# Comparing Brand Perception Through Exploratory Sentiment Analysis in Social Media

Mario Cichonczyk and Carsten Gips

Bielefeld University of Applied Sciences, Minden, Germany
firstname.lastname@fh-bielefeld.de

**Abstract.** The presented student project outlines a natural language processing pipeline for brand metric comparison in the Twitter ecosystem. Sentiment calculation for an unlabeled data set is demonstrated and calibrated using the statistical Central Limit Theorem as a guidance to anchor the sentiment indicator in a homogeneous market. The process is evaluated by comparing the sentimental market performance of three leading German logistics companies. A support for the value of sentiment analysis for automated customer feedback analysis in real-time is concluded.



## 1 Introduction

The brand philosophy behind a business is usually a driving principle of the entrepreneurial actions it follows. Ideally, these actions accumulate in a brand strategy and a finely tuned marketing mix to acquire market share and establish brand awareness and perception. The success of these marketing efforts can be measured by evaluating the time-delayed return on investment of associated profit margins. This approach has a deficit in explanatory power as it lacks a fine-grained insight into the complex effects of diversely faceted, multi-channel marketing and brand positioning methods. Therefore, marketing research relies on qualitative and quantitative analyses and surveying techniques for a more sophisticated evaluation of marketing investment impact. Targeted studies with resource expenditure are employed to answer specific questions of subjective brand perception. With technological development and progress, new approaches may be introduced to increase the efficiency of effect monitoring and thereby reducing inertia in strategic realignment according to market feedback. [18]
Since social media is getting more established in everyday life, intelligence can be gathered through a new and essentially cost-free feedback channel [12]. While social media marketing is concerned with the public relations effort in direction to the customer, the same platforms allow for an inversion of communication from consumer to business. The presented work explores how brand perception can be measured and compared by making use of natural language processing

in the Twitter ecosystem. To achieve this objective, the hashtag space of leading German logistics companies was analyzed and collected as an exemplary use case. An analytics pipeline was constructed and applied to the dataset to better understand the customer base and approximate overall consumer sentiment towards the logistics brands, linking a scalar metric to consensus opinion as outlined in [3]. This approach can show the value of sentiment analysis in social media analysis for customer satisfaction research of single companies and it will also allow for contextualization in regard to competitors. As social media is essentially a means of open many-to-many communication, the same pipeline can be applied to feedback aimed towards other brands and will then allow a more empirical comparison.

## 2 Approach

The main contribution of this work aims to show how current ideas in natural language processing may be applied to take advantage of automated social media brand perception analysis. We would like to outline our course of action and the thought process used in this applied research project as it may inspire other research towards the stated problem case. For this purpose, two typical marketing query questions were choosen as an example to demonstrate the approach:

1. How do leading German consumer logistics brands rank in Twitter user approval?
2. How can Twitter meta information augment user approval analysis?

Related works predominantly make use of machine learning on labeled data with the primary goal of method testing and evaluation [28][15][26]. These approaches perform well regarding categorization in sentiment classes, but they cannot be directly transferred to the investigated problem without manual data annotation as a pre-step for model training. Real-world data is unlabeled and seed datasets do not yet exist for the highly specific domain of German logistics, hence seemingly producing an unsupervised clustering problem with feature engineering and selection in focus.
The specific characteristics of brand perception analysis allows for a third option. Munoz & Kumar [23] point out that perceptional metrics help to gauge the effectiveness of brand-building activities across points of customer interaction. Brand profiling can be achieved by fixing an indicator within a metric of a market and then comparing this indicator to a companies brand and its competitors [23]. Therefore, the aim of the presented approach is to acquire such a polarity indicator. Tools, models and methods from both supervised and unsupervised natural language classification and processing become available with this modified problem definition. The constructed data pipeline adheres to this premise and is presented in detail in the following paragraphs.

### 2.1 Collection and Transformation

Data acquisition was implemented using the Tweepy[1] library for the Python programming language to interact with the commercially available Twitter API[2] for developers. Tweepy was chosen for its native ability to handle rate limited free access tokens and dynamic adjustment of traffic bandwidth. To stay within these limits and to focus on *customer* opinion, only the top three logistics companies for the courier, express and parcel services in Germany where selected, which as of 2018 are DHL, Hermes and DPD [13]. Therefore, the Tweepy filter

```
"#dhl OR #hermes OR #dpd OR dhl OR hermes OR dpd"
```

was constructed and 10594 tweets where collected over 3 weeks in the winter of 2018, limited to tweets with a set "German" language flag. The Twitter API returns tweets in JSON format, containing a large number of partly redundant data fields. All JSON data was parsed and attributes of interest to the research question where selected and named accordingly. The selected and renamed fields where "usr_id", "tweet_id", "usr_followers", "timestamp", "favorites", "retweets", "client", "hashtags" and the "text" itself. After transformation, the tweets where stacked as rows to form a 10594x9 matrix $\mathbf{M}$.

### 2.2 Data Exploration, Filtering and Feature Construction

First exploration and inspection of the dataset had shown that the Tweepy filter was not sufficient to separate the communications concerning the three selected logistics brands in a satisfactory manner. The data contained tweets which did not directly relate to the domain of interest. It became clear that a more in-depth analysis of the communication topics was necessary to further sanitize the results. Tweet topic modeling through utilization of keyword annotations - better known as "hashtags" - was identified to hold valuable advantages for the solution of this problem [31][29]. Based on the work presented by Wang, Wei & Zhang [30], a graph model was defined by the set of all hashtags $\mathcal{H} = \{h_0, h_1, ..., h_m\}$ contained within the dataset, where each hashtag $h_i$ represents a node weighted by its global occurence count and is associated with a set of tweets $\mathcal{T}_k = \{\tau_0, \tau_1, ..., \tau_n\}$ in which it occurs. The set of edges $\mathcal{E}$ consists of a link between two hashtags if they co-occur in the same tweet. The weight of an edge $e_{ij}$ between $h_i$ and $h_j$ is incremented for each co-occurence of $h_i$ and $h_j$. The graph model $\mathbf{HG}= \{\mathcal{H}, \mathcal{E}\}$ was used to isolate the logistics subgraph of interest which reduced $\mathcal{T}$, and thereby the rows of $\mathbf{M}$, to only hold tweets relevant to the research.
The "hashtags" column of $\mathbf{M}$ was transformed to $\mathbf{HG}$ and stored in GEXF-format[3]. This step made it possible to import the hashtag graph into existing graph analysis software[4], providing access to pre-optimized methods. The Yifan

[1] https://github.com/tweepy/tweepy
[2] https://developer.twitter.com/
[3] https://networkx.github.io/documentation/stable/reference/readwrite/gexf.html
[4] https://gephi.org/

Hu multilevel layout algorithm was chosen for its speed and good quality on large graphs [11] to achieve topic modeling. With this technique, all node and edge weights are used as force simulations of attraction and repulsion, creating a dispersed planar graph projection. Clusters of hashtags in frequent co-occurence form topic subgraphs while unrelated topics drift apart. After the embedding step, topics weakly connected to logistics can be visually identified.
The collective topics "dhl", "dpd" and "hermes" were found to form a distinct subgraph with minor outliers regarding "jobs" topics. Stronger interrelations to unrelated themes existed for the single topic "hermes". This effect can be attributed to the ambiguity of the term. Subgraphs concerning "fashion" and "export politics" where intertwined with tweets about the logistics company. The set of hashtags identifying those outlying tweets was added as a filter to remove non-logistics rows from **M**, resulting in an on-topic dataset.
Tweets which did not resemble a consumer opinion where filtered out, e.g. advertisements and news where undesirable data as they would distort the sentiment analysis. Therefore, the column "client" was investigated. Under the assumption that consumer opinion is predominantly voiced through consumer client software, other agents can be ignored. All user agents of the "client" column were manually mapped to the categories "Android", "iOS", "Desktop", "Other" and "Professional". "Other" resembles multi-platform clients which cannot directly be associated with a single platform. The "Professional" label identifies all user agents known to be developed for commercial usage, e.g. tweet automation or social media management software. All "Professional" rows where removed. Commercial users who rely on consumer client software are exempt from this pruning step. These were further analyzed by the count of their followers. For all rows in **M**, the unique "usr_id" fields where collected and then ranked by their "usr_followers" value. Since the Twitter follower count adheres to the power law [22], manually inspecting the top followed profiles was sufficient to mark opinion bearing commercial accounts for removal and neutralize their influence on global sentiment. For professional tweets which may have been remaining in the data after all filters, it was assumed that their opinion influx was significantly outweighed by the now superior consumer class tweets.
To finally conclude filtering and feature construction, **M** was extended by the columns "is_dhl", "is_dpd", "is_hermes", "hour" and "weekday". With these extra columns, searching and querying the dataset for the following steps is faster and more convenient. Finding tweets associated with specific brands is done by simple boolean masking, resulting in desired subsets of **M** without having to repeatedly parse and search tweet content for each query. Aggregation over time windows is sped up by pre-splitting the complex timestamp format of the Twitter API.
After the outlined data sanitation treatment, actual text analysis was possible.

### 2.3 Text Preprocessing

The Natural Language Tool Kit (NLTK) offers the basic functionalities necessary for natural language processing [2]. As such, it was used to pre-process the

tweet content information formulated by Twitter users. The "text" column was tokenized and all single tokens of interest were added as a new "tokens" column, containing a list of tokens for each tweet. All tokens were stripped of non-textual information, URLs removed, umlauts converted and otherwise undesirable information filtered out. The sanitized tokens represent all German words of a tweet, but not all words hold analytic information. NLTK provides a list of stop words for several languages, including German. Accordingly, all tokens where checked against the German NLTK stop word list and removed if they where considered a match. Afterwards, the term frequency of all tokens was calculated to identify other possible stop words. Several high ranking tokens were found to lack value for analysis:

```
"dhl, dpd, hermes, paket, mal, dass, schon, kommt, immer, seit,
                              fuer"
```

These were added to the stop word list and also removed. Brand name tokens are redundant as their information is already stored in **M** (see Paragraph 2.2). With the token list constructed, semantic polarity estimation can be examined on a word level.

### 2.4 Token Polarity

Since algorithms do not by default have any understanding of the emotional impact of word semantics, sentiment analysis relies on human consensus opinion. Databases with annotated word polarities between [-1, 1] for negative and positive sentiment respectively are used to look up a scalar value for a given token. For the German language, such a dictionary exists in the form of the SentiWS [27] project. As a first, naive approach, the pipeline's sentiment resolver tries to annotate the tokens of **M** directly through a query to SentiWS. If the word is present in the dictionary, no further search is required and its sentiment can be returned. SentiWS contains about 3500 basic forms and 34000 inflections. Despite this size, less than 15% of all tweet tokens where found in SentiWS. This result is not surprising given the combination of language syntax complexity, a raw tweet tokens count of more than 80000 and the effort involved in dictionary building. It was expected that only a low number of entries would directly match. Therefore, each token that could not be resolved in SentiWS is forwarded to the NLTK German stemmer. Stemming is the process of reducing a complex word, which might be an inflection, to its basic form [19]. This step can be understood as a simplification of the token, or in a more technical sense even a functional projection. The goal is to project the token space sample in such a way that its transformation aligns with the SentiWS target space. Thereby, a morphological alternative is potentially found and might allow for a sentiment look up. Increasing the number of token morphemes using this method also increased polarity coverage.
If the stemmer was not able to find a morpheme in SentiWS, syntactical alternative search is exhausted. Therefore, the actual *meaning* of the token can

be used to find synonyms whose sentiment is known. Liebeck [17] found the introduction of semantic equivalence to be of advantagous value for more thorough sentiment analysis and referred to synonyom search in synset databses like GermaNet [10]. Mohtarami [21] et al. explored and alternatively suggested the use of vector-based approaches for the same purpose. They observed that WordNet [20] - and therefore GermaNet as a descendant - perform satisfactory when *semantic* synonyms are searched but lack in accuracy when *sentimental* equivalence is the primary metric. To accomodate this problem, they introduced emotional features of words to construct their vectors. The key insight for the presented work is the necessity of a more general human-like language intelligence to identify word alternatives beyond pure semantic synonyms. Webber [32] came to the same conclusion and presented a proof of concept disambiguation and language analysis system trained on half a million Wikipedia articles instead of a domain specific corpus. The results suggested superior context-based general language processing capabilities. Therefore, a similar approach was chosen. Instead of synonym search in statically assembled synsets, a Word2Vec model trained on Wikipedia articles was used to find alternatives for tokens which neither themselves nor their stemmed variants could be resolved through SentiWS. Word2Vec was chosen due to its proven performance [8] regarding this purpose and because a large (650 million words), pre-trained, general Wikipedia knowledge model already exists[5] for the gensim[6] Word2Vec implementation. Training a similarly large model would not have been possible for the purposes of this student project.

The lookup process was constructed as follows. Gensim is used to retrieve a vector space embedding for the token, e.g. the word "house". As stated above, this vector representation shares a topological vicinity with its contextual synonyms, e.g "building" or "home". The aim is to find a spatially close synonym which can be associated with an entry in SentiWS. Starting from the "house" embedding, its neighbouring entries are probed in order of increasing distance. For example, "building" may highlight as the closest approximation to "house". The word "building" is therefore chosen as an alternative candidate. This candidate is then tested against SentiWS and if a polarity value could be retrieved, the candidate is selected. In this case, "building" may not have been a valid alternative, is rejected and the process continues with the next neighbour in increasing distance, which is "house". The new candidate is tested in the same manner and can be successfully matched with a sentiment, leading to its selection as a valid alternative to "house". The algorithm encapsulates the same process a human would employ in thinking about other phrasings of the same utterance. Figuratively speaking, the amount of imagination necessary to come up with another phrase is continually incremented until a suitable rephrasement is found. This can of course lead to misleading levels of synonymity if done ad infinitum. Therefore, it was decided to penalize the retrieved sentiment by the distance between the original token vector and its alternative, similarly to Kim & Shin [14].

[5] https://github.com/devmount/GermanWordEmbeddings
[6] https://radimrehurek.com/gensim/models/word2vec.html

The complete algorithm to resolve the sentiment value for a token vector $t$ can be defined as follows:

$$sentiment(t) = \max_{s \in \text{SentiWS} \cap \text{Word2Vec}} \frac{\text{SentiWS}(s)}{D(s,t)} \tag{1}$$

Note that this algorithm implicitly neutralizes alternatives if they are too broadly associated synonyms:

$$\lim_{D(s,t) \mapsto \infty} sentiment(t) = 0 \tag{2}$$

The end behaviour of $sentiment(t)$ ensures the absence of polarity distortion in synonym search.
All tokens in the data were labeled using this process.

### 2.5 Sentiment Weighting and Analysis

After the tokens of **M** were given an emotional weight as outlined in Section 2.4 on a word level, further analysis on a sentence level can proceed. Fang & Zhan [5] summarize that every word of a sentence has its syntactic role which defines how the word is used. These roles, also known as part of speech, have significant impact on the importance of their underlying sentiment for the polarity of the complete sentence. For example, words like pronouns usually do not contain any sentiment and are therefore neutral. In contrast, verbs or adjectives can hold different weights respectively [5]. Part of speech taggers are used to classify words according to their syntactic role. The NLTK tagger class has been extended for the German language and trainend on the TIGER [4] corpus in a different project[7], achieving an accuracy of 98% as stated by the authors. This effective tagger was chosen for the presented work due to its good performance, generalization capabilities and fast integration in NLTK. After POS processing, all tokens were labeled according to the STTS [33] tag system. Nichols & Song [25] have examined the relationship between scalar sentiment, part of speech and overall sentence polarity. They empirically compared the influence of POS strengths on classifier performance and approximated an optimal solution. Their exhaustive search for the set $POS = \{noun, verb, adjective, adverb\}$ and the strength weights $str(POS_i) \in \{1, 2, 3, 4, 5\}$ has shown that the best performance for purposes of sentiment analysis was achieved with the following scalar weight vector:

$$(str(noun) = 2, str(verb) = 3, str(adv) = 4, str(adj) = 5) \tag{3}$$

A mapping from the STTS tags to the categories utilized by Nichols & Song was introduced to ensure compatibility between the German tagger and their weighting approach. Afterwards, all tokens were accentuated according to their syntactical sentence function, resulting in increased sentiment variance and therefore more expressive overall tweet polarity.
As a last step before the culminating conflation of all individual token polarities per tweet, negations need to be handled as they significantly influence the

[7] https://github.com/ptnplanet/NLTK-Contributions/tree/master/ClassifierBasedGermanTagger

calculated emotion by their valency scopes [6]. Two primary ways for negation handling were tested: syntax scope analysis and a heuristic approach. Carrillo [1] et al. proposed that superior performance is achieved if the negation scope is determined by examining the valence subtree of the negation token based on part of speech association. After successfully labeling each word with a POS tag, the grammatical syntax reveals the subtree which is supposed to be negated and therefore inversely influential on sentiment. While this approach would grant realistic language sentiment, it presupposes that the syntax tree is immaculate. Especially for Twitter, this is rarely the case. Gui [9] et al. found that the Twitter culture of mutual communication is inherently comprised of non-standard orthography and reconstructing an approximately valid syntax tree requires substantial effort. Their findings where confirmable and therefore opposed the syntactical negation handling as proposed by Carrillo [1] et al. for practical appliance in the presented project. For this reason, the more widely [6] used heuristics solution was employed. The German tagger was able to reliably identify the negation token itself (e.g. "nicht") and labeled it accordingly with the fitting STTS tag. This label gave an anchor to which a rule-based negation heuristic could be expediently attached. Inspired by the syntactical solution, the heuristic successively searches for the next token with a sentiment that has been weighted by its tag (see the beginning of this section). The rationale behind the algorithm is that the sentiment bearing successor feature is assumed to be the most likely target for negation. Samples suggested that this heuristic rule performs sufficiently in relation to the goals of analysis.
After all negations were handled, it was finally possible to propose an estimated polarity per tweet. Similarly to the work of Kumar & Sebastian [16], sentiment was calculated by summing the weighted and - if necessary - negated token polarity scalars. The resulting values were added as new column to **M**.

### 2.6 Scale Calibration

Having calculated a value which one might consider "sentiment" is not enough for actual market analyses due to two reasons:

1. The scale - while argumentative coherent and grounded in the outlined rationale - can be understood as a valid indicator, it is still arbitrarily defined. Its definition is sound, but given that the scale is supposed to measure levels of human emotion, it needs validation. Such a test would require human evaluation, altering the problem to a supervised interpretation.
2. As Section 2 stated, Munoz & Kumar [23] emphasize indicator fixation and anchoring within the metric of a market to achieve brand profiling. Only then is empirical comparison to the calibrated indicator, and thereby competitive brands, feasible.

These two problems seemingly demand further research and evaluation. Contradistinctively, it is argued that the *combination of both* allows for a use-case specific solution if the fundamental nature of the underlying data is exploited by

utilizing the broad scope of opinions being uttered on Twitter. This characteristic permits the introduction of the established statistical Central Limit Theorem (CLT). The CLT is the observation of the convergence behaviour of probability distributions of an increasing number of one- or multi-dimensional random variables to a normal distribution [7]. For a public opinion surveying purpose as presented, the CLT leads to a beneficial conclusion: if a sufficiently large number of unrelated, random sample opinions are gathered from a sample population, the overall sample mean will be normally distributed around the population mean. Furthermore, if the opinion distribution is limited to the interval $[-1, 1]$ by the pre-conceived sentiment constraint and additionally, baseline polarity is assumed to be neutral, all essential properties of the expected opinion distribution are therefore known in advance without human intervention for validation. To exploit this reasoning for the calibration of the proposed polarity estimation process, a second Twitter dataset **GT** (for **G**round **T**ruth) was collected using the Tweepy filter

```
"#2018 OR #2019 OR #december OR #january"
```

and is processed through the same pipeline as **M**, leading to a broad, dispersed set of tweets unrelated to any specific topic. As these tweets cover a wide range of *independent* themes and conversational domains, it can be reasoned that the global population sentiment characteristics behave according to the CLT. This theory resembles the missing link between the "arbitrarily" constructed polarity estimation pipeline and actual market sentiment, resulting in the desired indicator described by Munoz & Kumar. Establishing the connection mathematically is possible in a multitude of ways, as long as the link adheres to the following formalism. Since the global sentiment distribution of the general dataset **GT** should at best follow the listed constraints, its histogram $\Phi(GT_{polarity})$ should resemble the normal distribution as close as possible. As such, the aim is to find a projection of $GT_{polarity}$ which minimizes the error between the histogram and the normal distribution. If such a projection is found, it acts as the calibration metric for the analytics pipeline. Thereupon, the calibrated projection can be used on the actual logistics dataset to infer class labels in relation to the opinion of the general population. If the distribution is discretized at the interquartile ranges (IQR), half of all opinions fall in the central area. These will be considered "neutral" and make up the overall majority. A quarter of all opinions fall left of the first IQR marker and will be considered negatively extreme. Their class label is "negative". And lastly, the remaining datapoints fall beyond the third IQR marker and are hence labeled "positive" as they express positively extreme sentiment. This baseline polarity will be the ground truth reference and **M** can be labeled in the same way, using the absolute sentiment boundaries dictated by **GT**. Subsequently, sentiment analysis and classification are concluded and evaluation of logistics opinions is finally possible.

## 3 Analysis

For evaluation, the questions put forward in Section 2 were answered using the constructed pipeline.

After discretization, the Twitter class label distribution is normally distributed and zero centered. The IQR markers force the calibration into the CLT assumption. Therefore, specialized tweet topics can be compared by calculating the relative distance between the class label tendencies.

### 3.1 "How do leading German consumer logistics brands rank in Twitter user approval?"

For the complete logistics dataset **M**, the class labels deviate from the Twitter baseline **GT**. Neutral sentiment is 11.92% less present in tweets relating to logistics while positive and negative sentiment are 4.58% and 7.34% above baseline respectively. This observation of increased variance shows that users communicate with higher emotional tendencies towards the topic. It can be concluded that opinions regarding the logistics domain are mostly stated more vigorously. To reduce selection bias, the logistics brands must therefore be compared exclusively within their domain. Otherwise, their relative ranking would be distorted by the overall preconceived notions of opinion. Appendix Figure 1 (top) visualizes this distortion. On first glance, all three brands perform with high emotional response, skewed towards negative feedback. This issue is the result of the distributional relationship to $\Phi(GT_{polarity})$. Drawing the conclusion that the three specific brands perform bad on Twitter is not precise as the entire domain *generally* provokes the shown response. Due to this implication, a more accurate baseline indicator for performance ranking is the sentiment histogram of the logistics domain. All brands react differently and more truthfully to the this metric and better conclusions can be drawn, as presented in Appendix Figure 1 (middle). Therein, it can be seen that the relative ranking is now increasingly expressive. DHL performs better than its competitors within the domain, having less negative class labels and more positive class labels than average. DPD and HERMES are negatively skewed beyond average expectation, performing worse than DHL. HERMES exclusively falls behind in both negative (more than average) and positive (less than average) opinions.

### 3.2 "How can Twitter meta information augment user approval analysis?"

The presented comparison solely relies on single tweet content and thus individual sentiment. The Twitter API grants access to information beyond pure textual data. Correlating the meta data to polarity can yield clarified insight. For example, one of the defining functionalities of Twitter is the ability to like and/or share ("retweet") opinions of other users. The implications for sentiment statistics are pivotal. Nagarajan, Purohit & Sheth [24] observed different levels of endorsement by peer users depending on tweet positivity. Hence, weighting tweet sentiment by

the amount of shared and approved peer affirmation links argumentative popularity to brand performance indication. Furthermore, if a large enough dataset is gathered, sentiment can be followed along the chain of retweets, forming an interesting graph traversal problem. It could be mapped out how positive and negative sentiment propagate through the Twitter ecosystem and how these multiplicative patterns differ for brands. For the presented project, the dataset is not large enough to reconstruct such patterns. Nonetheless, each entry of **M** does contain the integer counts of likes and retweets and these values hold analytic value. Reasoned by the arguments above, the two integers were summed per row and added as a new column labeled "Propagation". The new value resembles the amount of persons sharing the view being expressed in the tweet and can be used as a weight vector for sentiment aggregation. The results of Nagarajan, Purohit & Sheth were observable afterwards aswell. Our observations indicate that tweets in the logistics domain are generally shared less often than baseline, but *if* they are shared, they are more likely negative in contrast to **GT**. Due to this finding, the class distributions change significantly if polarity propagation is factored into relational brand metrics as shown in Appendix Figure 1 (bottom). User approval leads to considerable amplification of disparity. DPD and HERMES shift their distributions to pronounced neutrality. They both reduce negative sentiment slightly but also exceedingly decrease in positive sentiment by a large factor. In contrast, DHL overwhelmingly profits from the shared user opinions. Negative and neutral sentiment labels are shifted to an 11.04% increase in positive sentiment.

In addition to likes and retweets, the data also contains the timestamp at which a tweet was written. Especially in the logistics business, time plays an important role and should therefore be targeted analytically. In Appendix Figure 2, a heatmap of aggregated class labels per day and hour, averaged in mean rows and columns, is shown. For aggregation, the labels were interpreted as $\{-1, 0, 1\}$ for $\{negative, neutral, positive\}$ respectively. In **GT**, it can be observed that negative sentiment correlates with business hours. Furthermore, negativity peaks towards the end of the business week. Intuitively, non-delivery at the beginning of the weekend may induce disappointment as the customer would have to wait past the work-free days until the next business day. Such a hypothesis could be further evaluated if domain knowledge is introduced.

For the individual brands, different observations stand out:

1. DHL: Generally, DHL performs above average on Mondays. Sentiment then declines with progression of the week. Negativity peaks at the weekend.
2. HERMES: Best performance is expected on Tuesdays. Daily peak negativity shifts with weekly progression. Customers tend to express negative feedback incrementally in later hours, peaking at 20:00 Hours. The relation may be connected to the longer business hours employed by HERMES.
3. DPD: No obvious pattern is present except a sentiment low on Mondays 14:00 Hours. Otherwise, the data correlates to the baseline.

The underlying tweet texts and their themes where investigated at the emphasized days and hours but did not reveal any apparent common denominator ex-

plaining their occurrence in addition to there mere temporal correlation. These time distributions can serve more appropriate analyses for research with added knowledge of the internal structures of the different companies. Then, more precise assumptions can be made about the cause of the observed patterns.
The different client agents were also investigated in relation to sentiment but did not show any discernible correlation.

## 4 Conclusion

The outlined process has shown how powerful Twitter can be for sentiment analysis in brand perception polling. Especially when meta data is introduced, insights far beyond classical polling techniques become apparent. The different logistics brands have shown distinguishable approval performance and the inclusion of Twitter meta data was beneficial to inquire into these variations. Acquiring a professional API access may be costly at first, but the increase in data volume and quality can add more capabilities, such as instant metrics or even precise geolocation, a variable directly linkable to geographical key performance indicators in logistics. As such, the real-time data pipeline could be integrated into business intelligence and monitoring systems for anomaly detection and cause-effect analysis.
Furthermore, natural language processing methods were employed to demonstrate their value for modern-day feedback evaluation. In-depth studies into the many different ways to approach language processing problems may highlight even more fruitful pipeline steps. Building a domain specific language corpus should be the first obstacle to take in that direction. The current lack thereof discouraged the usage of many techniques like supervised machine learning. The same holds for the limited size of the tested data set. If more comprehensive collection was possible, access to other ways of statistical description and model building opens up. Especially in regard to finer recognition of irony and sarcasm, further improvement of the pipeline is required. In its current form, sarcasm was not adequately recognized. Relative brand metric comparison is still valid, as all brands suffered from this deficit in the same way. Sarcasm seems to be inherently linked to strongly opinionated utterances, which is a reason why the proposed workflow did not rely on emoticon recognition as other works suggest for label inference. Data exploration has shown that emoticons played a predominant role in sarcasm emphasis. This observation may hold value for further research but discouraged their importance for the current student project.
Summarizing, the work supports the assumption that modern social media can have a vital contribution to fast refinement of a brands marketing mix for strategic realignment in real-time. This is a benefit to ensure positive reception of corporate philosophies directly as a reaction to automated analyses of innovative, digital consumer feedback channels, using contemporary research in natural language processing.

# Appendix


**Fig. 1.** Twitter sentiment compared to logistics brands.

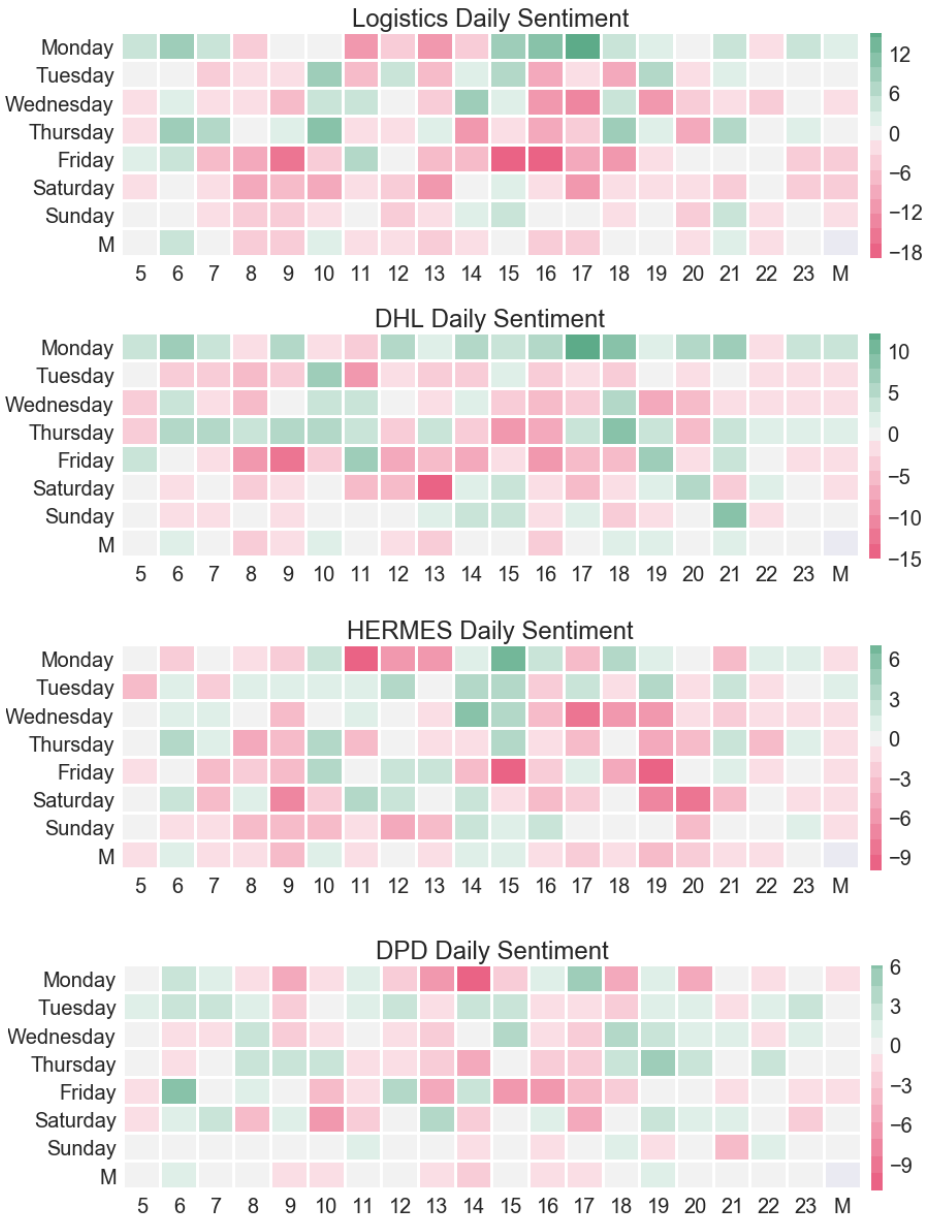


**Fig. 2.** Daily logistics sentiment.

## References


1. Carrillo de Albornoz, J., Plaza, L., Diaz, A., Ballesteros, M.: Ucm-i: A rule-based syntactic approach for resolving the scope of negation. In: *SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012). pp. 282–287. Association for Computational Linguistics (2012), http://aclweb.org/anthology/S12-1037
2. Bird, S., Loper, E.: Nltk: The natural language toolkit. In: Proceedings of the ACL 2004 on Interactive Poster and Demonstration Sessions. ACLdemo '04, Association for Computational Linguistics, Stroudsburg, PA, USA (2004). https://doi.org/10.3115/1219044.1219075
3. Culotta, A., Cutler, J.: Mining brand perceptions from twitter social networks. Marketing Science **35**(3), 343–362 (2016). https://doi.org/10.1287/mksc.2015.0968
4. Dipper, S., Kübler, S.: German Treebanks: TIGER and TüBa-D/Z, pp. 595–639. Springer Netherlands, Dordrecht (2017)
5. Fang, X., Zhan, J.: Sentiment analysis using product review data. Journal of Big Data **2**(1), 5 (Jun 2015). https://doi.org/10.1186/s40537-015-0015-2
6. Farooq, U., Mansoor, H., Nongaillard, A., Ouzrout, Y., Qadir, M.A.: Negation handling in sentiment analysis at sentence level. Journal of Computers **12**, 470–478 (01 2016)
7. Fischer, H.: A history of the central limit theorem: From classical to modern probability theory. Springer Science & Business Media (2010)
8. Goldberg, Y., Levy, O.: word2vec explained: deriving mikolov et al.'s negative-sampling word-embedding method. CoRR **abs/1402.3722** (2014), http://arxiv.org/abs/1402.3722
9. Gui, T., Zhang, Q., Huang, H., Peng, M., Huang, X.: Part-of-speech tagging for twitter with adversarial neural networks. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. pp. 2411–2420 (2017)
10. Hamp, B., Feldweg, H.: Germanet-a lexical-semantic net for german. Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications (1997)
11. Hu, Y.: Efficient and high quality force-directed graph drawing. Mathematica Journal **10**, 37–71 (01 2006)
12. Jansen, B.J., Zhang, M., Sobel, K., Chowdhury, A.: Twitter power: Tweets as electronic word of mouth. JASIST **60**, 2169–2188 (2009)
13. Jansen, B.J., Zhang, M., Sobel, K., Chowdhury, A.: Umsatzverteilung im kependkundenmarkt in deutschland nach anbietern im geschäftsjahr 2017/18. Handelsblatt **134**, 16 (2018)
14. Kim, Y., Shin, H.: A new approach for measuring sentiment orientation based on multi-dimensional vector space. CoRR **abs/1801.00254** (2018), http://arxiv.org/abs/1801.00254
15. Kouloumpis, E., Wilson, T., Moore, J.D.: Twitter sentiment analysis: The good the bad and the omg! Icwsm **11**(538-541), 164 (2011)
16. Kumar, A., Sebastian, T.: Sentiment analysis on twitter. International Journal of Computer Science Issues **9**, 372–378 (07 2012)
17. Liebeck, M.: Aspekte einer automatischen meinungsbildungsanalyse von onlinediskussionen. In: Ritter, N., Henrich, A., Lehner, W., Thor, A., Friedrich, S., Wingerath, W. (eds.) Datenbanksysteme für Business, Technologie und Web (BTW 2015) - Workshopband. pp. 203–212. Gesellschaft für Informatik e.V., Bonn (2015)

18. Löffler, R., Wittern, H.: Markenwahrnehmung und marken-differenzierung im zeitalter des web 2.0. In: Markendifferenzierung, pp. 359–375. Springer (2011)
19. Manning, C.D., Raghavan, P., Schütze, H.: Introduction to Information Retrieval, p. 32. Cambridge University Press, New York, NY, USA (2008)
20. Miller, G.A.: Wordnet: a lexical database for english. Communications of the ACM **38**(11), 39–41 (1995)
21. Mohtarami, M., Amiri, H., Lan, M., Tran, T.P., Tan, C.L.: Sense sentiment similarity: An analysis. In: Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence. pp. 1706–1712. AAAI'12, AAAI Press (2012), http://dl.acm.org/citation.cfm?id=2900929.2900970
22. Mueller, J., Stumme, G.: Predicting rising follower counts on twitter using profile information. CoRR **abs/1705.03214** (2017), http://arxiv.org/abs/1705.03214
23. Munoz, T., Kumar, S.: Brand metrics: Gauging and linking brands with business performance. Journal of Brand Management **11**(5), 381–387 (2004)
24. Nagarajan, M., Purohit, H., Sheth, A.P.: A qualitative examination of topical tweet and retweet practices. In: ICWSM (2010)
25. Nicholls, C., Song, F.: Improving sentiment analysis with part-of-speech weighting. vol. 3, pp. 1592 – 1597 (08 2009)
26. Pak, A., Paroubek, P.: Twitter as a corpus for sentiment analysis and opinion mining. In: LREC (2010)
27. Remus, R., Quasthoff, U., Heyer, G.: Sentiws - a publicly available german-language resource for sentiment analysis. In: LREC (2010)
28. Schweidel, D.A., Moe, W.W., Boudreaux, C.: Social media intelligence: Measuring brand sentiment from online conversations (2011)
29. Steinskog, A., Therkelsen, J., Gambäck, B.: Twitter topic modeling by tweet aggregation. In: Proceedings of the 21st Nordic Conference on Computational Linguistics. pp. 77–86. Association for Computational Linguistics (2017), http://aclweb.org/anthology/W17-0210
30. Wang, X., Wei, F., Liu, X., Zhou, M., Zhang, M.: Topic sentiment analysis in twitter: a graph-based hashtag sentiment classification approach. In: CIKM (2011)
31. Wang, Y., Liu, J., Qu, J., Huang, Y., Chen, J., Feng, X.: Hashtag graph based topic model for tweet mining. In: 2014 IEEE International Conference on Data Mining. pp. 1025–1030 (Dec 2014). https://doi.org/10.1109/ICDM.2014.60
32. Webber, F.D.S.: Semantic folding theory and its application in semantic fingerprinting. CoRR **abs/1511.08855** (2015), http://arxiv.org/abs/1511.08855
33. Westpfahl, S., Schmidt, T., Jonietz, J., Borlinghaus, A.: Stts 2.0. guidelines fuer die annotation von pos-tags fuer transkripte gesprochener sprache in anlehnung an das stuttgart tuebingen tagset (stts) (2017)