Rude Waiter but Mouthwatering Pastries! An Exploratory Study into Dutch Aspect-Based Sentiment Analysis

Orphée De Clercq and Véronique Hoste
LT^3, Language and Translation Technology Team
Ghent University
Groot-Brittanniëlaan 45, 9000 Ghent, Belgium
firstname.lastname@ugent.be

Abstract

The fine-grained task of automatically detecting all sentiment expressions within a given document and the aspects to which they refer is known as aspect-based sentiment analysis. In this paper we present the first full aspect-based sentiment analysis pipeline for Dutch and apply it to customer reviews. To this purpose, we collected reviews from two different domains, i.e. restaurant and smartphone reviews. Both corpora have been manually annotated using newly developed guidelines that comply to standard practices in the field. For our experimental pipeline we perceive aspect-based sentiment analysis as a task consisting of three main subtasks which have to be tackled incrementally: aspect term extraction, aspect category classification and polarity classification. First experiments on our Dutch restaurant corpus reveal that this is indeed a feasible approach that yields promising results.

Keywords: user-generated content, aspect-based sentiment analysis, semantic processing

1. Introduction

In today's information society, it cannot be ignored that large parts of our lives are spent and shared online. The arrival of Web 2.0 sites allowed site visitors to add content, called user-generated content (Moens et al., 2014). Typical for user-generated content is that it contains a lot of subjective material. As the amount of online information has grown exponentially, so has the interest in new text mining techniques to handle and analyze this growing amount of subjective text.

One of the main research topics is sentiment analysis, also known as opinion mining. The objective of sentiment analysis is the extraction of subjective information from text, rather than factual information. Originally, it focused on the task of automatically classifying an entire document as positive, negative or neutral (Liu, 2012). More recently, the focus has shifted from coarse-grained to fine grained sentiment analysis, where sentiment has to be assigned at the clause level (Wilson et al., 2009).

Often, users are not only interested in people's general sentiments about a certain product, but also in their opinions about specific features, i.e. parts or attributes of that product. The task of automatically detecting all sentiment expressions within a given document and the concepts and aspects (or features) to which they refer is known as aspect-based or feature-based sentiment analysis, i.e. ABSA (Pontiki et al., 2014). Such systems do not only try to distinguish the positive from the negative utterances, but also strive to detect the target of the opinion, which comes down to a very fine-grained sentiment analysis task.

In this paper we present two Dutch domain-specific corpora annotated for ABSA and the first pipeline for automatically performing this task on Dutch customer reviews. In Section 2 we describe which two corpora were collected, one comprising restaurant reviews and another comprising smartphone reviews, and how these have been manually annotated using newly developed guidelines that comply to standard practices in the field. In Section 3 we present our pipeline which consists of three incremental steps: aspect term extraction, aspect category classification and aspect polarity classification. For each step we report results of first experiments that were performed on our restaurants dataset. Section 4 concludes this paper and offers prospects for future work.

2. Dutch ABSA corpora

Aspect-based sentiment analysis has proven important for mining and summarizing opinions from online reviews (Gamon et al., 2005; Titov and McDonald, 2008; Pontiki et al., 2014). Several benchmark datasets have been made publicly available, such as the product reviews dataset of Hu and Liu (2004) or the restaurant reviews dataset of Ganu et al. (2009). More recently, parts of these two datasets were extracted and re-annotated for two SemEval shared tasks on aspect-based sentiment analysis (Pontiki et al., 2014; Pontiki et al., 2015). For Dutch, to our knowledge, no such benchmark datasets exist.

2.1. Data collection

We created two domain-specific corpora: one comprising restaurant (REST) and another comprising smartphone reviews (SMART). We are pleased to inform that both datasets have been made available for research purposes in the framework of SemEval 2016 task 5, the focus of which is multilingual ABSA (Pontiki et al., 2016).

All reviews were crawled from online user platforms. For the restaurant reviews we relied on TripAdvisor[1] and for the smartphone reviews on the online store Bol.com[2]. On both platforms reviews can only be submitted after having created a personal user profile, but every review is publicly available. In total, 670 reviews were crawled: 400 restaurant reviews and 270 smartphone reviews.

[1] www.tripadvisor.nl
[2] www.bol.com

	MAIN	ATTRIBUTES
REST	Ambience, Drinks, Food, Location, Restaurant, Service	*General, Prices, Style & Options, Quality, Miscellaneous*
SMART	Battery, Company, CPU, Devices, Display, Hard Disc, Hardware, Keyboard, Memory, OS, Ports, Power supply, Shipping, Smartphone, Software, Support	*Connectivity, Design & Features, General, Operation &Performance, Miscellaneous, Price, Quality, Usability*

Table 1: Main aspect categories and attributes within each domain.

2.2. Annotations

In order to create a gold-standard suitable for aspect-based sentiment analysis, annotation guidelines had to be developed. In order to ensure consistency in the field, we relied on the SemEval 2014 guidelines for the restaurant reviews corpus. The smartphone reviews were annotated in close collaboration with the SemEval 2016 shared task organizers. Consequently, the Dutch guidelines form an integral part of the ABSA 2016 guidelines[3].

The guidelines allow to distinguish the different aspects related to a restaurant visit or the purchase of a smartphone, viz. the opinions expressed towards specific entities (e.g. *pizza margarita, gin tonic, iPhone6*) and/or their attributes (e.g. *quality, price, design*) and the polarity expressed towards each of these aspects.

To be more precise, every review was first split into sentences and a sentence was only further annotated with aspect terms, categories and polarity when subjective. If we consider the following examples:

(1) Dit is het vuilste, slechtste restaurant ooit!
EN: This is the dirtiest, worst restaurant ever!

(2) Gaan eten bij Mama Mia deze namiddag.
EN: Went to lunch at Mama Mia this afternoon.

In the first example sentiment is definitely expressed whereas the second sentence is clearly just a factual sentence, which means that only Example 1 would receive further annotations.

The annotation process itself consists of three incremental steps. First, all targets, also known as **aspect expressions**, are indicated. The target is the word or the words referring to a specific entity or aspect (typically nouns, named entities or multi word expressions such as *restaurant, La Cucina, battery life, display, ...*). Important to note is that only explicit aspect expressions are indicated as targets. Implicit referrals or when a pronoun is used to refer to a target can also evoke an aspect, and are annotated as 'NULL' targets.

Once all explicit and implicit aspect terms have been indicated, they are assigned to predefined clusters or **aspect categories**. These categories each time consist of a main category (e.g. Food, Ambience, Service, Battery, Hardware, Phone,...) together with a more specific attribute (e.g. *Quality*, *Price*). In both domains different main categories and attributes are defined, see Table 1 for an overview. For a more detailed description and an overview of which main–attribute combinations are possible we refer to the previously mentioned guidelines.

In the third and final step of the annotation process, the **polarity** of the sentiment expressed towards every annotated aspect expression/category is indicated. Three main polarities are distinguished: positive, neutral and negative. The neutral label applies to mildly positive or negative sentiment or when two opposing sentiments towards one feature expression occur within one sentence.

Annotations were performed using BRAT[4], the brat rapid annotation tool (Stenetorp et al., 2012). It takes UTF8-encoded text files as input, and stores the annotations in a proprietary standoff format. An example is presented below: in the first step of the annotation process the aspect expression *wachttijd* has been indicated, next it was assigned to the category *Service–General*, and finally the polarity of the sentiment expressed towards this aspect has been labeled as *negative*.

Figure 1: ABSA annotation using BRAT.
EN: Long waiting time.

2.3. Annotation Statistics

All 670 reviews have been manually annotated by a trained linguist. These annotations were verified by another linguist and disagreement was resolved through mutual consent. Our participation as Dutch data providers in the SemEval2016 ABSA task allowed for all data to be checked on inconsistencies one final time (Pontiki et al., 2016).

Data statistics for both corpora are presented in Table 2 and in the next sections we will have a closer look at the actual annotations.

Domain	# revs	# sents	# toks
REST	400	2297	28289
SMART	270	1697	23444

Table 2: Data statistics of the two domain-specific corpora.

2.3.1. Restaurant domain

Out of the 2297 sentences, 76% (n = 1767) were considered as subjective, whereas 24% (n = 530) as not opinionated at all. The opinionated sentences were further anno-

[3]Guidelines available at http://goo.gl/wOf1dX

[4]Available at http://brat.nlplab/org

tated. In total, 2445 aspect expressions were annotated, ranging from sentences including one to twelve individual expressions. If we consider the six main aspect categories of the restaurants domain, we notice that three are mentioned more often, i.e. *Food*, *Restaurant* and *Service*, as visualized in Figure 2.

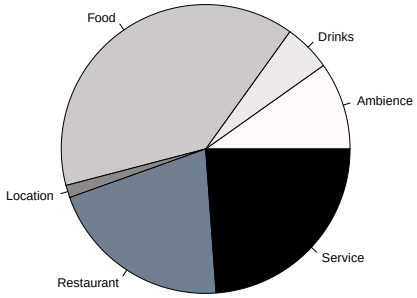

Figure 2: Main category distribution in the REST corpus.

When we investigate the distribution of the opinions expressed towards each of the main features, as visualized in Figure 3, we clearly notice that in our corpus there are overall more positive opinions. Especially when people refer to more general aspects such as the *Ambience* (70% positive) in a restaurant or the *Location* (71% positive), people tend to make positive remarks in our corpus. Only for the aspect category *Service* do we observe a different tendency in that slightly more negative (49%) than positive (45%) sentiments are expressed.

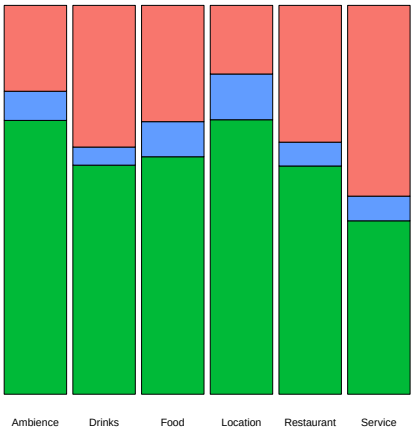

Figure 3: Heat barplots visualizing the amount of positive (green), neutral (blue) and negative (red) opinions expressed within each main REST category.

2.3.2. Smartphone domain

Out of the 1697 sentences, 75.5% (n = 1281) were considered as subjective, whereas 24.5% (n = 416) as not opinionated at all. Which is in line with the REST corpus. In total, 1281 targets were annotated, ranging from sentences including one to seven individual targets.

In the SMART domain no less than sixteen main aspect categories could be indicated. If we look at the main category distribution, however, we notice that six categories are referred to most often, i.e. *Battery*, *Display*, *Multimedia devices*, *OS*, *Phone* and *Software*. With an absolute maximum of 989 references to the main category *Phone*, i.e. 77% of all annotated targets. This is visualized in Figure 4. In this pie chart the category *Other* comprises the ten other main smartphone categories, which are much more specific and, as a consequence, also more sparse in our corpus.

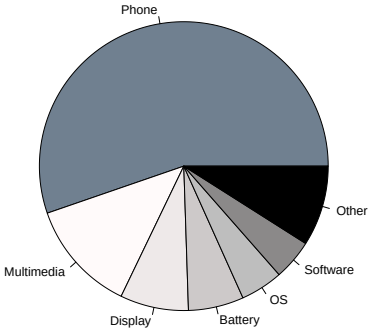

Figure 4: Main category distribution in the SMART corpus.

When we investigate the distribution of the opinions expressed towards each of the main categories, as visualized in Figure 5, we clearly notice that in this dataset most of the time positive sentiments are expressed by the users. Overall, there are a lot more positive utterances in the SMART corpus.

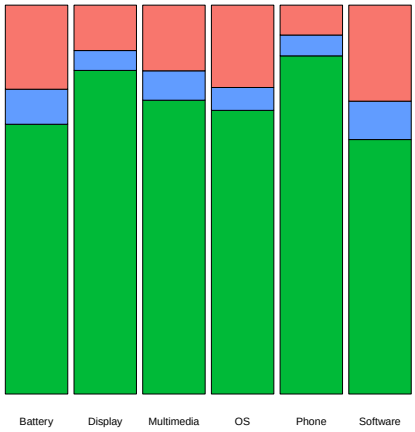

Figure 5: Heat barplots visualizing the amount of positive (green), neutral (blue) and negative (red) opinions expressed within the most frequent main SMART categories.

Especially when people refer to more general aspects such as the *Phone* itself or design features (*Display*), they express very positive thoughts, respectively 87% and 83% of positive utterances. When it comes to describing more specific features such as the *OS* or *Software* slightly more negative sentiments are expressed, i.e. 22% in both cases.

Intuitively, the difference in sentiment between both domains could be explained by the amount of premeditation preceding both actions. Choosing a restaurant is probably a much more spontaneous decision than buying a new (and often expensive) smartphone.

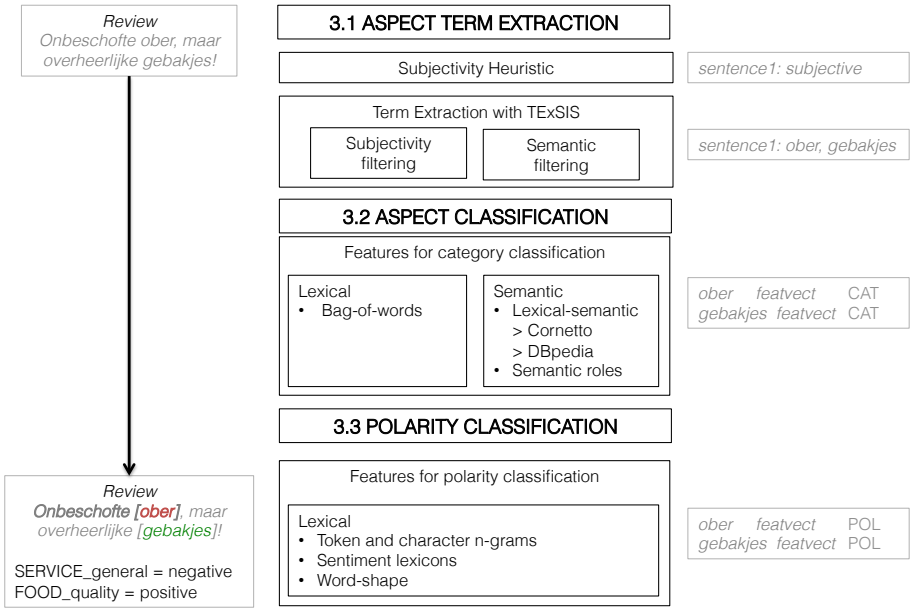

Figure 6: Architecture of our Dutch ABSA pipeline illustrated with an example.
EN: Rude waiter, but mouthwatering pastries!

3. ABSA pipeline

Following Pavlopoulos and Androutsopoulos (2014) and the SemEval ABSA subtask classification (Pontiki et al., 2014; Pontiki et al., 2015), we discern three individual subtasks when it comes to performing aspect-based sentiment analysis automatically: aspect term extraction (Section 3.1), aspect term aggregation (Section 3.2) and aspect term polarity estimation (Section 3.3). Following similar experiments we performed on English data (De Clercq et al., 2015), we present a pipeline for Dutch which tackles these three subtasks incrementally.

Figure 6 visualizes the architecture that was developed in order to perform the task of aspect-based sentiment analysis. In this section we describe the pipeline in closer detail and report results on first experiments that were performed on the REST corpus. To this purpose, the corpus was split in a development set comprising 300 reviews and a held-out test set comprising 100 reviews, data statistics are presented in Table 3 below.

Datasets	# revs	# sents	# toks
Development	300	1722	24894
Held-out	100	575	7652

Table 3: Experimental datasets statistics.

3.1. Aspect Term Extraction

The first part of an ABSA system requires that candidate terms are automatically extracted, these terms are typically nouns, noun phrases or multiword expressions and they should be related to the restaurants domain. Moreover, terms can only be extracted when they are part of a subjective statement. In order to determine this subjectivity, we performed a lexicon (Jijkoun and Hofmann, 2009) lookup on both the surface forms and lemmas. To this purpose, all reviews were first linguistically preprocessed using the LeTs Preprocess toolkit (Van de Kauter et al., 2013). Based on the resulting tokens and lemmas the lookup was performed. We only proceeded to the next steps when subjectivity was found.

3.1.1. Information sources

To extract candidate terms, we applied the hybrid terminology extraction system TExSIS (Macken et al., 2013). Whereas TExSIS was developed as a generic terminology-extraction system, we used a reduced version focussing mainly on the linguistic noun phrase extraction (see De Clercq et al. (2015) for similar experiments on English). In an additional step, we applied domain-specific heuristics for subjectivity and semantic filtering. For the former, we relied on the same subjectivity lexicon as mentioned above, and for the latter we relied on Cornetto (Vossen et al., 2013) and DBpedia (Lehmann et al., 2013). Let us consider the following example:

(3) Na een goede aperitief bestelde ons mama een pizza margherita, die was heerlijk!
EN: After a good appetizer our mother ordered a pizza margherita, which was divine!

In this sentence the TExSIS system will indicate *good appetizer*, *our mother* and *pizza margherita* as candidate terms. After subjectivity filtering, the positive word *good* will be stripped, leaving the term *appetizer*. The semantic filtering using Cornetto and DBpedia will lead to the conclusion that *pizza margherita* has more semantic links with the restaurant domain than the term *our mother*, which means that in the end only *appetizer* and *pizza margherita* will and should be extracted as aspect terms by our system.

3.1.2. Experiments and results

We evaluated the performance of the aspect term extraction by comparing the TExSIS system as such (TExSIS) with a system where subjectivity filtering was also included (TExSIS + subj) and a system where both subjectivity and semantic filtering were included (TExSIS + subj + sem). In order to perform these experiments, the development set of 300 reviews was split in a 250 document train set (devtrain) and a 50 document test set (devtest). To evaluate we calculated precision, recall and F-1. Finally, the best setting was tested on the held-out test set.

	Precision	Recall	F-1
TExSIS	24.78	39.61	30.48
TExSIS + subj	29.15	**66.18**	40.47
TExSIS + subj + sem	**37.85**	59.42	**46.24**
Held-out	35.87	58.18	44.38

Table 4: Results of the ATE experiments.

The results in Table 4 show that the subjectivity filtering improves especially the recall, whereas the semantic filtering is better for precision. The best overall F-1 is achieved with both filters (TExSIS + subj + sem). This setting was thus used to test on the held-out data. The results on this held-out test set are lower. Represented in absolute numbers, our held-out test set contains 373 explicit aspect term expressions, of which 217 were found by our system, leading to the recall of 58.18%. In total, however, our system predicted 605 explicit target mentions, leading to a moderate precision of 35.87%.

These results underline the difficulty of this first step, extracting the correct aspect terms is a challenging task and it would definitely benefit from additional optimization experiments. That is why for the next two subtasks, we decided to rely on gold-standard aspect terms, allowing us to focus more on optimizing the following two classification tasks (Section 3.2 and 3.3).

3.2. Aspect Category Classification

The second step in an ABSA system consists in classifying the list of possible candidate terms into broader aspect categories. For the restaurants domain, this comes down to a fine-grained multiclass classification task. If we look back at Table 1, we see that there exist six main categories and five attributes. With our system we aim to classify these in one go, which corresponds to classifying 13 different labels.[5]

3.2.1. Information sources

Such a fine-grained classification task requires a system that is able to grasp subtle differences between the various categories (e.g. *Food–General* versus *Food–Prices* versus *Food–Quality* versus *Food–Style&Options*). To create such a system we first of all extracted typical lexical **bag-of-words** features based on the sentence in which an aspect term occurs.

An analysis of the top-performing system of the SemEval2015 Task 12, however, revealed that besides lexical features, features in the form of clusters derived from a large reference corpus of restaurant reviews and thus capturing semantic information, are very useful (Toh and Su, 2015). For Dutch, we did not have such a large reference corpus available, but we did derive other semantic features. Based on the two semantic information sources that we also used for the semantic filtering of our TExSIS system, i.e. Cornetto and DBpedia, we derived various **lexical-semantic** features. To be more precise, for Cornetto this translated to six features, each representing a value indicating the number of (unique) terms annotated as aspect terms from that category that (1) co-occur in the synset of the candidate term or (2) which are a hyponym/hypernym of a candidate term in the synset. For DBpedia, this translated to eighteen unique DBpedia categories that were included as binary features, each of which could be used to generalize to the restaurant domain (e.g. DB_food, DB_gastronomy,...).

We also introduced a novel type of semantic information to our system, namely **semantic role** features. We hypothesized that additional information about agents and entities' semantic roles could provide semantic evidence with regard to the more fine-grained labels. In this respect, the predicates evoking certain roles, for example, should constitute an added value on top of the bag-of-words features when it comes to discerning the different attributes (e.g. *The food **tasted** good* vs *The food just **cost** too much*). To this purpose every review was processed with SSRL (De Clercq et al., 2012). For the semantic role feature construction, we retrieved the position of every aspect term and derived whether it evokes a semantic role or not. This information was stored in 19 binary features, each representing a possible semantic role label. The predicate token itself was also included as a separate feature.

3.2.2. Experiments and results

Again, we evaluated the performance of our aspect category classification by comparing different runs of our system and gradually adding more feature information: first bag-of-words (bow) alone were used, then the lexical-semantic (lexsem) features were added, and finally also the novel semantic role features (srl). Again the development set was used for testing which setting yielded the best result (by each time performing ten-fold cross validation experiments), and this best setting was then tested on the held-out test set. For all experiments we used LibSVM[6] as our classifier.

Important to note is that we performed two different rounds of experiments. In *Round 1* the added value of adding more features was empirically verified by gradually adding more features to our system. In *Round 2*, however, we also wanted to take into account the different interplays between features and investigate whether changing LibSVM's hyperparameters might also influence the performance. To this purpose we applied genetic algorithms (Desmet et al., 2013) to jointly search the hyperparameter and feature space (Joint optimization). For the features we looked once

[5] For the exact combinations of all main–attribute combinations we refer to the guidelines.

[6] Available at http://www.csie.ntu.edu.tw/cjlin/libsvm

at the level of feature groups (feat groups) and once at the individual features (indfeats). For our baseline, consisting solely of the bag-of-words features, only hyperparameter optimization was performed in *Round 2*. For all experiments, we report classification accuracy.

	Round 1	*Round 2*	
bow	53.28	54.69	
		Joint optimization	
		featgroups	indfeats
bow + lexsem	**60.72**	**62.94**	63.16
bow + srl	54.80	56.16	56.70
bow + lexsem + srl	60.01	62.89	**63.27**
Held-out			**66.42**

Table 5: Results of the aspect category classification experiments.

As shown in Table 5, both semantic information sources improve the performance when compared to the bag-of-words baseline. Whereas the semantic role features allow for a mild improvement of 1.47 points in the *Round 1* experiments, the lexical-semantic Cornetto and DBpedia features allow for an improvement of 7 points.
In the *Round 2* experiments, we go from a best score of 60.72 using the default settings and only the lexical-semantic features to one of 63.27 where both the hyperparameters and all semantic features have been optimized individually. The best overall results are achieved with the individual feature selection experiments. In this best setup, both lexical-semantic and semantic role features are included, resulting in an accuracy of 63.27. Contrasted with the other results, however, we can conclude that the added value of including lexical-semantic features is more outspoken. In a final step, the optimal setting was used to retrain the classifier on the development set and test it on the held-out test set, which leads to an accuracy of 66.42%, a promising result.

3.3. Polarity Classification

This brings us to the final step of our ABSA pipeline. Given a list of possible candidate terms and given that these were classified into one of the aspect categories, the final step consists in classifying the polarity expressed towards these aspects into one of the three possible polarity labels (positive, negative or neutral). We developed a first prototype of such a classifier for English in the framework of SemEval 2014 Task 9 (Van Hee et al., 2014) and it has also proven effective for this third subtask in an ABSA setting (De Clercq et al., 2015). We adapted the system to deal with Dutch text for the research presented here.

3.3.1. Information sources

We performed the three-way classification task by relying solely on the following lexical features:

Token and character n-gram features: binary values for each token unigram, trigram and bigram, as well as character n-gram features for each character trigram and fourgram in the training data.

Sentiment Lexicon features: the number of positive, negative and neutral words extracted from the Dutch Duoman (Jijkoun and Hofmann, 2009), and Pattern lexicons (De Smedt and Daelemans, 2012), all averaged over sentence length, as well as the sum of the polarity scores of all detected sentiment words.
Word-shape: numeric and binary features capturing the characteristics of a review sentence, such as features which indicate character or punctuation flooding in a review as this might hint at intense sentiment, e.g. 'cooooooool !!!!!'. We furthermore check whether the last token contains punctuation and count how many capitalized tokens are present within one sentence.

3.3.2. Experiments and results

For the experiments presented here, we again first optimized on the development data, after which this optimal setting was tested on the held-out test data. We performed ten-fold cross validation experiments on the development data using LibSVM and evaluated by calculating accuracy. In order to derive the optimal settings, we compared a setting with all features and the default LibSVM settings using a linear kernel (Default) to a setting where both the parameters and features were jointly optimized using genetic algorithms (Joint optimization).

	Default	Joint optimization
All features	76.40	**79.06**
Held-out		**81.23**

Table 6: Results of the polarity classification experiments.

From the results of Table 6, we observe that our system using only lexical features benefits from this joint optimization and goes from an accuracy of 76.40 to one of 79.06. For the experiments on our held-out test data, we achieve a top accuracy of 81.23.

4. Conclusion and Future Work

In this paper, we presented two Dutch domain-specific corpora annotated for the task of aspect-based sentiment analysis and the first pipeline for automatically performing this task on Dutch customer reviews. We first described how two domain-specific corpora have been collected, one comprising restaurant reviews and another comprising smartphone reviews, and how these have been manually annotated using newly developed guidelines that comply to standard practices in the field. We explained how both corpora differ in that different aspect categories have to be assigned. Moreover, the annotation statistics revealed that even though the sentiment expressed in both corpora is mostly positive, the proportion of positive aspects is much higher in the smartphone corpus. A possible explanation for this might be that purchasing a new smartphone is a much more informed decision than choosing a restaurant to have lunch or dinner.
In the second part of this paper, we explained in close detail the three different steps of our proposed ABSA pipeline: aspect term extraction, aspect category classification and

polarity classification. For the first step we applied a basic terminology extraction system and added two types of filtering: subjectivity filtering and semantic filtering. We revealed that although this filtering already helps to improve the performance. It remains a challenging task. On our held-out test set we achieve a moderate performance. We believe that further optimizing our terminology extraction system and adding other filtering techniques to better recognize domain-specific terms are interesting avenues for future work.

For the second step of aspect category classification, we incorporated two additional semantic information sources into our classifier, i.e. lexical-semantic information and semantic role information. We showed that performing optimization experiments using genetic algorithms in which our classifier's hyperparameters and the individual semantic features are jointly optimized leads to the best results. At the same time we noticed that especially the lexical-semantic features contributed to this added performance. On our held-out test set we achieved a promising accuracy of 66.61.

For the final step, polarity classification, we revealed that relying on a classifier using lexical features alone leads to good results, including a top accuracy of 81.23 on our held-out test set.

We are fully aware that relying on gold-standard aspect terms for performing both the aspect category and polarity classification tasks is an artificial setting. That is why for our future work we aim to improve the first step of aspect term extraction. This should enable us to evaluate the fully-automatic ABSA pipeline in one go without worrying about disproportionate error percolation. Other future work includes retraining and testing our pipeline on the smartphone corpus and comparing cross-domain performance. Also performing the ABSA task on multiple languages is something we would like to further explore in future work.

5. Acknowledgements

The work presented in this paper has been partly funded by the PARIS project (IWT-SBO-Nr. 110067). We would like to thank our annotators and the task organizer of SemEval 2016 Task 5 for their valuable help in annotating and double-checking the restaurant and smartphone reviews.

6. Bibliographical References

De Clercq, O., Monachesi, P., and Hoste, V. (2012). Evaluating automatic cross-domain semantic role annotation. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC-2012)*, pages 88–93, May.

De Clercq, O., Van de Kauter, M., Lefever, E., and Hoste, V. (2015). LT3: Applying hybrid terminology extraction to aspect-based sentiment analysis. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval-2015)*, pages 719–724.

De Smedt, T. and Daelemans, W. (2012). Vreselijk mooi! Terribly beautiful: a subjectivity lexicon for Dutch adjectives. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC-2012)*, pages 3568–3572.

Desmet, B., Hoste, V., Verstraeten, D., and Verhasselt, J. (2013). Gallop Documentation. Technical Report LT3 13-03, Ghent University.

Gamon, M., Aue, A., Corston-Oliver, S., and Ringger, E. (2005). Pulse: Mining customer opinions from free text. In *Proceedings of the 6th International Conference on Advances in Intelligent Data Analysis (IDA-2005)*, pages 121–132.

Ganu, G., Elhadad, N., and Marian, A. (2009). Beyond the stars: improving rating predictions using review text content. In *Proceedings of the 12th International Workshop on the Web and Databases (WebDB-2009)*, pages 1–6.

Hu, M. and Liu, B. (2004). Mining and summarizing customer reviews. In *Proceedings of the 10th International Conference on Knowledge Discovery and Data Mining (KDD-2004)*, pages 168–177.

Jijkoun, V. and Hofmann, K. (2009). Generating a non-English subjectivity lexicon: Relations that matter. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL-2009)*, pages 398–405.

Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P. N., Hellmann, S., Morsey, M., van Kleef, P., Auer, S., and Bizer, C. (2013). DBpedia – A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia. *Semantic Web Journal*, 6:167–195.

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

Macken, L., Lefever, E., and Hoste, V. (2013). TExSIS: Bilingual Terminology Extraction from Parallel Corpora Using Chunk-based Alignment. *Terminology*, 19(1):1–30.

Moens, M.-F., Li, J., and Chua, T.-S. (2014). *Mining user generated content*. Chapman and Hall/CRC.

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-2014)*, pages 44–52.

Pontiki, M., Galanis, D., Pavlopoulos, J., Papageorgiou, H., Androutsopoulos, I., and Manandhar, S. (2014). Semeval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval-2014)*, pages 27–35.

Pontiki, M., Galanis, D., Papageorgiou, H., Manandhar, S., and Androutsopoulos, I. (2015). Semeval-2015 task 12: Aspect based sentiment analysis. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval-2015)*, pages 486–495.

Pontiki, M., Galanis, D., Papageorgiou, H., Androutsopoulos, I., Manandhar, S., AL-Smadi, M., Al-Ayyoub, M., Zhao, Y., Qin, B., De Clercq, O., Hoste, V., Apidianaki, M., Tannier, X., Loukachevitch, N., Kotelnikov, E., Bel, N., Jiménez-Zafra, S. M., and Eryiğit, G. (2016). SemEval-2016 task 5: Aspect based sentiment analysis. In *Proceedings of the 10th International Workshop on*

Semantic Evaluation.

Stenetorp, P., Pyysalo, S., Topic, G., Ohta, T., Ananiadou, S., and Tsujii, J. (2012). BRAT: a web-based tool for NLP-assisted text annotation. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL-2012)*, pages 102–107.

Titov, I. and McDonald, R. (2008). A joint model of text and aspect ratings for sentiment summarization. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics (ACL-2008)*, pages 308–316.

Toh, Z. and Su, J. (2015). NLANGP: Supervised machine learning system for aspect category classification and opinion target extraction. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval-2015)*, pages 496–501, June.

Van de Kauter, M., Coorman, G., Lefever, E., Desmet, B., Macken, L., and Hoste, V. (2013). LeTs Preprocess: The multilingual LT3 linguistic preprocessing toolkit. *Computational Linguistics in the Netherlands Journal*, 3:103–120.

Van Hee, C., Van de Kauter, M., De Clercq, O., Lefever, E., and Hoste, V. (2014). LT3: Sentiment classification in user-generated content using a rich feature set. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval-2014)*, pages 406–410, August.

Vossen, P., Maks, I., Segers, R., van der Vliet, H., Moens, M., Hofmann, K., Sang, E. T. K., and de Rijke, M. (2013). Cornetto: a lexical semantic database for Dutch. In Peter Spyns et al., editors, *Essential Speech and Language Technology for Dutch, Theory and Applications of Natural Language Processing*, pages 165–184. Springer.

Wilson, T., Wiebe, J., and Hoffman, P. (2009). Recognizing contextual polarity: an exploration of features for phrase-level sentiment analysis. *Computational Linguistics*, 35(3):399–433.