Towards an integrated pipeline for aspect-based sentiment analysis in various domains

Orphée De Clercq[1], **Els Lefever**[1], **Gilles Jacobs**[1], **Tijl Carpels**[2] **and Véronique Hoste**[1]
[1] LT[3], Language and Translation Technology Team, Ghent University, Belgium
[2] Hello Customer, Belgium
{orphee.declercq, els.lefever, gillesm.jacobs, veronique.hoste}@ugent.be
tijl@hellocustomer.com

Abstract

This paper presents an integrated ABSA pipeline for Dutch that has been developed and tested on qualitative user feedback coming from three domains: retail, banking and human resources. The two latter domains provide service-oriented data, which has not been investigated before in ABSA. By performing in-domain and cross-domain experiments the validity of our approach was investigated. We show promising results for the three ABSA subtasks, aspect term extraction, aspect category classification and aspect polarity classification.

1 Introduction

With the rise of web 2.0 applications, customers have been given a new platform to express their opinions in the form of reviews on designated websites. At the same time many companies proactively collect direct customer feedback after an interaction, such as a store visit, a client meeting or online purchase. Both information types have in common that besides quantitative data ("How would you rate the overall shopping experience on a scale from one to ten") also qualitative data ("Why did you assign this score") is being collected. A fine-grained analysis of this qualitative textual feedback offers companies valuable detailed insights into the strong and weak aspects of their products and services and allows them to strengthen their offer.

Extracting this information automatically is known as the task of aspect-based sentiment analysis (ABSA). ABSA systems (Pontiki et al., 2014) focus on the detection of all sentiment expressions within a given document and the concepts and aspects (or features) to which they refer. 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 and "almost all real-life sentiment analysis systems in industry should be based on this level of analysis" (Liu, 2015, p10).

This fine-grained sentiment analysis task received special attention in the framework of three SemEval shared tasks: SemEval 2014 Task 4 (Pontiki et al., 2014) and SemEval 2015 Task 12 (Pontiki et al., 2015), which focussed on English customer reviews, and SemEval 2016 Task 5 (Pontiki et al., 2016) where seven other languages were also included. Each time the idea was to perform three subtasks: (i) extract all aspect expressions of the entities, (ii) categorize these aspect expressions into predefined categories and (iii) determine whether an opinion on an aspect is positive, negative or neutral.

In this paper, we discuss a fine-grained sentiment analysis pipeline to deal with qualitative Dutch feedback data coming from three different domains: banking, retail, and human resources. This paper presents a collaboration between academia and industry to create a proof-of-concept, the pipeline is currently in production at Hello Customer. In the framework of the SemEval shared tasks, similar methodologies have been investigated, but the research presented here differs in two ways. First, the main focus has always been on customer reviews of experiences (restaurants, hotels, movies) or tangible products (laptops, smartphones). Besides product-oriented data, we move towards more service-oriented data coming from financial institutions and human resources agencies. Second, the various ABSA subtasks have always been tackled and evaluated separately in the framework of SemEval. In reality, however, all steps have to be performed sequen-

Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pages 136–142
Copenhagen, Denmark, September 7–11, 2017.

tially, entailing error percolation from one step to the other. In this paper we present such an integrated pipeline for each domain and also perform cross-domain experiments.

The remainder of this paper is organized as follows. Section 2 describes the data we have collected and annotated. Next, in Section 3 we present the pipeline that has been developed for performing this task and in Section 4 we discuss the results. We end this paper with a conclusion and suggestions for future work.

2 Datasets and Annotations

In the past, ABSA datasets have been annotated comprising movie reviews (Thet et al., 2010), reviews for electronic products(Hu and Liu, 2004; Brody and Elhadad, 2010), and restaurant reviews (Brody and Elhadad, 2010; Ganu et al., 2009). As mentioned above, in the framework of three SemEval shared tasks (Pontiki et al., 2014, 2015, 2016), several benchmark review datasets coming from various domains (electronics, hotels, restaurants, and telecom) and languages (English, Dutch, French, Arabic, Chinese, Spanish, Turkish and Russian) have been made publicly available.

For the work presented here, direct customer feedback data written in Dutch was collected in three domains: banking, retail and human resources (HR). The data provider for the first domain, **banking**, is a large Belgian financial institution offering basic financial products (e.g. loans, insurances) and services (e.g. investing or financial advice). The second domain, **retail**, comprises data coming from a large clothing company with offline stores all over Belgium and an online webshop. Data for the third domain, **HR**, comes from two data providers who are active in the recruiting sector, namely employment agencies.

For all domains, data was collected by asking customers two things: (i) assign a NPS score[1] to the company and (ii) provide textual feedback for this score. This feedback is referred to as a *verbatim*, which can vary from one short sentence to various sentences discussing various aspects. Table 1 presents an overview of all data that has been collected and annotated in the three domains, expressed in number of verbatims and tokens.

[1]Net Promotor Score, a customer loyalty business metric. Customers are asked: How likely is it that you would recommend [company] to a friend or colleague? Trademark of Bain & Company, Inc and Fred Reichheld.

Domain	# verbatims	# tokens
Banking	1700	15870
Retail	1500	15796
HR	1000	11960

Table 1: Verbatims and tokens in each domain.

For the actual annotations, see Figure 1 for a visualization, we annotated each aspect term and assigned it to a predefined aspect category (CatEx). These aspect categories are domain-dependent and consist of a main category (e.g. Personnel) and subcategory (e.g. quality)[2]. For banking there are 22 such possible combinations, for retail 24 and for HR 23. Table 2 gives an overview of the three largest main categories per domain.

In a next step, sentiment bearing words were selected, assigned a polarity: positive, negative or neutral (OpinEx), and linked to the appropriate aspect term (is_about arrow). All annotations were carried out with the BRAT rapid annotation tool (Stenetorp et al., 2012).

Figure 1: Annotation (*EN: Friendly service*).

For all three domains, we went through the same annotation process to ensure consistency. First, a preliminary aspect category typology was devised after which 50 verbatims were annotated by two annotators independently from each other. These annotations were discussed, inconsistencies were resolved and the typology was altered, if necessary. Next, an inter-annotator agreement study was conducted on 50 new verbatims, which were again annotated by two independent annotators. The annotations were compared to the annotations of a third, more experienced annotator who also received more time to complete the task. Accuracy was calculated on two levels: the consistency of the annotated **category** expressions (cat) and the consistency of the annotated **polarity** expressions (pol).

As can be observed in Table 3, the IAA was high for all three domains. For the remainder of the annotation work, the same two annotators performed all annotations and frequently checked and discussed their work to ensure consistency.

[2]We were inspired by the SemEval ABSA annotation guidelines available at `http://bit.ly/2t0EkaB`.

Domain	cat	#	cat	#	cat	#
Banking	BANK	317	PERSONNEL	903	PRODUCT	168
Retail	STORE	306	PERSONNEL	682	COLLECTION	1191
HR	HR	129	PERSONNEL	637	SERVICES	230

Table 2: Typology of the three main aspect categories and occurrences per domain.

	Banking		**Retail**		**HR**	
	cat	pol	cat	pol	cat	pol
Annot 1	94	94	92	94	94	96
Annot 2	86	98	97	97	93	97

Table 3: IAA, expressed in accuracy (%).

3 Methodology

A pipeline was developed in order to perform the three incremental ABSA subtasks relying on supervised machine learning techniques. For the actual experiments, all datasets were split in a 90% train and a 10% held-out test set.

3.1 Aspect Term Extraction

Approaching the task of aspect term extraction as a sequential IOB labeling task has proven most successful (Liu, 2012). The two systems achieving top performance on English reviews for SemEval 2015 were a classifier using Conditional Random Fields (CRF) (Toh and Su, 2015) and a designated Named Entity Recognizer (San Vicente et al., 2015). Both systems implemented typical named entity features, such as word bigrams, trigrams, token shape, capitalization, name lists, etc. For SemEval 2016, subsequent work by Toh and Su (2016) found that using the output of a Recurrent Neural Network as additional features is beneficial for the labeling tasks.

We relied on a sequential IOB labeling approach using CRF as implemented in CRFSuite (Okazaki, 2007). For each token, and its two neighbouring tokens, the following features were extracted: (1) **token shape features**, based on whether the token contains capitalization, digits, or exclusively alphanumeric characters, as well as the final two and three characters as an approximate suffix; (2) **lemma**, (3) CGN **part-of-speech** (PoS) tag, (4) **syntactic chunk**, and (5) **Named Entity** label as provided by the LeTs preprocessing toolkit (Van de Kauter et al., 2013). Both full labels and coarse super-category for PoS, chunk, and NE labels were included as features.

For the experiments, CRF models with the LBFGS (Nocedal, 1980) optimization function were first trained on each domain separately and, next, all training data was combined, leading to four models in total. Hyper-parameters were optimized by randomized search with 500 iterations in 10-fold cross-validation. The models with winning hyper-parameters as determined by flat F1-score (weighted macro-averaging) were subsequently tested on the held-out test sets in three setups: in-domain (e.g. trained on banking and tested on banking), cross-domain (e.g. trained on banking and tested on retail) and all domain (e.g. trained on all training data and tested on banking).

To evaluate, we calculated flat (i.e. non-sequence) precision, recall, and F1-scores.

3.2 Aspect Category Classification

The aspect category classification subtask requires a system able to label a large variety of classes, in our case 22, 24 and 23 categories. The two systems achieving the best results for SemEval 2015 both used a classification approach (Toh and Su, 2015; Saias, 2015). Furthermore, especially lexical features in the form of bag-of-words have proven successful. The best system (Toh and Su, 2015) also incorporated lexical-semantic features in the form of clusters learned from a large corpus of reference review data, whereas the second-best (Saias, 2015) applied filtering heuristics on the classification output and thus solely relied on lexical information for the classification. For SemEval 2016 Toh and Su (2016) discovered that when the probability output of a Deep Convolutional Neural Network (Severyn and Moschitti, 2015) was added as additional features, the performance increased.

For the experiments presented here, classifiers were built using LibSVM (Chang and Lin, 2011). Our feature space includes lexical information by relying on **bag-of-word** features in the form of token unigrams. Because for Dutch no large reference review datasets are available in the var-

Train \ Test	Banking			Retail			HR		
	Prec	Rec	F-1	Prec	Rec	F-1	Prec	Rec	F-1
Banking	94.8	95.1	94.9	89.6	90.9	89.2	95.2	95.4	95.0
Retail	93.0	93.9	93.2	95.6	95.5	95.6	94.9	95.1	94.4
HR	93.4	94.2	93.4	91.0	91.4	89.7	96.5	96.8	**96.4**
All training	95.1	95.4	**95.2**	95.8	95.8	**95.8**	95.9	96.2	95.9

Table 4: Precision, recall, and F-1 scores for aspect term extraction on held-out test sets.

ious domains, we were inspired by the work of De Clercq and Hoste (2016) to also include lexical semantic features derived from Dutch **WordNet** information, viz. Cornetto (Vossen et al., 2013) and **DBpedia** (Lehmann et al., 2013) for the aspect terms available in the training data for each of the domains.

After training our models, these are tested on the held-out test set. Important to note is that for this setup we do not work with gold standard aspect terms, but rely on the output from the aspect term extraction step. Since each verbatim can be labeled with zero, one or more categories that are not mutually exclusive, we decided to use Hamming score, a multi-label evaluation metric that divides the number of correct labels by the union of predicted and true labels.

3.3 Aspect Polarity Classification

Machine learning approaches to sentiment analysis make use of classification algorithms, such as Naïve Bayes or Support Vector Machines trained on a labeled dataset (Pang and Lee, 2008). Current state-of-the-art approaches model a variety of contextual, lexical and syntactic features (Caro and Grella, 2013), allowing them to capture context and the relations between the individual words. Though deep learning techniques have also been applied to this subtask, mainly in the form of word embeddings (Mikolov et al., 2013), for SemEval 2016 the best performing system relied solely on (advanced) linguistic features (Brun et al., 2016).

We followed a supervised approach and built SVM classifiers using LibSVM. As we conceived ABSA as an integrated task, the input for the polarity classification includes the detected aspect term (result of step 1) and category (result of step 2), together with the preprocessed sentence in which the aspect term occurs. As a result, error percolation between the different steps impacts the performance of the polarity classification system. As information sources, we implemented the following features: (1) **bag-of-words**: binary token unigram features, (2) **lexicon** lookup features based on domain-specific lexicons extracted from the training data, as well as existing sentiment lexicons for Dutch, i.e. Pattern (De Smedt and Daelemans, 2012) and Duoman (Jijkoun and Hofmann, 2009), (3) **negator**: flips the value of negated lexicon matches and (4) the **predicted** category of the aspect term. For these experiments, we also envisaged the three different setups: in-domain, cross-domain, and all domain. It is important to mention that for sentiment prediction, the entire sentence is considered for the construction of the features. As a result, conflicting sentiments will be ruled out. In future work, we intend to limit the context window of the detected aspect term. As the polarity detection takes into account the output of the previous two steps, this task was also evaluated by means of the hamming score metric (cfr. 4.3).

4 Results

4.1 Aspect Term Extraction

In Table 4 the results are presented for the different experiments training on in-domain data (underlined scores), cross-domain data, and a combination of all training data. We observe good results for aspect term extraction for all three domains. In-domain scores are slightly better than cross-domain scores, except for retail. This might be explained by the fact that retail has very different aspect targets than the other two domains, which are both more services-oriented. In addition, the target extraction scores show that training on all data improves scores slightly for the banking and the retail domain, but decreases for HR.

4.2 Aspect Category Classification

To evaluate, we report hamming scores for (i) a classifier taking the in-domain predictions for aspect terms as input (*In-domain*) and (ii) the pre-

dictions of the classifier trained on all training data from the various domains for the aspect term extraction (*All training*).

	In-domain	All training
Banking	58.1	57.4
Retail	67.0	68.5
HR	46.6	46.8

Table 5: Aspect category classification results.

As can be seen in Table 5, the score difference between both setups is small. Overall, we observe that predicting the correct aspect categories is much more challenging for HR than for the other two domains. A qualitative analysis revealed that a lot of errors are caused by error percolation from the previous step. For HR more in particular, there is a lot of confusion between closely-related categories such as PERSONNEL_service and PERSONNEL_availability.

4.3 Aspect Polarity Classification

We report hamming scores for the classifiers taking the aspect terms derived from the aspects terms that were extracted in the *All training* setup[3].

Train \ Test	Banking	Retail	HR
Banking	84.5	83.3	67.1
Retail	**86.8**	**88.9**	**86.7**
HR	86.0	86.1	86.1
All	84.5	86.8	85.4

Table 6: Aspect polarity classification results.

Table 6 shows satisfactory results for polarity classification based on automatically predicted aspect terms. The results show that training polarity classifiers on all domains results in lower classification scores than in-domain training, which confirms the intuition that sentiment expressions are often ambiguous and domain-dependent. Although the HR data set is rather limited (1000 verbatims), cross-domain training on HR also results in consistently good polarity prediction for the other domains. Training on banking, however, results in bad polarity prediction for the HR aspect terms. A qualitative analysis revealed that the HR polarity classification relies on more general sentiment expressions also occurring in other domains (e.g. *vriendelijk (EN: friendly), super (EN: excellent)*), but also on very HR-specific sentiment words (e.g. *nauwkeurig (EN: accurate), doeltreffend (EN: effective)*). Remarkably, retail has the best cross-domain performance, it even outperforms the in-domain results for banking and HR. This is because the retail model always predicts the positive class for these two test sets, making this a hard to beat majority baseline.

[3]Experiments revealed no difference in performance when relying on the *in-domain* aspect terms.

5 Conclusion

In this paper we presented an ABSA pipeline that implements an integrated approach for the three ABSA subtasks, which have been performed and evaluated separately in previous research. We collected and annotated qualitative user feedback in three domains: banking, retail and HR. Especially the banking and HR data are novel in that they comprise service-oriented customer feedback.

By performing in-domain and cross-domain experiments we show promising classification results for all three subtasks. Considering the aspect term extraction task, it seems that training on all available training data is beneficial for the banking and retail domain. The HR domain, however, benefits most from in-domain training data. For the aspect category classification, again the HR domain reveals a different result than the other domains, in that it is much more harder to classify. The polarity classification experiments reveal that for all domains it is better to train on small domain-specific datasets instead of combining training data from different domains. Strikingly, the retail domain generalizes best to the other domains, though these results should be corroborated on larger datasets.

As we address the ABSA task incrementally, we observed error percolation in each step. We believe, however, that only an incremental approach reflects how ABSA is performed in a real-world setting. In future work, we will explore the viability of domain adaptation for ABSA on larger and different datasets and with other languages.

Acknowledgments

We wish to thank the annotators and the anonymous reviewers. This work has been supported by the Flanders Innovation & Entrepeneurship government agency (VLAIO).

References

Samuel Brody and Noemie Elhadad. 2010. An unsupervised aspect-sentiment model for online reviews. In *Proceedings of the 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-2010)*, pages 804–812.

Caroline Brun, Julien Perez, and Claude Roux. 2016. XRCE at SemEval-2016 Task 5: Feedbacked Ensemble Modeling on Syntactico-Semantic Knowledge for Aspect Based Sentiment Analysis. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 277–281.

Luigi Di Caro and Matteo Grella. 2013. Sentiment analysis via dependency parsing. *Computer Standards & Interfaces*, 35(5):442–453.

Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27.

Orphée De Clercq and Véronique Hoste. 2016. Rude waiter but mouthwatering pastries! An exploratory study into Dutch Aspect-Based Sentiment Analysis. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 23–28.

Tom De Smedt and Walter Daelemans. 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.

Gayatree Ganu, Noemie Elhadad, and Amélie Marian. 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.

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

Valentin Jijkoun and Katja Hofmann. 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.

Marjan Van de Kauter, Geert Coorman, Els Lefever, Bart Desmet, Lieve Macken, and Véronique Hoste. 2013. LeTs Preprocess: The multilingual LT3 linguistic preprocessing toolkit. *Computational Linguistics in the Netherlands Journal*, 3:103–120.

Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N. Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick van Kleef, Soren Auer, and Christian Bizer. 2013. DBpedia – A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia. *Semantic Web Journal*, 6:167–195.

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

Bing Liu. 2015. *Sentiment Analysis - Mining Opinions, Sentiments, and Emotions*. Cambridge University Press.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119.

Jorge Nocedal. 1980. Updating quasi-Newton matrices with limited storage. *Mathematics of Computation*, 35(151):773–782.

Naoaki Okazaki. 2007. CRFsuite: a fast implementation of Conditional Random Fields (CRFs).

Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval.*, 2(1-2):1–135.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad AL-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, Véronique Hoste, Marianna Apidianaki, Xavier Tannier, Natalia Loukachevitch, Evgeniy Kotelnikov, Núria Bel, Salud María Jiménez-Zafra, and Gülşen Eryiğit. 2016. Semeval-2016 task 5: Aspect based sentiment analysis. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 19–30.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Suresh Manandhar, and Ion Androutsopoulos. 2015. Semeval-2015 task 12: Aspect based sentiment analysis. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval-2015)*, pages 486–495.

Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. Semeval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval-2014)*, pages 27–35.

José Saias. 2015. Sentiue: Target and aspect based sentiment analysis in SemEval-2015 Task 12. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval-2015)*, pages 767–771.

Iñaki San Vicente, Xabier Saralegi, and Rodrigo Agerri. 2015. EliXa: A Modular and Flexible ABSA Platform. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval-2015)*, pages 748–752.

Aliaksei Severyn and Alessandro Moschitti. 2015. UNITN: Training Deep Convolutional Neural Network for Twitter Sentiment Classification. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval-2015)*, pages 464–469.

Pontus Stenetorp, Sampo Pyysalo, Goran Topic, Tomoko Ohta, Sophia Ananiadou, and Junichi Tsujii. 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.

Tun Thura Thet, Jin-Cheon Na, and Christopher S.G. Khoo. 2010. Aspect-based Sentiment Analysis of Movie Reviews on Discussion Boards. *Journal of Information Science*, 36(6):823–848.

Zhiqiang Toh and Jian Su. 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.

Zhiqiang Toh and Jian Su. 2016. NLANGP at SemEval-2016 Task 5: Improving Aspect Based Sentiment Analysis using Neural Network Features. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 282–288.

Piek Vossen, Isa Maks, Roxane Segers, Hennie van der Vliet, Marie-Francine Moens, Katja Hofmann, Erik Tjong Kim Sang, and Maarten de Rijke. 2013. Cornetto: a lexical semantic database for Dutch. In Peter Spyns and Jan Odijk, editors, *Essential Speech and Language Technology for Dutch, Theory and Applications of Natural Language Processing*, pages 165–184. Springer.