BEATnIk: an algorithm for the automatic generation of educational description of movies

Vinicius Woloszyn[1], Guilherme M. Machado[1], José Palazzo[1], Horacio Saggion[2], Leandro Krug Wives[1]

[1]Instituto de Informática – Universidade Federal do Rio Grande do Sul
Caixa Postal 15.064 – 91.501-970 – Porto Alegre – RS – Brazil

[2]Universitat Pompeu Fabra
Barcelona, Spain

{vwoloszyn,gmedeiros,palazzo,wives}@inf.ufrgs.br,

{first.last}@upf.edu

Abstract. *Teachers have increasingly employed different methods to enrich the learning of a subject in class, drive other assignments, and meet curriculum standards. One of such methods is the use of movies as an alternative educational experience to support class discussions. In this sense, websites such as TeachWithMovies [1], arise as a valuable support to the creation of lesson plans. In this website, each movie is described as a lesson plan targeting the learning of a subject. However, the creation of such lesson plan or even a simple educational description of the movie can demand much work and time, since the text describing the teaching plan must consider educational aspects of the movie. In this work, we propose BEATnIk (**B**iased **E**ducational **A**utomatic **T**ext summar**I**zation), which is an unsupervised algorithm to automatically generate movies' summaries. Such algorithm favors educational aspects from the text to generate a biased educational summary. The experiments conducted show that our approach statistically outperforms a baseline in precision, recall, and f-score.*

1. Introduction

The use of extracurricular learning material is a common practice inside a classroom. Teachers have been increasingly using movies, software and other kinds of learning objects that can support the teaching of the class subject, and some examples of such practices can be found in [Giraffa et al. 2015, Oliveira et al. 2016, Castro et al. 2016]. The use of movies is one of the simplest ways to support teaching because it is easily available and is a time-controlled experience inside the classroom. In this sense, websites such as TeachWithMovies[1], arise as a valuable support to the creation of lesson plans. In this website, a set of movies is described by teachers to be used as learning objects inside a classroom. Each movie description contains at least the movie's benefits and possible problems, a helpful background, a discussion; besides, with some descriptions, there are also questions to be used in class. The preparation of such type of material is a time-consuming activity, and we believe that an educational summary can help in the elaboration of a longer movie-based lesson plan.

Several works address the challenge of extracting specific aspects from users' reviews to compose a summary about a movie or a product. Most of those works rely

[1] http://www.teachwithmovies.org/index.html

DOI: 10.5753/cbie.sbie.2017.1377

on supervised algorithms such as classification and regression [Xiong and Litman 2011, Zeng and Wu 2013, Yang et al. 2015]. However, the quality of results produced by supervised algorithms is dependent on the existence of a large, domain-dependent training dataset. In this sense, semi-supervised and unsupervised methods are an attractive alternative to avoid the labor-intense and error-prone task of manual annotation of training datasets.

Considering such context we describe, in this paper, BEATnIk (Biased Educational Automatic Text Summarization) an unsupervised algorithm to generate biased summaries that cover educational aspects of movies from users' reviews. By utilizing BEATnIk we hope to help teachers in providing educational descriptions for movies. So, the paper's main contributions are: a) the description of a tool to assist teachers in the creation of lesson plans from the movies' reviews; and b) an unsupervised algorithm which outperforms the baseline, imitating the human educational description of the movie. BEATnIk can also be employed in other domains, it would only require small modifications to be able to generate, for instance, a biased summary that covers the personal user's aspect of interest about products on Online Collaborative Reviews Websites. We would also like to highlight that BEATnIk is open source and it is available on Internet [2].

The rest of this paper is organized as follows. Section 2 discusses the related works. Section 3 present the datasets employed on this work. Section 4 present details of BEATnIk algorithm. Section 5 describes the design of our experiments, and Section 6 discusses the achieved results. Section 7 summarizes our conclusions and presents future research directions.

2. Related work

Automatic Text Summarization (ATS) techniques have been successfully employed on user-content to highlight the most relevant information among the documents [Erkan and Radev 2004, Radev et al. 2004, Ganesan et al. 2010, Saggion and Poibeau 2013, Ramos et al. 2017]. Regarding the techniques employed, several works have explored unsupervised methods based on graph centrality. For instance MRR [Woloszyn et al. 2017], that combines the centrality scores and human explicit feedbacks to produce a ranking of relevant documents. Other example is presented by [Wu et al. 2011] that combines the centrality scores of each sentences with the documents' length to produce an overall centrality score for each review. However, this method does not scale well due the chosen centrality granularity, which implies dual use of PageRank, and requires pre-processing to identify specific textual features (e.g. nouns, adjectives).

There are also studies using supervised learning strategies to predict the text relevance [Xiong and Litman 2011, Zeng and Wu 2013, Yang et al. 2015]. Additionally, the use of regression algorithms consistently improves the prediction of helpfulness [Wan 2013]. However, a common drawback of supervised learning approaches is that the quality of results is heavily influenced by the availability of a large, domain-dependent annotated corpus to train the model. Unsupervised learning techniques are attractive because they do not imply the cost of corpus annotation either training. Therefore, it is described in this paper an unsupervised biased algorithm to extract educational aspects from movies' reviews with the goal to assist teachers on the task of creating movie-based lesson plans.

[2] http://xx.yy.zz

3. Datasets Employed

As the goal of our approach was to build a biased summarizer for educational purposes, we used two datasets to perform the experiments. The first served as a word thesaurus to implement the educational bias, and it was collected from an educational website [3] TeachWithMovies (TWM) where a set of movies are described by teachers with the goal to use them as learning objects inside a classroom. The second dataset is Amazon Movie Reviews (AMR) [McAuley and Leskovec 2013] which provides user comments about a large set of movies. Since we were interested in movies that appeared in both datasets, a filter was applied, and we ended up with 256 movies to perform our evaluation. Next, we describe with more details each dataset.

3.1. Teaching with Movies

The TeachWithMovies dataset was collected through a crawler developed by us. Different teachers described the movies on the website, but each movie has only one description, this was a challenge while collecting the data because the information was not standardized or had associated metadata.

However, we have noticed that some movies presented common information: i) movie description; ii) rationale for using the movie; iii) movie benefits for teaching a subject; iv) movie problems and warnings for young watchers; and v) objectives of using this movie in class. The developed crawler extracted such information, and we have used the movie description since it contains the greatest amount of educational aspects. In the end, 408 unique movies and video clips were extracted, but after matching with the Amazon dataset, we could use 256 movies.

3.2. Amazon Movie Reviews

The Amazon Movie Reviews was collected with a timespan of more than ten years and consists of proximately 8 millions of reviews that include product and user information, ratings, and a plain text review.In Table 1 is shown some statistics about the data.

Table 1. Amazon Movie Reviews Statistics

Dataset Statistics	
Number of reviews	7,911,684
Number of users	889,176
Expert users (with >50 reviews)	16,341
Number of movies	253,059
Mean number of words per review	101
Timespan	Aug 1997 - Oct 2012

4. BEATnIk Algorithm

In BEATnIk, a complete graph is constructed for each movie. In this graph, each sentence extracted from the Amazon's dataset becomes a node, and each edge's weight is defined by a similarity measure applied between sentences. An adapted cosine equation assesses the similarity. The algorithm then employs PageRank [Page et al. 1999] to compute the centrality of each node. The intuition behind this approach is that central sentences highlight aspects frequently mentioned in a text. Also, BEATnIk takes into account keywords

[3] http://www.teachwithmovies.org/index.html

extracted from the lesson plans of TWM (used as a bias) to compute the importance of each sentence. The final educational summary is based on the centrality score of the sentences weighted by the presence of educational keywords.

Let S be a set of all sentences extracted from the R user's reviews about a single movie, BEATnIk builds a graph representation $G = (V, E)$, where $V = S$ and E is a set of edges that connect pairs $\langle u, v \rangle \in V$. The score of each node (that represent a sentence) is given by the harmonic mean between its centrality score on the graph given by PageRank, and the sum of the frequencies of its education keywords (stated in equation 2). The pseudo-code of BEATnIk is displayed in Algorithm 1, where G is represented as the adjacency matrix W.

Algorithm 1 - BEATnIk Algorithm (S,B): O

- Input: a set of sentences extracted from the Amazon's reviews R, and a corpora B used as bias and
- Output: a extractive biased summary O based on reviews R.

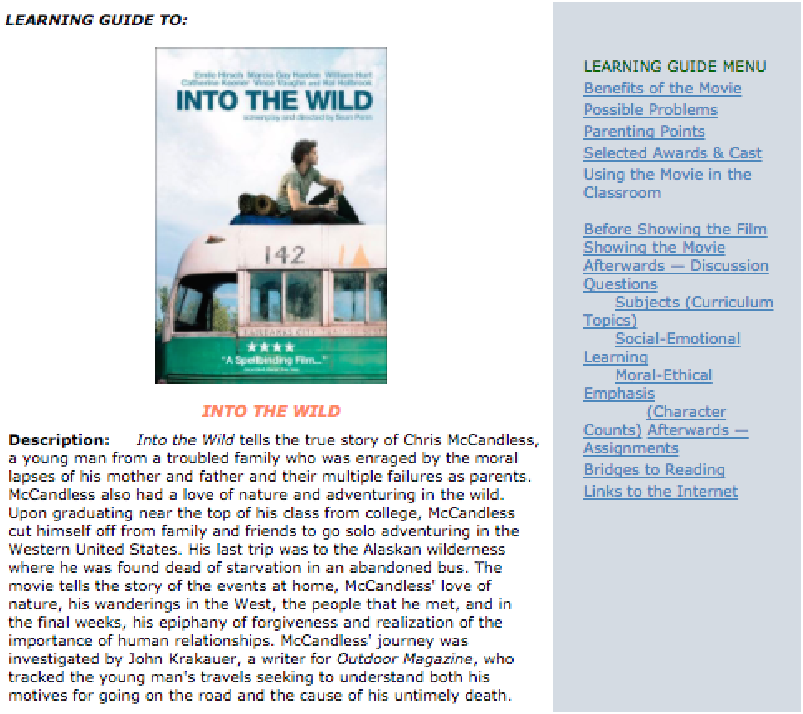
```
 1: for each $u, v \in S$ do
 2:     $W[u, v] \leftarrow$ idf-modified-cosine(u,v)
 3: end for
 4: for each $u, v \in S$ do
 5:     if $W[u, v] \geq \beta$ then
 6:         $W'[u, v] \leftarrow 1$
 7:     else
 8:         $W'[u, v] \leftarrow 0$
 9:     end if
10: end for
11: $P \leftarrow PageRank(W')$
12: for each $u \in S$ do
13:     $K \leftarrow$ sim-keyword(u, $B$)
14:     $O[u] \leftarrow \frac{\|S\|P_uK}{P_u+K}$
15: end for
16: Return $O$
```

The main steps of the BEATnIk algorithm are: (a) it builds a similarity graph (W) between pairs of reviews of the same product (lines: 1-3); (b) the graph is pruned (W') by removing all edges that do not meet a minimum similarity threshold, given by the parameter β[4] (lines 4-10); (c) using PageRank, the centrality scores of each node is calculated (line 11); (d) using the educational corpora, each sentence is scored according the presence of educational keywords (line 13); (e) The final importance score of each node is given by the harmonic mean between its centrality score on the graph, and the sum of its education keywords frequencies (line 14).

To get the similarity between the two nodes we define an adapted metric, that is the cosine difference between two corresponding sentence vectors [Erkan and Radev 2004]:

$$\text{idf-modified-cosine}(x, y) = \frac{\sum_{w \in x,y} \text{tf}_{w,x}\text{tf}_{w,y}(\text{idf}_w)^2}{\sqrt{\sum_{x_i \in x}(\text{tf}_{x_i,x}\text{idf}_{x_i})^2} \times \sqrt{\sum_{y_i \in y}(\text{tf}_{y_i,y}\text{idf}_{y_i})^2}} \quad (1)$$

[4] The best parameter obtained in our experiments is $\beta = 0.1$



LEARNING GUIDE TO:

INTO THE WILD

Description: *Into the Wild* tells the true story of Chris McCandless, a young man from a troubled family who was enraged by the moral lapses of his mother and father and their multiple failures as parents. McCandless also had a love of nature and adventuring in the wild. Upon graduating near the top of his class from college, McCandless cut himself off from family and friends to go solo adventuring in the Western United States. His last trip was to the Alaskan wilderness where he was found dead of starvation in an abandoned bus. The movie tells the story of the events at home, McCandless' love of nature, his wanderings in the West, the people that he met, and in the final weeks, his epiphany of forgiveness and realization of the importance of human relationships. McCandless' journey was investigated by John Krakauer, a writer for *Outdoor Magazine*, who tracked the young man's travels seeking to understand both his motives for going on the road and the cause of his untimely death.

LEARNING GUIDE MENU
Benefits of the Movie
Possible Problems
Parenting Points
Selected Awards & Cast
Using the Movie in the Classroom

Before Showing the Film
Showing the Movie
Afterwards — Discussion Questions
Subjects (Curriculum Topics)
Social-Emotional Learning
Moral-Ethical Emphasis
(Character Counts) Afterwards — Assignments
Bridges to Reading
Links to the Internet

Figure 1. A summarized snapshot of “Into the Wild” lesson plan

where $\text{tf}_{w,s}$ is the number of occurrences of the word w in the sentence s. We employed the approach described by [Mihalcea and Tarau 2004] to extract the keywords from the educational corpora. The similarity between the sentences and the keywords extracted from the TWM lesson plans are given by the following equation:

$$\text{sim-keyword}(x, B) = \sum_{w \in x} \text{tf}_{w \in keywords(B)} \tag{2}$$

The comparison of our approach to TextRank [Mihalcea and Tarau 2004], which is also a Graph-based Automatic Text Summarization, revealed that BEATnIk generates summaries closer to the educational description of the movies in TWM (details are presented in the next section).

5. Experiment Design

This section presents the experimental setting used to evaluate BEATnIk. It describes the method employed as the baseline for comparison, the educational plans adopted as Gold-standard and the metric applied for evaluation, as well as details of the experiment, performed to assess BEATnIk.

5.1. The baseline

The results obtained from our proposed approach are compared with Textrank [Mihalcea and Tarau 2004] algorithm. Textrank was chosen because it is also a graph-based ranking algorithm and has been widely employed in Natural Language tools [Řehůřek and Sojka 2010].

Textrank essentially decides the importance of a sentence based on the idea of “voting” or “recommending”. Considering that in this approach each edge represents a vote, the higher the number of votes that are cast for a node, the higher the importance of the node (or sentence) in the graph. The most important sentences compose the final summary.

5.2. Gold-Standard

The lesson plans found on the TWM website were used as a gold-standard to assess BEATnIk summaries. An English-speaking teacher describes each lesson plan and takes into consideration the educational aspects of the movie.

The lessons are categorized by movie genre, learning discipline, recommended age (from 3 years-old to college level), and alphabetical order. Inside the lesson plans, there is also some learning goals regarding the movie, such as the learning subject, the social-emotional learning, and the ethical emphasis.

Taking, for instance, the summary of “Into the Wild” lesson plan presented in Figure 1, we can observe in the “Description” section that the teacher has focused on discussing learning the importance of human relationships. At the top right, it is found the structure of the whole lesson available online [5]. In the remaining of the lesson, the teacher still presents some benefits of the movie, such as risky behavior can have fatal consequences and relationships with people are an essential part of life.

TWM provided a well-described educational dataset, and despite the lack of standardization of lessons plans, we could use it successfully as a gold-standard to perform our experiments.

5.3. Evaluation Metric

The evaluation was performed by applying ROUGE (Recall-Oriented Understudy for Gisting Evaluation) [Lin 2004], which is a metric inspired on the Bilingual Evaluation Understudy (BLEU) [Saggion and Poibeau 2013].

Specifically, we used ROUGE-N in the evaluation, this version of ROUGE makes a comparison of n-grams between the summary to be evaluated and the “gold-standard”; in our case, BEATnIk summaries and TWM lesson plans, respectively. We evaluated the first 100 words of the summaries obtained by our approach and the baseline since it corresponds to the median size of the gold-standard. ROUGE was chosen because it is one of most used measures in the fields of Machine Translation and Automatic Text Summarization [Poibeau et al. 2012].

5.4. BEATnIk’s bias

The set of lesson plans extracted from TMW was used as an educational bias for BEATnIk algorithm. When generating a biased summary for a specific movie, BEATnIk does not take in consideration such movie lesson plan. Instead, it builds a graph using all other movies information, excepting the movie to be summarized. This strategy avoids any positive influence on the performance of the predictive model.

The retrieved corpus was composed of 991 sentences and 2,811 unique tokens. In Table 2 we describe the first 20 keywords extracted from TWM corpus.

[5] http://www.teachwithmovies.org/guides/into-the-wild.html

Table 2. Keywords extracted from the lesson plans in TWM

Keywords	Frequency	Keywords	Frequency
film	0.01390	class	0.00354
movi	0.01062	famili	0.00345
children	0.00475	bulli	0.00345
benefit	0.00457	parent	0.00336
father	0.00440	boy	0.00319
use	0.00414	help	0.00311
stori	0.00406	point	0.00311
discuss	0.00388	live	0.00285
question	0.00362	life	0.00276
child	0.00362	time	0.00276

6. Results

In this section, we present BEATnIk's evaluation regarding the adopted baselines concerning precision, recall, and f-Score obtained by using ROUGE-N.

The gold-standard utilized in the experiments, as already stated in Section 5, is the educational description extracted from the TWM website. Table 3 shows the mean Precision, Recall, and F-Score, considering both BeatnIk and Textrank (the gold-standard used as the baseline).

The results presented in Table 3 show that BEATnIk outperformed the baseline in all measurements carried out. Regarding Precision, the differences range from 4.9 to 11.9 percentage points (pp) on all ROUGE-N analyzed, where N is the size of the n-gram used by ROUGE. Using Wilcoxon statistical test with a significance level of 0.05, we verified that BEATnIk is statistically superior when compared to the baseline. Regarding recall, the differences are also in favor of BEATnIk, ranging from 4.7 to 11.5 pp when compared to the baseline.

Table 3. Mean of ROUGE results achieved by BEATnIk and the Baseline

ROUGE-n	*Baseline*	*BEATnIk*	p-values
Precision-1	0.65615	**0.77028**	< 0.05
Recall-1	0.65003	**0.75611**	< 0.05
F_score-1	0.65283	**0.76296**	< 0.05
Precision-2	0.22394	**0.34350**	< 0.05
Recall-2	0.22192	**0.33744**	< 0.05
F_score-2	0.22284	**0.34037**	< 0.05
Precision-3	0.06313	**0.11268**	< 0.05
Recall-3	0.06387	**0.11102**	< 0.05
F_score-3	0.06347	**0.11182**	< 0.05

Regarding the distribution of Rouge's results, in Fig 2 it is shown a boxplot indicating that BEATnIk results are not only better in mean, but also in terms of lower and upper quartiles, minimum and maximal values.

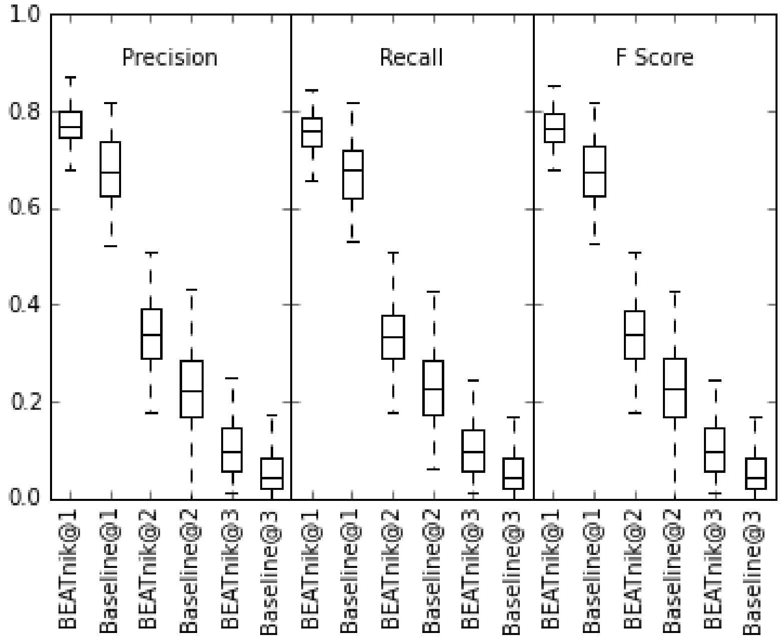


Figure 2. Distribution of Rouge results.

To illustrate the differences between the BEATnIk and a generic text summarizer on the task of extracting the educational aspects from the movie's reviews, consider the snippet of summaries about the movie 'Conrack' at table 4. In this example, while BEATnIk highlights the educational aspects such as *method lesson, teaching, and children*, the generic text summarizer used as baseline highlights the aspects frequent mentioned in the reviews, such as related to the *screenplay* and the *director*.

Table 4. Snippets of the summaries generated by BEATnIk and the Baseline about movie 'Conrack'

BEATnIk	Baseline
As well as being a method lesson in teaching, it is also a good personal film, and even if you don't warm to Jon Voight's character immediately, you will love the little children. [...]	*The director achieved a glimmering one in this hidden gem adapted from author Pat Conroy's novel The Water Is Wide.* [...]

7. Conclusion

In this paper, we presented BEATnIk, an algorithm for the identification of relevant reviews based on the concept of node centrality. The intuition behind BEATnIk is that central sentences extracted from users' movies reviews are closer to a human-made educational description of the movie if such central sentences contain "educational keywords". We proved this assumption and showed that BEATnIk achieved statistically superior results than Textrank (a general summary algorithm).

The main contributions of this paper are: a) the design and presentation of BEATnIk, a tool to assist teachers in the creation of lesson plans based on movies' reviews; and b) the methodology of experiments designed to assess BEATnIk, which outperformed the baseline, imitating the human educational description of the movies.

Finally, it is also important to state that we found out a considerable number of highly helpful sentences with low centrality indexes which lead us to consider the investigation of other techniques to select the most relevant sentences to compose the movies' educational description. In particular, the Segmented Bushy Path which is widely explored in text summarization performs a segmentation of the graph in portions that correspond to the topics of the text. Such segmentation may allow other relevant sentences to be considered in the output summary.

Acknowledgments

We would like to acknowledge CAPES and CNPq, Brazil; for partially financing this work.

References

Castro, M. C., Werneck, V., and Gouvea, N. (2016). Ensino de Matemática Através de Algoritmos Utilizando Jogos para Alunos do Ensino Fundamental II. page 1039.

Erkan, G. and Radev, D. R. (2004). Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, 22:457–479.

Ganesan, K., Zhai, C., and Han, J. (2010). Opinosis: a graph-based approach to abstractive summarization of highly redundant opinions. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 340–348. Association for Computational Linguistics.

Giraffa, L., Muller, L., and Moraes, M. C. (2015). Ensinado Programação apoiada por um ambiente virtual e exercícios associados a cotidiano dos alunos: compartilhando alternativas e lições aprendidas. In *Anais dos Workshops do Congresso Brasileiro de Informática na Educação*, volume 4, page 1330.

Lin, C.-Y. (2004). Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out: Proceedings of the ACL-04 workshop*, volume 8. Barcelona, Spain.

McAuley, J. J. and Leskovec, J. (2013). From Amateurs to Connoisseurs: Modeling the Evolution of User Expertise Through Online Reviews. In *Proceedings of the 22Nd International Conference on World Wide Web*, WWW '13, pages 897–908, New York, NY, USA. ACM.

Mihalcea, R. and Tarau, P. (2004). Textrank: Bringing order into texts. Association for Computational Linguistics.

Oliveira, M. V., Rodrigues, L. C., and Queiroga, A. (2016). Material didático lúdico: uso da ferramenta Scratch para auxílio no aprendizado de lógica da programação. page 359.

Page, L., Brin, S., Motwani, R., and Winograd, T. (1999). The pagerank citation ranking: bringing order to the web.

Poibeau, T., Saggion, H., Piskorski, J., and Yangarber, R. (2012). *Multi-source, Multilingual Information Extraction and Summarization*. Springer Science & Business Media.

Radev, D., Allison, T., Blair-Goldensohn, S., Blitzer, J., Celebi, A., Dimitrov, S., Drabek, E., Hakim, A., Lam, W., Liu, D., et al. (2004). Mead-a platform for multidocument multilingual text summarization.

Ramos, A. M. S., Woloszyn, V., and Wives, L. K. (2017). An experimental analysis of feature selection and similarity assessment for textual summarization. In *Colombian Conference on Computing*, pages 146–155. Springer.

Řehůřek, R. and Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA. `http://is.muni.cz/publication/884893/en`.

Saggion, H. and Poibeau, T. (2013). Automatic text summarization: Past, present and future. In *Multi-source, multilingual information extraction and summarization*, pages 3–21. Springer.

Wan, X. (2013). Co-regression for cross-language review rating prediction. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 526–531, Sofia, Bulgaria. Association for Computational Linguistics.

Woloszyn, V., dos Santos, H. D., Wives, L. K., and Becker, K. (2017). Mrr: an unsupervised algorithm to rank reviews by relevance. In *Proceedings of the International Conference on Web Intelligence*, pages 877–883. ACM.

Wu, J., Xu, B., and Li, S. (2011). An unsupervised approach to rank product reviews. In *Fuzzy Systems and Knowledge Discovery (FSKD), 2011 Eighth International Conference on*, volume 3, pages 1769–1772. IEEE.

Xiong, W. and Litman, D. (2011). Automatically predicting peer-review helpfulness. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 502–507, Portland, Oregon, USA. Association for Computational Linguistics.

Yang, Y., Yan, Y., Qiu, M., and Bao, F. (2015). Semantic analysis and helpfulness prediction of text for online product reviews. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics*, pages 38–44, Beijing, China. Association for Computational Linguistics.

Zeng, Y.-C. and Wu, S.-H. (2013). Modeling the helpful opinion mining of online consumer reviews as a classification problem. In *Proceedings of the IJCNLP 2013 Workshop on NLP for Social Media (SocialNLP)*, pages 29–35, Nagoya, Japan. Asian Federation of Natural Language Processing.