



Using language models to improve opinion detection

Faiza Belbachir, Mohand Boughanem

► To cite this version:

Faiza Belbachir, Mohand Boughanem. Using language models to improve opinion detection. Information Processing and Management, 2018, 54 (6), pp.958-968. 10.1016/j.ipm.2018.07.001 . hal-02279437

HAL Id: hal-02279437

<https://hal.science/hal-02279437>

Submitted on 5 Sep 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Open Archive Toulouse Archive Ouverte

OATAO is an open access repository that collects the work of Toulouse researchers and makes it freely available over the web where possible

This is an author's version published in:

<http://oatao.univ-toulouse.fr/22484>

Official URL

DOI : <https://doi.org/10.1016/j.ipm.2018.07.001>

To cite this version: Belbachir, Faiza and Boughanem, Mohand *Using language models to improve opinion detection*. (2018) Information Processing & Management, 54 (6). 958-968. ISSN 0306-4573

Any correspondence concerning this service should be sent to the repository administrator: tech-oatao@listes-diff.inp-toulouse.fr

Using language models to improve opinion detection

Faiza Belbachir^{*,a}, Mohand Boughanem^b

^a *Institut Polytechnique des Sciences Avancées (IPSA), France*

^b *Institut de Recherche en Informatique de Toulouse (IRIT), France*

A B S T R A C T :

Keywords:

Information retrieval

Opinion detection

Blog

Language model

Opinion mining is one of the most important research tasks in the information retrieval research community. With the huge volume of opinionated data available on the Web, approaches must be developed to differentiate opinion from fact. In this paper, we present a lexicon-based approach for opinion retrieval. Generally, opinion retrieval consists of two stages: relevance to the query and opinion detection. In our work, we focus on the second state which itself focusses on detecting opinionated documents. We compare the document to be analyzed with opinionated sources that contain subjective information. We hypothesize that a document with a strong similarity to opinionated sources is more likely to be opinionated itself. Typical lexicon-based approaches treat and choose their opinion sources according to their test collection, then calculate the opinion score based on the frequency of subjective terms in the document. In our work, we use different open opinion collections without any specific treatment and consider them as a reference collection. We then use language models to determine opinion scores. The analysis document and reference collection are represented by different language models (i.e., Dirichlet, Jelinek-Mercer and two-stage models). These language models are generally used in information retrieval to represent the relationship between documents and queries. However, in our study, we modify these language models to represent opinionated documents. We carry out several experiments using Text REtrieval Conference (TREC) Blogs 06 as our analysis collection and Internet Movie Data Bases (IMDB), Multi-Perspective Question Answering (MPQA) and CHESLY as our reference collection. To improve opinion detection, we study the impact of using different language models to represent the document and reference collection alongside different combinations of opinion and retrieval scores. We then use this data to deduce the best opinion detection models. Using the best models, our approach improves on the best baseline of TREC Blog (baseline4) by 30%.

1. Introduction

The large volume of opinionated data on the Web has caused a recent increase in a number of online phenomena, such as online shopping and online elections. These opinionated data need to be manipulated in order to analyze, deduce or predict users choices in a variety of domains. Unlike traditional topic-based retrieval, the documents returned by opinion mining should not only be relevant to the topic but contain opinions about it.

While blogs are a rich source of opinions, they makes opinion detection more difficult because bloggers have a specific language that incorporates emoticons and does not respect grammatical rules. In 2006, TREC (Voorhees, 2006) debuted a special track with the

* Corresponding author.

E-mail addresses: phdups@gmail.com (F. Belbachir), Mohand.Boughanem@irit.fr (M. Boughanem).

main task of gathering opinions on various topics. Over the years, several research groups at TREC have developed different approaches for opinion retrieval. Their aim was to retrieve a set of opinionated documents for a given set of topics. These approaches ranked the retrieved list of opinionated documents in three different ways: using machine learning-based classifiers (Liu, 2015; Mullen & Collier, 2004; Pang & Lee, 2004; Riloff & Wiebe, 2003), lexical sentiment dictionaries (Hannah, Macdonald, Peng, He, & Ounis, 2007; Lafferty & Zhai, 2001; Song, Qin, Shi, Lin, & Yang, 2007) or probabilistic models (Huang & Croft, 2009; Mei, Ling, Wondra, Su, & Zhai, 2007a; Zhang & Ye, 2008). Most opinion retrieval approaches are designed to work in two phases: topic-relevance retrieval and opinion retrieval. During the topic-relevance retrieval phase, a list of relevant documents is retrieved and ranked according to documents relevance to the given topic. In the opinion retrieval phase, this list is then ranked again by combining each documents relevance and opinion scores. Different works combine both scores (relevant and opinion) and have reported some improvements in opinion detection. TREC organizers have published baselines to encourage participants to experiment further with TREC approaches. In this study, we focus on the opinion retrieval phase. We therefore use the strongest provided TREC baseline (baseline4) Ounis, Macdonald, and Soboroff (2009) for determining topic relevance, while relying on language modeling techniques for opinion retrieval. To determine if a document is opinionated or not, we match its language model with the language models of documents containing subjective information (i.e., our reference collection). If a document is determined to be similar to our collection, we conclude that it is opinionated. The novelty and effectiveness of this approach rely on the following key features:

- We use various open and available subjective resources;
- We use different methods to calculate opinion scores;
- We adapt the language models used for opinion detection;
- We compare different language models for representing the analysis document and reference collection in order to find the models that most improve opinion detection as compared to the best baseline provided by TREC (baseline4).

This paper is organized as follows: In Section 2, we describe related work on the opinion retrieval phase. In Sections 3 and 4, we describe our proposed approach in detail. In Section 5, we discuss the experiments and the obtained results. Finally, we conclude the paper and summarize our findings.

2. Related work

There are several extant studies in the field of opinion mining. Some of these use approaches with a classifier that takes data as its input and produces output against testing data. With these approaches, difficulties include determining the features that represent opinionated documents, identifying the best classifiers for a better result, and choosing the training collection that represents the most subjective words. The features used to represent opinionated documents include, for example, the number of adjectives, verbs and adverbs (Bifet & Frank, 2010; Li, Mukherjee, Si, & Liu, 2015; Pang, Lee, & Vaithyanathan, 2002; Wang, Sun, Mukhtar & Rohini, 2008; Yang, Callan, & Si, 2006; Zhang, 2006). Some studies, meanwhile, examine the grammatical relations between different terms in a document. These studies suggest that subjectivity can only be measured in context, because some words, such as “like”, are considered subjective, when they are used in other sentences they may no longer express an opinion, such as in the sentence “it looks like a cat” (Agarwal, Xie, Vovsha, Rambow, & Passonneau, 2011; Liu, Liu, Zhang, Kim, & Gao, 2016; Saif, He, Fernandez, & Alani, 2014; Saif, He, Fernández, & Alani, 2016; Seki & Uehara, 2009; Wang, Chen, & Liu, 2016). Some studies, such as the work of Aidan, Kushmerick, and Smyth (2002), classify features of opinionated documents into two categories: those that depend on the query and incorporate relevance and opinion into the learning phase (Saif et al., 2014; Seki, Kino, Sato, & Uehara, 2007), and those that use characteristics independent of the topic and do not incorporate relevance into the learning phase. Furthermore, while some studies use a single classifier like support vector machine (SVM), naive bayes or logistic regression to return opinionated documents, others use multiple different classifiers to compare their impacts on opinion detection (Balahur, 2016; Balahur & Jacquet, 2015; Bauman, Liu, & Tuzhilin, 2016; Fu, Abbasi, Zeng, & Chen, 2012; Lu, Mamoulis, Pitoura, & Tsaparas, 2016; Mullen & Collier, 2004; Pang & Lee, 2004; Riloff & Wiebe, 2003; Seki et al., 2007; Tu, Cheung, Mamoulis, Yang, & Lu, 2016). Finally, some pre-existing approaches use internal collections built directly from the collection to be analyzed for collections training, while others use external collections built from independent collections of the analyzed collection (Aidan et al., 2002; Baccianella, Esuli, & Sebastiani, 2010; Bifet & Frank, 2010; Pak & Paroubek, 2010; Seki et al., 2007). Due to their reliance on machine learning, these approaches are dependent on learning the data, features and choices of the classifier being used. Other studies use dictionaries of subjective words to identify opinionated documents, considering documents that contain many subjective words to be opinionated. Sometimes, these dictionaries are directly prepared from the test data collection. Other times, ready-made lexical dictionaries like General Inquiry (Stone, Dunphy, Smith, & Ogilvie, 1966) or SentiWordNet (SWN) (Esuli & Sebastiani, 2006) are used, although some studies report that these are less effective than lexicons extracted from the test data collection itself (Andreevskaia & Bergler, 2006; Lafferty & Zhai, 2001). While many approaches use individual words for processing semantic information, some approaches use natural language processing techniques that consider not only the word but the entire sentence (Agarwal et al., 2011; Castro-Espinoza, Gelbukh, & González-Mendoza, 2013; Liu et al., 2016; Tang, Tan, & Cheng, 2009; Thelwall, Buckley, & Paltoglou, 2012; Thelwall, Buckley, Paltoglou, Cai, & Kappas, 2010). These approaches are typically based on the presence or absence of a subjective term in a document, and do not incorporate the frequency of the subjective term in question. However, a term that reoccurs several times in a document suggests a stronger opinion than one that appears only once. This issue helps to explain why relatively few studies employ language models to identify opinionated documents. Language modeling has proved its worth in the field of information retrieval (Song & Croft, 1999) for ad-hoc information retrieval tasks based on probability to represent a model of document and a model of query.

Similarly, researchers have exploited the benefits of language modeling in opinion mining as well. In particular, the authors of [Huang and Croft \(2009\)](#) propose a unified opinion-retrieval model based on the Kullback–Leibler (KL) divergence between the two probability distributions of the opinion relevance model and the document model. The authors divide sentiment words into query-dependent and query-independent categories by using sentiment expansion techniques, and then integrate them into a mixed model. In order to extract sentiment words, Cornell movie review datasets, the MPQA corpus¹ and TREC Blog 06 ([Macdonald & Ounis, 2006](#)) are used as opinionated collections. The authors of [Zhang and Ye \(2008\)](#) propose a generative model to unify topic relevance and opinion scores. The authors use a language model approach with smoothing to couple retrieval and opinion scores. Different sentiment lexicons like Wornet, General Inquiry and HowNet are used to calculate opinion scores, while binary independent retrieval is used to calculate topic relevance scores. The authors of [Mei, Ling, Wondra, Su, and Zhai \(2007b\)](#) also create a probabilistic model to capture the mixture of topics and sentiments simultaneously present in consumer products. The opinion score for a given product is a mixture of a variety of facets. However, associating a sentiment with products and facets is challenging. This approach has been tested with the small-scale collection OpinMind.² The authors of [Mei et al. \(2007a\)](#) propose an opinion retrieval model using the framework of generative language modeling. They model a collection of natural language documents or statements, each containing topic-bearing and sentiment-bearing words. Sentiment is either represented by a group of predefined seed words or extracted from a training sentiment corpus. This model has been demonstrated to be effective on the MPQA corpus. Our approach is based on using language models to identify opinionated documents. In contrast to pre-existing work, we use different open, subjective resources without any prior treatment; adapt the language model of information retrieval in opinion detection and analyze the different factors that impact opinions in order to best represent our documents.

3. Our opinion finding approach

In our approach, we consider that a document contains an opinion if it is generated by a language model associated with opinionated documents. To construct a model of opinionated language, we rely on a reference collection containing opinionated documents. Similarly, we also model the analysis document in the form of a language model. In order to measure the degree of subjectivity (i.e., the presence of opinions) in the document, we measure the similarity between the two models. In [Fig. 1](#), we describe the steps of our proposed lexical approach. These steps are as follows:

- *Step 1: Determine relevance to the query:* The process of opinion detection must return documents containing opinions that are relevant to the given query.
- *Step 2: Perform opinion detection:* To identify opinionated documents, we use different subjective resources, such as IMDB, CHESLY, ROTTEN and MPQA. We call these resources the reference collection, and consider it our opinion collection. We then represent the reference and documents collections in a variety of different ways using several language models.

The aim is to represent the collection and the document to be analyzed as accurately as possible. After modeling the analysis document and the reference collection, we measure the similarity of the two models using their opinion scores. The focus of our work is therefore the accurate modeling of the document and reference collection and the calculation of opinion scores. In the following sections, we detail our approach.

3.1. Language model to represent document

There are different language models used in the field of Information retrieval. We adapt some of them to the opinion detection by introducing the notion of opinion. For that we consider four models to represent the test document : Maximum Likelihood (ML), Dirichlet (DIR), Jelinek Merker (JM) and Two Stage (ST).

1. **Maximum likelihood (ML)** ([Jelinek, 1997](#)): In this method we use a simple language model to represent the document θ_{D_ML} (see [Eq. \(1\)](#)).

$$\theta_{D_ML} = P_{ML}(w|D) = \frac{\#c(w, D)}{|D|} \quad (1)$$

where $\#c(w, D)$ denotes the number of times the word w occurs in document D , and $|D|$ refers to the number of words in the document D .

To introduce the opinion concept in this model, we smooth the model of document with the model of reference collection.

2. **Jelinek Mercer (JM)** ([Bahl, Jelinek, & Mercer, 1983](#)): In this method, we interpolate the model of Maximum Likelihood $P_{ML}(w|D)$ with the language model of reference collection $P_{ML}(w|R)$ using a coefficient λ_d fixed to control smoothing for these two models, where $P_{ML}(w|R)$ represents the probability that a term w appears in the test collection R . This method is expressed by the [Eq. \(2\)](#).

$$\theta_{D_JM} = P_{JM}(w|D) = \lambda_d * P_{ML}(w|D) + (1 - \lambda_d) * P_{ML}(w|R) \quad (2)$$

¹ <http://www.cs.pitt.edu/mpqa/>.

² <https://techcrunch.com/tag/opinmind/>.

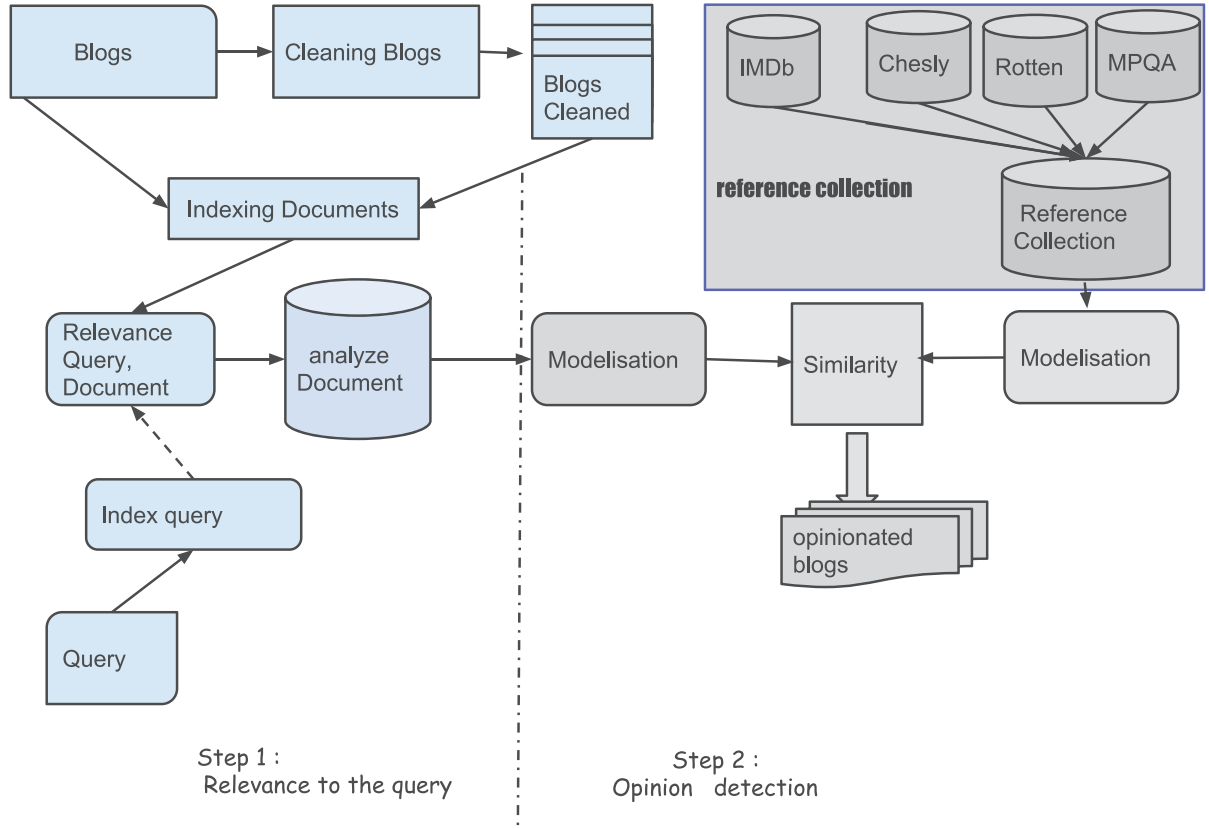


Fig. 1. Opinion finding approach.

It is clear that if coefficient $\lambda_d = 1$ the model of D is an ML model, and if $\lambda_d = 0$ it is the same as reference collection's model $P(w|R)$. For any other value of this coefficient, the document's model is generated by the model of the reference collection. This is used to boost the documents containing the terms of reference collection.

3. **Dirichlet (DIR)** (Zhai & Lafferty, 2001): This method also interpolates the ML model with a model of reference collection to encourage holders of opinion documents, with a dynamic smoothing parameter λ which changes depending on the length of the document. The model is represented by Eq. (3) where $\lambda = \frac{|D|}{|D| + \mu}$ and μ is a fixed parameter. This means that if the document is long, then the smoothing is low.

$$\theta_{D_DIR} = P_{DIR}(w|D) = \frac{|D|}{|D| + \mu} P_{ML}(w|D) + \frac{\mu}{|D| + \mu} P_{ML}(w|R) \quad (3)$$

4. **Two Stage (TS)** (Zhai & Lafferty, 2002): In this method, a combination of both smoothing models is made, this combination improves the representation of the document θ_D . First the document is estimated using Dirichlet smoothing with a dynamic coefficient $\lambda = \frac{|D|}{|D| + \mu}$ depending on the size of the document D .

But this model penalize longer documents. To remedy for this problem a second smoothing method based on Jenlinek Mercer used with a fixed smoothing coefficient γ_D . This combination of two smoothing models can smooth any kind of document, regardless of its size, with the language model of the reference collection. Another advantage of this double smoothing is that the document view is doubly boosted due to the insertion of the R model (of opinion model) in the model Dirichlet and the model of Jenlinek Mercer on a same time. This method is represented by Eq. (4).

$$\theta_{D_TS} = P_{TS}(w|D) = \gamma_D \frac{\#c(w|D) + \mu P_{ML}(w|R)}{|D| + \mu} + (1 - \gamma_D) P_{ML}(w|R) \quad (4)$$

To reinforce this concept of subjectivity, we also propose to consider the subjectivity a priori of a term based on the lexical resource SentiWordNet (SWN) (Esuli & Sebastiani, 2006).

SWN assigns three numerical scores (Obj(s), Pos(s), Neg(s)) to each synset of the WordNet describing how objective, positive or negative the terms within a synset are. The range of three scores lies in interval $[0, 1]$ and sum of all the scores equals to 1. A template of SWN is shown in Fig. 2. Template of SentiWordNet with first column: Parts of Speech (POS) of the Synset, 2nd column: Offset of the Synset in WordNet, 3rd Column: Positive Score of the Synset, 4th Column: Negative Score of the Synset, 5th Column: Entries of a Synset Quantitative analysis of the glosses of the synsets is performed to obtain three scores. The idea behind the creation of SWN was that different senses of a term might have different semantic orientations. For example, the term estimable is

POS	Offset	PosScore	NegScore	SynsetTerms
a	1000003	0.0	0.125	form-only#a#1
a	1000159	0.25	0.0	dress#a#1 full-dress#a#
a	1000307	0.0	0.125	titular#a#5 nominal#a#6
a	1000440	0.0	0.125	prescribed#a#4 positive#a#5

Fig. 2. SentiWordNet.

objective (i.e. $\text{Obj}(\text{estimable}) = 1.0$ with its $\text{Pos} = \text{Neg} = 0.0$) corresponding to its sense may be computed or estimated and SWN scores for the same term become as $\text{Obj}(\text{estimable}) = 0.25$, $\text{Neg}(\text{estimable}) = 0$ and $\text{Pos}(\text{estimable}) = 0.75$ when its sense deserving of respect.

A simple way to measure the subjectivity of a term in SWN is taking the average subjectivity (positive and negative) of synsets in which the term appears. In fact, this approach is rather simplistic; it makes no disambiguation of terms. We calculate the subjectivity score of a term by summing the positive and negative scores in all synsets of the term and dividing the total by the number of synset (see Eq. (5)).

$$\text{Subj}(w) = \sum_{si \in \text{sens}(w)} \frac{(\text{Neg}(si) + \text{Pos}(si))}{|\text{sens}(w)|} \quad (5)$$

Where respectively $\text{Neg}(si)$ or $\text{Pos}(si)$ is the negative/positive score of sense (or synset) si of term w as found in SentiWordNet SWN and $|\text{sens}(w)|$ is the total number of senses for term w in SWN.

Considering this subjectivity, the model of document θ_D will be represented by Eq. (6).

$$\theta_{D_M_Subj} = P_M(w|D) * \text{Subj}(w) \quad (6)$$

With $M \in \{ML, DIR, JM, TS\}$

After representation the model of document, we proceed to modelling the reference collection.

3.2. Language model to represent reference collection

The model of reference collection is also estimated in different ways. Eqs. (7)–(11) refers respectively to Eqs. (1)–(4) and (6) where some parameters are changed. We explain all equations as following:

The first model is based on a simple probability frequency of terms in the reference collection (see Eq. (7)).

$$\theta_{R_ML} = P_{ML}(w|R) = \frac{\#c(w, R)}{|R|} \quad (7)$$

where $\#c(w, R)$ denotes the number of times the word w occurs in the reference collection R and $|R|$ refers to the number of words in the reference collection.

If a term of the document does not appear in the reference collection then it will have a probability equal to zero. This model does not boost the terms that belong to the test collection. Thus three smoothing models are used (the same as used previously in the modeling of the document), these models combine the model of the reference collection (R) with the model of the test collection (C).

The first smoothing model is based on the model of Jenlinek Mercer represented by the Eq. (8).

$$\theta_{R_JM} = P_{JM}(w|R) = \lambda_r * P_{ML}(w|R) + (1 - \lambda_r) * P_{ML}(w|C) \quad (8)$$

Where λ_r is smoothing parameter and C is the test collection. This smoothing change with the value given to the variable λ_r . This means that it favours large documents containing the terms of the collection.

The second model is based on the Dirichlet smoothing model, represented by the Eq. (9), with a smoothing factor equal to $\frac{|R|}{|R| + \mu}$ where μ is a fixed parameter.

$$\theta_{R_DIR} = P_{DIR}(w|R) = \frac{|R|}{|R| + \mu} P_{ML}(w|R) + \frac{\mu}{|R| + \mu} P_{ML}(w|C) \quad (9)$$

Where $P_{ML}(w|C)$ is the probability that term w is in the test collection C .

These two models (JM and DIR) solved the problem to end up with a zero probability when a document does not contain the terms of reference collection by adding the probability $P_{ML}(w|C)$.

The third proposed model, called Two Stage, is a combination of the two modeles represented by Eq. (10).

$$\theta_{R_TS} = P_{TS}(w|R) = \gamma_R P_{ML}(w|R) + (1 - \gamma_R) \frac{\#c(w, C) + \mu P_{ML}(w|R)}{|C| + \mu} \quad (10)$$

γ_R is coefficient fixed to control smoothing models. This model has the advantage of boosting doubly opinionated documents

because it inserts twice probability of $P(w|R)$ in the ML model and in the DIR model. To booster advantage of subjective words, we propose to weight each term with its subjectivity score, as was done for the model document. The model of reference collection is represented by the general Eq. (11).

$$\theta_{R_M_Subj} = P_M(w|R) * Subj(w) \quad (11)$$

With $M \in \{ML, DIR, JM, TS\}$

After modeling the document and the reference collection, we define in the following different opinion scores that we have used to re-rank opinionated documents.

4. Opinion score

We propose to calculate the score of opinion at document level in various functions. We first use, the Kullback–Leibler divergence ($Score_KL_R(D, R)$) (Zhai & Lafferty, 2001).

$$Score_KL_R(D, R) = \sum_{w \in D} \theta_D * \log \frac{\theta_D}{\theta_R} \quad (12)$$

Where θ_D and θ_R are the language models of respectively the document and the collection of opinions as they were defined in the previous sections. This opinion score function computes the divergence between different probability distributions. When the score is lower it means that the document is similar to the reference collection.

A second way to calculate the opinion score is to assess the joint probability of all terms of the document. This amounts to compute the product of the probabilities of weighted terms of the entire document (see Eq. (13)):

$$Score_prod_R(D) = \prod_{w \in D} P(w|D) \quad (13)$$

The intuition here is the following: the distribution of words in the document is supposed to model the importance of words in the document and boost the opinion words. This means that more $Score_prod_R(D)$ is higher more the document contains opinions. The third way to measure the score is given by the following equation:

$$Score_mixte_R(D) = \sum_{w \in D} Opinion(w) * P(w|D) \quad (14)$$

This score takes into account the opinion score represented by $Opinion(w)$ and the frequency of words in the document represented by $P(w|D)$.

The opinion score is expressed by the terms of the document that must be in the reference collection $P(w|R)$ and in the lexicon SentiWordNet ($Subj$).

$$Opinion(w) = P(w|R) * subj(w) \quad (15)$$

All scores calculated in this section are not referring to the relevant concept. We recall that our aim is to study the opinion dimension only. Regarding the relevance dimension, we considered that it is calculated also using any information retrieval model. Therefore, to return a list of relevant documents to the query and expressing an opinion about it, we propose to combine in different ways relevance score and opinion score. This is discussed in the last part of the experimental section.

5. Experimentation results

5.1. Data collections

To evaluate our proposition, we conducted experiments using the TREC Blog 2006 data set as our test collection and the IMDB, CHESLY, MPQA and ROTTEN resources as our reference collection data.

5.1.1. Test collection: TREC Blogs track

The TREC Blog 2006 data collection (Voorhees, 2006) consists of more than 3.2 million blogs crawled for a period of 11 weeks from December 2005 to February 2006. The Text REtrieval Conference proposes a set of subjects (50 topics per year) and relevant judgements (qrels) tagged according to the following scheme: 0 for irrelevant blogs, 1 for relevant blogs, 2 for blogs with negative opinions, 3 for those with mixed opinions and 4 for those with positive opinions. In order to facilitate direct comparison between systems, five relevance retrieval baselines have been provided by TREC organizers, selected from the best performing retrieval runs submitted by participating groups. The best baseline is baseline4,³ and we accordingly compare our results with this baseline.

5.1.2. Reference opinion data collection

We used different corpora of opinion collections as our reference collection. The first corpus is provided by Pang and Lee (2004)

³ <http://trec.nist.gov/data/blog08.html>.

and contains 1000 positive and 1000 negative full-text movie reviews taken from the IMDB archive.⁴ The second corpus constitutes 50,000 subjective sentences from Rotten Tomatoes customer review snippets.⁵ The third is the MPQA opinion corpus,⁶ which contains 10,657 sentences across 535 documents. These documents are extracted from 187 different foreign and U.S. news sources from June 2001 and May 2002. The last corpus, CHESLEY (Chesley, Vincent, Xu, & Srihari, 2006), is a manually annotated dataset of objective and subjective documents developed by Chesley et al. (2006). It contains 496 subjective and 580 objective documents. Before using these corpora as our test collection and reference opinion data collection, we tagged them for parts of speech (POS-tagging) using CRFTagger (Phan, 2006). In keeping with several studies (Bruce & Wiebe, 1999; Wiebe, Bruce, & O'Hara, 1999), we considered only the adjectives, adverbs, verbs and nouns including subjective information.

5.2. Results

As previously mentioned, the purpose of our study is to investigate the opinion dimension of opinion retrieval exclusively. In order to evaluate the performance of our approach, we used the list of relevant documents supplied by TREC assessors, TREC Topics 2006. We then reordered the documents based on their opinion scores. Evaluation results are presented in terms of average precision (AP) for queries and mean average precision (MAP) for sets of queries and documents graded 10 for accuracy (i.e., P@10). We conducted some preliminary experiments and set smoothing parameters λ to 0.6 and μ to 0.1.

We evaluated our approach in different ways:

- the impact of SWN-calculated subjectivity on opinion detection;
- the impact of the reference collection on opinion detection;
- the impact of the document model on opinion detection;
- the impact of opinion scores on opinion detection;
- the impact of the method used for combining opinion score and relevance score on opinion detection.

5.2.1. Impact of SentiWordNet on opinion detection

In this experiment, we aimed to measure the impact of SWN calculated subjectivity on opinion detection. In order to do so, we compared the obtained results from the two models (document and reference) when incorporating subjectivity ($\theta_{D_ML_Subj}$, $\theta_{R_ML_Subj}$) and when not incorporating subjectivity (θ_{D_ML} , θ_{R_ML}). We restricted the study to ML language models due to our focus on the role of SWN-calculated subjectivity on models not already improve by smoothing techniques. The opinion scores were calculated according to the three scores ($Score_prod_R(D)$, $Score_KL_R(D)$, $Score_mixte_R(D)$). The best results, shown in Table 1, were obtained using the $Score_KL_R(D)$. Table 1 shows the MAP and P@10 values obtained for all 50 queries.

The results reveal that the model incorporating subjectivity was better than the model that did not (16% improvement on the MAP and 143% on P@10). We conclude that the SWN lexicon is a rich source of opinion words, even in blogs.

5.2.2. Impact of the reference collection model on opinion detection

We also studied the impact of the language model of the reference collection on opinion detection, with the aim of discovering the best model for representing reference collections. In order to do so, we considered the model of the document as a representation (θ_{D_ML}) so as not to bias the results and reference model of the three models mentioned previously ((θ_{R_DIR}) , (θ_{R_JM}) and (θ_{R_TS})). We calculated the opinion score using three scores, the best results shown in Table 2, were obtained using the $Score_KL_R(D, R)$. Table 2 details the results obtained in terms of the MAP and P@10 for all requests.

We note that the best result was produced by the pair (ML, TS). Specifically, the best representation of the reference collection was a two-stage model, with an improvement of over 39.26 in the MAP and 73.43 in the P@10 compared to other representations. A smoothing reference model with a model of the test document therefore improves opinion detection.

5.2.3. Impact of the document model on opinion detection

In order to study the impact of the test documents language model on opinion detection, we considered the model of the document in the three performances mentioned previously ((θ_{D_DIR}) , (θ_{D_JM}) and (θ_{D_TS})). The model that best represents the reference collection (θ_{R_TS}) was then deduced from the previous study. The opinion scores considered are those obtained by the $Score_KL_R(D, R)$ (see Table 3), because they give better results than the other scores. Table 3 displays the MAP and P@10 results obtained on all requests. The results demonstrate that the two stage model was the best model for representing opinionated documents, with an improvement of more than 4.88% in the MAP and 6.35% in accuracy compared to other representations.

The improvement of the result is until less because we use the best representation of the reference model that is smoothed with the test document model. We can conclude that using a reference model in the model of the test document improves the opinion score.

5.2.4. Impact of opinion scores on opinion detection

Finally, we compared the three scoring functions proposed in the previous section, namely the score based on the KL divergence

⁴ <http://www.cs.cornell.edu/people/pabo/movie-review-data/>.

⁵ <https://github.com/cacoderquan/Sentiment-Analysis-on-the-Rotten-Tomatoes-movie-review-dataset>.

⁶ <http://www.cs.pitt.edu/mpqa/>.

Table 1

Result of MAP and P@10 for models that take-do not take a subjectivity SWN.

Configuration	MAP	P@10
with Subjectivity	0.1488	0.2896
without Subjectivity	0.1279	0.1187

Table 2

Results of MAP and P@10 for different models to represent reference collection.

Model(document, reference)	MAP	P@10
(<i>ML</i> , <i>DIR</i>)	0.1231	0.1375
(<i>ML</i> , <i>JM</i>)	0.1782	0.3646
(<i>ML</i> , <i>TS</i>)	0.2098	0.4354

Table 3

Result of MAP and P@10 for different model to represent test document.

Model (document, reference)	MAP	P@10
(<i>DIR</i> , <i>TS</i>)	0.2122	0.4417
(<i>JM</i> , <i>TS</i>)	0.2260	0.4750
(<i>TS</i> , <i>TS</i>)	0.2298	0.4875

($Score_KL_R(D, R)$), the score based on the product of weighted terms ($Score_Prod_R(D)$) and the mixed score ($Score_Mixte_R(D)$). We made use of optimal representations for our document and reference collection, as derived from our previous study (θ_{D_TS} and θ_{R_TS}).

The results are represented in Table 4. We find that the KL divergence produces much better results than those obtained by the ($Score_Prod_R(D)$) and ($Score_Mixte_R(D)$) functions.

This is explained by the way in which the KL divergence explicitly expresses the concept of opinion in two different ways: in the model of the document θ_d and in the reference model θ_R . While the score for which is based on the product, the concept of opinion is expressed only in the probability $P(w|D)$, explaining its low performance. The mixed score, meanwhile, also expresses the concept of opinion in two ways with the probabilities $P(w|R)$ and $Subj(w)$ but differs from the KL divergence by including the concept of frequency words, expressed by the probability $P(W|D)$. The poor results obtained by the mixed score as compared to the similarity score reflect the fact that the repetition of a slightly subjective term several times in a document does not guarantee that this document is highly subjective, while a subjective term repeated only once in the document may make it highly opinionated.

6. Combining relevance and opinion scores

Due to our focus on assessing the opinion dimension exclusively, the results detailed above do not take into account the relevance of documents. In order to find relevant documents expressing an opinion, we propose combining the opinion score $Opinion(D)$ of a document with its relevance score. The opinion score of a document is calculated on $Score_KL_R(D)$, based on the best performances in previous studies (θ_{D_TS} and θ_{R_TS}). The relevance score of a document to a query $Retrieval(D, q)$ is provided by TREC. We combined the relevance and opinion scores in two different ways: The first score was based on a linear combination given by Eq. (16).

$$Score_Final_Linear(D) = \alpha * Retrieval(D, q) + (1 - \alpha) * Opinion(D) \quad (16)$$

The second was based on a product combination given by Eq. (17).

$$Score_Final_Product(D) = Retrieval(D, q) * Opinion(D) \quad (17)$$

$Retrieval(D, q)$ represents the retrieval score of the document D for query q , $Opinion(D)$ represents the opinion score of the document D calculated with the best measure score of opinion detection ($Score_KL_R(D, R)$), represented in Eq. (12) and α represents a smoothing parameter.

Table 4

Result of MAP and P@10 for different opinion score.

Opinion score	MAP	P@10
Similarity $Score_KL_R(D)$	0.2298	0.4875
Product $Score_Prod_R(D)$	0.1131	0.1333
Mixte $Score_Mixte_R(D)$	0.1654	0.3125

Table 5
Result of MAP and P@10 for product, linear and *baseline4* opinion score.

Configuration	MAP	P@10
Product	0.3145	0.6354
Linear	0.3453	0.6729
Baseline 4	0.3022	0.5240

Table 5 lists the results of MAP and P@10 scores of both configurations (Produced and Linear) and the best *baseline* TREC (*baseline 4*)⁷.

These results demonstrate that both configurations improve *baseline4*. The lowest result for these two configurations was produced by the product combination. This is most likely explained by the product combinations weighting of the relevance score and the opinion score, as compared to the linear combination, which promotes more opinionated documents by assigning $\alpha = 0.1$. There was an improvement of more than 14.26% in the MAP and 28.48% in the P@10 to linear score compared to the product score. We also observe that the configuration based on the linear score produced an improvement of more than 10.65% in the MAP and 24.84% in the P@10 compared to the best result produced by TREC, *baseline4*. The significance of these improvements on the baseline were validated with a *t*-test (with $p < 0.05$).

We also compared our work with other approaches, such as the work of Ellen Voorhees and al. Song et al. (2007), which obtained a MAP of 0.1885 and a P@10 of 0.5120 using TREC evaluation. Our approach improved on this work by more than 43.51% at the MAP level and 24.84% at the P@10 level. In addition, we compared our results against studies that did not participate in TREC but made use of the same collection and topics. In particular, we compared our work with the work of Seki and Uehara (2009), who obtain the best results with a MAP of 0.3221 but no P@10 calculation. With our approach, we improved on the MAP of these results by 3.67%. Compared to the work of Missen, Boughanem, and Cabanac (2010), who obtained a MAP of 0.3303 and a P@10 of 0.6340, our results show an improvement of 1.22% and 3.08%, respectively.

7. Conclusion

In this article, we propose an approach for identifying opinionated documents. In particular, we assume that a document contains opinions if it is similar to the opinionated collection (or reference collection). In order to measure this, we model the document and the reference collection using different language models. We adapt language models used in information retrieval to represent the query and document for opinion detection. In order to do so, we introduce a document model in the reference collection and a reference model in the document model using several smoothing methods. Our study further extends from modelling and improving opinion detection to include an examination of the impact of different factors that influence opinion retrieval. For example, we investigate whether the SWN lexicon is adapted to the language of blogs or contains opinion words that appear in blogs. We also evaluate different language models to identify the best representation of a document or reference collection. Finally, we calculate various scores for re-ranking opinion documents. The conducted experiments validate our hypothesis on the use of opinionated collections without analysis (i.e., subjective words extraction). We also conclude that our smoothing models best represent opinionated documents and scores based on similarity using KL divergence. Our results represent a significant improvement on the obtained TREC baseline, as well as other configurations. Our future work will focus on two main points. Firstly, we will investigate a better method for modeling opinion sources. In our current approach, all the terms in the reference collection are used; it would be valuable to discover a method to boost more subjective terms. Secondly, we will extend our language model to incorporate polarity detection. More specifically, the ability to identify whether the opinion expressed in a document is positive, negative or neutral would enable us to obtain additional information on the given subject. We will also focus on the use of various document information such as date, gender, age, profile and subject category for results validation and recommendations. It may also be valuable to investigate the existence of influencing factors in order to improve opinion detection. In a discussion about makeup, for example, a likely influencing factor is gender; when discussing phones, however, the likely factor is age. Once such factors are determined, their level of confidence will need to be determined.

Acknowledgement

All this work has been done at the IRIT Laboratory.

References

- Aidan, F., Kushmerick, N., & Smyth, B. (2002). *Genre classification and domain transfer for information filtering*. *European conference on information retrieval (ECIR) Lecture Notes in Computer Science, Springer* 2291. *European conference on information retrieval (ECIR)* 353–362.
- Agarwal, A., Xie, B., Vovsha, I., Rambow, O., & Passonneau, R. (2011). *Sentiment analysis of twitter data*. *Proceedings of the workshop on languages in social media* LSM'11 Stroudsburg, PA, USA: Association for Computational Linguistics 30–38.
- Andreevskaia, A., & Bergler, S. (2006). *Semantic tag extraction from WordNet glosses*. *Proceedings of the 5th conference on language resources and evaluation (LREC-2006)*,

⁷ <http://trec.nist.gov/pubs/trec17/papers/BLOG.OVERVIEW08.pdf>.

Genova, Italy.

- Baccianella, S., Esuli, A., & Sebastiani, F. (2010). Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In N. C. C. Chair), K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, & D. Tapias (Eds.). *Proceedings of the seventh international conference on language resources and evaluation (LREC'10)*. Valletta, Malta: European Language Resources Association (ELRA).
- Bahl, L. R., Jelinek, F., & Mercer, R. L. (1983). A maximum likelihood approach to continuous speech recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 5(2), 179–190. <https://doi.org/10.1109/TPAMI.1983.4767370>.
- Balahur, A. (2016). *Sentiment analysis - What are we talking about? Proceedings of the 7th workshop on computational approaches to subjectivity, sentiment and social media analysis, WASSA@NAACL-HLT 2016, June 16, 2016, San Diego, California, USA1*.
- Balahur, A., & Jacquet, G. (2015). Sentiment analysis meets social media - challenges and solutions of the field in view of the current information sharing context. *Information Processing and Management*, 51(4), 428–432. <https://doi.org/10.1016/j.ipm.2015.05.005>.
- Bauman, K., Liu, B., & Tuzhilin, A. (2016). Recommending items with conditions enhancing user experiences based on sentiment analysis of reviews. In T. Bogers, M. Koolen, C. Musto, P. Lops, & G. Semeraro (Vol. Eds.), *CEUR workshop proceedings, CEUR-WS.org*. : 1673. *Proceedings of the 3rd workshop on new trends in content-based recommender systems co-located with ACM conference on recommender systems (RecSys 2016), Boston, MA, USA, September 16, 2016*. (pp. 19–22).
- Bifet, A., & Frank, E. (2010). Sentiment knowledge discovery in twitter streaming data. In B. Pfahringer, G. Holmes, & A. Hoffmann (Eds.). *Discovery science: 13th international conference, DS 2010, Canberra, Australia, October 6–8, 2010. Proceedings* (pp. 1–15). Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-16184-1_1.
- Bruce, R. F., & Wiebe, J. M. (1999). Recognizing subjectivity: A case study of manual tagging. *Natural Language Engineering*, 5, 187–205.
- Advances in soft computing and its applications – 12th Mexican international conference on artificial intelligence, MICAI 2013, Mexico City, Mexico, November 24–30, 2013, proceedings, part II. In F. Castro-Espinoza, A. F. Gelbukh, & M. González-Mendoza (Vol. Eds.), *Lecture Notes in Computer Science: 8266*Springer<https://doi.org/10.1007/978-3-642-45111-9>.
- Chesley, P., Vincent, B., Xu, L., & Srihari, R. (2006). Using verbs and adjectives to automatically classify blog sentiment. *AAAI symposium on computational approaches to analysing weblogs (AAAI-CAAW)27–29*.
- Esuli, A., & Sebastiani, F. (2006). Sentiwordnet: A publicly available lexical resource for opinion mining. In *proceedings of the 5th conference on language resources and evaluation (LREC06)417–422*.
- Fu, T., Abbasi, A., Zeng, D., & Chen, H. (2012). Sentimental spidering: leveraging opinion information in focused crawlers. *ACM Transactions on Information and Systems*, 30(4), 24:1–24:30. <https://doi.org/10.1145/2382438.2382443>.
- Hannah, D., Macdonald, C., Peng, J., He, B., & Ounis, I. (2007). *University of glasgow at trec 2007: Experiments in blog and enterprise tracks with terrier. TREC*.
- Huang, X., & Croft, W. B. (2009). A unified relevance model for opinion retrieval. *CIKM947–956*.
- Jelinek, F. (1997). *Statistical methods for speech recognition*. Cambridge, MA, USA: MIT Press.
- Lafferty, J., & Zhai, C. (2001). Document language models, query models, and risk minimization for information retrieval. *Proceedings of the 24th annual international ACM SIGIR conference on research and development in information retrievalSIGIR '01*New York, NY, USA: ACM111–119. <https://doi.org/10.1145/383952.383970>.
- Li, H., Mukherjee, A., Si, J., & Liu, B. (2015). Extracting verb expressions implying negative opinions. *Proceedings of the twenty-ninth AAAI conference on artificial intelligenceAAAI'15*AAAI Press2411–2417.
- Liu, B. (2015). *Sentiment analysis - Mining opinions, sentiments, and emotions*. Cambridge University Press.
- Liu, Q., Liu, B., Zhang, Y., Kim, D. S., & Gao, Z. (2016). Improving opinion aspect extraction using semantic similarity and aspect associations. *Proceedings of the thirtieth AAAI conference on artificial intelligence, February 12–17, 2016, Phoenix, Arizona, USA2986–2992*.
- Lu, Z., Mamoulis, N., Pitoura, E., & Tsaparas, P. (2016). Sentiment-based topic suggestion for micro-reviews. *ICWSM. AAAI Press231–240*.
- Macdonald, C., & Ounis, I. (2006). *The TREC Blog06 collection: Creating and analysing a blog test collection*Technical Report. Scotland, United Kingdom: Department of Computing Science, University of Glasgow.
- Mei, Q., Ling, X., Wondra, M., Su, H., & Zhai, C. (Ling, Wondra, Su, Zhai, 2007a). Topic sentiment mixture: modeling facets and opinions in weblogs. *Proceedings of the 16th international conference on world wide web, WWW '07*. New York, NY, USA: ACM171–180. <https://doi.org/10.1145/1242572.1242596>.
- Mei, Q., Ling, X., Wondra, M., Su, H., & Zhai, C. (Ling, Wondra, Su, Zhai, 2007b). Topic sentiment mixture: modeling facets and opinions in weblogs. *Proceedings of the 16th international conference on world wide web, WWW 2007, Banff, Alberta, Canada, May 8–12, 2007*171–180. <https://doi.org/10.1145/1242572.1242596>.
- Missen, M. M. S., Boughanem, M., & Cabanac, G. (2010). Using passage-based language model for opinion detection in blogs (poster). *Annual ACM symposium on applied computing (SAC), Sierre, Switzerland, 22/03/2010-26/03/2010*. ACM1821–1822 <http://www.acm.org/>.
- Mullen, T., & Collier, N. (2004). Sentiment analysis using support vector machines with diverse information sources. *Proceedings of the conference on empirical methods in natural language processing (EMNLP)412–418* Poster paper.
- Ounis, I., Macdonald, C., & Soboroff, I. (2009). Overview of the TREC-2008 blog track. In *proceedings of TREC-2008*.
- Pak, A., & Paroubek, P. (2010). Twitter as a corpus for sentiment analysis and opinion mining. In N. C. C. Chair), K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, & D. Tapias (Eds.). *Proceedings of the seventh international conference on language resources and evaluation (LREC'10)*. Valletta, Malta: European Language Resources Association (ELRA).
- Pang, B., & Lee, L. (2004). A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts. *Proceedings of the 42nd annual meeting on association for computational linguisticsACL 04*Stroudsburg, PA, USA: Association for Computational Linguistics<https://doi.org/10.3115/1218955.1218990>.
- Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up?: Sentiment classification using machine learning techniques. *Proceedings of the ACL-02 conference on empirical methods in natural language processing – Vol. 10EMNLP '02*Stroudsburg, PA, USA: Association for Computational Linguistics79–86. <https://doi.org/10.3115/1118693.1118704>.
- Phan, X.-H. (2006). CRFTagger: CRF English POS Tagger. <http://crftagger.sourceforge.net/>.
- Riloff, E., & Wiebe, J. (2003). Learning extraction patterns for subjective expressions. *Proceedings of the 2003 conference on empirical methods in natural language processingEMNLP '03*Stroudsburg, PA, USA: Association for Computational Linguistics105–112. <https://doi.org/10.3115/1119355.1119369>.
- Saif, H., He, Y., Fernandez, M., & Alani, H. (2014). Semantic patterns for sentiment analysis of twitter. In P. Mika, T. Tudorache, A. Bernstein, C. Welty, C. Knoblock, D. Vrandečić, P. Groth, N. Noy, K. Janowicz, & C. Goble (Eds.). *The semantic web – ISWC 2014: 13th international semantic web conference, Riva del Garda, Italy, October 19–23, 2014. Proceedings, Part II* (pp. 324–340). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-11915-1_21.
- Saif, H., He, Y., Fernández, M., & Alani, H. (2016). Contextual semantics for sentiment analysis of twitter. *Information Processing and Management*, 52(1), 5–19. <https://doi.org/10.1016/j.ipm.2015.01.005>.
- Seki, K., Kino, Y., Sato, S., & Uehara, K. (2007). *TREC 2007 Blog track experiments at Kobe University. Proceedings of the sixteenth text retrieval conference (TREC)*.
- Seki, K., & Uehara, K. (2009). Adaptive Subjective Triggers for Opinionated Document Retrieval. *Proceedings of the second ACM international conference on web search and data miningWSDM '09*New York, NY, USA: ACM25–33. <https://doi.org/10.1145/1498759.1498805>.
- Song, F., & Croft, W. B. (1999). A general language model for information retrieval. *Proceedings of the eighth international conference on information and knowledge managementCIKM '99*New York, NY, USA: ACM316–321. <https://doi.org/10.1145/319950.320022>.
- Song, R., Qin, T., Shi, D., Lin, H., & Yang, Z. (2007). Dutir at trec 2007 blog track. In E. M. Voorhees, & L. P. Buckland (Eds.). *Proceedings of the sixteenth text retrieval conference, TREC 2007, Gaithersburg, Maryland, USA, November 5–9, 2007*National Institute of Standards and Technology (NIST), vol. Special Publication 500-274.<http://trec.nist.gov/pubs/trec16/papers/dalianu.blog.final.pdf>.
- Stone, P. J., Dunphy, D. C., Smith, M. S., & Ogilvie, D. M. (1966). *The general inquirer: A computer approach to content analysis*. Cambridge, MA: MIT Press.
- Tang, H., Tan, S., & Cheng, X. (2009). A survey on sentiment detection of reviews. *Expert Systems With Applications*, 36(7), 10760–10773. <https://doi.org/10.1016/j.eswa.2009.02.063>.
- Thelwall, M., Buckley, K., & Paltoglou, G. (2012). Sentiment strength detection for the social web. *Journal of American Society for Information Science and Technology*, 63(1), 163–173. <https://doi.org/10.1002/asi.21662>.
- Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., & Kappas, A. (2010). Sentiment in short strength detection informal text. *Journal of American Society for Information Science and Technology*, 61(12), 2544–2558. <https://doi.org/10.1002/asi.v61.12>.

- Tu, W., Cheung, D. W., Mamoulis, N., Yang, M., & Lu, Z. (2016). *Investment recommendation using investor opinions in social media. Proceedings of the 39th international ACM SIGIR conference on research and development in information retrieval, SIGIR 2016, Pisa, Italy, July 17–21, 2016*881–884. <https://doi.org/10.1145/2911451.2914699>.
- Voorhees, E. M. (2006). *Overview of the TREC 2006. Proceedings of the fifteenth text retrieval conference, TREC 2006, Gaithersburg, Maryland, USA, November 14–17, 2006*.
- Wang, S., Chen, Z., & Liu, B. (2016). *Mining aspect-specific opinion using a holistic lifelong topic model. Proceedings of the 25th international conference on world wide web WWW '16*Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee167–176. <https://doi.org/10.1145/2872427.2883086>.
- Wang, J., Sun Y., Mukhtar O., Rohini K.S... TREC 2008 at the University at Buffalo: Legal and Blog Track. Proceedings of the seventeenth text retrieval conference, (TREC). National Institute of Standards and Technology (NIST). 2008.
- Wiebe, J. M., Bruce, R. F., & O'Hara, T. P. (1999). Development and use of a gold-standard data set for subjectivity classifications.
- Yang, H., Callan, J., & Si, L. (2006). Knowledge transfer and opinion detection in the trec 2006 blog track. In E. M. Voorhees, & L. P. Buckland (Vol. Eds.), *TREC. Special Publication 500-272*National Institute of Standards and Technology (NIST).
- Zhai, C., & Lafferty, J. (2001). *A study of smoothing methods for language models applied to ad hoc information retrieval. Proceedings of the 24th annual international ACM SIGIR conference on research and development in information retrieval SIGIR '01*New York, NY, USA: ACM334–342. <https://doi.org/10.1145/383952.384019>.
- Zhai, C. X., & Lafferty, J. (2002). *Two-stage language models for information retrieval. Proceedings of the 25th annual international ACM SIGIR conference on research and development in information retrieval SIGIR '02*New York, NY, USA: ACM49–56. <https://doi.org/10.1145/564376.564387>.
- Zhang, E., & Zhang, Y. (2006). *UCSC on REC 2006 Blog Opinion Mining. Proceedings of the fifteenth text retrieval conference (TREC)*. National Institute of Standards and Technology (NIST) vol. Special Publication 500-272.
- Zhang, M., & Ye, X. (2008). *A generation model to unify topic relevance and lexicon-based sentiment for opinion retrieval. Proceedings of the 31st annual international ACM SIGIR conference on research and development in information retrieval SIGIR '08*New York, NY, USA: ACM411–418. <https://doi.org/10.1145/1390334.1390405>.