Feature Based Sentiment Analysis of Tweets in Multiple Languages

Maike Erdmann, Kazushi Ikeda, Hiromi Ishizaki,
Gen Hattori, and Yasuhiro Takishima

KDDI R&D Laboratories,
Fujimino Ohara 2-1-15, 356-8502 Saitama, Japan
{ma-erdmann,kz-ikeda,ishizaki,gen,takisima}@kddilabs.jp

Abstract. Feature based sentiment analysis is normally conducted using review Web sites, since it is difficult to extract accurate product features from tweets. However, Twitter users express sentiment towards a large variety of products in many different languages. Besides, sentiment expressed on Twitter is more up to date and represents the sentiment of a larger population than review articles. Therefore, we propose a method that identifies product features using review articles and then conduct sentiment analysis on tweets containing those features. In that way, we can increase the precision of feature extraction by up to 40% compared to features extracted directly from tweets. Moreover, our method translates and matches the features extracted for multiple languages and ranks them based on how frequently the features are mentioned in the tweets of each language. By doing this, we can highlight the features that are the most relevant for multilingual analysis.

Keywords: Feature based sentiment analysis, Twitter, multilingual.

1 Introduction

Nowadays, many people express their satisfaction or dissatisfaction with purchased products on the Internet. This information is invaluable for consumers, product developers, marketing analysts and many others. Since it is impossible to analyze the enormous amount of data manually, sentiment analysis, also known as opinion mining, has become a very popular research area. Especially feature based sentiment analysis is very promising, since it estimates not only the overall sentiment towards a product, but assigns a separate sentiment score for each of the product's features. For instance, "battery" is a typical feature of a smartphone, whereas "engine" is a typical feature of a car.

Traditionally, sentiment analysis is conducted for product review Web sites, which are comparatively easy to analyze, since the reviews are structured well and written in relatively formal language. Since only a small minority of consumers write review articles, some attempts have been made to perform basic sentiment analysis using messages posted on social networking services such as Twitter. An analysis of Twitter messages revealed that a product or brand is

B. Benatallah et al. (Eds.): WISE 2014, Part II, LNCS 8787, pp. 109–124, 2014.

mentioned in about 19% of all tweets [6]. The advantage of sentiment analysis using Twitter is that those messages are up to date and represent the sentiment of a large population on a huge variety of products in many different languages. When a product is sold in multiple countries, it is also interesting to understand the difference of customer satisfaction in those countries. This can be achieved by translating the product names and features of each product and comparing the sentiment analysis results of all languages.

In this research, we conduct feature based sentiment analysis of Twitter messages written in multiple languages. Since the extraction of product features from short and informally written tweets is very difficult, we extract features from online review articles first, and then collect tweets containing both the product name and one of the identified features. In that way, we combine the accurate extraction of features using formally written review texts with up to date sentiment analysis using social networking content.

Besides, we match the features extracted for each language by translating them and arranging them into groups of synonym features. After that, we rank the feature groups to highlight interesting features. Frequently mentioned features are ranked higher than infrequently mentioned features. Features that are mentioned much more frequently in one language than in other languages are emphasized as well.

2 Related Work

A lot of research on sentiment analysis has been described in the last decade. Liu and Zhang [7] as well as Pang and Lee [12] give a comprehensive overview of related work in that area.

The topic of feature based sentiment analysis has also received a lot of attention. Recently, Eirinaki et al. [3] have proposed a method for identifying features by extracting all nouns in the review texts and ranking them by the numbers of adjectives surrounding them. Naveed et al. [9] identify product features from topic keywords created through topic classification with LDA (Latent Dirichlet Allocation) [1] and then estimate sentiment for each product feature separately. Sentiment analysis using dedicated product review Web sites has the disadvantage that those reviews are not necessarily up to date, since most reviews are written shortly after the purchase of a product. Furthermore, only a small minority of consumers review their purchases using review Web sites, so the reviews do not represent all customers' opinions.

Therefore, sentiment analysis using social networking systems is also increasing in popularity. Sentiment Analysis in Twitter was one of the tasks of the International Workshop on Semantic Evaluation (SemEval) [11] in 2013. For instance, Mohammad et al. [13] proposed the usage of hashtags containing opinion bearing terms and emoticons. Günther and Furrer [4] preprocessed the tweets to facilitate their analysis. However, when conducting sentiment analysis using Twitter or other networking services, it is very difficult to obtain accurate results due to the short length, informal writing style and lack of structure of these texts. Moreover, Twitter messages are not composed with the purpose of reviewing a product, but only to share emotions with others in the social network.

Another problem of sentiment analysis using Twitter is that product features are not frequently mentioned explicitly in tweets and are therefore difficult to extract automatically. Therefore, as far as we know, only basic sentiment analysis, not feature based sentiment analysis, has been described for social networking services.

Not much research has been conducted in the area of cross-language sentiment analysis, and most research in that area focuses on creating a language independent sentiment analysis system [2] or applying machine translation in order to reuse a sentiment analysis system created for one language to analyze sentiment in a different language [14].

Nagasaki et al. [10] extract characteristic terms of texts on controversial topics (e.g. the term "extinction" for the topic "whaling") in multiple languages, which is similar to extracting product features. In addition to calculating the frequency rate of each n-gram in the respective language, they translate the n-grams and calculate a cross-lingual frequency rate, i.e. they compare the frequency of the n-gram to the frequency of its translation. However, the authors do not analyze sentiment, but only rank Weblogs according to their relevance to the topic.

Guo et al. [5] are the first researchers to combine cross-language sentiment analysis with feature based sentiment analysis. The authors introduce a method called Cross-lingual Latent Semantic Association which groups semantically similar product features across languages. However, they conduct sentiment analysis on well structured online review articles, which is much easier than analyzing Twitter messages. Besides, they do not attempt to analyze the differences in sentiment in each language beyond listing the sentiment scores side by side.

3 Feature Based Sentiment Analysis

In this section, we describe how we conduct feature based sentiment analysis on Twitter in multiple languages. The system structure is visualized in Figure 1. First, the product name is translated or transliterated[1] from the source language (L1) into the target language (L2) using bilingual dictionaries or machine translation. After that, feature based sentiment analysis is conducted for each language independently, as described in the following subsections. Since the extraction of product features from short and informally written tweets is very difficult, we extract features from online review articles first, and then collect tweets containing both the product name and one of the identified features. Finally, the sentiment analysis results of each language are matched and ranked as described in Section 4.

[1] Transliteration is the process of replacing the letters of a word with corresponding letters in a different alphabet.

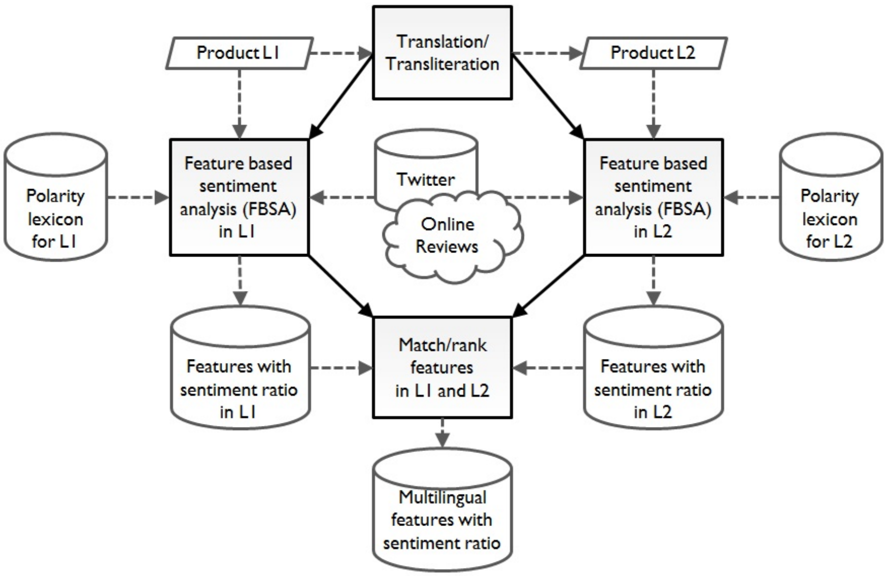

Fig. 1. Overview of the System Structure

3.1 Extraction of Product Features

The feature extraction process is visualized in Step 1a of Figure 2. First, a list of relevant review articles is collected using a Web search engine with a search query consisting of the product name in combination with one or more manually selected keywords, such as “review”, “features” or “specification” in the respective language. The plain text in the collected Web sites is crawled using software to remove all markup language tags as well as text that is not part of the review article (e.g. navigation, advertisement).

For the set of review articles texts, product features are extracted using LDA (Latent Dirichlet Allocation) [1]. In the second step, the top ranked topic keywords for each topic are ranked based on the co-occurrence of sentiment bearing terminology. The more often a topic keyword is accompanied by a sentiment bearing term, i.e. a term identified as having either positive or negative meaning, the more likely the topic keyword can be used as a product feature.

A polarity lexicon is used to decide which terms are sentiment bearing terms. For each occurrence of a topic keyword *tk* with a term in the polarity lexicon *plt*, the feature probability score *fps* increases as follows:

$$fps = fps + \left(\frac{1}{\text{distance of } tk \text{ and } plt} \right) \tag{1}$$

The top ranked topic keywords of each product in each language are then selected as product features.

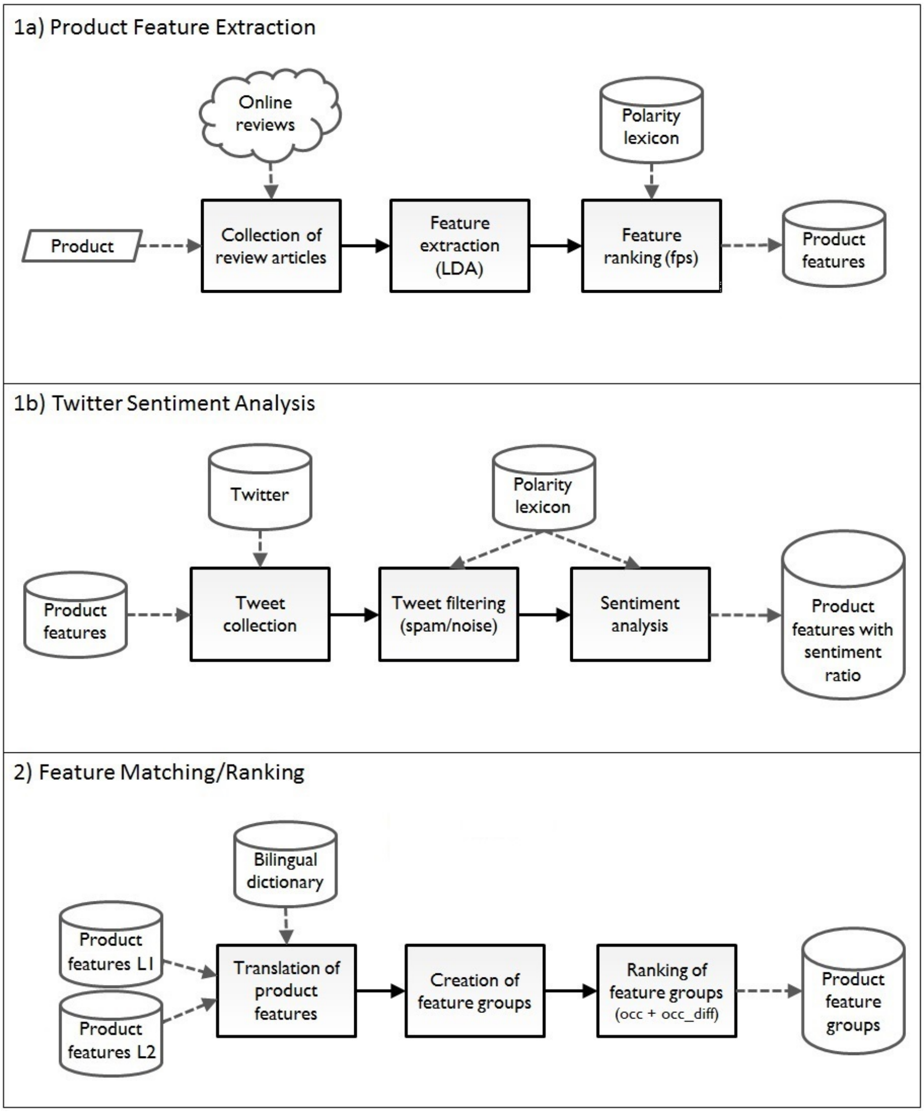

Fig. 2. Flowchart of the Proposed System

3.2 Performing Sentiment Analysis on Tweets

For the top ranked topic keywords estimated to be suitable product features, we collect tweets that contain both the product title and one of the product features as visualized in Step 1b of Figure 2. However, we filter out spam tweets such as advertisements by removing all tweets containing URLs to Web content other than pictures. Besides, we filter out noisy tweets that do not contain any sentiment bearing terminology and thus are not useful for sentiment analysis. The output of this step is a set of tweets for each feature.

Due to the short length of Twitter messages, almost all tweets contain at most one product feature. Therefore, we simplified the sentiment analysis task by assuming that the sentiment of a feature corresponds to the sentiment of the whole tweet. In that way, we can apply a standard sentiment analysis system without having to match the sentiment bearing terminology with the corresponding features. The output of the system is the sentiment ratio for each product feature, i.e. the percentage of positive, negative and neutral tweets.

4 Merging the Results of All Languages

After having extracted all product features and obtained their sentiment ratios in each language, we can translate the features using bilingual dictionaries or machine translation. Then, we arrange them into groups of synonym features, as visualized in Step 2 of Figure 2, and assign a ranking score to each product feature group, that is calculated as follows.

For estimating how frequent the features in a feature group appear in the tweets of the respective languages, the score *occ* calculates the occurrence frequency of a product feature *pf* in the tweets of product p in n languages. This is important, since high ranked features extracted from review article texts do not necessarily appear in the tweets in equally high frequency. Feature groups with high occurrence rates are scored higher than those with low occurrence rates.

$$occ(p, pf) = \left(\frac{tr(p, pf, l_1) + tr(p, pf, l_2) + \ldots + tr(p, pf, l_n)}{|n|}\right) \tag{2}$$

The function *tr(p,pf,l)* calculates the ratio of tweets for product p in language l containing product feature *pf*.

The score *occ_diff* calculates the difference in occurrence ratios among languages by subtracting the tweet ratio of the language with the lowest occurrence rate from the tweet ratio of the language with the highest occurrence rate. Feature groups with high occurrence rates in one language but low occurrence in other languages score higher than those with similar occurrence in all languages.

$$occ\_diff(p, pf) = max\_tr(p, pf) - min\_tr(p, pf) \tag{3}$$

The function max_tr(p,pf) calculates the tweet ratio *tr(p,pf,l)* for the language l that has the largest tweet ratio within the feature group. Correspondingly, the

Table 1. Products Used in Experiment

Category	Products
Smartphones	iPhone 5S, iPhone 5, iPhone 4S, Nexus 5, XPeria Z1, Galaxy S4
Cars	Prius, Lexus, Corolla, Nissan GT-R, Infiniti, Lancer Evolution, Impreza

function min_tr(p,pf) calculates the tweet ratio *tr(p,pf,l)* for the language l with the smallest tweet ratio within the feature group.

Finally, the overall score is calculated as the weighted sum of the two scores *occ* and *occ_diff* and the highest ranked feature groups are displayed as results.

5 Experimental Results

In this section, we explain and discuss an experiment in which we analyzed the sentiment for 13 products (6 smartphones and 7 cars) expressed in English and Japanese tweets.

5.1 Preparation of the Experiment

In order to extract features from tweets, we collected all tweets containing one of the 13 product titles in the time between December 2013 and March 2014. The products are listed in Table 1. For both categories, we selected products that are popular in the USA as well as in Japan. This limited us to only Japanese car brands, since the market share of foreign car brands in Japan is marginal. The diversity of foreign smartphones sold in Japan is also small, thus half of the smartphones in the experiment are Apple products.

Of course, it would also be interesting to select products that are not popular in one country, in order to find out the reasons why the product is not sold well. However, in order to evaluate our feature extraction method accurately, we had to make sure that we can collect enough data in both languages for each product.

Before starting the experiment, we created a large polarity lexicon by combining the data of several existing polarity lexicons[2,3]. Since polarity labeling is very subjective, i.e. the polarity of terms often depends on the context, terms were sometimes labeled differently in one lexicon than in other lexicons. We resolved these conflicts by choosing the most frequent label for each term. In that way, we were able to not only create a large scale polarity lexicon, but also improved the labeling quality.

In the first part of the experiment, we compared features extracted from online review articles to features extracted directly from tweets to show that features extracted from review articles are significantly more accurate than features extracted from tweets.

[2] English polarity lexicons: MICRO-WNOP, SentiWordNet, MPQA, AFINN, DeRose, McDonald, Avaya.

[3] Japanese polarity lexicons: Inui-lab, Avaya.

Table 2. Examples of Removed Tweets

Product	URL	No Sentiment	Example Tweet	Filter
iPhone 5S	no	no	Should I get an IPhone 5S?? Because there is just so many nice iPhone 5S covers.	Keep
	no	yes	I don't know if I should get an iPhone 5S or a Galaxy Note 3...#thestruggle	Remove
	yes	no	Best Free Games for the iPhone 5S: http://t.co/cioZSBHaLB via @youtube	Remove
	yes	yes	Here's YOUR CHANCE to WIN the NEW iPhone 5S! http://t.co/lCs9hGjQeT	Remove
Prius	no	no	Prius' sound like spaceships but that's pretty much the only cool thing about them.	Keep
	no	yes	I think @aaabbbiiii is the only person in our school with a Prius	Remove
	yes	no	Great deals, everday low price on Prius #Prius Share this http://t.co/BSjvJpf1YQ	Remove
	yes	yes	Used 2010 #Toyota #Prius, 82,617 miles, listed for $16,000 under used cars http://...	Remove

We noticed that about 80% of the collected tweets were either spam tweets such as advertisement or noise, i.e. tweets containing no sentiment, that interfere with the identification of accurate product features. Therefore, we decided to remove all tweets containing URLs to Web content other than pictures (spam tweets) and all tweets not containing any sentiment bearing terminology (noisy tweets). A few example tweets filtered out by these two rules are shown in Table 2. After filtering, approximately 1,250,000 English tweets and 250,000 Japanese tweets remained for the smartphone category. For the car category, about 260,000 English and 290,000 Japanese tweets remained.

In order to extract features from review article texts, we collected the top 10 search engine results for each of the product names in combination with the keyword "review". Then, we applied the boilerpipe[4] software to remove all markup language tags and parts of the Web site that are not part of the review article (e.g. navigation, advertisement).

5.2 Extraction of Product Features

We extracted topic keywords from both tweets (baseline) and reviews articles (proposed method) using tf-idf, df-idf and LDA. We decided to extract one set of features for the smartphone category and one set of features for the car category, since the accuracy is slightly higher than for keywords extracted for each product separately. We applied the feature extraction method described in Section 3.1. The top 100 topic keywords extracted of each method were ranked again according to their feature probability score *fps*. For keyword extraction

[4] http://code.google.com/p/boilerpipe/

Table 3. Precision of Top 20 Features

	Smartphones		Cars		
	English	Japanese	English	Japanese	Average
tweets	0.15	0.3	0	0.05	0.125
reviews	0.55	0.45	0.65	0.5	0.538

Table 4. Precision of Top 50 Features

	Smartphones		Cars		
	English	Japanese	English	Japanese	Average
tweets	0.24	0.14	0.04	0.08	0.125
reviews	0.5	0.42	0.54	0.38	0.46

using LDA, we set the number of topics to 10 in the same way as described in related work [5], and extracted the top 10 keywords for each topic.

All extracted topic features were manually evaluated. Since this is a very subjective decision, each topic keyword was evaluated by three judges and the evaluation of the majority of them was used to decide whether a topic keyword is a suitable product feature. Generally, a keyword is considered to be a suitable feature if it meets two criteria. First, the keyword needs to be a part of the product, thus it must be possible to describe the relationship in a "has a" expression. For instance, the "iPhone 5S" has a "camera", but the "iPhone 5S" does not have an "Apple". Second, a feature must directly impact the product quality. The quality of the "Prius" car, for example, is affected by the quality of the "engine", but it is not affected by the quality of the "road".

LDA ranked with the feature probability score *fps* achieved the best overall performance. Therefore, we show only the results of that method. Table 3 shows the precision of the top 20 features and Table 4 shows the results of the top 50 features. As the results show, the precision of our proposed method (features extracted from reviews) performed significantly better than the baseline method (features extracted from tweets). While the performance of the proposed method is certainly not optimal, the precision increased by about 40% for the top 20 and by about 30% for the top 50 features. Besides, even if we replace the feature extraction method by a different method, such as the method proposed by Eirinaki et al. [3] or Naveed et al. [9], the proposed method is likely going to perform better than the baseline method.

The top ranked features for each category are shown in Table 5 (tweets) and Table 6 (reviews). The Japanese keywords are translated into English. The plus and minus signs next to the keywords indicate whether the keyword is a suitable product feature. Notable is that the features extracted from English tweets contain many common terms such as "love", "day" or "people". We believe that this is because only a small percentage of tweets explicitly mention features of a product. The features extracted from Japanese tweets often contained terms

Table 5. Examples of Features Extracted from Tweets

(+) = suitable feature (−) = unsuitable feature

Smartphones		Cars	
English	Japanese	English	Japanese
(+) charger	(+) battery	(−) car	(−) chan
(−) Samsung	(+) cover	(−) love	(−) follow
(+) case	(−) Nokia	(−) day	(+) wheel
(−) love	(+) case	(−) people	(−) fugue
(−) day	(+) ios	(−) LOL	(−) Super
(−) shit	(−) SIM	(−) Mom	(−) crown
(−) year	(−) Docomo	(−) shit	(−) work
(−) time	(−) MNP	(−) girl	(−) Gran Turismo
(−) people	(−) running	(−) bitch	(−) legacy
(−) Christmas	(−) debut	(−) gt	(−) taxi

that are somehow related but not suitable as product features. (e.g. “Nokia”, “SIM”, “Super”).

In the next step, we confirmed that the product features extracted from review articles correspond to the terms that are frequently used in tweets describing the products. While we did not conduct a formal evaluation, we identified only a small number of mismatches. The product feature “display” was rarely found in smartphone related tweets and the product features “interior” and “seat” were rarely found in car related tweets, although these terms frequently appear in review article texts. On the other hand, the terms “charger” and “price” for smartphones and the terms “wheel” and “exhaust” for cars frequently appear in tweets but not in review articles, thus they were not identified as product features. However, the majority of product features seems to match the terms that are commonly used in tweets.

Table 6. Examples of Features Extracted from Reviews

(+) = suitable feature (−) = unsuitable feature

Smartphones		Cars	
English	Japanese	English	Japanese
(−) phone	(−) phone	(−) car	(−) car
(−) Apple	(+) camera	(−) drive	(+) engine
(+) camera	(+) app	(+) front	(−) model
(+) display	(+) Android	(+) interior	(+) hybrid
(+) screen	(+) LTE	(−) ride	(−) corner
(+) app	(−) user	(+) engine	(+) power
(−) feature	(+) display	(+) seat	(+) tire
(+) ios	(+) battery	(+) rear	(+) brake
(+) design	(+) mail	(+) handling	(−) gasoline
(−) Sony	(+) size	(−) model	(+) seat

5.3 Sentiment Analysis Results

Before conducting sentiment analysis for the extracted product features, we calculated the number of features per tweet. Almost all tweets did not contain any features. Of the tweets in which features were detected, about 89.6% contained only one feature, 9.2% contained two features, and only 1.2% contained three or more features. Therefore, we simplified the process of sentiment analysis of tweets by regarding the sentiment of a feature as equivalent to the sentiment of the whole tweet.

After that, we ranked all feature groups (pairs of English and Japanese features) according the sum of their *occ* and *occ_diff* scores (see Section 4), to prioritize the ones that are most interesting for the user. Again, we did not formally evaluate the results, but they corresponded well to our intuitive ranking of the features. The top ranked feature groups created from the top 100 English and top 100 Japanese feature are shown in Table 7 (smartphones) and 8 (cars). Missing tweet ratio scores indicate that the feature was not extracted in the corresponding language. The plus and minus signs next to the features indicate whether they are suitable product features. In the smartphone category, 10 feature groups (59%) are suitable features, whereas 9 feature groups (45%) in the car category are suitable features.

Only for the top ranked feature groups that were considered useful features, we analyzed the sentiment using a simple lexicon based sentiment analysis system [8]. For each product feature, we extracted the newest 1,000 tweets containing both the product title and the product feature. The results of two example products, “iPhone 5S” and “Prius”, are shown in Figure 3. The left bar of each feature group represents the English and the right bar the Japanese sentiment.

For the “iPhone 5S”, the sentiment expressed in the English tweets is generally more negative than the sentiment expressed in the Japanese tweets. This is not surprising, given that the market share of the “iPhone 5S” is significantly higher in Japan than in the USA and other English speaking countries. The highest

Table 7. Top 20 Ranked Feature Groups (Smartphones)

(+) = suitable feature (−) = unsuitable feature

feature	tweet ratio (en)	tweet ratio (jp)	*occ*	*occ_diff*	score
(−) Docomo	–	0.347	0.174	0.347	0.521
(−) phone	0.283	–	0.146	0.283	0.429
(+) Android	0.018	0.100	0.059	0.082	0.141
(−) day	0.083	–	0.043	0.083	0.126
(−) data	–	0.079	0.039	0.079	0.118
(−) year	0.076	–	0.039	0.076	0.115
(+) case	0.067	–	0.034	0.067	0.101
(−) people	0.064	–	0.033	0.064	0.097
(−) user	–	0.060	0.030	0.060	0.091
(+) app	0.016	0.063	0.040	0.047	0.087
(+) LTE	–	0.051	0.026	0.051	0.077
(+) battery	0.022	0.054	0.038	0.033	0.071
(+) size	0.005	0.048	0.027	0.044	0.070
(+) mail	–	0.042	0.021	0.042	0.063
(−) Apple	0.038	–	0.020	0.038	0.058
(−) upgrade	0.031	–	0.016	0.031	0.047
(+) screen	0.031	–	0.016	0.031	0.047
(+) camera	0.035	0.030	0.033	0.005	0.039
(−) game		0.021	0.011	0.021	0.032
(+) picture	0.021	–	0.011	0.021	0.032

difference can be observed for the features "app" and "battery". For "app", only 29% of the English tweets are positive whereas 71% of the Japanese tweets are positive. When analyzing the corresponding tweets manually, we discovered that many English speaking users state that applications crash frequently. Japanese users also state problems with the new operating system, but also praise specific applications, such as the new fingerprint recognition application and the new camera application. For "battery", 40.6

For some features of the "iPhone 5S", sentiment in only one of the languages could be obtained. The feature "LTE", for instance, appears only in Japanese tweets. The reason for that might be that many Japanese people use smartphones while commuting in the train or even replace their home computer with a smartphone. Therefore, they rely on fast Internet access technology more than users in other countries and mention it more frequently in their tweets.

For the car "Prius", the sentiment of English and Japanese tweets is slightly more positive than the sentiment of Japanese tweets. The most notable difference is the sentiment for the feature "engine", for which English tweets express a much more positive sentiment than Japanese tweets (62.4% positive tweets for English, 25.6% positive tweets for Japanese). Deeper analysis of the tweets revealed that while users in both languages complain about the rather small engine power of the Prius, only Japanese tweets declare the small engine sound of the "Prius" as a safety hazard. Streets in Japan are very narrow and often lack sideways, thus pedestrians are afraid of colliding with a car that they cannot hear approaching.

left bar = English sentiment, right bar = Japanese sentiment

Fig. 3. Examples of Sentiment Analysis Results

Table 8. Top 20 Ranked Feature Groups (Cars)

(+) = suitable feature (−) = unsuitable feature

feature	tweet ratio (en)	tweet ratio (jp)	*occ*	*occ_diff*	score
(−) car	0.388	0.037	0.216	0.357	0.573
(−) Aqua	–	0.316	0.158	0.316	0.473
(−) people	0.141	–	0.071	0.143	0.214
(+) engine	0.012	0.109	0.061	0.097	0.158
(+) hybrid	–	0.101	0.050	0.101	0.151
(+) brake	–	0.084	0.042	0.084	0.126
(−) road	0.053	–	0.027	0.054	0.081
(−) year	0.052	–	0.026	0.052	0.079
(−) drive	0.045	–	0.023	0.046	0.068
(+) front	0.045	–	0.023	0.045	0.068
(+) motor	–	0.038	0.019	0.038	0.057
(+) tire	0.016	0.043	0.030	0.027	0.057
(−) mile	0.030	–	0.015	0.030	0.045
(−) gasoline	–	0.029	0.015	0.029	0.044
(−) fun	0.028	–	0.014	0.029	0.043
(+) MPG	0.028	–	0.014	0.028	0.043
(−) lot	0.027	–	0.014	0.028	0.042
(+) acceleration	0.001	0.028	0.015	0.027	0.041
(+) steering wheel	–	0.024	0.012	0.024	0.036
(−) silver	–	0.022	0.011	0.022	0.032

The "Prius" sentiment analysis results also show that features in different languages should be matched when they are semantically related and not only when they are translations of each other. For instance, the English feature "MPG" (miles per gallon) could be grouped with the Japanese feature "hybrid", because they are both related to the concept of fuel efficiency. Moreover, the results of the features "engine" and "motor" should be merged, since they are interchangeable.

6 Conclusion and Future Work

In this paper, we introduced a method for performing feature based sentiment analysis on Twitter messages in multiple languages.

Sentiment analysis using tweets is invaluable, since Twitter users express sentiment towards a huge variety of products in many different languages, and because sentiment expressed on Twitter is more up to date and represents the sentiment of a larger population than review articles.

Since the extraction of product features from short and informally written tweets is very difficult, we extracted features from online review articles first and then collected tweets containing both the product name and one of the identified features. In an experiment with English and Japanese Twitter messages on 6 smartphones and 7 cars, feature extraction from review articles increased the precision of the extracted features by about 40% for the top 20 and by about 30% for the top 50 features.

Moreover, we proposed a system to highlight the features that are most relevant for multilingual sentiment analysis. We translated the English and Japanese features and arranged them into groups of synonym features. After that, we ranked the feature groups according to how frequent the features in a feature group appeared in the tweets of the respective languages and according to the difference in occurrence ratios among languages.

Our proposed system allows consumers, product developers and marketing specialists of internationally operating companies to track the sentiment of their products world-wide.

In the future, we want to improve the precision of feature extraction, since only about half of the extracted features in our experiment were useful for sentiment analysis, and combine sentiment analysis of semantically similar features (e.g. "screen" and "display"). Furthermore, we are planning to perform a deeper analysis of the sentiment results in each language, since it is very important to understand not only how sentiment differs among languages but also why.

Apart from that, we are interested in performing feature based sentiment analysis not only on products but on e.g. news, services or events. Nagasaki et al. [10] extracted characteristic terms of English and Japanese texts on controversial topics such as "whaling" or "organ donations" and detected very interesting differences in the terminology used in the texts of each language. Combining the extraction of characteristic terms with sentiment analysis can help us why people from different countries have different opinions on many topics.

References

1. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. The Journal of Machine Learning Research 3, 993–1022 (2003)
2. Cui, A., Zhang, M., Liu, Y., Ma, S.: Emotion tokens: Bridging the gap among multilingual twitter sentiment analysis. In: Salem, M.V.M., Shaalan, K., Oroumchian, F., Shakery, A., Khelalfa, H. (eds.) AIRS 2011. LNCS, vol. 7097, pp. 238–249. Springer, Heidelberg (2011)
3. Eirinaki, M., Pisal, S., Singh, J.: Feature-based opinion mining and ranking. Journal of Computer and System Sciences 78(4), 1175–1184 (2012)
4. Günther, T., Furrer, L.: Gu-mlt-lt: Sentiment analysis of short messages using linguistic features and stochastic gradient descent. In: Proceedings of the International Workshop on Semantic Evaluation (SemEval), pp. 328–332 (2013)
5. Guo, H., Zhu, H., Guo, Z., Zhang, X., Su, Z.: Opinionit: A text mining system for cross-lingual opinion analysis. In: Proceedings of the ACM International Conference on Information and Knowledge Management (CIKM), pp. 1199–1208 (2010)
6. Jansen, B.J., Zhang, M., Sobel, K., Chowdury, A.: Twitter power: Tweets as electronic word of mouth. Journal of the American Society for Information Science and Technology (JASIST) 60(11), 2169–2188 (2009)
7. Liu, B., Zhang, L.: A survey of opinion mining and sentiment analysis. In: Mining Text Data, pp. 415–463 (2012)

8. Tofiloski, M., Voll, K., Taboada, M., Brooke, J., Stede, M.: Lexicon-based methods for sentiment analysis. Computational Linguistics 37(2), 267–307 (2011)
9. Naveed, S.S.N., Gottron, T.: Feature sentiment diversification of user generated reviews: The freud approach. In: Proceedings of the International AAAI Conference on Weblogs and Social Media (ICWSM), pp. 429–438 (2013)
10. Nakasaki, H., Kawaba, M., Utsuro, T., Fukuhara, T.: Mining cross-lingual/cross-cultural differences in concerns and opinions in blogs. In: Proceedings of the International Conference on Computer Processing of Oriental Languages. Language Technology for the Knowledge-based Economy (ICCPOL), pp. 213–224 (2009)
11. Nakov, P., Kozareva, Z., Ritter, A., Rosenthal, S., Stoyanov, V., Wilson, T.: Semeval-2013 task 2: Sentiment analysis in twitter. In: Proceedings of the International Workshop on Semantic Evaluation (SemEval), pp. 312–320 (2013)
12. Pang, B., Lee, L.: Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval 2(1-2), 1–135 (2008)
13. Saif, X.Z., Mohammad, M.: Svetlana Kiritchenko. Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets. In: Proceedings of the International Workshop on Semantic Evaluation (SemEval), pp. 321–327 (2013)
14. Wan, X.: Co-training for cross-lingual sentiment classification. In: Joint conference of the Annual Meeting of the ACL and the International Joint Conference on Natural Language Processing of the AFNLP (ACL-IJCNLP), pp. 235–243 (2009)