Sentiment Analysis in Social Networks

Bachelorarbeit

Zur Erlangung des akademischen Grades “Bachelor of Science (B.Sc.)” im Studiengang Wirtschaftswissenschaft der Wirtschaftswissenschaftlichen Fakultät der Leibniz Universität Hannover

Vorgelegt von

Name: Podszus Vorname: Simona Natalia

[redacted] [redacted] [redacted] [redacted]

Prüfer: Prof. Dr. M. H. Breitner

Hannover, der 09.08.2016

Contents

Figures

Tables

Abbreviations

Abbreviation	Meaning
API	Application Programming Interface
BI	Business Intelligence
Cf.	Compare
DWH	Data Warehouse
ETL	Extraction, Transformation, Loading
HAC	High Adjective Count
IR	Information Retrieval
KDD	Knowledge Discovery from Data
MOS	Maximum Opinion Score
NLP	Natural Language Processing
OLAP	Online Analytical Processing
POS	Part-Of-Speech

R&D	Research and Development
SA	Sentiment Analysis
SMM	Social Media Monitoring
UGS	User-Generated Content
WWW	World Wide Web

1 Introduction

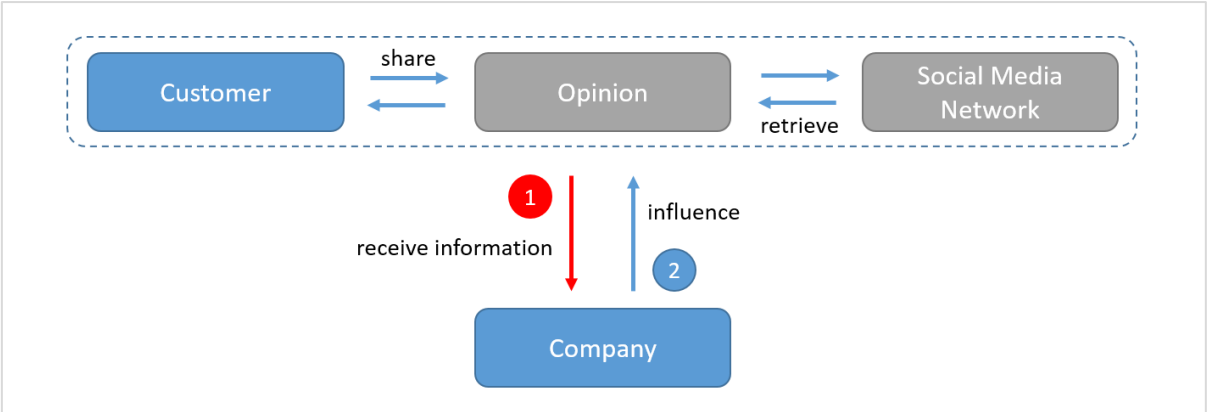
1.1 Motivation and Relevance

"Opinion is the medium between knowledge and ignorance." (Plato, 380 BC)[1]

What Plato said, is still valid today: Opinions are rarely expressing the perfect truth about objects in terms of facts. Nevertheless, the opinion of others is important to people (Pang & Lee, 2008). The knowledge base and access to facts provided by the internet has not changed that. On the contrary, more and more consumers are feeling comfortable with the web and share their opinions online (Hu & Liu, 2004), for instance, they are expressing themselves by tweeting or commenting on YouTube[2] videos about products (Smith et al., 2012). The number of customer reviews is growing rapidly since e-commerce is becoming more and more popular (Hu & Liu, 2004). According to a study[3] published by the Nielson Company (2015), 66% of customers trust consumer opinions posted online for obtaining information. This is the second most-trusted source for product recommendations after the personal contact.

That is why, companies depend on a high amount of recommendations for their products or services and in case, that they are not recommended, on a feedback. In order to gain influence about the word-of-mouth marketing, companies are interested in their customers' thoughts on their products and the product's features. While personal contacts can hardly be influenced by companies, happenings in the World Wide Web (WWW) are observable more easily. Therefore, knowledge about the digital product recommendation process is very valuable for companies' success and a chance for a higher customer-orientation (see Figure 1).

Figure 1: Companies and the Digital Product Recommendation Process



[1] The original source is Plato's Republic, for translation and interpretation cf. Lee (2010) or Annas (1981).
[2] www.twitter.com, www.youtube.com
[3] The Nielsen Global Trust in Advertising Survey polled 30,000 online respondents in 60 countries to measure consumer sentiment about 19 paid, earned and owned advertising mediums.

In order to gain information out of this high amount of unstructured data, companies observe social networks in the sense of a Social Media Monitoring (SMM). SMM is about analyzing the prevailing customer's mood and react appropriately. Since a manual SMM effort bears a lot of time and costs, intelligent instruments are necessary. To identify a technique that may lead to these desired results, this work focuses on Sentiment Analysis (SA), a sub-section of Text Mining which is a computer-based analysis of opinion and mood in text databases. The main advantage of a SA to companies are that they enable them to receive information about their customers' sentiment automatically. Moreover, a SA gives them the chance to understand swarm mechanism in the digital product recommendation process and to influence customers by using these mechanisms actively.

1.2 Research Objectives and Structure

In the first part of this work, the theoretical background is established as a basis for the further chapters. In order to reach a common understanding, definitions of Data Mining, Knowledge Discovery from Data (KDD), Text Mining and Natural Language Processing (NLP) are set (2.1). Within a review of related work, the status quo of the research about a SA regarding product features is shown. Moreover, existing research approaches and current gaps in the field of feature-based SA in social networks are considered.

The second and qualitative part deals with the phenomenon, that more and more companies are trusting less in expert's reviews for business decisions, but rather focus on the opinion of crowds meaning different stakeholders such as customers (Mollik & Nanda, 2015). In order to receive this customer feedback, different instruments can be used such as SA as an advanced form of SMM. Therefore, the second part of this work is divided into three parts. Firstly, social media networks and especially YouTube as an opinion-rich data sources are examined with regard to a feature-based SA. Secondly, different use cases are developed along a companies' value chain, especially regarding R&D, marketing, and sell. Thereby, the opportunities and limits of a SA of social network data are discussed. Also, social networks pose several questions for sentiment analysis, that are considered in a third part. The research question (RQ) is as following:

> RQ 1: What are the chances and challenges of a SA in Social Media Networks with regard to product recommendations?

The third and quantitative section describes the proposal and implementation of a text mining system, called Feature Extractor. Its main objective is to extract the product features from YouTube comments which are the most important for the users. The Feature Extractor can be seen as a ground work for a Sentiment Extractor that aims to aggregate and rank these features using a sentiment score. The proposal also covers, how an interface that is presenting the

The Feature Extractor builds the preliminary work for an Opinion Extractor. While the main objective of the Feature Extractor is to extract high discussed features from social network data, the downstream Sentiment Extractor aims to aggregate and rank the features using a sentiment score that determines how positive or negative the features are described by users of the social network. In this application features broadly mean product attributes or functions of a product (Hu & Liu, 2004). In the following, important features are defined as the nouns, for which reviewers express a lot of opinions. Following Chapter 2.1, our KDD process can be separated into four main steps: data selection, data preparation, data mining and interpretation. In order to select the proper data sets, the task of the KDD process has to be described. It is defined as following: Retrieve information from users' opinions about the most important features of the Samsung Galaxy S7 expressed as YouTube comments.

4.1.2 Data Selection

Having a defined task, the selection of suitable data is essential. The object of observation are YouTube comments relating to the Samsung Galaxy S7. Chapter 3.2 lays focus on the discussion, why social networks and especially YouTube data sets are a valuable data source for the defined task. The object of observation are the YouTube users, who are not influenced by this research and therefore behave naturally. That is why a high validity is given. The analysis for this purpose is limited on the headlines of the videos to classify the associated comments as relevant and the comments in the form of text data. Sentiments expressed in video and audio form, as well as meta-ratings, are excluded from the KDD process. For the download of comments, YouTube provides an application programming interface[10] (API) that allows access to comments of individual videos automatically and guidance documents, such as example queries. The corpus for this study consists of 1301 comments from top English-language product related videos, which can be considered as a high sample size that allows representative results for the evaluation of product features.

Besides a dataset of YouTube comments, different packages that store functions and datasets in R are necessary (Venables & Smith, 2007. For this application, four packages are selected. The Natural Language Processing Infrastructure (NLP) package offers basic classes and methods for NLP (Hornik, 2016). With the Text Mining (TM) package, basic text mining applications can be applied (Feinerer, 2015). The openNLP package is a machine learning based toolkit for the NLP in Java and supports the algorithm during the part-of-speech (POS) tagging (Hornik, 2016). Furthermore, the SnowballC package is necessary for the usage of porter's word stemming algorithm in the preparation (Bouchet-Valat, 2014).

[10] https://www.youtube.com/yt/dev/de/api-resources.html

4.1.3 Data Preparation

Within the preparation different tasks are executed to improve the quality of the text data and the text mining algorithm. The data preparation and the text mining steps of the Feature Extractor follow the High Adjective Count (HAC) pseudocode of Eirinaki et al. (2012) that is adopted in Figure 10. The accuracy of their algorithm turns out to be very high on structured online reviews and they, therefore, recommend extensions for other datasets from social networks.

Figure 10: High Adjective Count Algorithm in Pseudocode (Eirinaki et al., 2012)

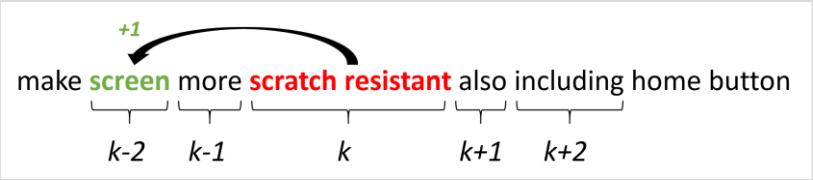
```
HighAdjectiveScores(reviews)
nouns_score_map <- {}
foreach review in reviews do
        Assign part of speech tags to the review
        Apply stemming
foreach line in the review do
foreach adjective in the line find the closest noun
nouns_score_map [noun]++
potential_features_map <- {}
foreach noun in noun_score_map
if noun_score_map[noun] > threshold
potential_features_map[noun] = threshold
return potential_features
```

Eirinaki et al. (2012) suggest stemming and POS tagging as preparation. This technique will add a few preparation steps, in order to improve the quality of the results. After the merger of text sources from different videos into one corpus, the morphologic analysis is executed. The extra whitespace is eliminated and multiple white space characters are collapsed to a single blank. Because upper cases offer no added value and may increase the complexity in this analysis, the whole corpus is converted to lower case. For the same reason, stop words are removed. Stemming the words improves the quality of single-word indexing by grouping words that have the same stem (Keikha et al., 2008). In our experiment, we use Porter's algorithm for stemming. As a last step of the morphologic analysis, any punctuation characters are removed. Afterward within the syntactical analysis, the POS tagging is executed by use of the openNLP package. Hereby, the text documents of the corpus are written as a string. Sentence and word token annotations are set and the tags for word tokens are distributed. For instance, "NN" indicates a noun and "JJ", "JJR", "JJS" indicate adjectives depending on if they are in basic form, comparative or superlative.

4.1.4 Text Mining

After the preparation, the main text mining algorithm can be performed. The main idea is to identify adjectives, because they are likely to express an opinion, and afterward, associate them with the closest noun. The assumption is that the less distance is between two words, the more likely a content-based connection is (Mikolov et al. 2013). The nouns that are associated with many adjectives can be interpreted as potential product features of interest. The proceed in R can be demonstrated on an exemplary comment, shown in Figure 11.

Figure 11: Association of an Adjective with a Noun



At the beginning of the text mining process, the noun score map, the list of found nouns and its associated scores, named opinion scores, is set to zero. Then, each word of the corpus is checked. If an adjective *k* is found, the algorithm checks the word before (*k-1*) and behind the adjective (*k+1*). In order to get more precise results, only adjectives with a target probability of 90% are taken into consideration. If none of them is a noun, the algorithm checks the next words in the front (*k+2*) and in the back (*k-2*) of the adjective. In the example, the word at position *k-2* is a noun. Now the algorithm checks, if the word "screen" is already part of the noun score map. If it is on the noun score map, its score is raised by one. If it is not on the map, the noun is added to the map and receives a score of one. For this time, the sentiment and the strength of the expressions that are associated with the screen are not relevant. It is shown, that the screen may be an important feature, disregarding with which sentiment polarity, positive or negative. For the case that two nouns are found ahead and behind the adjective simultaneously, both nouns are associated with the adjective and both scores are raised. This proceed is executed for every word in the corpus. Finally, the analysis results in a noun score map with nouns and their score ordered descending. In order to exhibit only the most important features, a score threshold may be set. The nouns that reach the threshold are copied and printed as potential features. An extract of the results can be found in the enclosed attachment.

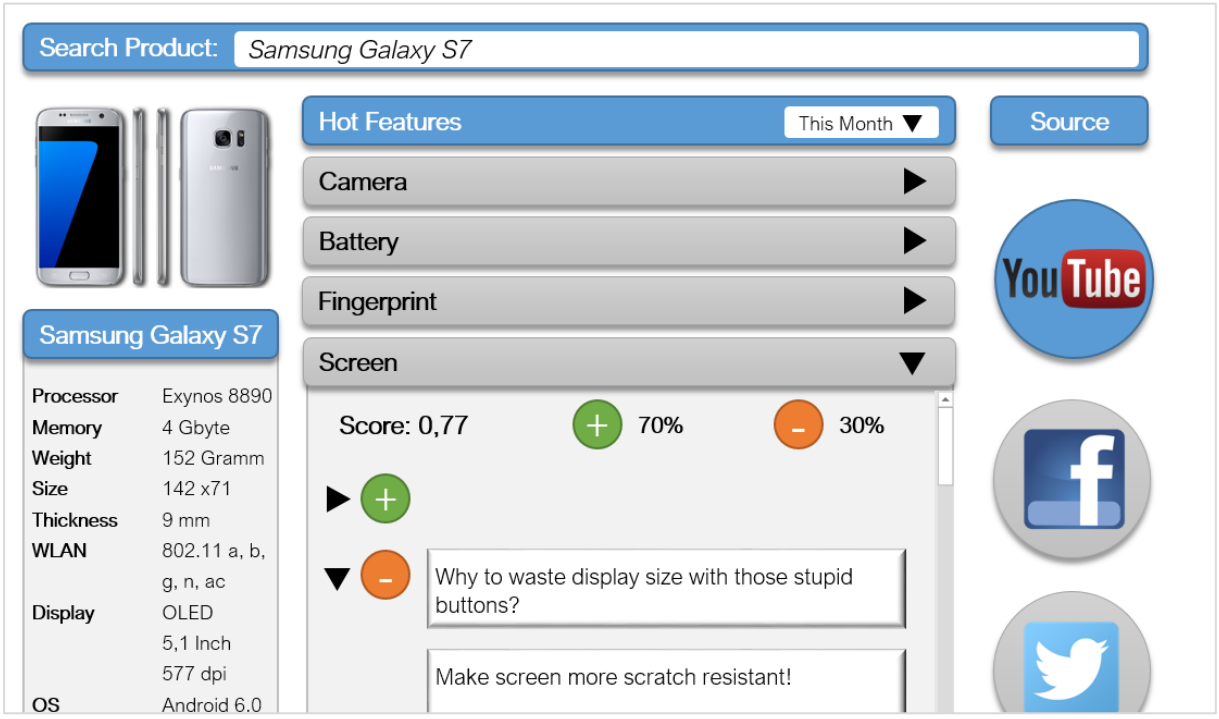
4.1.5 User Interface

Within the interpretation, the results are evaluated by the user who formulated the task of the KDD process. At the user interface that is between the Feature Extractor and the Sentiment Extractor, the user can decide which nouns that are exhibited as potential features are considered and which can be excluded for the further process. For instance, the word "smartphone" was exhibited as a potential feature, whereby it is no attribute, but the device

itself. Moreover, scores of synonym attribute like “battery” and “power” can be aggregated.

Having the features of interest, the Sentiment Extractor is ranking the extracted features using the opinion scores assigned in the previous step. Hereby, one can follow “The Max Opinion Score Algorithm” of Eirinaki et al. (2012). The found adjectives are compared to a list of opinion words. Subsequently, each adjective is assigned a score that indicates the sentiment polarity. For instance, “brilliant” gives a relatively high positive number added to the score, while “catastrophic” subtracts a relatively high number from the score of the associated feature. Finally, the scores are aligned to a number between very negative (-1) and very positive (+1) and printed for each feature.

Figure 12: Proposed User Interface for the Feature and Sentiment Extractor



Within the scope of the subsequent interpretation, interesting patterns are transmitted and visualized to the user of the system. A search engine may be a useful interface for the Feature and Sentiment Extractor, in order to enable the user to issue a query such as a product name, as visualized in Figure 12. Besides, the user can choose the social network he wants to use as a data source. As a side information, the user receives the main facts about the product. The focus of the interface lays on the interactive graphic representation of the most important attributes meaning highly discussed features. The interface ranks the features descending in their importance. For each feature, the polarity score s is ($-1<s<1$) is shown. Moreover, the user can see the shares of positive and negative comments of single features, and make them visible by clicking. The added value of the interface may lay in the complexity reduced overview and an interactive exploration of the products’ sentiments expressed in various social network sources.

advertisement due to their commercial overtone by the community itself (Weinberg et al., 2010). Users are following influencers' opinions because they put more trust into real life or digital product recommendations than an indirect advertisement. Therefore, it is useful for companies to be careful by trying to influence the community. Too obvious or aggressive advertisement as a form of secret infiltration of customer communities can destroy the credibility of a brand and hampers the building up of a valuable network (Weinberg et al., 2010). In a worst case scenario, the consequence of aggressive advertisement is that the swarm mechanisms actively turn the crowd against a company, known as "shitstorms". Therefore, the engagement of companies in social networks should be limited to observation and a respectful dialog (Li & Bernoff, 2009). In this way, the company demonstrates that the customers' opinions are respected and the customers' reliability is stated as a valuable asset. Further research may consider historical social media marketing strategies of companies. A visualization of a social network and the sentiment of its users regarding a certain product on a timeline could be an informative and educationally project. Different influencers and their raising interest could be shown, followed by the sequential reaction of other connected users.

Another point of criticism is the shift in the balance of power due to SA-based decision making. Relying on users' posts for important business decisions increases the impact of customers' opinions for a company. On the one hand that proceed could be called customer-oriented. On the other hand, customers may abuse their power. Small and medium-sized companies already suffer from the power of customers' ratings because of their importance for further sales. Current cases of blackmailing a company by threatening to review products badly, demonstrate that customers start to realize their power (Bock & Seidenspinner, 2016). This shift in the balance of power needs to be explored, in order to develop protective mechanisms. Concrete recommendations, on how further experiments could improve the proposed KDD process practically, was discussed in Chapter 4.2.

Besides companies a summarized and feature-based information carried out with a SA could also be useful from the customer's point of view. According to Rainie and Horrigan (2007), more than half of the American internet users feel frustrated, confused or overwhelmed by the amount of information they have found online while doing online shopping or research.

6 Conclusion

More and more customers are sharing their opinions about products in social media, for instance, by commenting on a tech YouTube video. Many others are seeking and using this advice in terms of product recommendations before making purchases. About two-thirds of customers trust other customers' reviews posted online in order to obtain information. That is why companies depend on a high amount of positive recommendations in social networks to sell their products or service. Thereby, information on the proceedings within the digital product

recommendation process becomes a high value to companies. This paper constitutes SA as a helpful instrument to get an overview of the sentiments expressed in social networks. Especially the feature-based mining of product reviews turned out to be a promising application of SA.

In the first part of this work, the theoretical background was set in order to reach a common understanding. The description of a general KDD process as a "task of discovering interesting patterns from large amounts of data" has been related to text mining and SA as specific KDD processes (Fayyad et al, 1996). The different steps from selection and preparation to mining and interpretation have been examined in order to apply them on the Feature Extractor. A systematic literature review covered research work about data mining in general, text mining and its application fields. The main focus was set on SA, whereby two techniques of SA have been extracted as the most frequent used. While the subjectivity analysis differs objective and subjective texts, the polarity analysis questions whether a text passage expresses a positive, negative or neutral mood. Different applications fields of SA have been considered. While twitter and amazon as data sources received a lot of attention, research about the feature-based SA of YouTube comments is a desideratum. That is why the purpose of this work was to close parts of this gap and to increase the interest in YouTube comments as a research worthy data source.

The second part focused on the chances and challenges of a SA in social networks in general and in particular of YouTube. For consumers, social networks offer an environment for sharing their thoughts, feelings, and opinions, and also for demonstrating their brand-based affiliation. This connection turned out to be very useful for a feature-based SA. Moreover, a high sample quantity due to many data sets and the topicality of data are the main arguments for using this data source. Social networks can be described as graphs of relationships, that are based on binding strength, the construct of social closeness, voluntariness, and frequency of contact. YouTube contains a high amount of community feedback and provides content rich data about community interests. Also, YouTube is well suited for the analysis of social dependencies due to its high level of networking and connections. Finally, the direct assignment of sentiments to product types facilitates the analysis of feature-based product information.

The chances of a SA in social networks have been described using different use cases along a companies value chain. When the structures of the communities have been clarified, a SA can be helpful to understand the feature-based opinion of the community about the own products and the products of competitors. Hereby, results cannot be treated as facts or as a blueprint for the next product development. It has been shown, that the social aspect in the dependencies between communities hampers the communities' swarm intelligence. Moreover, individuals have incentives, to revise their opinions toward conformity and consensus. The overestimation of the swarm's current opinion could follow a collective error. Consequently, the communication process has to take the specifics of social media networks as high dynamics and

rich interdependencies into account and especially use them. Companies can try to move the swarm by adjusting prices and using long-lasting opinion trends for R&D. Another profitable application can be seen in the identification of cross- and upselling potential and the cancellation prevention regarding existing customers. Moreover, they could take part of the communities' interaction directly, to set a valuable network. Otherwise, the company can try to influence the swarm indirectly by connecting to influencers who may trigger their own contacts. A number of tools on the market that support companies regarding SA in social media demonstrates that some companies already started to see potential and efficiency in this kind of information retrieval.

However, the greatest advantages of social media are also the most challenging tasks regarding the KDD-process. The processing of the huge data volume of social networks requires suitable men-, hard- and software. Additionally, high dynamics, meaning continually updated, changed and deleted contents make a real-time SA more difficult. Also, extracting high-value content is the main challenge, since noise like spam or typos has to be dealt with. These challenges point out how important further research in this field is to get more reliable results.

The third and quantitative part described the proposal and implementation of the Feature Extractor. The proceed is transferable to different sources and product. The task this work issued was to retrieve information from users' opinions about the most important features of the Samsung Galaxy S7 expressed as YouTube comments. The data preparation and text mining algorithm followed a pseudocode of Eirinaki et al. (2012). The window- and count-based proceed was described and a user interface was designed. The proceed had several limitations that have been considered afterward. One main limitation was the available data which has not been used, such as meta-ratings, timestamps, and authors. Also, the association of an adjective and a noun turned out to have a lack of considering the semantical context. General problems with NLP systems have been considered. For further experiments, the continuous bag-of-words model and the skip-gram model according to Mikolov et al. 2013 have been suggested to recognize synonyms automatically.

The critical review made the value of crowds' opinions to the subject of the discussion. Social independencies motivate people to revise their opinions toward conformity and, thereby, make a collective error more likely. Besides, SA can even falsify the actual existing common opinion. Limitations are that not all customers are using social networks. And the people who are commenting on products are rather dissatisfied customers. Moreover, comments may be left in the heat of the moment and they are rarely adjusted or removed afterward. However, companies should not try to influence the community too obvious or aggressive. In a worst case scenario, swarm mechanisms could actively turn the crowd against a company and the companies' credibility could be damaged. Therefore, the engagement of companies in social networks should be limited to observation and a respectful dialog. Another point of criticism is the shift

in the balance of power due to SA-based decision making. Since customers are beginning to recognize their power, some may have the intention to abuse it.

In conclusion, it can be said that SA in social networks can improve the communication between customers and companies. Companies have many opportunities to use this opinion-rich data source in their value chain. Especially information to feature-based sentiments gives companies the opportunity to interact with the community and its digital recommendation process actively. The proposed approach demonstrated a simple extraction of features of interest in communities. Thereby, many needs of research have been revealed, especially regarding YouTube contents.

Finally, it can be said, that opinions that are extracted from social networks can be very helpful for companies, especially by replacing manual effort. Otherwise, results of a SA are not replacing neutral numbers and facts. They require an interpretation and a discussion in detail before they are used as the basis for decisions. They are useful as a support for the communication between the company and the customers' web community and should be a complemented part of an integrated R&D, sales, and marketing strategy.