Learning rules for automatic identification of implicit aspects in Portuguese

Mateus Tarcinalli Machado[1], **Thiago Alexandre Salgueiro Pardo**[1],
Evandro Eduardo Seron Ruiz[2], **Ariani Di Felippo**[3]

[1]Núcleo Interinstitucional de Linguística Computacional (NILC)
Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo

mateusmachado@usp.br, taspardo@icmc.usp.br

[2]Faculdade de Filosofia, Ciência e Letras de Ribeirão Preto, Universidade de São Paulo

evandro@usp.br

[3]Núcleo Interinstitucional de Linguística Computacional (NILC)
Departamento de Letras, Universidade Federal de São Carlos

ariani@ufscar.br

Abstract. *This sentiment analysis work is focused on the task of identifying aspects, emphasizing the so-called implicit aspects, i.e., those that are not explicitly mentioned in the texts. For this, we analyzed frequency-based methods, adapted rules from the English language to Portuguese, and developed a method that learns new rules through corpus analysis.*

Resumo. *Este trabalho de análise de sentimentos está focado na tarefa de identificação de aspectos, dando ênfase aos chamados aspectos implícitos, ou seja, aqueles que não são mencionados explicitamente nos textos. Para isso, analisamos métodos baseados em frequência, adaptamos regras da língua inglesa para o português e desenvolvemos um método que aprende novas regras por meio de análise de corpus.*

1. Introduction

Sentiment analysis is an area of applied computing research related to natural language processing, which aims to analyze people's opinions, feelings, assessments, attitudes, and emotions concerning entities such as products, services, organizations, individuals, issues, events, topics, and their attributes [Liu 2012]. It presents a wide range of studies and is often referred to in the literature by slightly different names and tasks, such as opinion mining, opinion extraction and sentiment mining, among others. According to [Taboada 2016], sentiment analysis is a growing field at the intersection between linguistics and computer science that aims to determine automatic sentiments in a text.

Many authors [Medhat et al. 2014, Rana and Cheah 2016, Yadollahi et al. 2017] divide sentiment analysis into three levels: document, sentence, and aspect. Each level aims to address different challenges, with aspect-based analysis being the most refined. At the document level, the goal is usually related to the identification of the polarity of the whole document, to determine whether its overall opinion is positive or negative. This level of analysis is important in social and psychological studies carried

out on social networks, consumer satisfaction, analysis of patients in medical settings, and many others [Yadollahi et al. 2017]. Sentiment analysis performed at the sentence level aims to find the polarity of a sentence. At this level of analysis, it is also important to identify whether the target sentence is subjective or objective, i.e., whether it contains an opinion. This level of sentiment analysis is usually influenced by the context around the sentence and is considered very important for applications dealing with texts from social networks, comments and short messages, among others. Finally, the aspect level deals with more refined sentiment analysis strategies, trying to determine people's opinion about a specific characteristic (aspect) of a product, a service, or an entity [Medhat et al. 2014, Rana and Cheah 2016, Yadollahi et al. 2017]. To carry out this level of analysis, it is essential to identify the entities mentioned and their respective aspects and the so-called words of opinion related to the aspects. From there, the polarity of opinions directed to each identified aspect is determined. The results of the aspect-level sentiment analysis can then be finally summarized and visualized, for example, in a table containing the found aspects and their respective polarities [Rana and Cheah 2016]. This level of processing can also be called aspect-based sentiment analysis [Liu 2012] and is very useful for language processing applications, as text summarization [López Condori and Pardo 2017].

In this paper, we focus on the aspect detection task, emphasizing the treatment of implicit aspects. An aspect is considered explicit if it is mentioned directly in a sentence, otherwise, it is considered implicit [Liu 2012]. For example, in the sentence "The image quality of this camera is great", the expression "image quality" is an explicit aspect term. On the other hand, in "This camera is expensive", the word "expensive" implicitly refers to the "price" aspect.

It is a fact that we have a smaller amount of works that deal with implicit aspects [Ravi and Ravi 2015, Rana and Cheah 2016]. For Portuguese language, in particular, this is still a challenge [Pereira 2020]. Several authors [Zhang and Zhu 2013, Panchendrarajan et al. 2016] point out that between 15% and 30% of the aspects of opinionated texts are mentioned implicitly, and, for some categories, these percentages increase significantly. The non-detection of these aspects leads to the loss of important information and, consequently, negatively affects the results of related applications. The identification of implicit aspects and the treatment of texts in the Portuguese language, despite being complex problems, are relevant tasks and have the potential to advance the state of the art in the area.

This paper presents an analysis of frequency and rule-based methods for aspect detection. We also develop a method that learns extraction rules from a corpus. The rest of this work is organized as follows: Section 2 presents the main related work, especially the ones for Portuguese language. Section 3 presents our methods and datasets. Section 4 describes the results obtained by the implemented methods. Finally, Section 5 discusses the strengths and weaknesses of these methods.

2. Related work

There are few works that explicitly try to characterize, model and identify implicit aspects in the sentiment analysis area, which may be explained by the difficulties of dealing with them. We briefly synthesize the main related work in what follows.

The authors of [Cai et al. 2020] formulated the task as a problem of hierarchical prediction of categories and sentiments, where first the algorithm identify categories of aspects in a sentence and then the sentiments related to each detected category. To do this, they used a convolutional network of hierarchical graphs (Hier-GCN). To codify the sentences and their attributes, the authors used representations of bidirectional transformer encoders (BERT) [Devlin et al. 2018], which is a pre-trained model and has achieved excellent results in many natural language processing tasks. Their results show that this type of modeling is suitable for the detection of the aspect category together with the classification of sentiments, achieving in its highest result 0.76 precision, and 0.73 recall.

In [Marcacini et al. 2018], the authors developed a method based on heterogeneous networks that combines features as labeled aspects, unlabeled aspects, and linguistic features. To perform the classification, they developed an algorithm that propagates the labels using linguistic features as a bridge for this propagation. As a result, the method obtained f-measure between 0.56 and 0.68, depending on the analyzed domain.

The work of [Balage Filho 2017] analyzed the following aspect extraction methods: frequency-based, relation-based, and machine learning-based ones. The frequency-based method is the simplest. It selects the most frequent nouns and noun phrases as aspects. The relation-based one analyzes the relationships between aspects and the opinion related to them. In machine learning, an annotated training set is used to create a model that identifies the aspects. In their experiments, the frequency-based method showed the best result with 0.49 f-measure.

In the work of [Costa and Pardo 2020], the authors analyzed lexical-based methods for extracting aspects from opinions written in Portuguese. In a first experiment, the authors made use of aspect ontologies, in order to identify explicit and implicit aspects. In a second experiment, the authors extended a frequency detection method of aspects using the distributional models Word2Vec [Mikolov et al. 2013a] in order to enrich the process. The experiment performed with ontologies achieved the best results, with 0.53 precision, and 0.44 recall.

3. Data and Methods

3.1. Methods

In this work, we applied four methods for aspect detection: two frequency-based and two rule-based methods. For the frequency-based ones, we implemented the Freq Baseline and Freq Baseline with Word2Vec methods to serve as comparative bases for other methods. Regarding the rule-based methods, we tried to translate and apply the rules discovered in [Poria et al. 2014] to texts in Portuguese, as well as to use an automatic method for learning new detection rules. Before detailing the methods, we introduce the dataset that we use.

3.2. Dataset

For the execution of the experiments, we worked with a corpus formed by opinionated texts in Portuguese, resulting from the work of [Vargas and Pardo 2018]. The corpus is one of the few existent corpora with identified implicit aspects, even considering works in English. It has 60 product reviews from 3 different domains: cameras, books, and smartphones. Table 1 shows the characteristics of the corpus. It is important to note

that the percentages of implicit aspects found, with values close to 15% in the domains of cameras and smartphones, demonstrate the relevance of the specific analysis of this category of aspects. We evaluated all experiments in this corpus.

Table 1. Corpus composition.

Domains	Reviews	Aspects	Explicits	% Explicits	Implicits	% Implicits
Cameras	60	352	299	84.94%	53	15.06%
Books	60	330	304	92.12%	26	7.88%
Smartphones	60	455	387	85.05%	68	14.95%

In the corpus we found different cases of implicit aspects. For example, in "Very compact and very beautiful", the size and design aspects can be identified by their qualifiers. In "Camera of small measurement", the size aspect appears as a semantically close word. Finally, in "Works anywhere", only with specific knowledge about the context, in this case the functioning of the smartphone, we can understand that the phrase refers to the signal aspect. These are just a few examples, which demonstrate how complicated the task of identifying implicit aspects can be.

In a pre-processing stage, we prepared the texts for the execution of the methods. We selected these processes according to the needs of the algorithms. In our case, we performed the following processes:

- **Dataset division into training and test sets**: with the aim of consistently comparing the implemented methods, we performed all experiments using the same training (for searching parameters or rules, as we will see later) and testing (to validate the methods) datasets;
- **Sentences tokenization:** split of texts into sentences;
- **Words tokenization:** split of sentences into tokens, which can be words, symbols and punctuation, among others;
- **PoS tagging:** identification of morpho-syntactic categories of the tokens;
- **Dependency analysis:** construction of sentence dependency trees, to understand the relationship between tokens.

To perform these tasks, except for the division in training and testing sets, we used spaCy [Honnibal et al. 2020] module with pt_core_news_lg model from the Python programming language.

3.3. Method 1: Freq-Baseline

This method selects nouns and noun phrases as aspect candidates and analyzes their frequencies in relation to the number of sentences in the analyzed dataset. The algorithm selects the most frequent candidates as aspects. It is a method with an easy implementation that achieves good results, which is why it is often used as a basis for comparison to evaluate other algorithms.

The definition of the most frequent aspects is performed by comparing the frequency of the candidates being analyzed with a pre-defined cutoff frequency. In [Hu and Liu 2004], the authors used a cutoff frequency of 1%, a value that ended up being used in other studies as well. As in [Machado et al. 2017], in this work, we varied

this cutoff frequency and analyze what would be the most appropriate value. We worked with cutoff frequencies in an interval of 0.01% and 10% with an increment of 0.01% at each execution of the algorithm. Thus, at each iteration, we performed the calculations of precision, recall, f-measure, and percentages of explicit and implicit aspects correctly detected.

3.4. Method 2: Freq-Baseline + Word2Vec

As in [Pavlopoulos and Androutsopoulos 2014] and [Machado et al. 2017], in this work we used the distributional model Word2Vec [Mikolov et al. 2013b] to exclude candidates for aspects that are not related to the domain under analysis. The algorithm compared each candidate aspect with two vectors formed by the centroid of the Word2Vec vectors of words related to the context under study and of the Word2Vec vectors of general domain words. The candidate aspects closest to this general centroid vector were discarded, thus remaining only words more related to the context of the dataset. As it is a method used only as a baseline, we will not go into further details about it. More information can be found in the mentioned references. As in the previous experiment, we searched for the best cutoff frequencies, which remained the same.

3.5. Method 3: Adapted Rules of [Poria et al. 2014]

In our first experiment with rules, we adapted the rules from the work of [Poria et al. 2014]. The method comprises handcrafted rules that analyze the sentence dependency trees together with two lexicons: one with a list of implicit aspects and another called SenticNet [Cambria et al. 2014] that is a concept-level knowledge base containing a set of semantics, sentics, and polarities associated to natural language concepts.

As we did not find an implicit lexicon for the Portuguese language, we created one based on the existent corpus of [Vargas and Pardo 2018]. To maintain consistency with the other experiments and do not introduce bias in the results, we used only the terms found in the training set.

We translated the sentences of the presented examples in [Poria et al. 2014] and adapted the rules to Portuguese, changing the order or type of some components. Finally, we executed them in the adopted dataset.

3.6. Method 4: Automatically Learned Rules

The results of the adaptations of the rules developed in [Poria et al. 2014] for Portuguese were below expectations. This motivated us to look for more adequate rules for the Portuguese language and with better detection of implicit aspects. For this, we used the dependency trees of the sentences of the training set.

We analyzed the elements that had some relationship of dependency on the aspects identified in the datasets and, together with the morpho-syntactic analysis performed, we created a set of rules. As an example, Figure 1 presents the dependency tree, obtained from spaCy, of the sentence "Camera with the best zoom.". In this sentence, the term "zoom" is marked as an aspect, therefore, for the formation of the rules, the terms that have some relationship with it were used.

In the first stage, the elements that have a direct relationship with the marked aspects were selected. The relations and morpho-syntactic characteristics of the elements

Figure 1. Example of dependency tree.
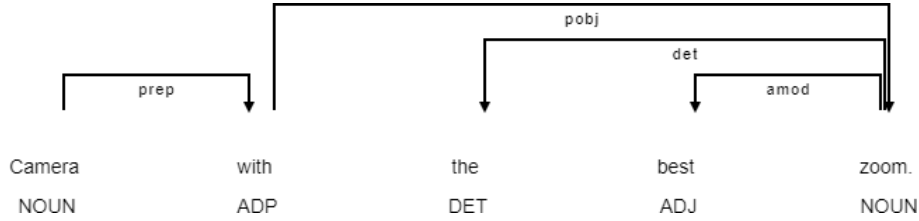
were combined, forming aspect extraction rules. For example, see the rules presented in Table 2. We can interpret the first rule of this table as follows: if a noun (NOUN) is in an adjectival modifier relationship (amod) with an adjective (ADJ), then the first token is extracted as aspect. This same process was repeated, but this time analyzing combinations of three, four, and five tokens, thus generating 1,200 rules for camera domain, 1,470 for books, and 1,564 for smartphones.

Table 2. Example of learned rules.

Tokens		Rules			
Token 1	Token 2	Class 1	Relation	Class 2	Aspect
zoom	best	NOUN	amod	ADJ	Token 1
		*	amod	ADJ	Token 1
		NOUN	*	ADJ	Token 1
		NOUN	amod	*	Token 1
		NOUN	*	*	Token 1
		*	amod	*	Token 1
		*	*	ADJ	Token 1
with	zoom	ADP	pobj	NOUN	Token 2
			⋮		

After the execution of the rule generation process with all the sentences in the corpus, the next step was the selection of the best rules. For that, the rules were created using the class DependencyMatcher of the module spaCy in the language Python, and each one was executed in the training set.

We calculated, for each rule, the Laplace precision and support metrics. We tested rules with a support value between 1 and 60 combined with a Laplace precision ranging between 0.1 and 1, with an increment of 0.1. For each combination of Laplace precision and support, we selected the rules that met these requirements, ran them on the test suite, and calculated the precision, recall, and f-measure metrics. To select the best set of rules, we choose the rules that got the highest result for f-measure.

4. Results

Our first experiments should serve as baseline results. Although simple, we already have a first contribution with the analysis of cutoff frequencies. As mentioned, several authors use the value of 1% as a cutoff. We find a cutoff frequency between 0.40% and 0.53% as a more appropriate value. Using the pruning mechanism with Word2Vec, as in other works, led to an increase in accuracy and a decrease in recall. The identification of implicit aspects, as expected because of the characteristics of the method, was practically non-existent. The results are summarized in Table 3.

Table 3. Results for Frequency-based methods.

Domain	Method	% Cutoff	Precision	Recall	F-measure	% Explicits	% Implicits
Camera	Freq	0.53	0.57	0.57	0.57	35.16	0.00
Camera	Freq+W2V	0.53	0.72	0.56	0.63	34.07	0.00
Book	Freq	0.42	0.47	0.67	0.55	23.17	0.00
Book	Freq+W2V	0.42	0.70	0.59	0.64	15.85	0.00
Smartphone	Freq	0.40	0.46	0.53	0.49	27.61	4.17
Smartphone	Freq+W2V	0.40	0.56	0.47	0.51	23.88	0.00

As mentioned, the results of the experiment with the rules adapted from [Poria et al. 2014] did not reach satisfactory results. Therefore, we focused on finding new rules that would be more appropriate for the Portuguese language. The results are summarized in Table 4. For the camera domain, we got the best f-measure result with 26 rules selected by a minimum Laplace precision of 0.30 and support of 6. Regarding the book domain, we only found 1 rule selected by the Laplace precision of 0.40 and support of 21. Finally, in the smartphone domain, we found 3 rules selected by the Laplace precision of 0.10 and support of 18.

Table 4. Results for learned rules.

Domain	Laplace[1]	Support	Rules	Precision	Recall	F-measure	% Explicits	% Implicits
Camera	0.30	6	26	0.50	0.58	0.54	36.26	15.38
Book	0.40	21	1	0.81	0.54	0.65	13.41	20.00
Smartphone	0.10	18	3	0.61	0.42	0.50	20.90	4.17

Analyzing the learned rules, we could observe that most of the rules show nouns as aspects, changing only the classes and relationships with other terms.

We also selected the common rules between the domains, to find rules that were theoretically less susceptible to the context. Table 5 shows the rules that were found, and, as we can see, there was a greater number of rules found between the domains of cameras and smartphones because of the greater similarity between them. The common rules between smartphone and book were the same as found between camera, smartphone, and book. The first rule can be interpreted as: any token in a nominal subject relation with a noun that is in a determiner relation with a determiner.

Table 5. Common rules between domains.

Sets	Rule	Aspect
camera ∩ smartphone	['NOUN', 'nsubj'], '>', ['DET', 'det']	Token 1
	['NOUN', 'obj'], '>', ['DET', 'det']	Token 1
	['NOUN', 'conj'], '>', ['PUNCT', 'punct']	Token 1
camera ∩ smartphone ∩ book	['NOUN', 'nsubj'], '>', ['DET', 'det']	Token 1

The results were inferior to the previous experiment, as we can see in Table 6. There was no improvement, even in terms of accuracy, that could have increased with a smaller number of rules. This result was probably due to the simplicity of the found rules, which, as previously mentioned, always ended up selecting only nouns as aspects, similar to what occurs in the frequency-based methods.

Table 6. Results for rules common to all the domains.

Domain	Rules	Precision	Recall	F-measure	% Explicits	% Implicits
Camera	Camera ∩ Smartphone	0.49	0.31	0.38	15.38	0.00
Camera	Camera ∩ Smartphone ∩ Book	0.62	0.12	0.21	5.49	0.00
Book	Camera ∩ Smartphone	0.53	0.57	0.55	17.07	20.00
Book	Camera ∩ Smartphone ∩ Book	0.81	0.54	0.65	13.41	20.00
Smartphone	Camera ∩ Smartphone	0.61	0.42	0.50	20.09	4.17
Smartphone	Camera ∩ Smartphone ∩ Book	0.73	0.24	0.36	10.45	0.00

Finally, Table 7 presents a summary of the results got by each method. We can observe that only the rules we found could find implicit aspects in all domains. Considering f-measure, in the camera domain, the frequency-based method with Word2Vec was superior and, in the other domains, it was practically equivalent to the results of the rules, although it was not able to find implicit aspects.

Table 7. Results by implemented method.

Method	Camera		Book		Smartphone	
	F-measure	% Implicits	F-measure	% Implicits	F-measure	% Implicits
Freq	0.57	0.00	0.55	0.00	0.49	4.17
Freq + Word2Vec	**0.63**	0.00	0.64	0.00	**0.51**	0.00
Poria et al. rule-based	0.10	0.00	0.14	0.00	0.14	**7.35**
Corpus-based rules	0.54	**15.38**	**0.65**	**20.00**	0.50	4.17

5. Conclusion

Regarding the frequency-based methods, we could observe that, despite being simple, they achieved interesting results. With the study of the cutoff frequency, we could still get a significant improvement in these results. Using Word2Vec also proved to be effective in eliminating some aspects that were wrongly identified, a fact observed by the increase in precision. On the other hand, it ended up eliminating the few implicit aspects that had been identified.

The adaptation of the rules from English to Portuguese proved inefficient. Although at first the analysis of the examples of the rules proved promising, in the experiments, the results were below expectations. The method for finding rules in our last experiment has shown promise. Although we did not get a vast improvement in the f-measure, the improvement in implicit aspect detection was substantial. In future work, we intend to consider additional elements for analysis and creation of rules, such as sentiment lexicons, and to separately search for rules for explicit and implicit aspects.

The interested reader may find more information at the web portal of the POeTiSA (*POrtuguese processing - Towards Syntactic Analysis and parsing*) project[2].

Acknowledgements

---

[2]`https://sites.google.com/icmc.usp.br/poetisa`

References

Balage Filho, P. P. (2017). *Aspect extraction in sentiment analysis for portuguese language*. PhD thesis, Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, Brazil.

Cai, H., Tu, Y., Zhou, X., Yu, J., and Xia, R. (2020). Aspect-category based sentiment analysis with hierarchical graph convolutional network. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 833–843, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Cambria, E., Olsher, D., and Rajagopal, D. (2014). Senticnet 3: a common and common-sense knowledge base for cognition-driven sentiment analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 28.

Costa, R. W. M. and Pardo, T. A. S. (2020). Métodos baseados em léxico para extração de aspectos de opiniões em português. In *Anais do IX Brazilian Workshop on Social Network Analysis and Mining*, pages 61–72. SBC.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Honnibal, M., Montani, I., Van Landeghem, S., and Boyd, A. (2020). spaCy: Industrial-strength Natural Language Processing in Python.

Hu, M. and Liu, B. (2004). Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177. ACM.

Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*, 5(1):1–167.

López Condori, R. E. and Pardo, T. A. S. (2017). Opinion summarization methods: Comparing and extending extractive and abstractive approaches. *Expert Systems with Applications*, 78:124–134.

Machado, M. T., Pardo, T. A. S., and Ruiz, E. E. S. (2017). Analysis of unsupervised aspect term identification methods for portuguese reviews. *Anais do XIV Encontro Nacional de Inteligência Artificial e Computacional (ENIAC), SBC*, pages 239–249.

Marcacini, R. M., Rossi, R. G., Matsuno, I. P., and Rezende, S. O. (2018). Cross-domain aspect extraction for sentiment analysis: A transductive learning approach. *Decision Support Systems*, 114:70–80.

Medhat, W., Hassan, A., and Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams engineering journal*, 5(4):1093–1113.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. *arXiv preprint arXiv:1310.4546*.

Panchendrarajan, R., Ahamed, N., Murugaiah, B., Sivakumar, P., Ranathunga, S., and Pemasiri, A. (2016). Implicit aspect detection in restaurant reviews using cooccurence of words. In *Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 128–136.

Pavlopoulos, J. and Androutsopoulos, I. (2014). Aspect Term Extraction for Sentiment Analysis: New Datasets, New Evaluation Measures and an Improved Unsupervised Method. In *Proceedings of the 5th Workshop on Language Analysis for Social Media (LASM)@ EACL*, pages 44–52.

Pereira, D. A. (2020). A survey of sentiment analysis in the portuguese language. *Artificial Intelligence Review*, pages 1–29.

Poria, S., Cambria, E., Ku, L.-W., Senticnet, C. G., and Gelbukh, A. (2014). A Rule-Based Approach to Aspect Extraction from Product Reviews. In *Proceedings of the second workshop on natural language processing for social media (SocialNLP)*, pages 28–37.

Rana, T. A. and Cheah, Y.-N. (2016). Aspect extraction in sentiment analysis: comparative analysis and survey. *Artificial Intelligence Review*, 46(4):459–483.

Ravi, K. and Ravi, V. (2015). A survey on opinion mining and sentiment analysis: Tasks, approaches and applications. *Knowledge-Based Systems*, 89:14 – 46.

Taboada, M. (2016). Sentiment analysis: An overview from linguistics. *Annual Review of Linguistics*, 2:325–347.

Vargas, F. A. and Pardo, T. A. S. (2018). Aspect Clustering Methods for Sentiment Analysis. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 11122 LNAI:365–374.

Yadollahi, A., Shahraki, A. G., and Zaiane, O. R. (2017). Current state of text sentiment analysis from opinion to emotion mining. *ACM Computing Surveys (CSUR)*, 50(2):1–33.

Zhang, Y. and Zhu, W. (2013). Extracting implicit features in online customer reviews for opinion mining. In *Proceedings of the 22nd International Conference on World Wide Web*, pages 103–104.