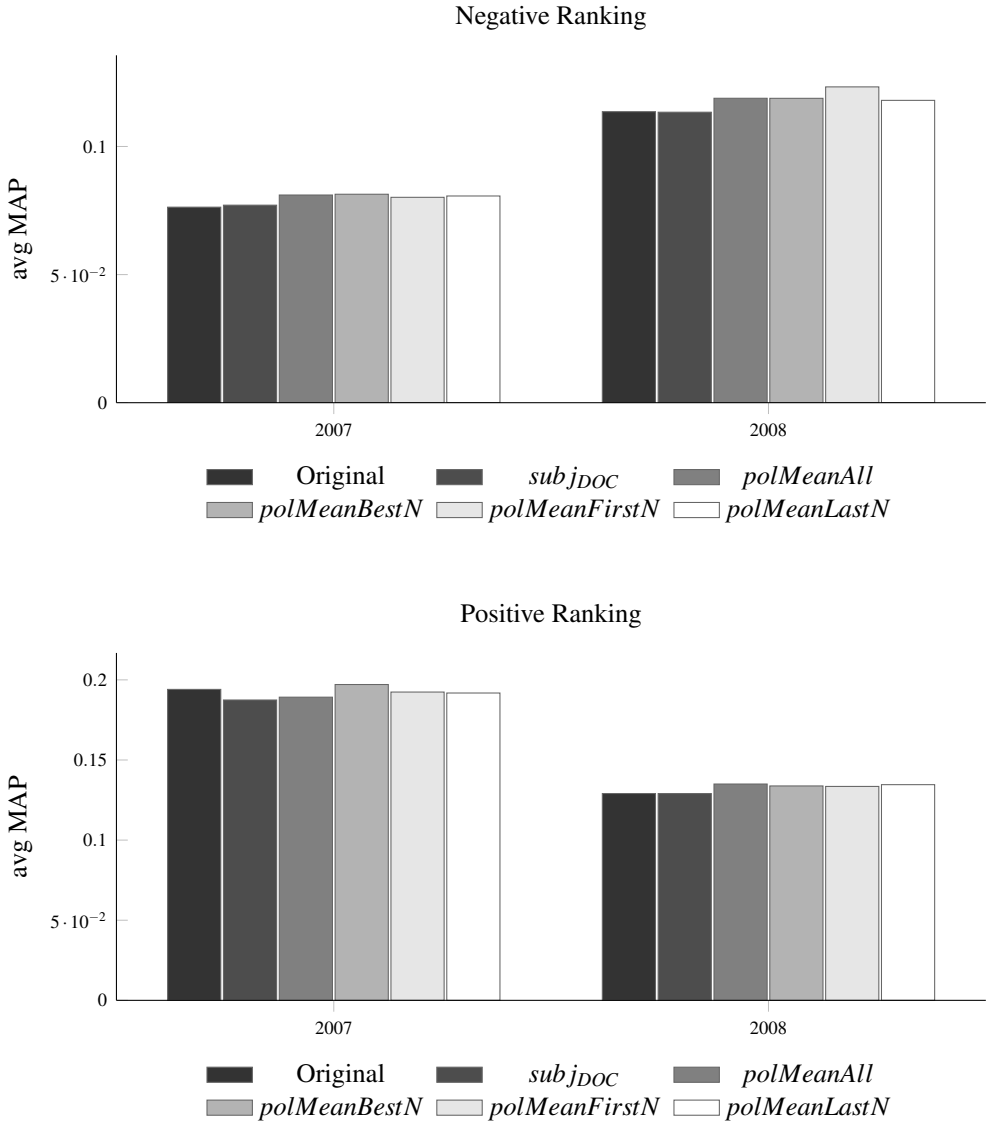

Figure 4.3: Average MAP (computed across the five baselines) obtained by the polarity methods. TREC 2007 and 2008 topics, and positive and negative rankings.

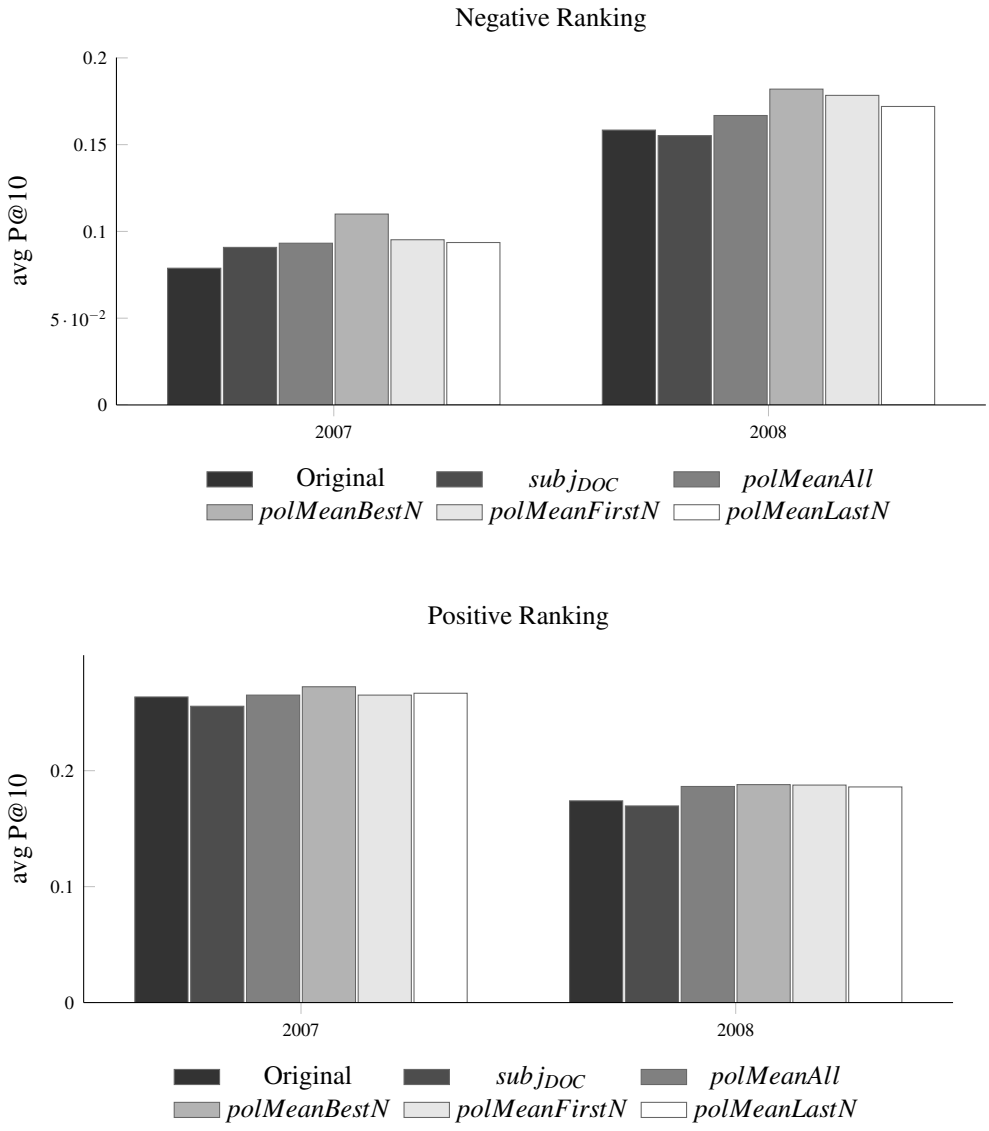

Figure 4.4: Average P@10 (computed across the five baselines) obtained by the polarity methods. TREC 2007 and 2008 topics, and positive and negative rankings.

	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 4.3: 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.

than the baselines[4]. The methods proposed here perform as well as the best TREC polarity approach (KLE, Pohang University of Science and Technology) [LhNK+08], showing better performance for some configurations. These results show that sentence-level methods are an effective strategy for polarity estimation, performing comparably to state-of-the-art TREC systems.

Table 4.4 reports the parameter values trained for each method. 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. 3.11. The value of this parameter indicates that the relevance component is much more important than the subjectivity component.

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. 4.1). This means that in positive rankings the polarity evidence is more important than content-match evidence. This

[4]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.

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

Table 4.4: Parameters trained.

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 the near future.

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 larger subset of sentences. From Table 4.4 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 sentence with the highest $pol(S,Q)$ as the best guidance to understand the overall sentiment of a blog. Observe that this behaviour is similar to the one obtained for subjectivity (i.e., the most subjective sentence was the best guidance to estimate whether or not a document is subjective). This finding can be useful, to build opinion summaries with positive and negative sentiments.

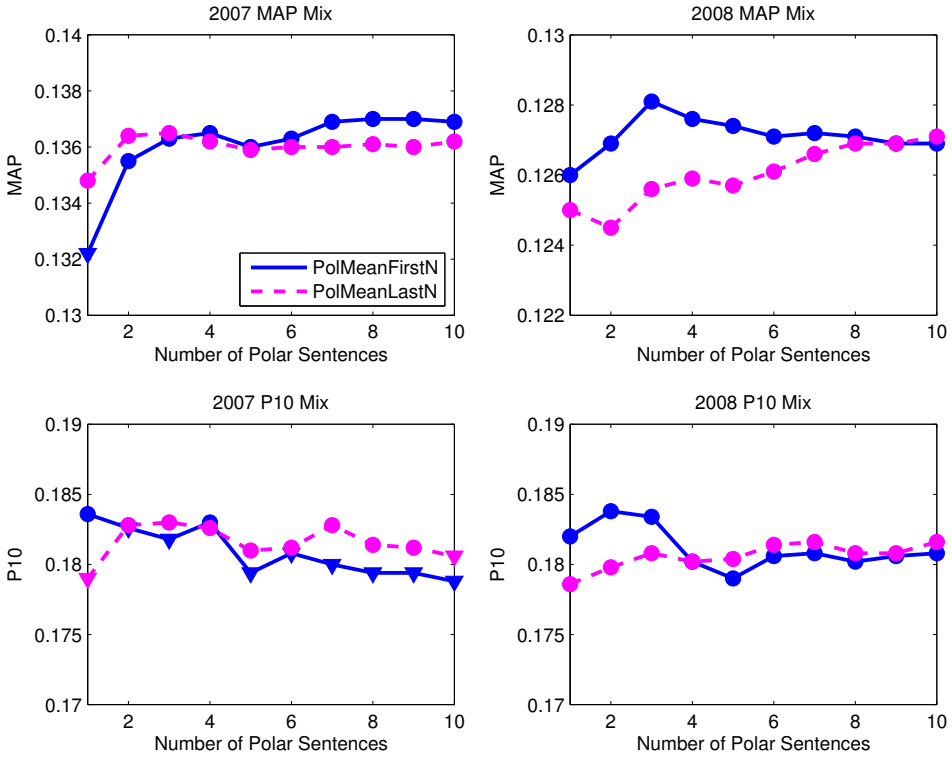

Figure 4.5: Performance of polarity methods against the number of sentences utilised. A ▼ indicates a significant decrease in performance over the PolMeanBestN method, while a • indicates a non significant difference in performance with respect to the PolMeanBestN method.

4.1.4 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 sentence with the highest $pol(S,Q)$ as a representation of the sentiments of the author. However, the methods based on sentences taken from specific document locations often work quite well. This is an interesting finding. Location-aware methods were not very effective for subjectivity purposes (Section 3.2) but they appear to be more robust for polarity estimation purposes.

Figure 4.5 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). This means that 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.

4.1.5 Effectiveness vs Efficiency

In the previous subsection 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 $subj_{DOC}$ need to classify all sentences in the post to compute the polarity score of a document. In contrast, the location-based methods only 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 [HMO08, HMHO08]. 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[5]. 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 above). 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

[5]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 spam documents, which tend to be large.

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

Table 4.5: Average time taken to classify complete documents vs the time taken to classify narrower subsets containing the first/last polar sentences.

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

This classification, which is required to compute $pol(S)$ (eq. 4.1), 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. 4.2) 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 [LhNK$^+$08]. 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.

4.2 Rhetorical Structure Theory for Polarity Estimation

In Chapter 3 we discussed the benefits of RST for subjectivity estimation. We argue that RST is even more promising for polarity detection. For instance, let us consider the sentence "*Although I like the characters, the book is horrible*". This sentence can be easily classified as subjective by straightforward keyword-based methods (e.g., sentiment lexicons). But determining whether this sentence is positive or negative is a more intriguing task that cannot be merely solved by counting sentiment-bearing terms. The core of the sentence, i.e., the nucleus, provides a negative sentiment with respect to a book ("*the book is horrible*"). The other segment is a satellite with contrasting information with respect to the nucleus, admitting to

some positive aspects of the book ("*Although I like the characters*"). For a human reader, the polarity of this sentence is clearly negative. However, in a classical (word-counting) sentiment analysis approach, all words would contribute equally to the total sentiment, thus yielding a verdict of a neutral or mixed polarity at best. Exploiting the information contained in the RST structure could result in the nucleus being given a higher weight than the satellite, thus shifting focus to the nucleus segment. In order to exploit the rhetorical relations distinct rhetorical roles of individual text segments should be treated differently when aggregating the sentiment conveyed by these text segments. This could be accomplished by assigning different weights to distinct rhetorical roles, quantifying their contribution to the overall sentiment conveyed by a text [HGH+11]. For instance, accounting for the rhetorical roles of text segments by means of a RST-based analysis has proven to be extremely useful when classifying the overall document-level polarity of a limited set of movie reviews [HGH+11].

4.2.1 Efficiency Issues

Applying RST-based analysis at large scale (involving millions of documents) is challenging because of the computational complexity of these linguistically advanced techniques. In this section, we study how we can mitigate this issue.

In Section 4.1 we have shown that the sentence with the highest $pol(S,Q)$ of a blog post post is highly indicative of the overall recommendation of the author. This finding is convenient as it permits us to apply RST on narrow parts of the blog posts. In this section we therefore analyse the structure of the discourse only for the passages selected by *polMeanBestN* (see Section 4.1.2). This is beneficial to avoid noisy chunks of text and it is also convenient from a computational complexity perspective.

The level of positivity (resp. negativity) of a sentence is computed as the ratio of positive (resp. negative) terms found in that sentence (see Equation 4.1). This is a very rough lexicon-based approach. We claim that a more elaborated polarity analysis can be done by including RST into the computation of $pol(S)$. We propose to compute $pol(S)$ as a weighted sum of the polar terms occurring in the nucleus and the satellite, respectively:

$$pol(S) = w_{nuc} \cdot pol_{nuc}(S) + w_{sat} \cdot pol_{sat}(S), \tag{4.3}$$

where *nuc* represents the nucleus of the sentence S, *sat* is the satellite of the sentence S, w_{nuc} is the weight for nucleus, w_{sat} is the weight for the concrete satellite; and $pol_{nuc}(S)$ and $pol_{sat}(S)$ represent the ratio of positive (resp. negative) terms tagged in the nucleus and

satellite, respectively, of sentence S. Observe that w_{sat} and w_{nuc} are free parameters that need to be trained for each type of rhetorical relation.

4.2.2 Experiments

In this section, we report the experiments designed to determine the usefulness of RST in a large-scale multi-topic domain. Concretely, we use again the BLOG track but we focus our attention on polarity. Spam detection, topic retrieval in blogs, and subjectivity classification are out of the scope of these experiments. Hence, we analyse the effect of RST on the set of subjective documents identified by the standard baseline runs. This means that the input to the polarity estimation methods is a set of opinionated documents with varied polarity orientations (positive, negative, or mixed polarity). The objective is to distinguish the type of polarity that every document has (i.e., search for positive, and search for negative documents). This polarity task, per se, is quite challenging because there are many off-topic passages and conflicting opinions.

For parameter training (i.e., w_{sat} and w_{nuc}) we applied Particle Swarm Optimisation (PSO), which has shown its merits for automatic parameter tuning in IR [PVS12]. PSO is a population-based stochastic optimisation technique, inspired by the social behaviour of bird flocking or fish schooling, and included in swarm intelligence techniques. The potential solutions, called *particles*, fly through the problem space following the current optimum particles. The movements of the particles are guided by the best known position of each particle in the search space as well as the entire swarm's best known position. The process is repeated until a satisfactory solution is discovered.

The basic PSO algorithm is summarised in Algorithm 1. Each particle i stores its current position x_i^t , velocity v_i and its best known position pb_i^t at time t. Moreover, the algorithm considers the best known position of the entire swarm (gb^t).

Table 4.6 shows the performance results of the polarity approaches. The best value in each column for each baseline is underlined. The symbols ▲ and ▼ indicate a significant improvement or decrease over the corresponding baseline. To specifically measure the benefits of RST we compare its performance against the performance achieved by the original *PolMeanBest-N*. We also report the average MAP and P@10 (computed across the five baselines) of the different alternatives in Figure 4.6 and Figure 4.7.

PolMeanBestN estimates the overall recommendation of a blog post by taking into account the on-topic sentence in the blog post that has the highest polarity score (e.g., the most

Algorithm 1 Particle Swarm Optimization Algorithm

Initialise all particles i with random positions x_i^0 in search space as well as random velocities v_i^0.
Initialise the particle's best known position (pb^0) to its initial position.
Calculate the initial swarm's best known position gb^0 .
repeat
  for all Particle i in the swarm **do**
    Pick random numbers: $rp, rg \in (0,1)$
    Update the particle's velocity: $v_i^{t+1} = a * v_i^t + b * r_p * (pb_i^t - x_i^t) + c * r_g * (gb^t - x_i^t)$
    Compute the particle's new position: $x_i^{t+1} = x_i^t + v_i^{t+1}$
    if $fitness(x_i^{t+1}) < fitness(pb_i^t)$ **then**
      Update the particle's best known position: $pb_i^{t+1} = x_i^{t+1}$
    end if
    if $fitness(pb_i^t) < fitness(gb^t)$ **then**
      Update the swarm's best known position: $gb^{t+1} = pb_i^{t+1}$
    end if
  end for
until termination criterion is met
return The best known position: gb.

controversial contents of the post). This configuration leads to a performance comparable to the best performing approach at the TREC 2008 Blog track (KLE system) [SMM+12]. The RST technique proposed here is an evolution over *PolMeanBestN*, in which the estimation of polarity is also done with the highest polarity sentence but we take into account its RST structure (eq. 4.3). The symbols $\triangle$ and $\triangledown$ indicate a significant improvement or decrease over *PolMeanBestN*.

The technique that performs the best across all different baselines is the RST-based method, showing usually significant improvements with respect to both the baseline and *PolMeanBestN*. Another observation is that the performance of negative document rankings is lower than the performance of positive document rankings. This may be caused by negative documents being harder to find. Additionally, the lexicon-based identification of negative documents may be thwarted by people having a tendency of using rather positive words to express negative opinions [HGH+11]. This is somehow addressed by RST, which does not apply a crude counting-based method.

Table 4.7 shows the weights learnt for different RST elements. The weight of the nucleus was fixed to one. Weights of satellites are real numbers in the interval $[-2,2]$. Having been

	Negative		Positive	
	MAP	P@10	MAP	P@10
baseline1	.2402	.2960	.2662	.3680
+*polMeanBestN*	.2408	.3000	.2698	.3720
+*polMeanBestN*(*RST*)	.2516	.3180△ ▲	.2733	.3740△ ▲
baseline2	.2165	.2780	.2390	.3340
+*polMeanBestN*	.2222	.2820	.2368	.3160
+*polMeanBestN*(*RST*)	.2261▲	.3100△ ▲	.2423△	.3560△ ▲
baseline3	.2488	.2840	.2758	.3500
+*polMeanBest*	.2524	.2760	.2755	.3420
+*polMeanBestN*(*RST*)	.2584△ ▲	.2820	.2770△	.3380▼
baseline4	.2636	.2740	.2731	.3580
+*polMeanBestN*	.2730	.2840	.2705	.3500
+*polMeanBestN*(*RST*)	.2825△	.3240△ ▲	.2716	.3620△ ▲
baseline5	.2238	.3000	.2390	.3600
+*polMeanBestN*	.2279	.3120	.2404	.3580
+*polMeanBestN*(*RST*)	.2393	.3420△ ▲	.2786△ ▲	.4380△ ▲

Table 4.6: RST Polarity Results. The best value in each column for each baseline is underlined. The symbols ▲ and ▼ indicate a significant improvement or decrease over the corresponding baseline. The symbols △ and ▽ indicate a significant improvement or decrease over *PolMeanBestN*.

assigned a weight of 1, nuclei are assumed to play a more or less important role in conveying the overall sentiment of a piece of natural language text. Yet, some types of satellites appear to play an important role as well in conveying the overall sentiment of a document. For instance, the most salient relations (highest percentage of appearance in the collection) in our training set are the *elaboration* and the *attribution* relation. For both positive and negative documents, satellite segments elaborating on matters presented in nuclei are typically assigned relatively high weights, exceeding those assigned to nuclei. Bloggers may, therefore, tend to express their sentiment in a more apparent fashion in elaborating segments rather than in the core of the text itself. A similar pattern emerges for *attribution* satellites as well as for persuasive text segments, i.e., those involved in *enablement* relations, albeit to a more limited extent (lower frequency of occurrence). Interestingly, however, the information in *attribution* satellites are more important in negative documents than in positive documents. Another important observation is that the sentiment conveyed by elements in contrast satellites gets a

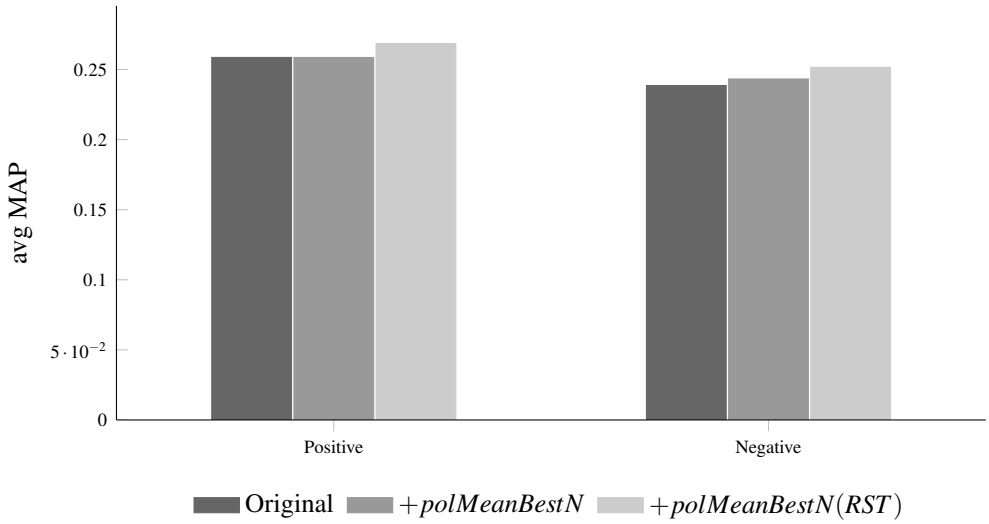

Figure 4.6: Average MAP performance obtained by *polMeanBestN* and *polMeanBestN*(*RST*) methods.

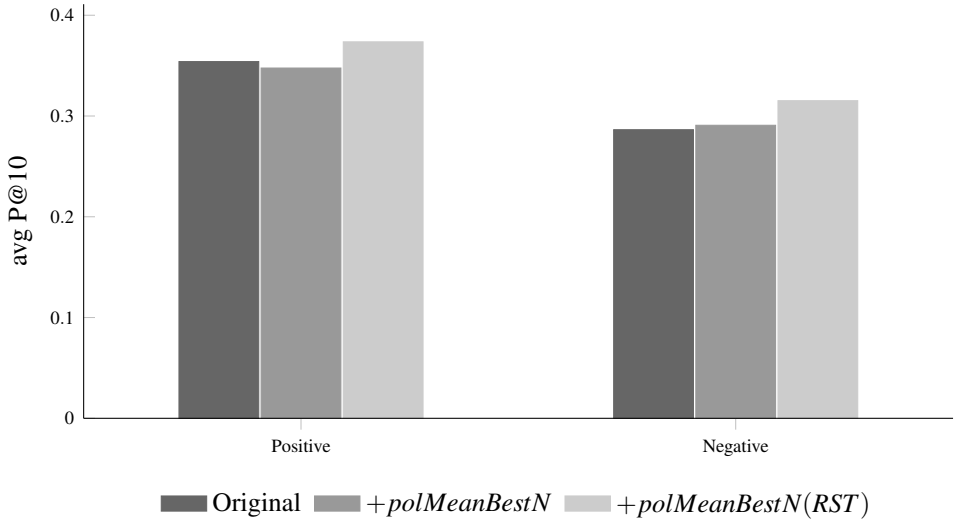

Figure 4.7: Average P@10 performance obtained by *polMeanBestN* and *polMeanBestN*(*RST*) methods.

	Positive		Negative	
Relation	% of occurrence	Weight	% of occurrence	Weight
attribution	.183	0.531	.177	2.000
background	.034	-0.219	.038	-2.000
cause	.009	1.218	.009	-0.011
comparison	.003	-1.219	.003	-2.000
condition	.029	-0.886	.025	-2.000
consequence	.001	0.846	.001	1.530
contrast	.016	-1.232	.017	-2.000
elaboration	.207	2.000	.219	2.000
enablement	.038	2.000	.038	1.221
evaluation	.001	0.939	.001	-2.000
explanation	.007	2.000	.008	2.000
joint	.009	-1.583	.010	1.880
otherwise	.001	-1.494	.001	-0.428
temporal	.003	-2.000	.003	-0.448

Table 4.7: Optimised weights for RST relation types trained with PSO over positive and negative rankings and the percentage of presence of different relations in the training set

negative weight. This permits to appropriately estimate the polarity of sentences such as the one we introduced in the beginning of this section ("*Although I like the characters, the book is horrible*").

4.3 Classification of Positive vs Negative Sentences

In the previous section we were concerned with polarity detection at document level. Basically, given a document, we analysed its sentences and their rhetorical discourse as a way to estimate the polarity of the document. However, the quality of the sentence-level polarity estimations was only indirectly analysed because the TREC Blog track lacks judgements associated to sentences or passages. The NTCIR07 English MOAT research collection, the Multi-Perspective Question Answering dataset (MPQA) and the Finegrained Sentiment Dataset (FSD)[6] supply polarity judgements at sentence level and we can use them to assess

[6]Note that the PL research collection –utilised for subjectivity classification experiments– do not provide polarity judgements.

the impact of positional, rhetorical and linguistic features on the classification of polar sentences. This section is dedicated to report these experiments.

4.3.1 Sentence Features

We use the same experimental procedure proposed for subjectivity classification of sentences (see Section 3.3). The only difference is that some features have been now divided into two independent features depending on polarity information (e.g., proportion of opinionated terms is now replaced by proportion of positive terms and proportion of negative terms) (see Table 4.8).

Set	Feature
Vocabulary	Unigrams and bigrams (binary)
Length	Length of the sentence Length of the nucleus Length of the satellite Length of the document that contains the sentence
Positional	Absolute position of the sentence in the document Relative position of the sentence in the document
POS	Number of occurrences of every POS tags (one feature for each POS tag, see Table 3.8)
Syntactic Patterns	The presence of a POS syntactic pattern (one binary feature for each pattern defined in Table 3.9)
Sentiment Lexicon	Number and proportion of positive terms in the sentence Number and proportion of negative terms in the sentence Number and proportion of exclamations and interrogations in the sentence
RST	Contains a satellite (binary) Contains specific satellite types (binary)
Sentiment RST	Number and proportion of positive terms in the nucleus Number and proportion of negative terms in the nucleus Number and proportion of positive terms in satellites Number and proportion of negative terms in satellites Number and proportion of exclamations and interrogations in the nucleus Number and proportion of exclamations and interrogations in satellites

Table 4.8: Sentence features for polarity classification. The features related to satellites are defined for each specific type of rhetorical relation mentioned in Table 3.10.

dataset	# positive sentences	# negative sentences	#unique unigrams	#unique bigrams
MOAT	179	417	2218	2812
MPQA	1626	3255	6463	9203
FSD	923	1320	1275	1996

Table 4.9: Test collections for experimentation in polarity classification. The tables include the number of unique unigrams and bigrams after pre-processing. We did not apply stemming and we did not remove common words. We only removed terms that appeared in less than four sentences.

4.3.2 Experiments

We are interested in the classification of positive vs. negative sentences. Therefore, we experiment with the positive and negative sentences presented from MOAT[7], MPQA and FSD (see the complete statistics of these collections in Table 4.9).

In Table 4.10, Table 4.11, Table 4.12 (and Figure 4.8, Figure 4.9 and Figure 4.10) we report the polarity classification performance on MOAT, MPQA and FSD, respectively. Again, vocabulary-based classifiers were regarded as baselines and we tested the incorporation of various combinations of features into the baseline classifiers.

A general trend that can be observed is that our best classifiers tend to have a bias towards negative classifications, which typically show a high recall and a somewhat lower precision. Positive sentences are typically identified with a higher precision than recall. This bias can be attributed to the polarity classes being unequally distributed in the data, which holds especially true for the MOAT collection.

One trend emerging from the experiments is the limited extent to which our considered length, positional, POS, POS linguistic patterns and RST features contribute to the overall sentence-level polarity classification performance. Some of these features were useful for detecting opinionated passages (see Section 3.3), but do not have much discriminative power in terms of the polarity of such opinionated passages. For instance, position and POS features were indicative of subjectivity but do not help here to estimate polarity. This makes sense because the position of a sentence could arguably be indicative of subjectivity (e.g., a news article might begin with factual content) but it is hardly a polarity cue. Similarly, some POS

[7] We have three assessors in the MOAT dataset and we used a voting system (majority rule) to obtain the polarity label of every sentence.

	Positive			Negative			microavg	micro
Features	Prec.	Rec.	F_1	Prec.	Rec.	F_1	F_1	sign test
Unigrams	.4389	.2200	.2931	.7289	.8818	.7981	.6859	
+ Length	.3933	.2381	.2966	.7253	.8456	.7808	.6658	$\sim$
+ Position	.3987	.2721	.3235	.7300	.8275	.7757	.6631	$\sim$
+ POS	.4703	.2517	.3279	.7368	.8808	.8024	.6946	$\sim$
+ Synt. Patterns	.5053	.2154	.3020	.7343	.9113	.8133	.7054	$\sim$
+ Sentiment Lexicon	.6505	.3039	.4143	.7609	.9314	.8376	.7456	$\gg$
+ RST	.4724	.2721	.3453	.7403	.8723	.8009	.6946	$\sim$
+ Sentiment RST	**.6809**	.2902	.4070	.7596	**.9428**	**.8413**	**.7497**	$\gg$
+ All	.5603	**.5374**	**.5486**	**.8088**	.8227	.8157	.7383	$\gg$
Uni and bigrams	.6154	.2177	.3216	.7414	.9428	.8301	.7282	
+ Length	.4829	.3515	.4069	.7553	.8418	.7962	.6966	$\ll$
+ Position	.3922	.3424	.3656	.7376	.7769	.7567	.6483	$\ll$
+ POS	.4976	.2381	.3221	.7373	.8990	.8102	.7034	$\ll$
+ Synt. Patterns	.4625	.2517	.3260	.7360	.8770	.8003	.6919	$\ll$
+ Sentiment Lexicon	.6634	.3084	.4211	.7626	.9342	.8397	.7490	$>$
+ RST	.5607	.2200	.3160	.7388	.9276	.8225	.7181	$\sim$
+ Sentiment RST	**.7439**	.2766	.4033	.7594	**.9600**	**.8480**	**.7577**	$\gg$
+ All	.5226	**.5760**	**.5480**	**.8137**	.7788	.7959	.7188	$\sim$

Table 4.10: Polarity classification results in the MOAT collection, in terms of precision, recall, and F_1 scores for positive and negative sentences. For each vocabulary representation (i.e., unigrams, or unigrams and bigrams), the best performance for each metric is bolded. The symbols $\gg$ and $>$ (resp. $\ll$ and $<$) indicate a significant improvement (resp. decrease) with respect to the vocabulary-based baselines, with $p < .01$ and $p < .05$, respectively. $\sim$ indicates that the difference was not statistically significant ($p > .05$).

features, e.g., the number of adjectives, are often indicative of subjectivity but do not reveal by themselves the orientation of the sentiments.

One of the best performing combinations was again the one that includes the sentiment lexicon features. In all cases, this configuration led to significant improvements with respect to the baselines. However, Sentiment RST was the only feature set whose inclusion into the baseline led to improvements with p-value always lower than .01. Sentiment RST features help to differentiate between discourse units, based on their rhetorical roles, when analysing the polarity of these segments. This yielded to polarity classifiers that were slightly more robust than those constructed from structure-unaware features (i.e., Sentiment Lexicon). These sentence-level polarity classification results validate the observed potential of RST-guided Sentiment Analysis in the large-scale polarity ranking of blog posts (see Section 4.2). Finally,

	Positive			Negative			microavg	micro
Features	Prec.	Rec.	F_1	Prec.	Rec.	F_1	F_1	sign test
Unigrams	.6570	.5824	.6175	.8021	.8477	.8243	.7592	
+ Length	.6585	.5893	.6220	.8046	.8470	.8253	.7610	$\sim$
+ Position	.6405	.5996	.6194	.8057	.8315	.8184	.7541	$\sim$
+ POS	.6563	.5854	.6188	.8030	.8465	.8242	.7593	$\sim$
+ Synt. Patterns	.6636	.5824	.6204	.8029	.8521	.8268	.7621	$\sim$
+ Sentiment Lexicon	**.6973**	.6554	.6757	.8325	.8575	.8448	.7901	$\gg$
+ RST	.6554	.5822	.6166	.8018	.8467	.8236	.7584	$\sim$
+ Sentiment RST	.6881	.6541	.6707	.8310	.8515	.8411	.7857	$\gg$
+ All	.6949	**.6662**	**.6802**	**.8362**	**.8535**	**.8448**	**.7910**	$\gg$
Uni and bigrams	.6770	.5463	.6047	.7928	.8695	.8294	.7616	
+ Length	.6771	.5542	.6095	.7953	.8676	.8299	.7630	$\sim$
+ Position	.6598	.5704	.6119	.7985	.8527	.8247	.7585	$\sim$
+ POS	.6665	.5389	.5959	.7893	.8649	.8254	.7561	$\ll$
+ Synt. Patterns	.6743	.5446	.6025	.7920	.8682	.8284	.7602	$\sim$
+ Sentiment Lexicon	.7171	.6362	.6742	.8276	**.8743**	.8503	.7948	$\gg$
+ RST	.6707	.5473	.6027	.7924	.8654	.8273	.7593	$\sim$
+ Sentiment RST	.7114	.6310	.6688	.8251	.8718	.8478	.7915	$\gg$
+All	**.7131**	**.6551**	**.6829**	**.8340**	.8680	**.8507**	**.7970**	$\gg$

Table 4.11: Polarity classification results in the MPQA collection, in terms of precision, recall, and F_1 scores for positive and negative sentences. For each vocabulary representation (i.e., unigrams, or unigrams and bigrams), the best performance for each metric is bolded. The symbols $\gg$ and $>$ (resp. $\ll$ and $<$) indicate a significant improvement (resp. decrease) with respect to the vocabulary-based baselines, with $p < .01$ and $p < .05$, respectively. $\sim$ indicates that the difference was not statistically significant ($p > .05$).

the combination of all features worked well, but was inferior to both Sentiment Lexicon and Sentiment RST.

4.3.3 Feature Weights

In this section we perform a feature weight analysis similar to the one presented in Section 3.3.3. To meet this aim, we selected the *unigrams & bigrams + All* classifier (MOAT) with the scores scaled between 0 and 1. In Table 4.13 we report the most discriminative features for sentence polarity classification. To avoid any reference to a company, brand or trademark, some features are listed in the table with the label ***removed***. A positive w_i indicates that high values of the feature go in favour of a positive classification. On the other hand, a negative

	Positive			Negative			microavg	micro
Features	Prec.	Rec.	F_1	Prec.	Rec.	F_1	F_1	sign test
Unigrams	.6596	.6175	.6379	.7302	.7647	.7471	.7021	
+ Length	.6451	.5195	.5755	.6897	**.7889**	.7360	.6745	$\ll$
+ Position	.6720	.6217	.6459	.7352	.7758	.7550	.7104	$>$
+ POS	.6740	.6221	.6470	.7359	.7777	.7562	.7116	$>$
+ Synt. Patterns	.6747	.6389	.6563	.7434	.7724	.7576	.7157	$\gg$
+ Sentiment Lexicon	**.6936**	.6717	**.6825**	.7630	.7808	**.7718**	**.7345**	$\gg$
+ RST	.6690	.6074	.6367	.7285	.7780	.7524	.7055	$\sim$
+ Sentiment RST	.6911	.6742	**.6825**	**.7636**	.7774	.7704	.7336	$\gg$
+ All	.6391	**.7011**	.6687	.7622	.7075	.7338	.7048	$\sim$
Uni and bigrams	.6801	.5872	.6302	.7231	.7960	.7578	.7073	
+ Length	.6618	.4590	.5421	.6742	.8268	.7427	.6705	$\ll$
+ Position	.6958	.5855	.6359	.7260	.8109	.7661	.7152	$>$
+ POS	.6883	.5792	.6291	.7218	.8063	.7617	.7098	$\sim$
+ Synt. Patterns	.6949	.5792	.6318	.7233	.8122	.7652	.7132	$\sim$
+ Sentiment Lexicon	.7149	.6578	**.6852**	.7614	.8063	**.7832**	**.7432**	$\gg$
+ RST	.6878	.5734	.6254	.7194	.8078	.7610	.7082	$\sim$
+ Sentiment RST	**.7153**	.6494	.6808	.7576	**.8091**	.7825	.7412	$\gg$
+ All	.6686	**.6902**	.6792	**.7656**	.7473	.7563	.7230	$>$

Table 4.12: Polarity classification results in the FSD collection, in terms of precision, recall, and F_1 scores for positive and negative sentences. For each vocabulary representation (i.e., unigrams, or unigrams and bigrams), the best performance for each metric is bolded. The symbols $\gg$ and $>$ (resp. $\ll$ and $<$) indicate a significant improvement (resp. decrease) with respect to the vocabulary-based baselines, with $p < .01$ and $p < .05$, respectively. $\sim$ indicates that the difference was not statistically significant ($p > .05$).

w_i makes that high values of the feature go in favour of a negative classification. Many highly discriminative features are those computed from opinion lexicon (e.g., *#pos nuc. norm*, *#pos*, *#neg,*, *#neg nuc.*). The highest weight is assigned to *#pos nuc. norm*, which represents the proportion of positive terms labelled in the nucleus of the sentence. The proportion of negative terms in the nucleus, and the proportion of positive terms in the satellite also appear in the top 20. The association of some unigrams (e.g., favourable, important, businesses, investors, forward) with the positive class seems natural, whereas the association of other unigrams (e.g., Mr, Instead, time, holds) with the positive class seems incidental. Similarly, some unigrams associated with the negative class (e.g., only, downturn) make sense but other unigrams seem to be there by chance. Regarding RST, having a comparative satellite or having a joint satellite seems to be indicative of positivity. However, this needs to be confronted against other sources

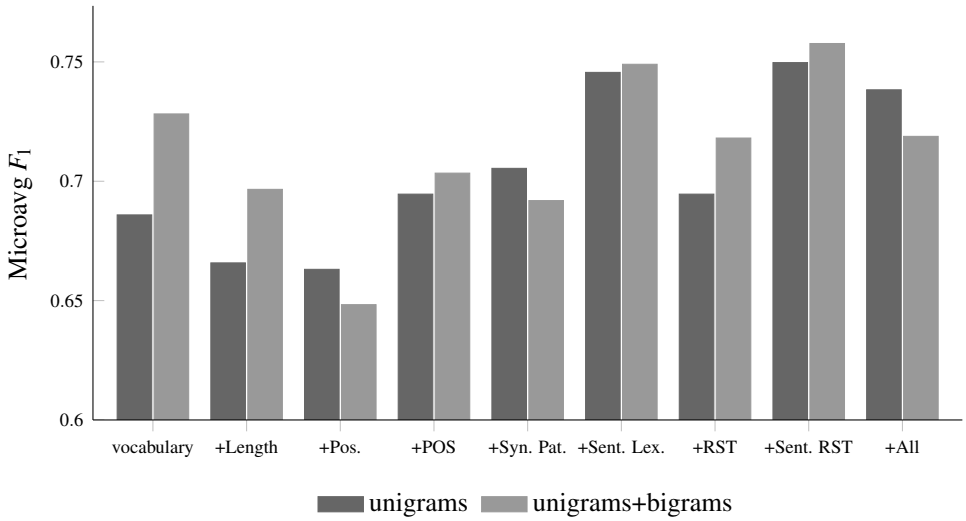

Figure 4.8: Microavg F_1 performance obtained by the different polarity classifiers in the MOAT collection.

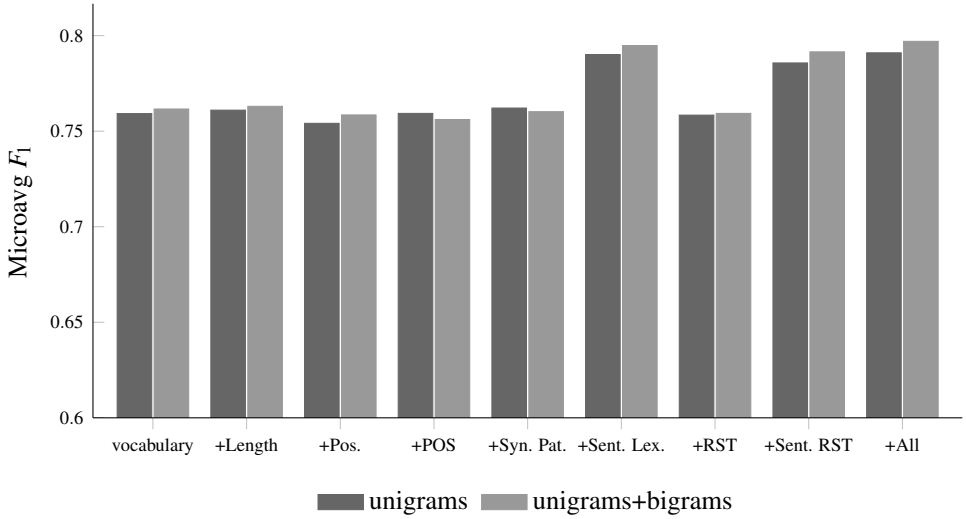

Figure 4.9: Microavg F_1 performance obtained by the different polarity classifiers in the MPQA collection.

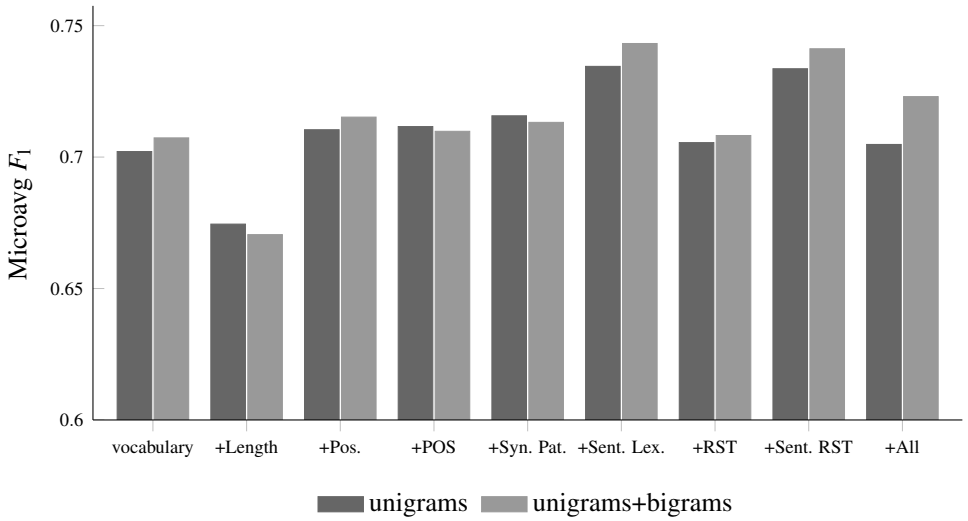

Figure 4.10: Microavg F_1 performance obtained by the different polarity classifiers in the FSD collection.

of data because some RST relations marginally occur in our test collection (see Table 4.14, where we can observe the absolute counts of every relation type in the subjective sentences).

Table 4.15 reports the ranking of weights for non-vocabulary features. Regarding POS features, exclamation marks (feature *#POS(EX)*) tend to be associated with the positive class. This seems to indicate that these marks are used to emphasise positive thoughts. Another intriguing finding is that the use of past tenses is highly correlated with negative sentences ($\#POS(VBN)$, $\#POS(VBD)$ features have a negative score), whereas the use of present tenses is associated with positive sentences ($\#POS(VB)$ has a positive weight). Comparative adjectives ($\#POS(JJR)$) are also indicative of positivity.

4.4 Conclusions

In this chapter we have proposed different methods of search for positive and negative opinions. First of all, we investigated the impact of sentence-level information in a challenging problem: polarity estimation of blog posts. To meet this aim, we have deeply studied different ways to aggregate sentence-level evidence. We have also assessed the impact of sentence

rank	w_i	feature	feature set	rank	w_i	feature	feature set
1	1.5946	#pos nuc. norm.	Sent. RST	26	0.7030	occurred	vocab.
2	1.3012	#pos	Sent. Lex.	27	0.7021	has satellite (Comp.)	RST
3	-1.2846	#neg	Sent. Lex.	28	0.7001	critical	vocab.
4	-1.2302	this	vocab.	29	0.6867	#pos sat. (Attr.)	Sent. RST
5	1.1911	Instead	vocab.	30	0.6788	#POS(EX)	POS
6	1.0088	**removed**	vocab.	31	-0.6674	downturn	vocab.
7	-0.9677	#neg nuc.	Sent. RST	32	-0.6499	India	vocab.
8	0.9390	#pos norm (Attr.)	Sent. RST	33	0.6449	world and	vocab.
9	0.8784	investors	vocab.	34	-0.6436	do	vocab.
10	-0.8762	only	vocab.	35	0.6408	seemed	vocab.
11	-0.8724	#neg nuc. norm.	Sent. RST	36	0.6396	important	vocab.
12	0.8605	forward	vocab.	37	-0.6379	enough	vocab.
13	0.8425	doubt the	vocab.	38	0.6361	There	vocab.
14	-0.8131	like	vocab.	39	0.6359	attacks	vocab.
15	0.8079	businesses	vocab.	40	-0.6321	To	vocab.
16	0.8044	favourable	vocab.	41	-0.6314	global	vocab.
17	0.8023	freedom	vocab.	42	0.6312	has satellite (Joint)	RST
18	-0.7820	long	vocab.	43	-0.6269	#neg norm (Attr.)	Op. RST
19	0.7813	Mr	vocab.	44	0.6220	#pos nuc.	Op. RST
20	-0.7746	about	vocab.	45	0.6184	policy	vocab.
21	-0.7336	even	vocab.	46	0.6136	shows	vocab.
22	-0.7296	to a	vocab.	47	0.6091	must	vocab.
23	0.7251	economic	vocab.	48	0.6088	returned	vocab.
24	0.7166	time	vocab.	49	-0.5833	#neg (Attr.)	Op. RST
25	0.7072	hotels	vocab.	50	0.5805	problems	vocab.

Table 4.13: List of the 50 features with the highest $|w_i|$ in the best(scaled) classifier. The features are ranked by decreasing $|w_i|$.

locations on polarity estimation and have evaluated performance in terms of effectiveness and efficiency.

We have found that location-aware polarity estimation yields 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 valuable for creating polarity-biased snippets.

We have also demonstrated that we can improve efficiency with no impact on effectiveness. Most of the methods proposed are based on simple combinations of polarity and topi-

relation type	total	pos	neg
Condition	14	7	7
Attribution	169	59	110
Manner	1	0	1
Background	17	9	8
Temporal	2	2	0
Comparison	6	4	2
Evaluation	2	0	2
Elaboration	95	24	71
Enablement	20	5	15
Explanation	11	1	10
Contrast	14	3	11
Cause	6	1	5
Joint	10	9	1

Table 4.14: Number of RST relationships found in the subjective sentences of MOAT.

cality. We are aware that there might be better and more formal ways to approach this combination of evidence (e.g. polarity and relevance might be combined using formal methods to learn query-independent weights [CRZT05]). 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.

In this chapter we have also studied the usefulness of a RST-based polarity analysis in the blogosphere. We have found that the use of discourse structure significantly improves polarity detection in blogs. We applied an effective and efficient strategy to select key opinion sentences within a blog post. By analysing these extracts, we found some interesting trends related to the way in which people express their opinions in blogs. There is a clear predominance of attribution and elaboration rhetorical relations; and bloggers tend to express their sentiment in a more apparent fashion in elaborating and attributing text segments rather than in the core of the text itself.

We are interested in applying more refined representations of the rhetorical relations (e.g., LMs [LLL12]). This will be subject to further research in the near future. Another problem we are aware of is that we have used only one sentence to estimate the polarity of the blog post. Under these conditions the benefits of applying rhetorical relations have some limitations

rank	w_i	feature	feature set
1	1.5946	#pos terms nuc. norm.	Sent. RST
2	1.3012	#pos terms	Sent. Lex
3	-1.2846	#neg terms	Sent. Lex
4	-0.9677	#neg terms nuc.	Sent. RST
5	0.9390	#pos terms sat. norm (Attribution)	Sent. RST
6	-0.8724	#neg terms nuc. norm.	Sent. RST
7	0.7021	has satellite (Comparison)	RST
8	0.6867	#pos terms sat. (Attribution)	Sent. RST
9	0.6788	#POS(EX)	POS
10	0.6312	has satellite (Joint)	RST
11	-0.6269	#neg terms sat. norm (Attribution)	Sent. RST
12	0.6220	#pos terms nuc.	Sent. RST
13	-0.5833	#neg terms sat. (Attribution)	Sent. RST
14	-0.5771	#sents document	Length
15	0.5137	has satellite	RST
16	0.5088	#POS(RBS)	POS
17	-0.4914	#POS(VBD)	POS
18	0.4712	#POS(VB)	POS
19	-0.4037	#POS(VBN)	POS
20	0.3860	position norm	Position
21	0.3501	#POS(RP)	POS
22	0.3251	#POS(JJR)	POS
23	0.3239	has satellite (Background)	RST
24	-0.3200	has satellite (Elaboration)	RST
25	0.3005	#pos terms sat. (Elaboration)	Sent. RST

Table 4.15: List of the top 25 non-vocabulary features with the highest $|w_i|$ in the best(scaled) classifier. The features are ranked by decreasing $|w_i|$.

(e.g., the selected sentence may not be a good representative for the blog post). In the near future, we plan to explore the benefits of discourse structure while taking more sentences into account in our analysis. Related to this, one of the core problems derived to the use of RST is the processing time required for identifying discourse structure in natural language text. Therefore, we would like to explore more efficient methods of identifying the discourse structure of texts.

In the last part of this chapter we proposed several classification methods to estimate whether a subjective sentence is positive or negative. We have studied classical features such

as n-grams or lexicon-based counts in combination with more advanced features such as location and rhetorical features. Among all features tested, length and binary RST-based features did not give any added value for polarity classification. These features hardly improved over the baselines and, often, led to performance decreases. We conclude that the presence of particular rhetorical relations or the length of the sentence were not indicative of polarity. Additionally, positional and POS features did not help much, while sentiment lexicon together with unigrams/bigrams were quite accurate. However, Sentiment RST features were slightly more robust than Sentiment Lexicon features and, therefore, unigrams/bigrams+Sentiment RST seems like a sensible choice for polarity classification at sentence level. This is interesting because sentiment RST features were slightly inferior to pure lexicon-based features for subjectivity estimation. For polarity estimation purposes, sentiment RST features can capture the specific relations between the different parts of the sentences and weight the polar terms accordingly. For instance, the contrast statement presented in the sentence *"The film was awful but it was nice going with her"* cannot be merely solved with lexicon-based approaches.

Chapter 5

Conclusions

The main motivation of this thesis was to explore different types of evidence to effectively determine on-topic opinions in texts. Our research has been oriented towards general and efficient techniques able to work in a wide range of conditions. Searching for documents conveying opinions related to query topics and studying the subjectivity and polarity of these documents have been two fundamental problems that we approached.

First, we demonstrated that some document features, related to the pattern of matching between query and the document's sentences, are valuable for finding on-topic blog posts. We worked with an effective BM25 parameter configuration (which outperforms the default BM25 configuration) and we re-ranked an initial retrieval baseline in a way that incorporates new document features based on sentence scores. We tested the effectiveness of our approach by combining four sentence-level features and evaluating them against two different datasets. Two of these features (ratio of peaks and median of unique terms) offer good performance and led to combined methods that outperform state-of-the-art models for blog topic retrieval.

Regarding opinion finding, we found that comments are very useful to move the query towards opinionated words. This finding led us to define a high-precision query expansion method for opinion retrieval. We observed that words extracted from comments potentially lead to a more accurate query-dependent opinion vocabulary. Our opinion finding experiments also showed that passage-level and location-aware evidence is useful to understand whether or not a document is subjective. One interesting finding was that the most subjective on-topic sentence provides a good opinion proxy for the document. This could lead to better opinion-biased summaries. Finally, we proposed several classification methods to estimate whether or

not a sentence is opinionated. We studied classical features such as n-grams or lexicon-based counts in combination with more advanced features such as location and rhetorical features. Among all features tested, length and binary RST-based features did not give any added value. Positional features worked well as a recall-oriented mechanism for detecting subjective sentences and POS features were valuable for subjectivity classification. Nevertheless, a score based on counting sentiment terms from a general-purpose vocabulary, i.e., Sentiment Lexicon, was at least as effective as accounting for POS labels. Other syntactic patterns were tested but they were only beneficial when the baseline classifier handled unigrams representations. With bigrams, syntactic patterns did not give any added value. Lexicon-based features consistently gave high performance in sentence subjectivity classification, being the best approach in combination with n-grams. Sentiment RST features were slightly inferior to pure lexicon-based features. Overall, classifying sentences based on sentiment lexicon scores and unigrams/bigrams is an effective and safe choice for subjectivity classification.

Chapter 4 has been centered on methods of search for positive and negative opinions. First, we deeply studied different ways to aggregate sentence-level evidence. We paid special attention to the impact of sentence location on polarity estimation and took into account both efficiency and effectiveness considerations. We have found that location-aware polarity methods yield state-of-the-art performance, which is robust across different topic-relevance baselines. 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 valuable for creating polarity-biased snippets and these snippets can be constructed by processing a small subset of the document's sentences. We have also studied the role of RST for polarity analysis in the blogosphere. We found that discourse structure significantly improves polarity detection in blogs. Based on this analysis, we found that there is a clear predominance of attribution and elaboration rhetorical relations. Bloggers tend to express their sentiment in a more apparent fashion in elaborating and attributing text segments rather than in the core of the text itself. In the last part of Chapter 4 we proposed several classification methods to estimate whether a subjective sentence is positive or negative. We considered classical features such as n-grams or lexicon-based counts in combination with more advanced features such as location and rhetorical features. Among all features tested, length and binary RST-based features did not give any added value for polarity classification. Similarly, positional and POS features were not very helpful. Two effective sets of features were Sentiment Lexicon features and Sentiment RST features. Sentiment

RST features were slightly more robust than Sentiment Lexicon features and, therefore, unigrams/bigrams+Sentiment RST seems like a sensible choice for polarity classification at sentence level. This is interesting because sentiment RST features were slightly inferior to pure lexicon-based features for subjectivity estimation. But Sentiment RST features are promising in polarity estimation because they can capture the specific relations between the different parts of the sentences and weight the polar terms accordingly.

Summing up, in this thesis we have identified a specific set of factors that are indicative of subjectivity and relevance and, therefore, could act as a valuable guidance to detect opinionated documents, to extract relevant opinions and to estimate their polarity. We have also proposed innovative methods and models able to combine different types of evidence –obtained at document and passage level– when determining on-topic opinions in texts. Besides effectiveness, efficiency has been a constant concern across the thesis. Some types of evidence, such as discourse structure, had only been tested in the literature against small collections from narrow domains (e.g., movie reviews). We have shown here how to incorporate this type of features into general-purpose opinion retrieval solutions that operate at large scale.

APPENDIX A

PUBLICATIONS

- **[JCR Journal]** Jose M. Chenlo, David E. Losada. *An Empirical Study of Sentence Features for Subjectivity and Polarity Classification*, Information Sciences (Impact factor in 2012: 3.643(Q1)), (*To appear, 2014*). In this paper we extended and combined the body of empirical evidence regarding sentence subjectivity classification and sentence polarity classification, and provided a comprehensive analysis of the relative importance of each set of features using data from multiple benchmarks. To the best of our knowledge, this is the first study that evaluated a highly diversified set of sentence features for the two main sentiment classification tasks.

- **[JCR Journal]** Jose M. Chenlo, Alexander Hogenboom, David E. Losada. *Rhetorical Structure Theory for Polarity Estimation: an Experimental Study*, Data & Knowledge Engineering (Impact factor in 2012: 1.519(Q2)), submitted in October 2013 (*Under review*). In this paper, we investigated the usefulness of Rhetorical Structure Theory in various Sentiment Analysis tasks on different types of information sources. First, we demonstrated how to perform a large-scale ranking of individual blog posts in terms of their overall polarity, by exploiting the rhetorical structure of a few key evaluative sentences. In order to further validate our findings, we additionally explored the potential of Rhetorical Structure Theory in sentence-level polarity classification of news and product reviews. Our most valuable polarity classification features turned out to capture the way in which polar terms are used, rather than the sentiment-carrying words per se.

- **[JCR Journal]** Jose M. Chenlo, Javier Parapar, David E. Losada, José Santos. *Finding a Needle in the Blogosphere: An Information Fusion Approach for Blog Distillation*

Search, Information Fusion (Impact factor in 2012: 2.838(Q1)), submitted in February 2014 (*Under review*). In this paper we proposed a group of textual and social-based signals, and apply different Information Fusion algorithms for a Blog Distillation Search task. Efficiency is an imperative here and, therefore, we focused not only on achieving high search effectiveness but also on designing efficient solutions. We analysed several existing optimisation methods (Line Search, Particle Swarm Optimisation and Differential Evolution), proposed two new hybrid methods, and compared all the alternatives following a standard methodology.

– **[LNCS]** Jose M. Chenlo, David E. Losada. *A Machine Learning approach for Subjectivity Classification based on Positional and Discourse Features*. Proceedings of IRFC2013, 6th Information Retrieval Facility Conference, Limassol, Cyprus, October 2013. Lecture Notes in Computer Science vol. 8201. *(ISBN:978-3-642-41056-7) (full paper, acceptance rate 50% (8/16))*. In this paper, we studied the role of structural features to guide sentence-level subjectivity classification. More specifically, we combined classical n-grams features with novel features defined from positional information and from the discourse structure of the sentences. Our experiments showed that these new features are beneficial in the classification of subjective sentences.

– **[LNAI]** Jose M. Chenlo, Javier Parapar, David E. Losada. *Comments-Oriented Query Expansion for Opinion Retrieval in Blogs* . Proceedings of CAEPIA 2013, 15th Conference of the Spanish Association for Artificial Intelligence, Madrid, Spain, September 2013. Lecture Notes in Artificial Inteligence vol. 8109. *(ISBN:978-3-642-40642-3) (full paper, acceptance rate 41% (27/66))*. We argued here that the comments are a good guidance to find on-topic opinion terms that help to move the query towards burning aspects of the topic. We studied the role of the different parts of a blog document to enhance blog opinion retrieval through query expansion. The proposed method does not require external resources or additional knowledge and our experiments showed that this is a promising and simple way to make a more accurate ranking of blog posts in terms of their sentiment towards the query topic. Our approach compared well with other opinion finding methods, obtaining high precision performance without harming mean average precision.

– **[LNCS]** Jose M. Chenlo, Alexander Hogenboom, David E. Losada. *Sentiment-based Ranking of Blog Posts using Rhetorical Structure Theory*. Proceedings of NLDB 2013,

18th International Conference on Applications of Natural Language to Information Systems, Manchester (UK), 2013. Lecture Notes in Computer Science vol. 7934. *(ISBN:978-3-642-38823-1) (full paper, acceptance rate 26% (21/80), CORE ERA 2008: C)*. We applied sentence-level methods to select the key sentences that convey the overall on-topic sentiment of a blog post. Then, we applied RST analysis to these core sentences in order to guide the classification of their polarity and thus to generate an overall estimation of the document's polarity with respect to a specific topic. Our results showed that RST provides valuable information about the discourse structure of the texts that can be used to make a more accurate ranking of documents in terms of their estimated sentiment in multi-topic blogs.

– Jose M. Chenlo, J. Atserias, C. Rodriguez, R. Blanco. *FBM-Yahoo! at RepLab 2012*. Proceedings of RepLab 2012 Lab, An evaluation campaign for Online Reputation Management Systems (within CLEF 2012), Rome (Italy), 2012. *(ISBN:978-88-904810-3-1)*. This paper describes our participation in the profiling task of RepLab 2012, which aimed at determining whether a given tweet is related to a specific company and, in if this being the case, whether it contains a positive or negative statement related to the company's reputation or not. We addressed both problems (ambiguity and polarity reputation) using Support Vector Machines (SVM) classifiers and lexicon-based techniques, building automatically company profiles and bootstrapping background data. Concretely, for the ambiguity task we employed a linear SVM classifier with a token-based representation of relevant and irrelevant information extracted from the tweets and Freebase resources. With respect to polarity classification, we combined SVM lexicon-based approaches with bootstrapping in order to determine the final polarity label of a tweet.

– **[ACM]** Jose M. Chenlo, David E. Losada. *Effective and Efficient Polarity Estimation in Blogs based on Sentence-Level Evidence*. Proceedings of ACM 20th Conference on Information and Knowledge Management, CIKM 2011, Glasgow (Scotland), 2011. ACM press. *(ISBN:978-1-4503-0717-8) (full paper, acceptance rate 15% (134/918), CORE ERA 2008: A)*. In this work we showed that we can successfully determine the polarity of blog post guided by a sentence-level analysis that takes into account topicality and the location in the post of the subjective sentences. Our experimental results showed that some of our proposed variants are both highly effective and computationally-lightweight.

– Jose M. Chenlo, David E. Losada. *Combining Document and Sentence Scores for Blog Topic Retrieval*. 1st Spanish Conference on Information Retrieval, Proceedings of CERI 2010, Madrid (Spain), Jun 2010. *(ISBN:978-84-693-2200-0) (full paper, acceptance rate 46% (20/43))*. In this paper we proposed some adjustments to effective blog retrieval methods based on the distribution of sentence scores. We hypothesized that we can successfully identify truly relevant documents by combining score features from document and sentences. This helped to detect right contexts related to queries. Our experimental results showed that some of our proposed variants can outperform state-of-the-art blog topic retrieval models.

Appendix B

Siguiendo el reglamento de los estudios de tercer ciclo de la Universidad de Santiago de Compostela, aprobado en la Junta de Gobierno el día 7 de abril de 2000 (DOG de 6 de marzo de 2001) y modificado por la Junta de Gobierno del 14 de noviembre de 2000, el Consejo de Gobierno del 22 de noviembre de 2003, del 18 de julio de 2005 (artículos 30 a 45), del 11 de noviembre de 2008 y del 14 de mayo de 2009; mostramos a continuación un resumen en español de la tesis.

Resumen

En esta tesis nos centramos en sistemas Minería de Opiniones y Análisis de Sentimientos y proponemos un análisis de grado fino de las opiniones vertidas en textos. Concretamente, la motivación principal de esta tesis es comprender cómo combinar diferentes tipos de evidencias para determinar de forma efectiva opiniones relevantes en textos de diferente índole. Para lograr dicho objetivo consideramos diferentes tipos de señales en los textos, desde evidencia de emparejamiento de contenido (obtenida a nivel de documento y de oración) hasta aspectos estructurales de los textos.

La tecnología actual de Minería de Opiniones sufre una serie de carencias que no la hacen apta para resolver las necesidades de información actuales. Un hecho que evidencia dichas carencias es que la gente suele utilizar motores de búsqueda convencionales, los cuales adolecen de capacidades avanzadas de búsqueda de opiniones, para buscar opiniones sobre sus intereses. Esto hace que el esfuerzo de determinar cuales son las opiniones relevantes clave recaiga en el usuario. La falta de aceptación en la actualidad de los sistemas de Minería de Opiniones viene motivada por las limitaciones de los modelos desarrollados, que son simplistas y ofrecen un rendimiento modesto. En esta tesis estudiamos un conjunto concreto de factores indicadores de subjetividad y relevancia y tratamos de entender cual es la mejor manera de combinarlos para detectar documentos con opiniones, extraerlas y estimar su polaridad. También se propondrán nuevos métodos y modelos capaces de incorporar diferentes tipos de señales –obtenidas a nivel de documento y pasaje– para determinar opiniones relevantes en textos. La intención de esta tesis es hacer aportaciones en diferentes áreas, incluyendo aquellas relacionadas con i) búsqueda de documentos con opiniones, ii) detección de subjetividad

a nivel de documento y pasaje, y iii) estimación de polaridad. Otro aspecto importante que guía esta investigación es la eficiencia. Algunos tipos de señales o evidencias, como es la estructura del discurso de los textos, han sido probadas con anterioridad sólo en colecciones pequeñas y en dominios muy reducidos (por ejemplo, críticas de películas). Esto es debido a su elevada complejidad computacional. A lo largo de la presente tesis se demostrará que estas características lingüisticas avanzadas –basadas en análisis de discurso– pueden conducir potencialmente a un mejor entendimiento de la manera de expresar subjetividad en los textos. Adicionalmente, se mostrará que este tipo de evidencia puede ser inyectada de manera eficiente en soluciones de búsqueda de opiniones de propósito general que operan con grandes volúmenes de datos (por ejemplo, la web).

Appendix C

List of Stopwords

a	after	allyou	andor	appeared
abaft	afterer	almost	anear	appearing
abafter	afterest	along	anent	appears
abaftest	afterward	alongside	another	appropriate
about	afterwards	already	any	appropriated
abouter	again	also	anybody	appropriater
aboutest	against	although	anyhow	appropriates
above	aid	always	anyone	appropriatest
abover	ain	amid	anything	appropriating
abovest	albeit	amidst	anywhere	are
accordingly	all	among	apart	ares
aer	aller	amongst	aparter	around
aest	allest	an	apartest	as
afore	alls	and	appear	ases

aside
asides
aslant
astraddle
astraddler
astraddlest
astride
astrider
astridest
at
athwart
atop
atween
aught
aughts
available
availabler
availablest
awfully
b
be
became
because
become
becomes
becoming
becominger
becomingest
becomings
been
before
beforehand
beforehander
beforehandest
behind
behinds
below
beneath
beside
besides
better
bettered
bettering
betters
between
betwixt
beyond
bist
both
but
buts
by
by-and-by
byandby
c
cannot
canst
cant
canted
cantest
canting
cants
cer
certain
certainer
certainest
cest
chez
circa
co
come-on
come-ons
comeon
comeons
concerning
concerninger
concerningest
consequently
considering
could
couldst
cum
d
dday
ddays
describe
described
describes
describing
despite
despited
despites
despiting
did
different
differenter
differentest
do
doe
does
doing
doings
done
doner
dones
donest
dos
dost
doth
downs

downward
downwarder
downwardest
downwards
during
e
each
eg
eight
either
else
elsewhere
enough
ere
et
etc
even
evened
evenest
evens
evenser
evensest
ever
every
everybody
everyone
everything
everywhere
ex
except
excepted
excepting
excepts
exes
f
fact
facts
failing
failings
few
fewer
fewest
figupon
figuponed
figuponing
figupons
five
followthrough
for
forby
forbye
fore
forer
fores
forever
former
formerer
formerest
formerly
formers
fornenst
forwhy
four
fourscore
frae
from
fs
further
furthered
furtherer
furtherest
furthering
furthermore
furthers
g
get
gets
getting
go
gone
good
got
gotta
gotten
h
had
hadst
hae
hardly
has
hast
hath
have
haves
having
he
hence
her
hereafter
hereafters
hereby
herein
hereupon
hers
herself
him
himself
his
hither
hitherer

hitherest
hoo
hoos
how
how-do-you-do
howbeit
howdoyoudo
however
huh
humph
i
idem
idemer
idemest
ie
if
ifs
immediate
immediately
immediater
immediatest
in
inasmuch
inc
indeed
indicate
indicated
indicates
indicating
info
information
insofar
instead
into
inward
inwarder
inwardest
inwards
is
it
its
itself
j
k
l
latter
latterer
latterest
latterly
latters
layabout
layabouts
less
lest
lot
lots
lotted
lotting
m
main
make
many
mauger
maugre
mayest
me
meanwhile
meanwhiles
midst
midsts
might
mights
more
moreover
most
mostly
much
mucher
muchest
must
musth
musths
musts
my
myself
n
natheless
nathless
neath
neaths
necessarier
necessariest
necessary
neither
nethe
nethermost
never
nevertheless
nigh
nigher
nighest
nine
no
no-one
nobodies
nobody
noes
none
noone
nor

nos
not
nothing
nothings
notwith-
standing
nowhere
nowheres
o
of
off
offest
offs
often
oftener
oftenest
oh
on
one
oneself
onest
ons
onto
or
orer
orest
other
others
otherwise
otherwiser
otherwisest
ought
oughts
our
ours
ourself
ourselves
out
outed
outest
outs
outside
outwith
over
overall
overaller
overallest
overalls
overs
own
owned
owning
owns
owt
p
particular
particularer
particularest
particularly
particulars
per
perhaps
plaintiff
please
pleased
pleases
plenties
plenty
pro
probably
provide
provided
provides
providing
q
qua
que
quite
r
rath
rathe
rather
rathest
re
really
regarding
relate
related
relatively
res
respecting
respectively
s
said
saider
saidest
same
samer
sames
samest
sans
sanserif
sanserifs
sanses
saved
sayid
sayyid
seem
seemed
seeminger
seemingest

seemings
seems
send
sent
senza
serious
seriouser
seriousest
seven
several
severaler
severalest
shall
shalled
shalling
shalls
she
should
shoulded
shoulding
shoulds
since
sine
sines
sith
six
so
sobeit
soer
soest
some
somebody
somehow
someone
something
sometime
sometimer
sometimes
sometimest
somewhat
somewhere
stop
stopped
such
summat
sup
supped
supping
sups
syn
syne
t
ten
than
that
the
thee
their
theirs
them
themselves
then
thence
thener
thenest
there
thereafter
thereby
therefore
therein
therer
therest
thereupon
these
they
thine
thing
things
this
thises
thorough
thorougher
thoroughest
thoroughly
those
thou
though
thous
thouses
three
thro
through
througher
throughest
throughout
thru
thruer
thruest
thus
thy
thyself
till
tilled
tilling
tills
to
together
too
toward
towarder

towardest
towards
two
u
umpteen
under
underneath
unless
unlike
unliker
unlikest
until
unto
up
upon
uponed
uponing
upons
upped
upping
ups
us
use
used
usedest
username
usually
v
various
variouser
variousest
verier
veriest
versus
very
via
vis-a-vis
vis-a-viser
vis-a-visest
viz
vs
w
was
wast
we
were
wert
what
whatever
whateverer
whateverest
whatsoever
whatsoeverer
whatsoeverest
wheen
when
whenas
whence
whencesoever
whenever
whensoever
where
whereafter
whereas
whereby
wherefrom
wherein
whereinto
whereof
whereon
wheresoever
whereto
whereupon
wherever
wherewith
wherewithal
whether
which
whichever
whichsoever
while
whiles
whilst
whither
whithersoever
whoever
whomever
whose
whoso
whosoever
why
with
withal
within
without
would
woulded
woulding
woulds
x
y
ye
yet
yon
yond
yonder
you
your
yours
yourself
yourselves
z
zillion

References

[AAB+08] Giambattista Amati, Edgardo Ambrosi, Marco Bianchi, Carlo Gaibisso, and Giorgio Gambosi. Automatic construction of an opinion-term vocabulary for ad hoc retrieval. In *Proceedings of the IR research, 30th European conference on Advances in information retrieval*, ECIR'08, pages 89–100, Berlin, Heidelberg, 2008. Springer-Verlag.

[AjAC+04] Nasreen Abdul-jaleel, James Allan, W. Bruce Croft, O Diaz, Leah Larkey, Xiaoyan Li, Mark D Smucker, and Courtney Wade. UMass at TREC 2004: Novelty and HARD. In *Proceedings of TREC-13*, NIST Special Publication. National Institute for Science and Technology, 2004.

[AMWZ09] Timothy G. Armstrong, Alistair Moffat, William Webber, and Justin Zobel. Has adhoc retrieval improved since 1994? In *Proc. 32nd Annual International ACM SIGIR Conference*, pages 692–693, 2009.

[AWB03] James Allan, Courtney Wade, and Alvaro Bolivar. Retrieval and novelty detection at the sentence level. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, SIGIR '03, pages 314–321, New York, NY, USA, 2003. ACM.

[AXV+11] Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow, and Rebecca Passonneau. Sentiment analysis of Twitter data. In *Proceedings of the Workshop on Languages in Social Media*, LSM '11, pages 30–38, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[BHMV04] Philip Beineke, Trevor Hastie, Christopher Manning, and Shivakumar Vaithyanathan. Exploring sentiment summarization. In *Proc. AAAI Spring Symposium on Exploring Attitude and Affect in Text Theories and Applications*, pages 12–15, 2004.

[BYRN08] Ricardo Baeza-Yates and Berthier Ribeiro-Neto. *Modern Information Retrieval*. Addison-Wesley Publishing Company, USA, 2nd edition, 2008.

[CD11] Carlos Castillo and Brian D. Davison. Adversarial web search. *Found. Trends Inf. Retr.*, 4(5):377–486, May 2011.

[CH79] Bruce Croft and David J Harper. Using probabilistic models of document retrieval without relevance information. *Journal of Documentation*, 35:285–295, 1979.

[Che12] Hui Chen. The impact of comments and recommendation system on online shopper buying behaviour. *JNW*, 7(2):345–350, 2012.

[CKP07] Deepayan Chakrabarti, Ravi Kumar, and Kunal Punera. Page-level template detection via isotonic smoothing. In *Proceedings of the 16th international conference on World Wide Web*, WWW '07, pages 61–70, New York, NY, USA, 2007. ACM.

[CL08] Yin-Wen Chang and Chih-Jen Lin. Feature ranking using linear SVM. *Journal of Machine Learning Research - Proceedings Track*, 3:53–64, 2008.

[CMS09] W. Bruce Croft, Donald Metzler, and Trevor Strohman. *Search Engines: Information Retrieval in Practice*. Addison-Wesley Publishing Company, USA, 1st edition, 2009.

[Cro93] W. Bruce Croft. Knowledge-based and statistical approaches to text retrieval. *IEEE Expert*, 8(2):8 –12, april 1993.

[Cro00] W. Bruce Croft. *Combining Approaches to Information Retrieval*, ir 1, pages 1–36. Kluwer Academic Publishers, 2000.

[CRZT05] Nick Craswell, Stephen E. Robertson, Hugo Zaragoza, and Michael J. Taylor. Relevance weighting for query independent evidence. In *SIGIR*, pages 416–423, 2005.

[Eve08] Stefan Evert. A lightweight and efficient tool for cleaning web pages. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, may 2008. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2008/.

[FCH+08] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. LIBLINEAR: A library for large linear classification. *J. Mach. Learn. Res.*, 9:1871–1874, June 2008.

[GCC09] Shima Gerani, Mark J. Carman, and Fabio Crestani. Investigating learning approaches for blog post opinion retrieval. In *Proceedings of the 31th European Conference on IR Research on Advances in Information Retrieval*, ECIR '09, pages 313–324, Berlin, Heidelberg, 2009. Springer-Verlag.

[GCC10] Shima Gerani, Mark James Carman, and Fabio Crestani. Proximity-based opinion retrieval. In *Proc. 33rd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '10, pages 403–410, New York, NY, USA, 2010. ACM.

[HC09] Xuanjing Huang and Bruce Croft. A unified relevance model for opinion retrieval. In *Proceedings of the 18th ACM conference on Information and knowledge management*, CIKM '09, pages 947–956, New York, NY, USA, 2009. ACM.

[HGH+11] Bas Heerschop, Frank Goossen, Alexander Hogenboom, Flavius Frasincar, Uzay Kaymak, and Franciska de Jong. Polarity analysis of texts using discourse structure. In *Proc. 20th ACM International Conference on Information and Knowledge Management. CIKM'11*, CIKM '11, pages 1061–1070. ACM press, 2011.

[HMHO08] Ben He, Craig Macdonald, Jiyin He, and Iadh Ounis. An effective statistical approach to blog post opinion retrieval. In *Proceedings of the 17th ACM conference on Information and knowledge management*, CIKM '08, pages 1063–1072, New York, NY, USA, 2008. ACM.

[HMO08] Ben He, Craig Macdonald, and Iadh Ounis. Ranking opinionated blog posts using OpinionFinder. In *SIGIR*, pages 727–728, 2008.

[HSL07] Meishan Hu, Aixin Sun, and Ee-Peng Lim. Comments-oriented blog summarization by sentence extraction. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, CIKM '07, pages 901–904, New York, NY, USA, 2007. ACM.

[JYZ08] Lifeng Jia, Clement T. Yu, and Wei Zhang. UIC at TREC 2008 blog track. In *TREC*, 2008.

[KFJ06] Pranam Kolari, Tim Finin, and Anupam Joshi. SVMs for the Blogosphere: Blog Identification and Splog Detection. In *AAAI Spring Symposium on Computational Approaches to Analysing Weblogs*. Computer Science and Electrical Engineering, University of Maryland, Baltimore County, March 2006. Also available as technical report TR-CS-05-13.

[KJF+06] Pranam Kolari, Akshay Java, Tim Finin, Tim Oates, and Anupam Joshi. Detecting spam blogs: A machine learning approach. In *2006. Proceedings of the 21st National Conference on Artificial Intelligence (AAAI)*, 2006.

[KZ01] Marcin Kaszkiel and Justin Zobel. Effective ranking with arbitrary passages. *Journal of the American Society for Information Science and Technology*, 52:344–364, 2001.

[LAD96] X Allan Lu, Maen Ayoub, and Jianhua Dong. Ad hoc experiments using EUREKA. In *Proceedings of TREC-5*, NIST Special Publication, pages 229–240. National Institute for Science and Technology, 1996.

[LC01] Victor Lavrenko and W. Bruce Croft. Relevance based Language Models. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR'01, pages 120–127, New York, NY, USA, 2001. ACM.

[LF73] F. Wilfrid Lancaster and Emily Gallup Fayen. *Information retrieval: on-line [by] F. W. Lancaster and E. G. Fayen*. Melville Pub. Co. Los Angeles,, 1973.

[LhNK+08] Yeha Lee, Seung hoon Na, Jungi Kim, Sang hyob Nam, Hun young Jung, and Jong hyeok Lee. KLE at TREC 2008 blog track: Blog post and feed retrieval. In *In Proceedings of TREC-08*, 2008.

[Liu11] Tie-Yan Liu. *Learning to Rank for Information Retrieval*. Springer, 2011.

[Liu12] Bing Liu. *Sentiment Analysis and Opinion Mining*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers, 2012.

[LLL12] Christina Lioma, Birger Larsen, and Wei Lu. Rhetorical relations for information retrieval. In *Proc. 35th international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '12, pages 931–940, New York, NY, USA, 2012. ACM.

[Los10] David E. Losada. Statistical query expansion for sentence retrieval and its effects on weak and strong queries. *Inf. Retr.*, 13(5):485–506, October 2010.

[LSC+07] Yu-Ru Lin, Hari Sundaram, Yun Chi, Junichi Tatemura, and Belle L. Tseng. Splog detection using self-similarity analysis on blog temporal dynamics. In *Proceedings of the 3rd international workshop on Adversarial information retrieval on the web*, AIRWeb '07, pages 1–8, New York, NY, USA, 2007. ACM.

[MFM02] Janez Brank Marko Grobelnik Natasa Milic-Frayling and Dunja Mladenic. Feature selection using support vector machines. In *In Proc. of the 3rd Int. Conf. on Data Mining Methods and Databases for Engineering, Finance, and Other Fields*, pages 84–89, 2002.

[MHOS08] Craig Macdonald, Ben He, Iadh Ounis, and Ian Soboroff. Limits of opinion-finding baseline systems. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '08, pages 747–748, New York, NY, USA, 2008. ACM.

[Mis07] Gilad Mishne. Using blog properties to improve retrieval. In *International Conference on Weblogs and Social Media 2007, Retrieved February 29, 2008 from: http://www.icwsm.org/papers/3–Mishne.pdf*, 2007.

[ML06] Yi Mao and Guy Lebanon. Sequential models for sentiment prediction. In *ICML Workshop on Learning in Structured Output Spaces*, 2006.

[MO06] Craig Macdonald and Iadh 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.

[MOS07] Craig Macdonald, Iadh Ounis, and Ian Soboroff. Overview of the TREC 2007 blog track. In *Proc. TREC 2007, the 16th Text Retrieval Conference*, Gaithersburg, United States, 2007.

[MOS09] Craig Macdonald, Iadh Ounis, and Ian Soboroff. Is spam an issue for opinionated blog post search? In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '09, pages 710–711, New York, NY, USA, 2009. ACM.

[MRS08] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schtze. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA, 2008.

[MSM93] Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. Building a large annotated corpus of english: The penn treebank. *COMPUTATIONAL LINGUISTICS*, 19(2):313–330, 1993.

[MT88] William C. Mann and Sandra A. Thompson. Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8(3):243–281, 1988.

[Nal04] Ramesh Nallapati. Discriminative models for information retrieval. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '04, pages 64–71, New York, NY, USA, 2004. ACM.

[NNLL09] Sang-Hyob Nam, Seung-Hoon Na, Yeha Lee, and Jong-Hyeok Lee. Diffpost: Filtering non-relevant content based on content difference between two consecutive blog posts. In Mohand Boughanem, Catherine Berrut, Josiane Mothe, and Chantal Soule-Dupuy, editors, *Advances in Information Retrieval*, volume 5478 of *Lecture Notes in Computer Science*, pages 791–795. Springer Berlin Heidelberg, 2009.

[OMdR+06] Iadh Ounis, Craig Macdonald, Maarten de Rijke, Gilad Mishne, and Ian Soboroff. Overview of the TREC 2006 blog track. In *Proc. TREC 2006, the 15th Text Retrieval Conference*. NIST, 2006.

[OMLS11] Iadh Ounis, Craig Macdonald, Jimmy Lin, and Ian Soboroff. Overview of the TREC 2011 microblog track. In *Proc. TREC 2011, the 20th Text Retrieval Conference*. NIST, 2011.

[OMS08] Iadh Ounis, Craig Macdonald, and Ian Soboroff. Overview of the TREC 2008 blog track. In *Proc. TREC 2008, the 17th Text Retrieval Conference*, Gaithersburg, United States, 2008. NIST.

[PB07] Javier Parapar and Álvaro 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.

[PC98] Jay M. Ponte and Bruce Croft. A language modeling approach to information retrieval. pages 275–281, 1998.

[PL04] Bo Pang and Lillian Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *42nd Annual Meeting of the Association for Computational Linguistics (ACL 2004)*, pages 271–280. Association for Computational Linguistics, 2004.

[PL05] Bo Pang and Lillian Lee. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the ACL*, 2005.

[PL07] Bo Pang and Lillian Lee. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135, 2007.

[PLCB10a] Javier Parapar, Jorge López-Castro, and Álvaro Barreiro. Blog posts and comments extraction and impact on retrieval effectiveness. In *1st Spanish Conference on Information Retrieval, CERI'12*, pages 5–16, Madrid, 2010.

[PLCB10b] Javier Parapar, Jorge López-Castro, and Álvaro Barreiro. Blog snippets: a comments-biased approach. In *Proceedings of the 33rd international ACM*

SIGIR conference on Research and development in information retrieval, SIGIR '10, pages 711–712, New York, NY, USA, 2010. ACM.

[PLV02] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up? sentiment classification using machine learning techniques. In *Pr. of the Conference on Empirical Methods in Natural Language Processing*, 2002.

[Por80] Martin F. Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, 1980.

[PVS12] Javiar Parapar, Maria M. Vidal, and Jose Santos. Finding the best parameter setting: Particle swarm optimisation. In *2nd Spanish Conference on Information Retrieval, CERI'12*, pages 49–60, 2012.

[Rij79] C. J. Van Rijsbergen. *Information Retrieval*. Butterworth-Heinemann, Newton, MA, USA, 2nd edition, 1979.

[RK13] Fiana Raiber and Oren Kurland. Using document-quality measures to predict web-search effectiveness. In *ECIR*, pages 134–145, 2013.

[RL03] Ian Ruthven and Mounia Lalmas. A survey on the use of relevance feedback for information access systems. *Knowl. Eng. Rev.*, 18(2):95–145, 2003.

[Rob05] Stephen E. 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.

[Roc71] J Rocchio. Relevance feedback in information retrieval. In G Salton, editor, *The SMART Retrieval System: Experiments in Automatic Document Processing*, pages 313–323. Prentice Hall, Inc., 1971.

[RSS+12] Lee Rainie, Aaron Smith, Kay Lehman Schlozman, Henry Brady, and Sidney Verba. Social media and political engagement. Pew Internet & American Life Project Report, October 2012.

[RW03] Ellen Riloff and Janyce Wiebe. Learning extraction patterns for subjective expressions. In *Proc. of Conference of Empirical Methods in Natural Language Processing*, pages Pages 105–112, 2003.

[RWJ+94] Stephen E. Robertson, Steve Walker, Susan Jones, Micheline Hancock-Beaulieu, and Mike Gatford. Okapi at TREC-3. In *Proc. of TREC-3, the 3th Text Retrieval Conference*, Gaithersburg, United States, 1994.

[RWW03] Ellen Riloff, Janyce Wiebe, and Theresa Wilson. Learning subjective nouns using extraction pattern bootstrapping. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003 - Volume 4*, CONLL '03, pages 25–32, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics.

[RZ09] Stephen E. Robertson and Hugo Zaragoza. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389, 2009.

[SAC07] Mark D. Smucker, James Allan, and Ben Carterette. A comparison of statistical significance tests for information retrieval evaluation. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, CIKM '07, pages 623–632, New York, NY, USA, 2007. ACM.

[SHMO09] Rodrygo L. T. Santos, Ben He, Craig Macdonald, and Iadh Ounis. Integrating proximity to subjective sentences for blog opinion retrieval. In *Proc. 31th European Conference on IR Research on Advances in Information Retrieval*, ECIR '09, pages 325–336, Berlin, Heidelberg, 2009. Springer-Verlag.

[SJ09] Jangwon Seo and Jiwoon Jeon. High precision retrieval using relevance-flow graph. In *Proc. 32nd Annual International ACM SIGIR Conference*, pages 694–695, 2009.

[SM03] Radu Soricut and Daniel Marcu. Sentence level discourse parsing using syntactic and lexical information. In *Proc. 2003 Conference of the North American Chapter of the ACL on Human Language Technology - Volume 1*, NAACL '03, pages 149–156, Stroudsburg, PA, USA, 2003. ACL.

[SMHM99] Craig Silverstein, Hannes Marais, Monika Henzinger, and Michael Moricz. Analysis of a very large web search engine query log. *SIGIR Forum*, 33(1):6–12, September 1999.

[SMK05] Tetsuya Sakai, Toshihiko Manabe, and Makoto Koyama. Flexible pseudo-relevance feedback via selective sampling. *ACM Transactions on Asian Language Information Processing (TALIP)*, 4(2):111–135, 2005.

[SMM+12] Rodrygo L. T. Santos, Craig Macdonald, Richard McCreadie, Iadh Ounis, and Ian Soboroff. Information retrieval on the blogosphere. *Found. Trends Inf. Retr.*, 6(1):1–125, January 2012.

[SNWG09] Swapna Somasundaran, Galileo Namata, Janyce Wiebe, and Lise Getoor. Supervised and unsupervised methods in employing discourse relations for improving opinion polarity classification. In *Proc. 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1*, EMNLP '09, pages 170–179, Stroudsburg, PA, USA, 2009. ACL.

[SWY75] Gerard M Salton, Andrew Wong, and ChungShu Yang. A vector space model for automatic indexing. *Commun. ACM*, 18(11):613–620, 1975.

[TM11] Oscar Täckström and Ryan McDonald. Discovering fine-grained sentiment with latent variable structured prediction models. In *Proceedings of the 33rd European Conference on Advances in Information Retrieval*, ECIR'11, pages 368–374, Berlin, Heidelberg, 2011. Springer-Verlag.

[Tur02] Peter D. Turney. Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 417–424, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics.

[VdSP+06] Karane Vieira, Altigran S. da Silva, Nick Pinto, Edleno S. de Moura, João M. B. Cavalcanti, and Juliana Freire. A fast and robust method for web page template detection and removal. In *Proceedings of the 15th ACM international conference on Information and knowledge management*, CIKM '06, pages 258–267, New York, NY, USA, 2006. ACM.

[WdMDR09] Wouter Weerkamp and de Maarten De Rijke. External query expansion in the blogosphere. In *Seventeenth Text REtrieval Conference (TREC 2008)*. NIST, NIST, 2009.

[WdR08] Wouter Weerkamp and Maarten de Rijke. Credibility improves topical blog post retrieval. In *Proceedings of ACL-08: HLT*, page 923–931, Columbus, Ohio, 2008. Association for Computational Linguistics, Association for Computational Linguistics.

[WHS+05] Theresa Wilson, Paul Hoffmann, Swapna Somasundaran, Jason Kessler, Janyce Wiebe, Yejin Choi, Claire Cardie, Ellen Riloff, and Siddharth Patwardhan. OpinionFinder: A system for subjectivity analysis. In *HLT/EMNLP*, 2005.

[WR05] Janyce Wiebe and Ellen Riloff. Creating subjective and objective sentence classifiers from unannotated texts. In *Proc. of 6th International Conference on Intelligent Text Processing and Computational Linguistics*, 2005.

[WWH05] Theresa Wilson, Janyce Wiebe, and Paul 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.

[YL99] Yiming Yang and Xin Liu. A re-examination of text categorization methods. In *22nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 1999)*, pages 42–49. Association for Computing Machinery, 1999.

[ZL04] Chengxiang Zhai and John Lafferty. A study of smoothing methods for Language Models applied to information retrieval. *ACM Trans. Inf. Syst.*, 22(2):179–214, April 2004.

[ZNSS11] Cacilia Zirn, Mathias Niepert, Heiner Stuckenschmidt, and Michael Strube. Fine-grained sentiment analysis with structural features. Number 12. Asian Federation of Natural Language Processing, 2011.

[Zob98] Justin Zobel. How reliable are the results of large-scale information retrieval experiments? In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '98, pages 307–314, New York, NY, USA, 1998. ACM.