Sentiment Analysis and Sentence Classification in Long Book-Search Queries

Amal Htait, Sébastien Fournier, and Patrice Bellot

Aix Marseille Univ, Université de Toulon, CNRS, LIS, Marseille, France.
{firstname.lastname}@univ-amu.fr

Abstract. Handling long queries can involve either reducing its size by eliminating unhelpful sentences, or decomposing the long query into several short queries based on their content. A proper sentence classification improves the functionality of these procedures. Can Sentiment Analysis have an effective role in sentence classification? This paper analyses the correlation between sentiment analysis and sentence classification in long book-search queries. Also, it studies the similarity in writing style between book reviews and sentences in book-search queries. To accomplish this study, a semi-supervised method for sentiment intensity prediction, and a language model based on book reviews are presented and used. In addition to graphical illustrations reflecting the feedback of this study, followed by interpretations and conclusions.

Keywords: sentiment intensity · language model · search queries · books · word embedding · seed-words · book reviews · sentence classification.

1 Introduction

Social cataloging web applications store and share book catalogs and various types of book metadata, while allowing users to search for books or seek recommendations. Its recommendation and search queries are usually destined to humans [1], what makes them often long, descriptive, and even narrative. Users may express their needs for a book, preferences in a type or genre of books, opinions toward certain books, describe content or event in a book, and even sometimes share personal information (e.g. *I am a teacher*).

Being able to differentiate types of sentences, in previously described long queries, can improve in different ways the automation of such book-search tasks. Detecting unhelpful to search sentences in the query (e.g. *Thanks for any and all help.*), can help in query reduction. And classifying sentences by the type of information within, can be used for adapted search. For example, sentences including good read experience, with a book title, can be oriented to a book similarity search, but sentences including a certain topic preference should be focusing on a topic search. And also, sentences including personal information can be used for personalised search.

[1] An example of a query in LibraryThing: https://www.librarything.com/topic/4920

In this work, sentence classification is studied on two levels: the helpfulness of the sentence by containing meaningful information for the search, and the type of information provided by the sentence. And three types of information are highlighted on: book titles and author names (e.g. *I read "Peter the Great His Life and World" by Robert K. Massie.*), personal information (e.g. *I live in a very conservative area*), and narration of book content or story (e.g *The story opens on Elyse overseeing the wedding preparation of her female cousin.*).

The default in text classification is using terms as features, and likewise for sentence classification. In this work, the possibility of introducing new features is tested. Since "Different types of sentences express sentiment in very different ways"[4], the correlation between sentiment in a sentence and its type is studied to test the possibility of introducing sentiment as a feature. For this task, sentiment intensity is calculated, for its capacity to distinguish between sentences of same polarity, using a semi-supervised method explained in Section 4.

In addition, sentences in a query can share similar writing style and subjects with book reviews. Below is a part of a long book-search query:
I just got engaged about a week and a half ago and I'm looking for recommendations on books about marriage. I've already read a couple of books on marriage that were interesting. ***Marriage A History talks about how marriage went from being all about property and obedience to being about love and how the divorce rate reflects this. The Other Woman: Twenty-one Wives, Lovers, and Others Talk Openly About Sex, Deception, Love, and Betrayal not the most positive book to read but definitely interesting. Dupont Circle A Novel I came across at Kramerbooks in DC and picked it up. The book focuses on three different couples including one gay couple and the laws issues regarding gay marriage ...***

In the query example, the part in bold represent a description of specific books content with books titles, e.g. "*Marriage A History*", and interpretations or personal point of view about the book with expressions like "*not the most positive book ... but definitely interesting*". These sentences seem as book reviews sentences. Therefore, calculating the similarity between sentences in a query and books reviews can be a possible feature for sentence classification, as it can help classifying sentences with book titles. To calculate that similarity in a general form, a reviews' statistical language model is used to find, for each sentence in the query, the probability of being generated from that model (and therefore its similarity to that model's training dataset of reviews).

This work covers an analysis of sentence's type correlation with its sentiment intensity and its similarity to reviews, and it is presented as below :

- Presenting the book-search queries used for this work.
- Extracting sentiment intensity of each sentence in the queries.
- Creating a statistical language model based on reviews, and calculating the probability for each sentence to be generated from the model.
- Presenting in graphs and analyzing the relation between language model scores, sentiment intensity scores and the type of sentences.

2 Related Work

For the purpose of query classification, many machine learning techniques have been applied, including supervised [9], unsupervised [6] and semi-supervised learning [2]. In book-search field, fewer studies covered query classification. Ollagnier et al. [11] worked on a supervised machine learning method (Support Vector Machine) for classifying queries into the following classes: **oriented** (a search on a certain subject with orienting terms), **non-oriented** (a search on a theme in general), **specific** (a search for a specific book with an unknown title), and **non-comparable** (when the search do not belong to any of the previous classes). Their work was based on 300 annotated query from INEX SBS 2014[2]. But the mentioned work, and many more, processed the query classification and not the classification of the sentences within the query. The length of book-search queries created new obstacles to defeat, and the most difficult obstacle is the variety of information in its long content, which require a classification at the sentence level.

Sentences in general, based on their type, reveal sentiment in different ways, therefore, Chen et al. [4] focused on using classified sentences to improve sentiment analysis with deep machine learning. In this work, the possibility of a reverse perspective is studied, which is the improvement of sentence classification using sentiment analysis.

In addition, this work is studying the improvement of sentence classification using language model technique. Language models (LM) have been successfully applied to text classification. In [1], models were created using training annotated datasets and then used to compute the likelihood of generating the test sentences. In this work, a new model is created based on book reviews and used to compute the likelihood of generating query sentences, as a similarity measurement between book reviews style and book-search query sentences type.

3 Book-search queries

The dataset of book-search queries, used in this work, is provided by CLEF - Social Book Search Lab - Suggestion Track [3]. The track provides realistic search queries, destined for humans and collected from LibraryThing[4].

Out of 680 user queries, from the 2014's dataset of Social Book Search Lab, 43 queries are randomly selected based on their length, since this work focus on long queries. These 43 queries have more than 55 words, stop-words excluded. Then, each query is segmented into sentences, which results a total of 528 sentences. These sentences are annotated, for this study, based on their helpfulness to the search, and on the information they provide as: book titles and authors names, personal information, and narration of book content. An example is shown in the below XML extraction at Figure 1.

[2] https://inex.mmci.uni-saarland.de/data/documentcollection.html

[3] http://social-book-search.humanities.uva.nl/#/suggestion

[4] https://www.librarything.com/

```
<sentences>
  <sentence helpful=”False” info=”Null” >
    Where is the time to go online and talk?
  <\sentence>
  <sentence helpful=”True” info=”General”>
    No sappy romance involved
  <\sentence>
  <sentence helpful=”True” info=”Personal_Info”>
    I am a sixth grade science teacher
  <\sentence>
  <sentence helpful=”True” info=”Book_Content”>
    Pierre becomes for a while a Mason
  <\sentence>
  <sentence helpful=”True” info=”Book_Title_Author”>
    I have one by Robert Fitzgerald Peter
  <\sentence>
<\sentences>
```

Fig. 1: An example of annotated sentences in book-search queries.

4 Sentiment Intensity

As part of this work, sentiment intensity is calculated for each sentence of the query. Sentiment intensity is chosen for its capability of capturing more accurately the sentiment in text, as for its capacity to distinguish between sentences of same polarity. The following method is inspired by a semi-supervised method for sentiment intensity prediction in tweets, and it was established on the concepts of adapted seed-words and word embedding [8]. To note that the seed-words are words with strong semantic orientation, chosen for their lack of sensitivity to the context. They are used as paradigms of positive and negative semantic orientation. And adapted seed-words are seed-words with the characteristic of being used in a certain context or subject. Also, word embedding is a method to represent words in high quality learning vectors, from large amounts of unstructured and unlabeled text data, to predict neighboring words.

In the work of Htait et al. [8], the extracted seed-words were adapted to micro-blogs. For example, the word *cool* is an adjective that refers to a moderately low temperature and has no strong sentiment orientation, but it is often used in micro-blogs as an expression of admiration or approval. Therefore, *cool* is considered a positive seed-word in micro-blogs. In this paper, book-search is the targeted domain for sentiment intensity prediction, therefore, the extracted seed-words are adapted to books domain, and more specifically, extracted from book reviews since the reviews has the richest vocabulary in the book domain.

Using annotated book reviews, as positive and negative, by Blitzer et al.[5] [3], the list of most common words in every annotation class is collected. Then,

[5] Book reviews from Multi-Domain Sentiment Dataset by http://www.cs.jhu.edu/ mdredze/datasets/sentiment/index2.html

after removing the stop-words, the first 43 most relevant to book domain words, with strong sentiment, are selected manually from each previously described list, as positive and negative seed-words. The following is an example of the extracted seed-words, adapted to book domain, as positive: $insightful$, *touching*, *masterpiece*, and as negative: *endless*, *waste*, *unnecessary*.

Word embedding, or distributed representations of words in a vector space, are capable of capturing lexical, semantic, syntactic, and contextual similarity between words. And to determine the similarity between two words, the measure of cosine distance is used between the vectors of these two words in the word embedding model. In this paper, a word embedding model is created based on more than 22 million Amazon's book reviews [7], as training dataset, after applying a pre-processing to the corpora (e.g. tokenization, replacing hyperlinks and emoticons, removing some characters and punctuation).

For the purpose of learning word embedding from the previously prepared corpora (which is raw text), Word2Vec [10] is used with the training strategy Skip-Gram (in which the model is given a word and it attempts to predict its neighboring words). To train word embedding and create the models, Gensim[6] framework for Python is used. And for the parameters, the models are trained with word representations of dimension 400, a context window of one and negative sampling for five iterations (k = 5) [8]. As a result, a model is created with a vocabulary size of more than 2.5 million words.

Then, and for each word in the sentence, the difference between average cosine similarity with positive seed-words and negative seed-words represent its sentiment intensity score, using the previously created model. For example, the word $confusing$ has an average cosine similarity with positive seed-words equals to 0.2073 and an average cosine similarity with negative seed-words equals to 0.3082, what makes its sentiment intensity score equals to -0.1008 (a negative score represent a negative feeling). And for the word *young* the sentiment intensity score equals to 0.0729, which is rather neutral, but closer to positive than to negative sentiment.

To predict the sentiment intensity of the entire sentence, first the adjectives, nouns and verbs are selected from the sentence using Stanford POS tagger [13], then the ones with high sentiment intensity are used by adding up their score to have a total score for the sentence. Note that the created tool Adapted Sentiment Intensity Detector (ASID), used to calculate the sentiment intensity of words, is shared by this work's researchers as open source [7].

5 Reviews' language model

The book reviews are considered a reference in sentence's characteristic detection, since a similarity in style is noticed between certain sentences of user queries and the reviews. To calculate this similarity in writing style, a statistical language

[6] https://radimrehurek.com/gensim/index.html

[7] https://github.com/amalhtait/ASID

modeling approach is used to compute the likelihood of generating a sentence of a query from a book reviews language model. Such method is unsupervised and does not require an annotated dataset.

The statistical language modeling were originally introduced by Collins in [5], and it is the science of building models that estimate the prior probabilities of various linguistic units [12]. It makes it possible to easily consider taking into account large linguistic units, like bigrams and trigrams. The model can be presented as $\theta_R = P(w_i|R)$ with $i \in [1, |V|]$, where $P(w_i|R)$ is the probability of word w_i in the reviews corpora R, and $|V|$ is the size of the vocabulary. And this model is used to denote the probability of a word according to the distribution as $P(w_i|\theta_R)$ [14].

The probability of a sentence W to be generated from a book reviews language model θ_R is defined as the following conditional probability $P(W|\theta_D)$ [14], which is calculated as following:

$$P(W|\theta_D) = \prod_{i=1}^{m} P(w_i|\theta_R) \tag{1}$$

where W is a sentence, w_i is a word in the sentence W, and θ_R represents the book reviews model.

The tool SRILM[8] [12] is used to create the model from book reviews dataset (as training data), and to compute the probability of sentences in queries to be generated from the model (as test data). The language model is created as a standard language model of trigram and Good-Turing discounting (or Katz) for smoothing, based on 22 million of Amazon's book reviews [7], as training dataset.

The tool SRILM offers details in the diagnostic output like the number of words in the sentence, the sentence likelihood to model or the logarithm of likelihood by $logP(W|\theta_R)$, and the perplexity which is the inverse probability of the sentence normalized by the number of words, as shown in Equation 2. In this paper, the length of sentences vary from one word to almost 100 words, therefore the score of perplexity seems more reliable for a comparison between sentences. To note that minimizing perplexity is the same as maximizing probability of likelihood, and a low perplexity indicates the probability distribution is good at predicting the sample.

$$perplexity = \sqrt[m]{\frac{1}{P(W|\theta_R)}} \tag{2}$$

with m as the number of words.

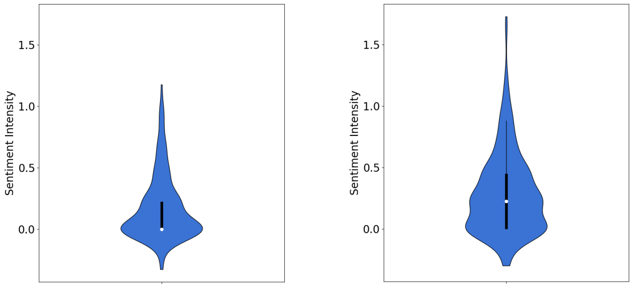
[8] http://www.speech.sri.com/projects/srilm/

6 Scores representation in graphs

As previously explained in Section 3, a corpora of 528 sentences from user queries is created and annotated as the examples in Figure 1. Then, for each sentence the sentiment intensity score and the perplexity score are calculated following the methods previously explained in Sections 4 and 5. To present the scores, Violin plots are used for their ability to show the probability density of the data at different values. Also, they include a marker (white dot) for the median of the data and a box (black rectangle) indicating the interquartile range.

6.1 Sentiment intensity, perplexity and helpfulness correlation

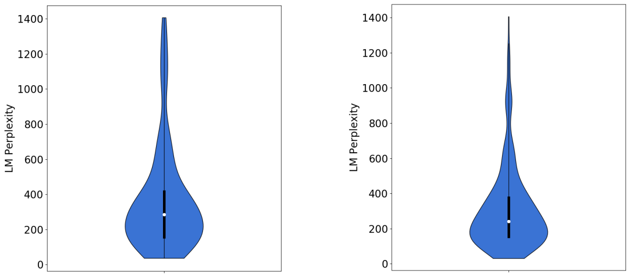
The graph in Figure 2 shows the distribution (or probability density) of **sentiment intensity** between two categories of sentences: on the right the sentences which are helpful to the search and on the left the sentences which are unhelpful to the search (noise). The shape on the left is horizontally stretched compared to the right one, and mostly dilated over the area of neutral sentiment intensity (sentiment score = 0), where also exist the median of the data. On the other hand, the shape on the right is vertically stretched, showing the diversity in sentiment intensity in the helpful to search sentences, but concentrated mostly in the positive area, at sentiment score higher than zero but lower than 0.5.



(a) Unhelpful (b) Helpful

Fig. 2: The distribution of sentiment intensity between two categories of sentences: on the right the sentences which are helpful to the search and on the left the sentences which are unhelpful to the search.

The graph in Figure 3 represent the distribution of **perplexity** between two categories of sentences: on the right the sentences which are helpful to the search and on the left the sentences which are unhelpful to the search (noise). Both shapes are vertically compressed and dilated over the area of low perplexity. But the graph on the right, of the helpful sentences, shows the median of the data on a lower level of score of perplexity, than the left graph. Explained by the slightly horisontal dilation of the left graph above the median level.



(a) Unhelpful

(b) Helpful

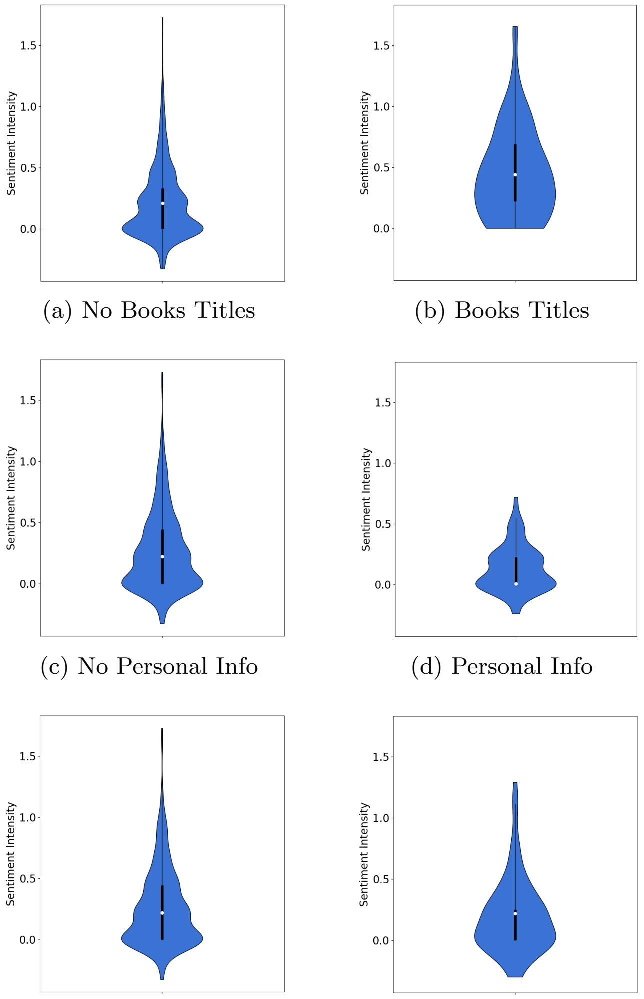
Fig. 3: The distribution of perplexity between two categories of sentences: on the right the sentences which are helpful to the search and on the left the sentences which are unhelpful to the search.

6.2 Sentiment intensity, perplexity and information type correlation

The graphs in Figure 4 shows the distribution of sentiment on sentences based on their type of information. The graphs are described below consecutively, from top to bottom, by information type:

- Book titles and authors names: on the right the sentences with books titles or authors names, and on the left the sentences without books titles and authors names. The graph on the right shows a high distribution of positive sentiment, but the left graph shows a high concentration on neutral sentiment with a small distribution for positive and negative sentiment. Also, It is noticed the lack of negative sentiment in sentences with books titles or authors names.
- Personal information: on the right the sentences containing personal information about the user, and on the left the sentences without personal information. The graph on the right shows a high concentration on neutral sentiment, where also exist the median of the data, and then a smaller distribution in positive sentiment. On the left, the graph shows a lower concentration on neutral sentiment, but it is noticeable the existence of sentences with extremely high positivism.
- Narration of book content: on the right the sentences containing book content or events, and on the left the sentences without book content. Both graphs are vertically stretched but have different shapes. The graph on the right shows a higher distribution of negative sentiment as for sentences with book content, and the graph on the left shows higher positive values.

The graphs in Figure 5 shows the distribution of perplexity between the informational sentences, consecutively from top to bottom: Book titles and authors names, Personal information and Narration of book content. When comparing the first set of graphs, of book titles and authors names, the left graph has its median of data on a lower perplexity level than the right graph, with a higher



(a) No Books Titles

(b) Books Titles

(c) No Personal Info

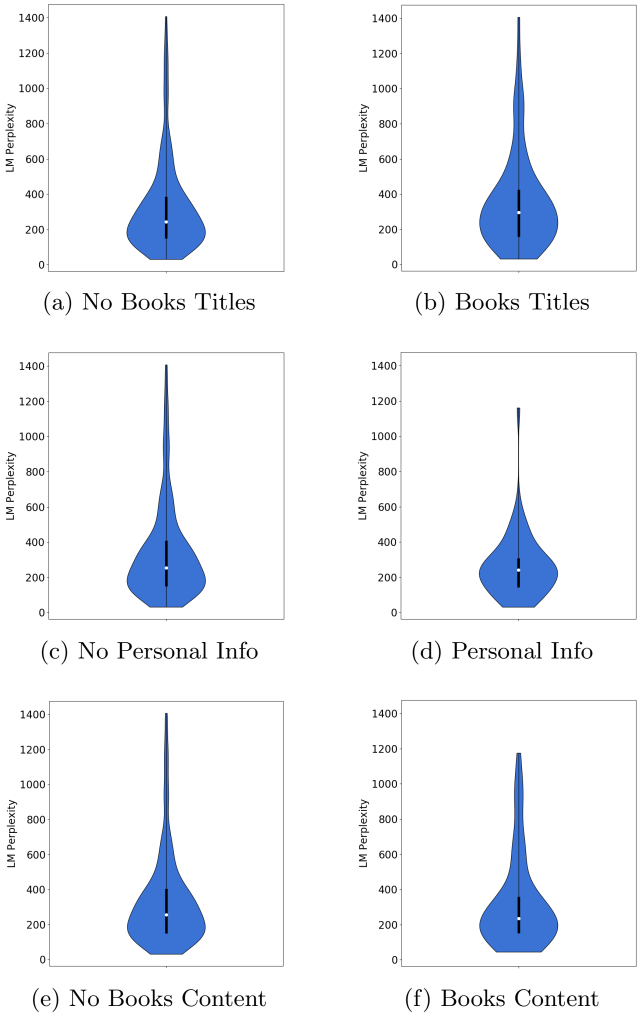
(d) Personal Info

(e) No Books Content

(f) Books Content

Fig. 4: The distribution of Sentiment between the informational categories of sentences: Books titles or authors names, Personal information and Narration of book content.

concentration of data in a tighter interval of perplexity. For the second sets of graphs, of personal information, the right graph shows a lower interquartile range than the left graph. As for the third set of graphs, of book content, a slight difference can be detected between the two graphs, where the left graph is more stretched vertically.



(a) No Books Titles

(b) Books Titles

(c) No Personal Info

(d) Personal Info

(e) No Books Content

(f) Books Content

Fig. 5: The distribution of perplexity between the informational categories of sentences: Books titles or authors names, Personal information and Narration of book content.

6.3 Graphs interpretation

Observing the distribution of data in the graphs of the previous sections, many conclusions can be extracted:

- In Figure 2, it is clear that unhelpful sentences tend to have high level of emotions (positive or negative), but unhelpful sentences (noise) are more probable to be neutral.

- The Figure 3 shows that sentences with high perplexity, which means they are not similar to book reviews sentences, have a higher probability of being unhelpful sentence than helpful.
- The Figure 4 gives an idea of sentiment correlation with sentences information: sentences with book titles or author names have a high level of positive emotions, but sentences with personal information tend to be neutral. And sentences with book content narration are distributed over the area of emotional moderate level, with a higher probability of positive than negative.
- The Figure 5 gives an idea of the correlation between the sentences information and their similarity to reviews: sentences with no book titles are more similar to reviews than the ones with book titles. Also, sentences with personal information tend to be similar to reviews. And sentences with book content narration show a slight more similarity with reviews sentences style than the sentences with no book content narration.

7 Conclusion and Future work

This paper analyses the relation between sentiment intensity and reviews similarity toward sentences types in long book-search queries. First, by presenting the user queries and books collections, then extracting the sentiment intensity of each sentence of the queries (using Adapted Sentiment Intensity Detector (ASID)). Followed by creating a statistical language model based on reviews, and calculating the probability of each sentence being generated from that model. And finally by presenting, in graphs, the relation between sentiment intensity score, language model score, and the type of sentences.

The graphs show that sentiment intensity can be an important feature to classify the sentences based on their helpfulness to the search. Since unhelpful sentences (or noise sentences) are more probable to be neutral in sentiment, than helpful sentences. Also, the graphs show that sentiment intensity can also be an important feature to classify the sentences based on the information within. It is clear in the graphs, that the sentences containing book titles are richer in sentiment and mostly positive compared to sentences not containing book titles. In addition, the graphs show that sentences with personal information tend to be neutral, in a higher probability than those with no personal information.

On the other hand, the graphs show that the similarity of sentences to reviews style can also be a feature to classify sentences by helpfulness and by their information content, but in a slightly lower level of importance than sentiment analysis. Similarity between sentences and book reviews style is higher for helpful sentences, for sentences with personal information and for sentences with narration of book content, but not for sentences containing book titles.

The previous analysis and conclusions gives a preview on the role that sentiment analysis and similarity to reviews can play in sentence classification of long book-search queries. The next task would be to test these conclusions by using sentiment analysis and similarity to reviews, as new features, in a supervised machine learning classification of sentences in long book-search queries.

Acknowledgments This work has been supported by the French State, managed by the National Research Agency under the "Investissements d'avenir" program under the EquipEx DILOH project (ANR-11-EQPX-0013).

References

1. Bai, J., Nie, J.Y., Paradis, F.: Using language models for text classification. In: Proceedings of the Asia Information Retrieval Symposium, Beijing, China (2004)
2. Beitzel, S.M., Jensen, E.C., Frieder, O., Lewis, D.D., Chowdhury, A., Kolcz, A.: Improving automatic query classification via semi-supervised learning. In: Data Mining, Fifth IEEE international Conference on. pp. 8–pp. IEEE (2005)
3. Blitzer, J., Dredze, M., Pereira, F.: Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In: Proceedings of the 45th annual meeting of the association of computational linguistics (2007)
4. Chen, T., Xu, R., He, Y., Wang, X.: Improving sentiment analysis via sentence type classification using bilstm-crf and cnn. Expert Systems with Applications pp. 221 – 230 (2017)
5. Collins, M.: Three generative, lexicalised models for statistical parsing. In: Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics. pp. 16–23. Association for Computational Linguistics (1997)
6. Diemert, E., Vandelle, G.: Unsupervised query categorization using automatically-built concept graphs. In: Proceedings of the 18th international conference on World wide web. pp. 461–470. ACM (2009)
7. He, R., McAuley, J.: Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In: proceedings of the 25th international conference on world wide web. pp. 507–517. International World Wide Web Conferences Steering Committee (2016)
8. Htait, A., Fournier, S., Bellot, P.: Lsis at semeval-2017 task 4: Using adapted sentiment similarity seed words for english and arabic tweet polarity classification. In: Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017). pp. 718–722 (2017)
9. Kang, I.H., Kim, G.: Query type classification for web document retrieval. In: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval. pp. 64–71. ACM (2003)
10. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
11. Ollagnier, A., Fournier, S., Bellot, P.: Analyse en dépendance et classification de requêtes en langue naturelle, application à la recommandation de livres. Traitement Automatique des Langues (2015)
12. Stolcke, A.: Srilm-an extensible language modeling toolkit. In: Seventh international conference on spoken language processing (2002)
13. Toutanova, K., Klein, D., Manning, C.D., Singer, Y.: Feature-rich part-of-speech tagging with a cyclic dependency network. In: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1. pp. 173–180. Association for Computational Linguistics (2003)
14. Zhai, C.: Statistical language models for information retrieval. Synthesis Lectures on Human Language Technologies **1**(1), 1–141 (2008)