# Sentiment Analysis for Social Media Comments

Christoph Emunds, Richard Polzin,
Dominik Nerger, Benedikt Heinrichs

June 28, 2017

**Abstract**

Analyzing customer experience offers valuable data for Business Intelligence. Especially in e-commerce, where users write reviews about products and services, the analysis of the customers' sentiment towards these entities yields significant insights into potential strengths and weaknesses of the product or service.

Our work focuses on evaluating customer experience through the application of aspect-based sentiment analysis on social media posts. The data set contains a year's worth of Facebook comments on the pages of the two supermarket chains Tesco and Sainsbury. The goal is to extract as many triplets $(e_i, a_{ij}, s_{ij})$ as possible, where $e_i$ is an entity (product or service), $a_{ij}$ is an aspect (performance, battery, politeness, etc.) of this entity , and $s_{ij}$ is the sentiment polarity label (negative, neutral, positive) for this entity-aspect combination.

To accomplish this task, many different subtasks need to be solved. After a rudimentary preprocessing routine, the posts need to be identify the part of speech (POS) category for every word. Then Named Entity Recognition (NER) must be applied and words containing sentiment need to be detected and evaluated. As a significant number of social media texts do not follow the grammatical rules or contain a lot of misspellings the complexity of those tasks is increased significantly. With the completed analysis we identify potential products and services, their corresponding aspects and sentiment words, which are scored and aggregated into opinions.

The results obtained from this analysis are visualized to get a better understanding of the customers' feelings towards these products, services, and their aspects. The visualization supports businesses in a wide range of decisions, such as product positioning and pricing. This can lead to better customer satisfaction, faster reactions to trends and overall revenue growth.

# Contents

**1 Introduction** **1**

**2 Related work** **1**
2.1 POS-Tagging . . . . . 1
2.2 NER-Tagging . . . . . 2
2.3 Lexicon-based sentiment analysis . . . . . 2

**3 Approach** **3**
3.1 Preprocessing . . . . . 3
3.2 Extracting entities and aspects . . . . . 3
3.3 Determining sentiment polarities . . . . . 4

**4 Results** **4**
4.1 Visualization . . . . . 5

**5 Conclusion** **6**

**References** **8**

# 1 Introduction

Sentiment analysis uses techniques from natural language processing (NLP), text analysis, and computational linguistics to extract subjective information from texts. In applications that involve analyzing customer experience, this subjective information typically represents the opinion of a person about a certain entity. Quantifying the feelings and opinions that customers possess towards products and services offers valuable information for the company.

Such information can be used to monitor reputation or improve understanding of the market's needs. In contrast to generic sentiment analysis, this information can only be acquired by precisely identifying the entities and aspects towards which a sentiment is expressed. This specific task is referred to as Aspect Based Sentiment Analysis (ABSA). The project presented in this document focuses on ABSA, finding and analyzing customers' opinions on products and their aspects from social media posts of a supermarket chain.

In today's world, social media is a big part of the social life. On social networks like Facebook, many companies represent themselves through a corporate account. This opens up the possibility for many people to interact with those companies. Everyone can leave a comment on the page of a company, celebrity or a sports teams. People can review or comment with a lot of context. General posts can be created, as well as reactions to news or a posting that concerns a specific product. These comments or reviews are often filled with positive or negative sentiments.

It should therefore be possible to collect the information in these posts and extract sentiments about products or aspects of the company. Our approach on a solution will be presented in this report, starting with a description of the related work that already exists. Afterwards, our approach and the single steps that are performed will be introduced. Then, the results that have been obtained will be discussed, as well as the visualization that has been implemented. To finish up the report, a conclusion will be drawn.

# 2 Related work

The related work presented in this section gives a quick overview over the main techniques that we used to identify entities and their aspects, as well as determining the sentiments.

## 2.1 POS-Tagging

POS-Tagging (part-of-speech tagging), also called grammatical tagging or word-category disambiguation, refers to the process of identifying particular parts of speech like nouns or verbs. Identifying the role of a certain word within a sentence is important for the task of sentiment analysis, as identifying entities and their corresponding aspects can be handled much easier relying on certain assumptions about the part of speech of words.

While being important POS-Tagging is also a complex problem. Word-forms in natural language are often ambiguous. For example the word 'dogs' is usually thought of as a plural noun, but can be used as a verb as well:

> The sailor dogs the hatch.

Due to the complexity, machine learning techniques are often applied in POS-Taggers. Popular approaches such as the Viterbi algorithm (Viterbi, 2006), the Brill tagger (Brill, 1992) or the Baum-Welch algorithm (Rabiner, 2012) work with techniques such as dynamic programming, supervised learning or hidden Markov models.

## 2.2 NER-Tagging

Named-Entity recognition (NER) is a task that seeks to locate and classify specific information in text. This information is called a named entity and can refer to categories such as the names of persons, locations, times or many others.

An annotated sentence could look like this :

> [Tim Cook]$_{[Person]}$ has a Net worth of [785 million USD]$_{[MonetaryValue]}$ as of [March 16. 2017]$_{[Time]}$.

While state-of-the-art NER-Taggers perform very well and produce near-human performance they are also brittle and do not perform well in domains they were not designed for (Thierry Poibeau, 2001).

## 2.3 Lexicon-based sentiment analysis

An approach to determine sentiment polarities is the lexicon-based sentiment analysis. The method created by (Ding, Liu, & Yu, 2008) can be broken down to the following four steps:

In the first step, all sentiment words and phrases in the text are marked with the help of a sentiment lexicon. This is done by assigning positive or negative values to those sentiments, e.g. [+1] for a positive or [-1] for a negative sentiment.

In the second step, sentiment shifters are applied. A sentiment shifter can be described as a word that negates the value that has been applied in the first step, changing its value from positive to negative or the other way around. Typically, these words are negation words like *not*.

In the third step, but-clauses and further contrary words are handled. With these contrary words, sentences can be split up into two parts. Depending on the value of the sentiments mentioned in the first part, the sentiments in the second part receive the opposite value.

In the fourth and last step, the opinions are aggregated according to the following function:

$$score(a_i, s) = \sum_{ow_j \in s} \frac{sw_j.so}{dist(sw_j, a_i)}$$

where $sw_j$ is a sentiment word in s, $dist(sw_j, a_i)$ is the distance between aspect $a_i$ and sentiment word $sw_j$ in $s$. $sw_j.so$ is the sentiment score of $sw_i$. Depending on the score, it can be determined whether the opinion on the aspect $a_i$ in $s$ is positive or negative. If the final score is neither positive or negative, it is a neutral aspect.

# 3 Approach

In the following section, the approach towards aspect based sentiment analysis will be presented. First preprocessing will be discussed. Then, we will show how the products and aspects are extracted. Afterwards, the NLTK module *VADER* is used to determine the sentiment polarities that are present in the Facebook posts.

## 3.1 Preprocessing

We preprocessed the data by extracting the actual text of the posts into separate files. We decided to exclude posts that are less than 20 characters long or include images. This is due to the fact that posts which are too short do not yield much information most of the time. Furthermore, posts that include images often refer to objects in the image, which makes it hard to understand the author's sentiment without analyzing the image content.

## 3.2 Extracting entities and aspects

For each post, the POS tags as well as different entities are obtained, with the entities being created through different Named Entity Recognition models. Both POS-tagging and Named Entity Extraction are accomplished with OpenNLP. The Named Entity Recognition in the English language is available for different entities, each with their own model. We utilized the following models: *person*, *location*, *organization*, *date* and *money*.

Because no pre-trained model exists for the Named Entity Recognition of products or services, we decided to use the OpenNLP model for organization to approximately tag these. Most nouns in the English language are written in lower case, but organizations as well as products often contain capital letters. Therefore, we consider everything that is tagged as *Organization* as a product or service.

Searching for aspects is done by iterating over the complete set of posts for each entity. If a post contains the product, we identify the sentences it appears in. We then look for other nouns in these sentences and count their frequency. From those potential aspects, we consider those that occur in at least 10% of the

posts that mentioned the product. To reduce the impact of noise, we decided to remove some of the products by hand.

After identifying a set of aspects for each entity, we merge similar entities and their aspects by tokenizing the entities' names and lemmatizing them. This results in a reduction of entities, e.g. *Customer Service* and *Customer Services* are considered as the same entity. The entities' aspects are merged accordingly. Moreover, all aspects that are part of the entity's name are removed from the aspect list. For instance, this removes the aspects *tesco* and *express* from the list of aspects for the entity *Tesco Express*.

If no aspects can be found for a specific entity, it is removed from the final list of entities. For every entity that is part of the final list, the aspect *general* is added as well.

### 3.3 Determining sentiment polarities

Liu (2012) defines an opinion as a quintuple $(e_i, a_{ij}, s_{ijkl}, h_k, t_l)$, where $e_i$ is the $i$'th entity and $a_{ij}$ is the $j$'th aspect of entity $i$ the opinion is expressed on. $s_{ijkl}$ is the sentiment polarity, which can take on the values *positive* or *negative*. $h_k$ describes the opinion holder and $t_l$ the time at which the opinion was expressed.

However, from the given data set, it is not possible to determine the person that wrote a post. Moreover, we did not focus on extracting the time at which the sentiment was expressed, as we did not aim to provide a temporal overview over the shifting of opinions towards the products and services.

Our approach aims to extract opinion triples $(e_i, a_{ij}, s_{ij})$. For each post we first identify the sentences a certain entity-aspect combination appears in. Afterwards, we score these sentences with VADER (*Valence Aware Dictionary and sEntiment Reasoner*) (Hutto & Gilbert, 2014), which is a tool for lexicon and rule-based sentiment analysis. It is included in the NLTK Python library and was specifically designed to be used on social media texts, e.g. by understanding emoticons and slang words.

After obtaining the opinion triples for each post, they are aggregated by entity and aspect to be more comprehensible. This creates an overview over the customers' sentiments towards the products and services. We also incorporate excerpts from the posts into the summarization, i.e. the sentences that an entity-aspect combination appears in. These excerpts are only used for the visualization later on, where the product, aspect and sentiment words are color coded.

## 4 Results

We were given a data set containing a year's worth of Facebook comments on the pages of the two supermarket chains Tesco and Sainsbury. However, we mainly focused on the Tesco data set, since it is bigger and the developed pipeline can easily be applied to other data sets.

Through the pre-processing we reduced the amount of posts by 40.6%, from 37701 original posts down to 22395. The entity extraction resulted in 1598 products. After removing some of them by hand, we ended up with 1379 products. Furthermore, products for which no aspect was found were removed. Finally, 1256 products that at least contain one aspect were identified, decreasing the product list by 21.5%.

We labeled 200 posts concerning the entity *Customer Service* by hand. From these 200 posts, we extracted 254 opinion triples, in which we filled in the sentiment manually. Comparison with VADER's results showed that 139 of the 254 triples (i.e. 54.72%) match.

## 4.1 Visualization

For the visualization, we use the aggregated opinions that are the result of executing the pipeline and the sentiment lexicon by (Hu & Liu, 2004). On the server side, we use Python with the library Flask. On the client side, we use the JavaScript libraries d3.js, vue.js and jQuery.

The data is visualized with each product having its own page. Each aspect of the product is visualized with a bar chart and the corresponding post excerpts. The bar chart can be seen in Figure 1. It shows the amount of negative, neutral and positive sentiments for the aspect *store* of the entity *Customer Service*.

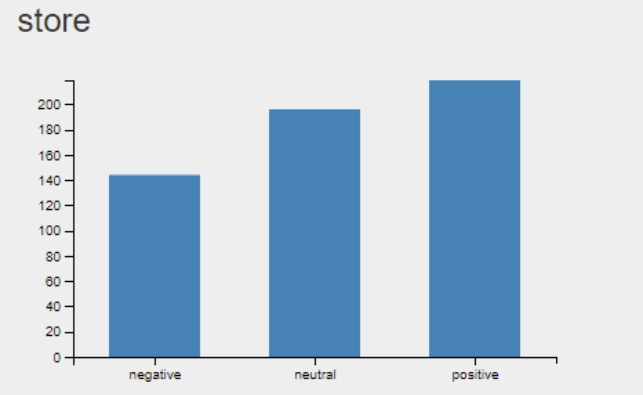


Figure 1: Bar chart displaying the sentiments regarding the customer service with aspect 'store'

An excerpt of posts regarding the combination of product and aspect is shown, with a maximum of 10 posts per aspect. In Figure 2, the highlighted posts can be seen. Positive and negative sentiments are highlighted using the sentiment lexicon. This particular sentiment lexicon includes misspellings, morphological variants, slang and social-media mark-up of 2006 positive and 4783 negative words. Positive and negative sentiments are colored green and red, respectively. Products are highlighted in orange and Aspects in cyan.

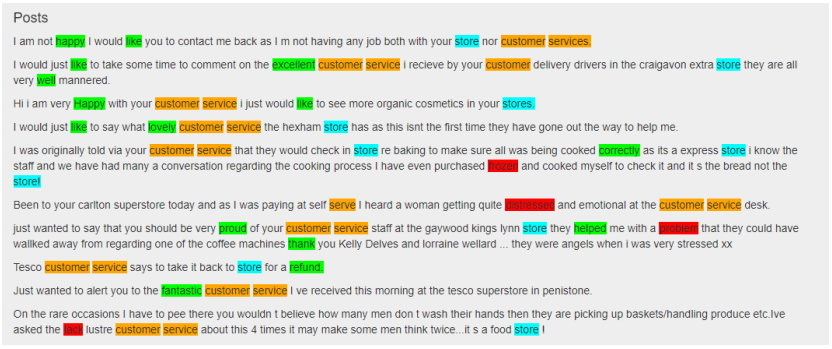


Figure 2: Highlighted posts regarding the customer service with aspect 'store'

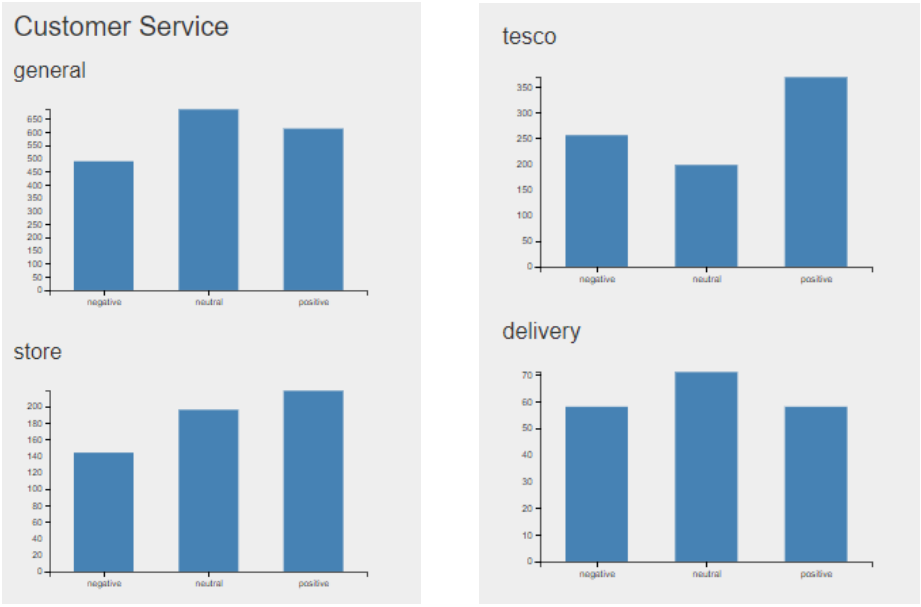


Figure 3: Sentiment charts for different aspects of Customer Service

For a subset of the sentiment charts for the entity *Customer Service* see Figure 3. The aspects 'general', 'store' and 'tesco' have a more positive sentiment, while the aspect 'delivery' has about the same amount of positive and negative sentiments. One potential insight might be that improving delivery should be prioritized over improvements of the store.

# 5 Conclusion

The accuracy of 54.72% as mentioned in Section 4 leaves a lot of potential for improvement. Several factors could be related. For example, the baseline for comparison only consists of 200 posts that were labeled by hand. These are by no means indicative of the actual data set. In the scope of this project,

each of the four contributors tagged 50 posts. To ensure a proper test set for validation, each contributor would have to label the same set of posts, such that the Kappa measure could be derived. Due to time constraints, this was not a priority.

Furthermore, the function for the sentiment polarity determination was not evaluated extensively. The different parameters were mainly set by intuition and no detailed analysis was conducted.

A general shortcoming of the proposed pipeline is that we do not handle multi token aspects. The developed system is based on identifying common nouns as aspects. These, by definition, do not span multiple tokens.

As mentioned in Section 3.2, we consider every organization found as product or service. This has many caveats. For one, it is required that services like *Customer Service* actually appear capitalized in our data set at least once to be found. Moreover, a lot of things that are not actually organizations are tagged as such. This could be improved by incorporating more knowledge regarding common nouns and noun phrases, creating our own Named Entity Recognition model.

Another potential source of improvement would be the incorporation of dependency parsing to determine the sentiment towards an entity-aspect combination on a more detailed level. Currently, we use the sentiment of the whole sentence as a foundation for the sentiment polarity determination.

# References


Brill, E. (1992). A simple rule-based part of speech tagger. In *Proceedings of the third conference on applied natural language processing* (pp. 152–155). Stroudsburg, PA, USA: Association for Computational Linguistics. Retrieved from http://dx.doi.org/10.3115/974499.974526 doi: 10.3115/974499.974526

Ding, X., Liu, B., & Yu, P. S. (2008). A holistic lexicon-based approach to opinion mining. In *Proceedings of the 2008 international conference on web search and data mining* (pp. 231–240). New York, NY, USA: ACM. Retrieved from http://doi.acm.org/10.1145/1341531.1341561 doi: 10.1145/1341531.1341561

Hu, M., & Liu, B. (2004). Mining and summarizing customer reviews. In *Proceedings of the tenth acm sigkdd international conference on knowledge discovery and data mining* (pp. 168–177). New York, NY, USA: ACM. Retrieved from http://doi.acm.org/10.1145/1014052.1014073 doi: 10.1145/1014052.1014073

Hutto, C. J., & Gilbert, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In E. Adar, P. Resnick, M. D. Choudhury, B. Hogan, & A. H. Oh (Eds.), *Icwsm*. The AAAI Press. Retrieved from http://dblp.uni-trier.de/db/conf/icwsm/icwsm2014.html#HuttoG14

Liu, B. (2012). *Sentiment analysis and opinion mining*.

Rabiner, L. (2012). *First-hand:the hidden markov model*. Retrieved from http://ethw.org/First-Hand:The_Hidden_Markov_Model

Thierry Poibeau, L. K. (2001). Proper name extraction from non-journalistic texts. *Language and computers*, *37*, 144–157.

Viterbi, A. (2006, September). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Trans. Inf. Theor.*, *13*(2), 260–269. Retrieved from http://dx.doi.org/10.1109/TIT.1967.1054010 doi: 10.1109/TIT.1967.1054010