An Evaluation Method for the Performance Measurement of an Opinion Mining System

Manuela Angioni, Andrea Devola, Mario Locci, Francesca Mura,
Franco Tuveri and Mirella Varchetta
Center for Research and Scientific Studies in Sardinia, Bld. 1, Piscina Manna, Pula, Cagliari, Italy

Keywords: Opinion Mining, User Profiling, Natural Language Processing.

Abstract: This paper proposes an evaluation method for the performance measurement of an Opinion Mining system, parameterized according to the reviewer's point of view. The work aims to highlight and resolve some issues resulting from previous activities in evaluating the goodness of the results obtained by the analysis of the reviews. The evaluation method is based on a model of Opinion Mining system able to identify and assess the aspects included in a collection of reviews and the weighted importance of such aspects for their authors. A user profiling system will work together with the Opinion Mining system, providing the set of parameters to associate with the aspects and allowing the Opinion Mining system to configure itself according to the user preferences. For the preliminary experiments, a narrower sub-set of Yelp dataset limited to restaurants has been used.

1 INTRODUCTION

In previous works (Angioni et al., 2015, Angioni et al., 2016) we have faced some problems related to the integration of Opinion Mining systems with recommendation systems.

In Angioni et al. (2016), we have proposed an ensemble of aspect-based Opinion Mining algorithms, using a lexicon-based approach, with a Matrix Factorization to improve the prediction of results in a recommendation system. In this context, the Opinion Mining system has been used to work on the textual reviews about restaurants extracted from the Yelp dataset (available at https://www.yelp.com/dataset/challenge) producing a set of ratings about the business activities to be compared with the Yelp ratings.

Beyond the recommendation systems, the present paper aims to highlight the objective difficulty encountered in evaluating the goodness of the results obtained by the analysis of the reviews through the Opinion Mining system.

In the mentioned works, the performance of the system has been carried out by two researchers that have manually evaluated a collection of 200 reviews to check the validity of the related ratings, using a common evaluation method. The comparison with the ratings manually assigned by the researchers allowed, also, the evaluation of the performance of the Opinion Mining methodology.

The observation of the rate associated by the Yelp users to their reviews pointed out that the assignment of the final rating is not always consistent with the content of the review. Moreover, the inconsistencies occurred frequently during the manual analysis of the reviews, sometimes evidencing strong differences between the rates assigned by Yelp users and the manual evaluation provided by evaluators.

Initially, it was thought to be a mistake in the allocation of the rate with respect to the content of the review. A more detailed examination highlighted that the assignment of the final rate is correct and simply reflects the user's preferences, other than those of the evaluators.

Instead, the aspect to point out is the differences of "taste" between the two Italian evaluators and the US American authors of the Yelp reviews, i.e. differences in food culture, “culture” in the broader sense, economic opportunities, and competences.

The influence of the cultural differences is particularly true in some cases, such as the choice of the food, less in others. However, the particular needs of each person affect the opinions expressed and the preferences in the choice of products and services. These differences definitely affect the assessments of products and services.

Angioni, M., Devola, A., Locci, M., Mura, F., Tuveri, F. and Varchetta, M.
An Evaluation Method for the Performance Measurement of an Opinion Mining System.
DOI: 10.5220/0006767404480455
In *Proceedings of the 20th International Conference on Enterprise Information Systems (ICEIS 2018)*, pages 448-455
ISBN: 978-989-758-298-1

These remarks have been found to be fundamental for a reasoning behind Opinion Mining systems, particularly as regards the performance evaluation of these systems.

In order to achieve this result, the Opinion Mining system, as well as the evaluator of the reviews, should be as similar as the user who writes the review. So, we are planning to adopt a user profiling system, applied to the user-related data, able to avoid the differences expressed by the evaluators and the reviewers in the evaluation of the reviews and in the rating expressed. It is thus more likely that evaluators and reviewers having similar profiles express similar ratings about the same reviews.

For this purpose, the paper describes an evaluation method for Opinion Mining systems, applied to an Opinion Mining model, parameterizable according to the reviewer profile.

The method is based on a model of Opinion Mining system able to identify and assess the aspects included in a collection of reviews and the weighted importance of such aspects for their authors. The considerations made in the previous works regard the definition of a method that is as objective as possible and able to provide the two evaluators with the most stringent criteria for evaluating the opinions expressed in the analyzed reviews. The defined criteria led to the definition of a set of 12 aspects related to the Yelp reviews about Restaurants, on which the evaluations were focused. Analyzing these aspects, we realized that the semantic analysis system we implemented is able to detect and effectively exploit 8 of these 12 aspects that will be considered in this paper. What we are interested in understanding in this phase of work is how much the customization of the Opinion Mining system based on the user profiling allows to improve its performance.

The rest of the paper is structured as follows. Section 2 presents the state of the art about Opinion Mining systems. The Section 3 introduces the proposed method, describing the analysis of textual resources and the tools involved to perform the chunk analysis, the feature evaluation and the identification of the users' profiles. Section 4 proposes some considerations about the work done and, finally, Section 5 discusses the conclusions and the future works.

2 STATE OF THE ART

Most websites, like Yelp, allow their customers to better explain their opinion of the product by more detailed textual reviews.

Some Opinion Mining systems are based only on users' overall ratings about items, but do not consider and do not work on the opinions expressed by the users about the different aspects of an item. As a result, the rate associated to a review does not wholly summarize the opinion of the users, maybe ignoring important information.

Most recommender systems do not use textual information to generate recommendations. The extraction of textual opinions is a significant extension. Some authors (Snyder and Barzilay, 2007) analyzed restaurants reviews to deduce the author's sentiments regarding some aspects related to the domain, as food and service. They were among the first to consider faceted opinions instead of the overall opinions. More recently (Homoceanu et al., 2011) extracted valuable information by means of faceted Opinion Mining able to help cluster reviews and to improve decision support. Jakob et al. (2009) obtained marginal improvements by adding opinions about predefined topics to star ratings.

Over the last few years some studies have been proposed in order to assess how the use of Opinion Mining systems together with the user profiling is useful to improve the performance of the recommendation systems.

Musal et al. (2017), with a view similar to of (Hariri et al., 2011), believe that the topics, mentioned in the reviews, are the most important information available in order to model a recommendation able to produce personalized profile rakings, by means of user interest profiles.

From the point of view of the Opinion Mining the most recent studies focus on detailing such information in order to gain knowledge more closely reflecting the complexity of businesses, products and services contexts.

As Pang and Lee affirm in (Pang and Lee, 2008) at least one related set of studies claims that "the text of the reviews contains information that influences the behaviour of the consumers, and that the numeric ratings alone cannot capture the information in the text" (Ghose and Ipeirotis, 2007).

A common problem to the user-generated reviews is usually related to the inconsistency in terms of length, content, treated aspects and usefulness because not every user writes about all the relevant aspects, which characterize a business activity. For this reason relevant information would be disregarded, causing a lack of useful data in the input of an Opinion Mining algorithm.

Considering the Opinion Mining lexicon-based approach, most available opinion lexicons list words

with their values, i.e. with a positive, negative or neutral polarity. Hence, the evaluation of the sentences has to be composed according to the different values of its words.

Several studies (Choi and Cardie, 2008), (Klenner et al., 2009), (Liu and Seneff, 2009) (Moilanen and Pulman, 2007), (Thet et al., 2010) have been carried out in recent years regarding the evaluation of the opinions, composed by the single words of the sentences.

Wogenstein et al. (2013) are among the few authors to propose a study about the evaluation of an aspect-based Opinion-Mining system, based on a lexicon approach. They used a phrase-based opinion lexicon for the German language, which directly includes negation and valence shifting in the phrases and applied a distance-based algorithm for linking the opinion phrases to the aspects related to the insurances, their products and services. Two persons, not involved in the project and not aware of the algorithm used for the opinion mining, performed the evaluation of the accuracy of the algorithm manually, tagging strong opinions expressed about aspects of insurances.

Opinion Mining systems continue to be of great interest, but we believe that it is necessary to start talking about the evaluation of their performances according to the reviewer's point of view and to evaluate the different approaches.

3 THE METHOD

The performance of the systems working on textual resources is strongly linked to the expressive forms used in the sentences analyzed.

An opinion can be expressed in a sentence, through a verb, e.g., *I like it,* an indirect expression, an idiomatic expression, with irony, or even emphasizing some words with capital letters.

Currently is a success even just to be able to deliver acceptable performance in analyzing the just mentioned examples. Building an Opinion Mining system that works on reviews, able to autonomously identify the individual aspects and the resulting views, is for now a very difficult goal to reach for any workgroup.

The idea we outline in this paper is that, in order to evaluate a system of Opinion Mining, it is necessary to structure it so that it would be parameterized according to the reviewer's point of view.

The problem is twofold. On the one hand there is the difficulty in implementing a good Opinion Mining system, on the other hand there is an obvious problem in its objective performance evaluation.

We therefore propose an evaluation method based on a model of an Opinion Mining system, shown in Figure 1, able to identify and evaluate the aspects present in a collection of reviews and the weighted importance of such aspects for the authors. The evaluation method exploits the specific structure of the Opinion Mining system appraising how the parameters associated to the aspects can change the performances of the proposed model.

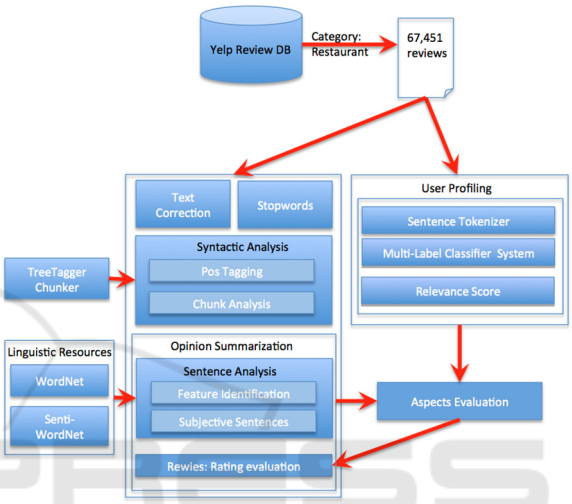

Figure 1: The model of the Opinion Mining system conditioned by the user profiling system.

User profiling, needed to identify the characteristics of the authors, is constructed taking into account the aspects discussed by the author in his reviews. For each profile, the weight associated with each aspect, derived from an evaluation of all a user's reviews, will modify the Opinion Mining system parameters and the results.

In this phase of study of the model we are developing two different methods for the aspects detection. In the first case we use a set of terms, extracted with a term frequency function (TF), mapped on the set of aspects we are interest in. In the second case a Machine Learning method, based on the sentence classification by means of binary classifiers, has been developed. The set of aspects and terms extracted with the first method will allow also the validation of the equivalent set extracted with the second method. The two methods for the aspects detection work together in the rating evaluation of the reviews, as depicted in Figure 1 that describes the model of Opinion Mining system conditioned by the user profiling system.

Further reviews can be easily inserted because the Machine Learning method is able to analyze them according to the aspects, profiling the user.

In order to accomplish the text analysis and the profiling of users and reviewers, a description of the tools and technologies used will be provided in the following sections.

3.1 Chunk Analysis

The text analysis is a key element of the whole process and is made through several tools, some of which has been specifically constructed. The applied linguistic approach is based on the use of the TreeTagger Chunker (TTC) (Schmid, 1994), a syntactic parser that analyzes the text and returns a chunk sequence.

An example of chunks subdivision for the phrase "The dog barks" is shown below:

Sentence: The dog barks

```
<NC>
   The DT the
   dog NN dog
</NC>
<VC>
   barks VBZ bark
</VC>
```

where the <NC> and the <VC> tags identify a noun chunk and a verb chunk, respectively.

The TreeTagger Chunker provides as input a text, in form of chunk sequences, to ANTLR (Another Tool for Language Recognition), which through a lexer and a parser performs a chunk analysis by returning a parse tree of the original text.

ANTLR (http://www.antlr.org/) is a parser generator for reading, processing, executing, or translating structured text or binary files, that allows to define grammars in both ANTLR syntax (similar to EBNF and YACC) and in a special abstract syntax for Abstract Syntax Tree (AST).

A grammar, defined through the set of lexer rules, determines exactly which character sequences are considered valid. Each character or group of characters collected in such a way is called token. The tokens, in our case, are text portions identified by keywords.

The parser organizes the received tokens in acceptable sequences, according to the rules defined by the grammar. The parser will thus be able to recognize the patterns that make up specific structures of the speech and to group them appropriately. If the parser encounters a token sequence that does not match any of the allowed sequences, it will raise an error.

The rules are defined by a EBNF-like (Extended Backus-Naur Form) notation, and outline the desired token patterns, exploiting the position of individual chunks and their contents.

A brief list of some of the defined rules is shown below as example. As usual `+` means 1 or more, `*` 1 or more time and `?` once or not at all.

The rule `document` identifies the whole document as a concatenation of sentences:

```
document: sentence+ (EOF) # start;
```

The rule `sentence` identifies the chunks inside a sentence:

```
sentence:
content?(nva_pc[...]navn|nvn|any_chunks
|SENT)+;
```

The `nva` rule identifies the adjectives referring to a noun i.e., `n_chunk` identifies a nominal chunk, `v_chunk` a verbal one, and `adjc_chunk` an adjective chunk:

```
nva: n_chunk v_chunk adjc_chunk+;
```

The rules `n_chunk`, `v_chunk` and `adjc_chunk` are defined in the following way:

```
n_chunk: NC (content)+ NC_END
content*;
v_chunk: VC (content)+ VC_END
content*;
adjc_chunk: ADJC (content)+ ADJC_END
content*;
```

The couple of symbols `NC` and `NC_END`, `VC` and `VC_END`, and `ADJC` and `ADJC_END` identifies the opening and the closure of the tags.

The rule `adjjr_chunk` permits to identify the comparative adjectives in a sentence:

```
adjjr_chunk:
ADJC content? TERM JJR (TERM|UNK)
ADJC_END | ADJC content TERM JJ
(TERM|UNK) ADJC_END ;
```

The list further includes over a dozen rules able to identify the various chunks or specializations of them, further nineteen rules to identify super patterns, including the above `nva` rule. Other support rules have been also defined to identify valid terms, the set of conjunctions, the set of POS tags and to identifies some lemmas that the TreeTagger Chunker can not solve.

The layer that deals with the parsing of the text in input is automatically generated by ANTLR starting from the defined grammar. It then generates `.tokens` files, containing the key-value references of the individual token generated by the lexer, and a set of *java* classes that identify the parser and the lexer.

The rules allow identifying how nouns, verbs, adjectives and adverbs are related in a sentence. So, for each noun detected by a given pattern, even if repeated, it is possible to collect adjectives, adverbs, verbs, or combinations related to it.

By providing the system with the ability to syntactically and semantically recognize sentences in natural language, it is possible to support the collection of subjective opinions related to the target nouns, that are the nouns we are interested in, with their relative polarity.

In the chunk analysis phase, the text of a review is pre-processed by the TTC. A listener developed in Java, interacting with the parse tree, identifies the patterns defined by the rules, and hence the sentences potentially expressing a polarity.

The sentences to work on are a subset of the total set of sentences composing the corpus of the reviews. Initially an analysis of the most commonly used terms is performed through a Term Frequency (TF) function.

The dataset chosen for the study is made available by the Yelp social network. An important feature of this particular data set is that it provides not only the star ratings (from 1 to 5 stars) assigned by the users, but also a textual review. The Restaurant category was chosen, because is the most representative in the dataset. Have been considered only the users giving a number of reviews greater than 9, as more reliable.

Targeting the Restaurants category, 67.451 reviews have been extracted.

The identification of the features for the Yelp reviews has been performed evaluating the nouns frequency in the text through a word counter. We first removed the stop words and then the cleaned text was tokenized obtaining as a result a collection of about 4000 words, including individual and compound words.

We condensed this set by only considering words with a frequency greater than 100, in order to test the potential of the proposed approach, to be extended in a future work. Finally, we identified 935 nouns as candidate features. The features were then manually validated and classified into eight aspects based on their topic: Ambience, Beverage, Dessert, Food, Food Variety, Price, Service, and Staff.

The use of a TF function on the corpus of the reviews has certainly resulted in a proliferation of features (Dong and Smyth, 2017) but their 8-aspects pooling provides us with two benefits.

The first is that fine-grained opinions about specific features of an item provide better information compared to the overall opinion. The second is that the amount of the identified nouns in the reviews provides a more accurate measure of the relevance of the aspects, which are part of the information useful to the user profiling system, currently under development.

3.2 Feature Evaluation

The previously described chunk analysis allows associating terms, identified as features, with attributes, adjectives, adverbs, and verbs.

Through the values that SentiWordNet (Baccianella et al., 2010) associates with each adjective and adverb linked to each of the features, the polarity value is obtained for each of them.

For each review, the polarity values for each of the (eight) aspects considered will be identified, based on the existing feature-aspect mapping.

The overall assessment of the importance of the single aspect is finally calculated for each reviewer.

3.3 User Profiling

The introduction of the eight weighted aspects, while certainly help to improve the method of analysis of the reviews, does not solve alone the cited discrepancies because it does not take into account the priorities and the preferences expressed by the reviewers according to their profiles.

As anticipated, the join of a user profiling system with the Opinion Mining system composes the model.

The user profiling system provides the set of parameters to associate with the aspects and it will allow the Opinion Mining system to configure itself according to the user preferences.

The analysis of the reviews has highlighted that users base theirs criticism on few aspects. In addition, different users focus on particular aspects rather than others. In our opinion, when a user is interested to a specific aspect spends more sentences to describe it. It is thus possible to capture the interest of a user by weighting their interests through the aspects he is talking about.

The profiles are obtained considering the reviews written in the past by the author.

We proposed a method based on a term-based approach built on sentence classification. The referred categories are the cited set Λ composed by eight aspects: Ambience, Beverage, Dessert, Food, Food Variety, Price, Service, and Staff.

3.3.1 Aspect Classification

The sentences of a review can include one, but usually two or three, different aspects, which means that more than a label could be associated to a single phrase. This problem can be considered a typical multi-label and multi-class classification task.

A commonly used approach to the multi-label classification is to break down the problem into one or more single label problems. In this way, a single-label classifier can be used to make single-label classifications. The output of each single classifier is codified to have a multi-label representation. The most widely used transformation method is based on a binary classifier. A multi-label problem has been transformed into a binary problem for each label.

The classifier was therefore built as an ensemble of specialized binary classifiers $b_1,\ldots,b_n$ one for each aspect $A_1,\ldots,A_n$, as shown in Figure 2. This was possible because the correlation between the aspects was not significant.

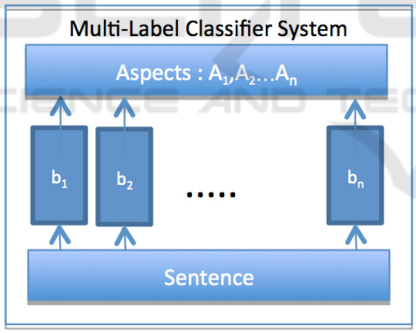

Figure 2: The Multi-Label Classifier System.

3.3.2 Binary Classifier

Naive Bayes, SVM and Decision Tree have been proven to be text classifiers with excellent performance. Related to the bag-of-words (bow) methods there are several problems, like the wide range of features or the loss of the position of the words. In fact, the bow ignores the context and the order of the words without fully capturing the semantics of the words.

The high-order n-gram model has been studied to capture contextual information. Low ranking accuracy has been due to the data sparsity (Post and Bergsma, 2013).

Recent studies in the field of the ANN (Artificial Neural Networks) have produced satisfactory results as RNN (Recurrent Neural Networks) and CNN (Convolutional Neural Networks) models that exploit the sequential nature of the text. In particular, the CNN model (Kim, 2014), using pre-trained word vectors for sentence-level classification tasks, has produced excellent results.

Considerable improvements have been made with a model based on RNNs and CNNs for sequential short-text classification (Lee and Dernoncourt, 2016).

In order to classify sentences, binary classifiers for the short-text classification have been used. In particular, LSTM (Long Short Term Memory), a recurrent neural network (RNN), has been used, according to the model proposed by Hochreiter and Schmidhuber (1997). LSTM is a specific RNN that models temporal sequences with long-range dependencies.

Figure 3 shows the proposed neural network model. The LSTM is unrolled for a time window T to process a sequence of vectors h_T, $h_{T-1},\ldots$, h_1. The words w_t are one-shot coded. An embedded layer M is the input from which $h_t = Mx_t$ is passed to the LSTM network at each time step. The output sequence is multiplied by the matrix O on which the softmax activation function is performed.

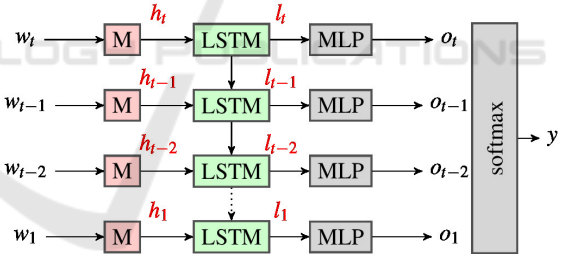

Figure 3: Model of the binary classifier.

The content of each review is classified and scored by relevance. Reviews are analyzed and split in phrases in order to identify the aspects, limited to the specific set Λ. To identify the relevance of an aspect into a review the algorithm calculates the amount of phrases about it.

The relevance $r_{i,a}$ of the aspect $a \in \Lambda$, related to the review i, is defined as:

$$r_{i,a} = \frac{\sum_{k=1}^{n} C_a(p_{k,i})}{n_i} \quad (1)$$

where n_i is the number of phrases, C_a is a binary classifier for the aspect a and $p_{k,i}$ the k-th phrase.

3.3.3 Set-up

For the experiments, a narrower sub-set of Yelp dataset limited to restaurants has been used.

A subset of the reviews included in the dataset has been manually tagged to build the training set. The reviews have been subdivided into sentences and analyzed using Lucene StandardTokenizer (https://lucene.apache.org/).

Each sentence has been associated with the aspects detected, with up to 3 aspects per sentence. In such a way a multi-label and multi-class training set was built. Later, from the built-in dataset, training-sets were built for every aspect.

The generic training set Ω_a of the aspect $a \in \Lambda$, is composed by the phrases extracted from a subset of the reviews. The phrases labelled with a are defined as positive examples, the other are the negative ones.

The obtained sets were balanced by the over-sampling method.

The set-up parameters of each binary classifier C_a and the learning parameters are shown in Table 1. The ADAM stochastic optimization method was used to optimize the network.

Table 1: The set-up parameters of binary classifiers.

Max sequence length	50
Number of units LSTM cell	20
Embedding size	300
Learning rate	0.001

4 SOME CONSIDERATIONS

The work presented in this paper will carry on through the integration of the two components: the text analysis module, composed by the Chunk Analysis and the Feature Evaluation, with the parameterization of results based on user profiles. We are working on the integration of the software modules that will lead to a test and to the evaluation of the results.

The evaluation phase involves the analysis of a collection of about 200 reviews according to the eight aspects mentioned before. The obtained results will be: the evaluation of each single aspect contained in the review, and the overall evaluation of the review on a discrete rating value between 1 and 5 stars.

The model assessment will be implemented taking into account the following conditions:

- Evaluation of reviews written by authors with a profile similar to the human evaluator. The rates produced by the model, parameterized on the profile of the review author, will be compared to those provided by the evaluator and the rate provided by the review author.
- Evaluation of reviews with a different profile from the human evaluator. The model is set with parameters based on the profile of the human evaluator. The resulting rates provided by the model will be compared to those provided by the evaluator and those of the review author.
- Evaluation of the collection of reviews used in the paper (Angioni et al., 2016). For each review of the collection, the model will be set on the profile of its author. The results will be compared to the results obtained in the previous model.

The provided average rating will be evaluated in terms of precision, recall, and F1-score and will be discussed in a next paper.

5 CONCLUSIONS AND FUTURE WORK

In this paper a method to evaluate the results of an Opinion Mining system parameterized according to the reviewer's point of view through a user profiling system, applied to the user-related data has been proposed.

In our expectations such method will be able to avoid the differences expressed by the human evaluators and the reviewers in the evaluation of the reviews and in the rating expressed.

Future work will address the interaction between the weights, associated with the eight aspects, calculated through the profiling system, and the evaluation of reviews. The rating calculated by the Opinion Mining model, applied to the same reviews, will therefore vary from profile to profile.

These further activities will permits us to validate the idea behind what has been so far presented also providing an assessment of the performance of the method used.

REFERENCES

Angioni, M., Clemente, M. L., Tuveri, F., 2015: Combining Opinion Mining with Collaborative Filtering. In *Proceedings of WEBIST 2015, 11th International Conference on Web Information Systems and Technologies*, Lisbon, Portugal.

Angioni, M., Clemente, M. L., Tuveri, F., 2016. Improving Predictions with an Ensemble of Linguistic Approach and Matrix Factorization. *Web Information Systems and Technologies*, Springer Valérie Monfort, Karl- Heinz Krempels, Tim A. Majc, pp 169--190 vol. 246, Lecture Notes in Business Information Processing.

Baccianella, S., Esuli, A., Sebastiani, F., 2010 SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. In *LREC 2010, 7th International Conference on Language Resources and Evaluation*, Malta, pp. 2200-2204.

Choi, Y. and Cardie, C., 2008. Learning with Compositional Semantics as Structural Inference for Subsentential Sentiment Analysis. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 793–801.

Dong, R. and Smyth, B., 2017. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence* (IJCAI-17).

Ghose, A., Ipeirotis, P. G., 2007: Designing novel review ranking systems: Predicting usefulness and impact of reviews. In: *International Conference on Electronic Commerce (ICEC)*.

Hariri, N., Mobasher, B., Burke, R., and Zheng, Y. 2011. Context-aware recommendation based on re- view mining. In *Proceedings of the 9th Workshop on Intelligent Techniques for Web Personalization and Recommender Systems (ITWP 2011)*, page 30, 2011.

Hochreiter, S. and Schmidhuber, J., 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780. MIT Press. 515–520 arXiv preprint arXiv:1603.03827.

Homoceanu, S., Loster, M., Lofi, C., and Balke, W.-T., 2011. Will I like it? providing product overviews based on opinion excerpts. In *Proceedings of IEEE Conference on Commerce and Enterprise Computing (CEC)*.

Jakob, N., Weber, S. H., Muller, M. C., and Gurevych, I., 2009. Beyond the stars: exploiting free-text user reviews to improve the accuracy of movie recommendations. In *Proceedings of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion*, TSA '09, pages 57–64, New York, NY, USA. ACM.

Kim, Y., 2014. Convolutional neural networks for sentence classification. In *Proceeding of EMNLP 2014*. Doha, Qatar. arXiv preprint arXiv:1408.5882.

Klenner, M., Petrakis, S., and Fahrni, A., 2009. Robust Compositional Polarity Classification. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing*, pages 180–184.

Lee, J. Y. and Dernoncourt, F. (2016). Sequential short-text classification with recurrent and convolutional neural networks. In *Proceedings of NAACL-HLT 2016*, pages

Liu, J. and Seneff, S., 2009. Review Sentiment Scoring via a Parse-and-Paraphrase Paradigm. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 161–169.

Moilanen, K. and Pulman, S., 2007. Sentiment Composition. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing*.

Musat, C.-C., Liang, Y., Faltings, B., 2013. Recommendation Using Textual Opinions. In *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence.* Beijing, China.

Pang, B., Lee L., 2008: Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval 2(1-2),* DOI: 10.1561/1500000011.

Post, M. and Bergsma, S., 2013. Explicit and implicit syntactic features for text classification. *51st Annual Meeting of the Association for Computational Linguistics* - Short Papers (ACL Short Papers 2013) Sofia, Bulgaria

Schmid, H., 1994: Probabilistic Part-of-Speech Tagging Using Decision Trees. *In Proceedings of the International Conference on New Methods in Language Processing*, pp. 44-49.

Snyder, B. and Barzilay, R., 2007. Multiple aspect ranking using the good grief algorithm. *In Proceedings of the Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, HLT-NAACL. 300–307.

Thet, T. T., Na, J.-C., and Khoo, C. S. G., 2010. Aspect-based sentiment analysis of movie reviews on discussion boards. *Journal of Information Science*, 36:823–848.

Wogenstein, F., Drescher, J., Reinel, D., Rill, S., and Scheidt, J., 2013. Evaluation of an algorithm for aspect-based opinion mining using a lexicon-based approach. In *Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining (WISDOM '13).* ACM, New York, USA DOI: 10.1145/2502069.2502074.