Gesellschaft für Informatik (Hrsg.): SKILL 2019,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2019 11

Improvement of automated social media sentiment analysis methods - a context-based approach

Dennis Debeye[1] Tim Eder[2] Paul Vincent Guigas [3] Viktoria Schuberth[4]

Abstract: The sentiment analysis of social media data increasingly gains importance in business and research. But still, topical algorithms cope with problems, since it is reasonably manageable to extract the tonality of a social media post, but not the authors attitude towards a given topic. However, in most cases, this is the relevant information users of social media analysis tools are looking for. To tackle this problem, we propose a context-based algorithm that not only focuses on isolated postings, but also takes the authors' earlier postings and their interactions with other users' posts into account to derive their actual opinion on a subject. To evaluate this approach, we implemented a test system and compared the algorithm's results to manually assessed sentiments.

Keywords: sentiment analysis; social media analysis; opinion mining

1 Introduction

At the present time, we find ourselves midway of a revolution of the way people communicate with each other: Instant messaging and social media increasingly gain importance, not only for private communication, but also for businesses and organizations. On social media, everybody is a potential content creator; every user has a channel to broadcast their opinion and thoughts on certain topics. This challenges marketing departments. They are forced to step away from traditional monodirectional campaigns and towards running social or even cross media campaigns, that involve continuous exchange with their digital communities. The communication has become bilateral.
Not only do users of social media talk *with*, but also *about* the organizations, their products, offered services or activities. And this makes social media a great opportunity for such organizations, since an analysis of these publicly available communication data from social media networks can help to improve marketing strategies and achieve better follower/customer engagement, better follower/customer service, better reputation management and brand awareness, product innovation, business process improvement and even the discovery

[1] HTWK Leipzig , IMN, Karl-Liebknecht-Str. 132, 04277 Leipzig, Germany dennis.debeye@stud.htwk-leipzig.de
[2] HTWK Leipzig , IMN, Karl-Liebknecht-Str. 132, 04277 Leipzig, Germany tim.eder@stud.htwk-leipzig.de
[3] HTWK Leipzig , IMN, Karl-Liebknecht-Str. 132, 04277 Leipzig, Germany paul_vincent.guigas@stud.htwk-leipzig.de
[4] HTWK Leipzig , IMN, Karl-Liebknecht-Str. 132, 04277 Leipzig, Germany viktoria.schuberth@stud.htwk-leipzig.de

of new business opportunities (cf. [HPH14]). Already today, these potentials are utilized in a wide range of application fields, e.g. in political election campaigns [AG17], the banking sector [Co15] or even on the stock market [HPH14]
There are many thinkable metrics and indicators that provide valuable information for organizations to improve the understanding for their communities, e.g. term frequency or demographic information. One of the key indicants surely is the *sentiment*, i.e. the information about how people feel about a certain topic. To find a communities sentiment, social media analysis (SMA) tools run *automated sentiment analyses* on a vast amount of original posts and comments.
This paper focuses exactly on these automated, computational sentiment analysis (SEA) procedures. Automated SEA usually utilizes methods of the fields *information retrieval*, *text mining*, *web mining* and *natural language processing* (NLP). The common problems most of these approaches have, is that they focus on the bare text of a single post to derive a sentiment from ist. No matter how good these approaches are, they will always only be able to reflect the sentiment of an isolated post. However, the information most data scientists and sociologists are looking for will most likely not be the sentiment of a *post* but the sentiment of a *user*. Therefore, in this paper, we will present an approach to a technique to enhance the accuracy of an existing automated SEA tools by taking into account the context (i.e. the users other interactions with the social community) of the analyzed social media postings. As application example, we chose the micro blogging platform *Twitter*. In the remainder of this paper, we will first examine the current state of the art regarding SEA in general, then depict the motivation for a new, context based approach and afterwards present our solution and evaluate and discuss our findings.

2 State of the Art

Long before the World Wide Web was invented, we have let our decisions be influenced by others point of view. *Can somebody recommend this washing machine? Can someone recommend a local restaurant?* The internet gives us a place to share opinions with others and spread it around the world. Depending on the platform and the count of potential readers, these viewpoints could reach thousands and millions of people. But not only recommendations about products and services can be shared, also political statements and information. Especially Twitter, the most popular microblogging platform, has a widespread community. Each day 500 million tweets are sent. 326 million people use twitter monthly [Ho19]. Sentiment Analysis is one of the first steps to extract the opinion out of an unstructured text with relatively low investment of time and effort [Pa18].

Over the time sentiment analysis became a very important topic for researchers and firms for evaluating opinions over a large user population. The number of papers related to this topic increases rapidly: according to [Mä18], almost 7000 papers have been published until 2016. We have also seen a massive increase in the number of papers focusing on sentiment analysis and opinion mining during the recent years. The most commonly used target classes

for the sentiment classification are *positive* and *negative*, sometimes extended by a third *neutral* class.

2.1 Lexical Approach

A tremendous amount of papers rely on the lexical approach. A lexicon of sentiments represents a list of words or lexical attributes and are labeled with their semantic polarization [Li12]. Manually creating such lexicons is a very costly - but robust - technique, so most of the approaches resort to preexisting lexicons [HG14].
Another way to generate opinion lexicons is the corpus-based approach. It's focused on conjunctions, like AND, OR, EITHER-OR and NEITHER-NOR. The conjoined word after AND has usually the same orientation like it's predecessor. People usually express the same opinion on both sides of the conjunction [Li12]. After BUT, we can expect a contrary opinion. In practice this approach is not always consistent. [Ho13] introduces a way to improve the lexical-based sentiment classification. They analyzed how emoticons provide sentiments and based on that, they made a manually created emoticon sentiment lexicon. A relevant approach for opinion mining and especially for this paper is described in [HG14]: they created a rule-based model for sentiment analysis. In a micro-blogging context they first construct and empirically validate a *goldstandard* list of lexical features combined with their sentiment value. Adding five general rules that contain grammatical and syntactical conventions to the lexical features, the sentiment intensity becomes even clearer.
Our approach relinquishes the costly work of creating a lexicon beforehand and we do not have do deal with grammatical or syntactical rules. Lexical approaches are not always consistent. In case of inconsistency, we use deepSEA to take a closer look on ambiguous tweets. We focus on the author of the tweet and his opinion towards the tweet's topic and not only on the occurring words and phrases.

2.2 Machine Learning

Manually creating lexicons is a time-consuming process. Since machine learning is becoming more and more relevant in the field of natural language processing, it's also a state-of-the-art approach in sentiment analysis. The Naive Bayes (NB) is a classifier based on the Bayes rule and the naive assumption that features' probabilities are independent of each other. This works well on text categorization.
Maximum Entropy models should prefer the most uniform models that satisfy a given constraint. These classifiers are used in the approach described in [MKP02]. In a sentiment analysis context this is useful for recognition of patterns in word usage between different classes to put it into these categories.
Support Vector Machines are non-probability classifiers. SVMs work by separating data points in space using hyperplanes [GBH09]. [PP10] presented a way by collecting a corpus of text with emoticons. These collected corpora were used to train a classifier to identify

positive and negative sentiments. Each of the described machine learning approaches require large sets of training data and always depend on them. It's also expensive in terms of memory consumption, CPU processing and classification time. With machine learning the sentiment assumptions are made by correlating large sets of data. This does not imply causation [Le19]. To prevent this possibility of errors, deepSEA takes a closer look at the user and her or his specific opinion to make the sentiment less ambigous. It also does not have the need of a large training set or special hardware requirements.

2.3 Semantic Approach

There is also a way to extract the sentiment out of an unstructured text with semantic web technologies. One approach is introduced in the book [SHA12]. The semantic values extracted from tweets can be used to measure the overall correlation of a group of entities (e.g. all APPLE products) with a given sentiment polarity. For example products of brands can be used as vocabulary, which are listed in the training data with occurrences in positive and negative tweets. The entities *iPad, iPod* and *Mac Book Pro* appeared preferably in tweets with positive polarity and they are all mapped to the semantic concept PRODUCT/APPLE. As a result, the tweet from the test set *Finally, I got my iPhone. What a product!* is more likely to have a positive polarity because it contains the entity *iPhone* which is also mapped to the concept PRODUCT/APPLE. It is also possible to give a statement about products, which are not mentioned in the training set (e.g. they are not released yet), because they are mapped to their brands. ArsEmotica - an application software for associating the predominant emotions to artistic resources of a social tagging platform - also uses the semantic approach. This is done by exploiting and combining available computational and sentiment lexicons with an ontology of emotional categories. Words with a relevant sentiment meaning get ranked by users feedback. So the emotional ontology hierarchy grows and is available within the linked data network [Ar12]. This approach calculates a sentiment relative to related topics, while the deepSEA approach tries to estimate a tweets sentiment using tweets relative to the same topic. This way we expect to get an even more accurate result

2.4 Context-based Approach

In a user-based context, we can extract additional information of the opinions about a specific topic. For example in conversations of users about this subject, the users attitude or likes and retweets. The approach described in [VCB14] goes hand in hand with the machine learning technologies: they are using kernel functions - integrated in SVMs - to capture specific aspects of sentiment connection between two tweets. In conversations on Twitter they disambiguate even very short messages and characterize them according to their authors. Their experimental evaluation proves that sequential tagging effectively embodies evidence about the contexts. They also improve detection accuracy to around 20%. In case of ambiguous phrases in tweets, deepSEA looks at the author and her or his

previous opinions on this topic. Strictly speaking, deepSEA does not analyze the sentiment of the tweet, it rather analyzes the sentiment of the user.

3 Problem motivation

Most previous approaches to sentiment analysis do not take a tweet's context into consideration, they merely look at the choice of words in a single tweet. While this is fine for tweets in a more literal style, it is more complicated to analyse tweets written in figurative language. Especially the detection of irony and sarcasm is a field, which faces a certain set of challenges. Key factors in detecting irony and sarcasm are aspects on the syntactic and textual level, the semantics and pragmatics as well as the discourse analysis [HR17]. While progress was made in the fields of the syntactical, lexical as well as the semantical level, the aspect of discours analysis still is a widely unknown territory. Key factor therefore is the analysis of the discussion itself, i.e. how utterances relate to each other within the Twitter ecosystem. It is therefore necessary to investigate the broader context of a tweet by taking a closer look at the authors and their intentions. The main motivation of this paper is, to achieve this analysis by taking Twitter's own metrics (e.g. number of likes and retweets on a tweet) into account.

4 Solution

4.1 deepSEA Approach

Sentiments get determined by our underlying analysis tool *VADER* (cf.[HG14]) with a confidence. If this confidence is below a set threshold, the tweet's sentiment gets marked as uncertain. To specify an uncertain sentiment, the users opinion on the subject is put in broader context. Based on the hypothesis, that there is a confirmation bias in a Twitter users hashtag usage [KL18], the users interaction within the Twitter social ecosystem is being analyzed. In this context, the initial tweet is as important as the set of the original authors interactions on the same subject. Therefore, the deepSEA approach is based on the following assumptions:

- If a user posts multiple tweets about a certain topic(represented by the usage of a subject hashtag), these tweets reflect the same sentiment towards that topic.[KL18]
- A like (vgl. [Tw]) shows the users appreciation for a tweet . Knowing, that the users hashtag usage is biased, the assumption can be made, that a like on a tweet using a hashtag the user has tweeted about before, reflects an approval for the opinion of certian tweet.

The tweets the author interacts with are used to calculate a second sentiment which considers the authors feeling about the subject itself. In case of an uncertainty in the initial tweet's

sentiment rating, these derived tweets are used to adjust the sentiment in the authors initial tweet.

4.2 deepSEA Logic

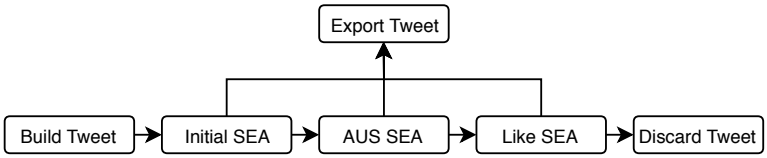

Fig. 1: The deepSEA rating process is controlled via the sentiment pipeline and is subdivided into rating steps.

The deepSEA logic is subdivided into four partial steps as seen in Fig. 1. The Twitter metadata is built into a readable data format and transported through a pipeline of different SEA steps. If the current step is successful, the tweet will get exported and the sentiment of the most recent SEA step will be regarded as the true or final sentiment of the tweet. The success of an analysis step depends on each step's individual definition. In the *Initial Sentiment Analysis* (Initial SEA) the initial tweet's sentiment gets determined by the sentiment analysis tool *VADER* [HG14]. If the confidence of the initial tweet's lexical analysis is above a fixed threshold, the analysis step succeeded. Otherwise, the tweet is presented to the *Advanced User Subject Sentiment Analysis* (AUS SEA). The AUS SEA rates the users subject based on derived tweets. By analyzing the sentiment of the users previous tweets on that subject with VADER, a ratio of the most frequently occurring sentiment in relation to all derived tweets is drawn. The most frequently occuring sentiment becomes the new sentiment of the tweet, whereas the ratio represents its confidence. If this confidence is below a set threshold, the tweet is passed to the *Like Sentiment Analysis* (Like SEA). The Like SEA step uses all tweets related to the analyzed subject that the user has liked as a basis to draw a ratio analogous to the AUS SEA. If the confidence of this ratio is also under a set threshold, the tweet is marked as not ratable and therefore discarded. After the automated sentiment-analysis processed is finished, tweets are rated in *Twinder* (a subsystem designed to manually rate tweets through a user interface) for further analysis and evaluation.

4.3 Test Architecture

For gathering the necessary data, a system with the architecture presented in Fig.2 was implemented. A logging microservice was realized to keep track of every step of the processed data. Using the stream processing framework Apache Kafka, a Twitter stream for a certain topic is subscribed and incoming data conveyed to the *deepSEA Logic* microservice, where its sentiment gets determined through the process described in section 4.1. The overall process is triggered by a REST call against the *deepSEA Ingest* service. Also, with

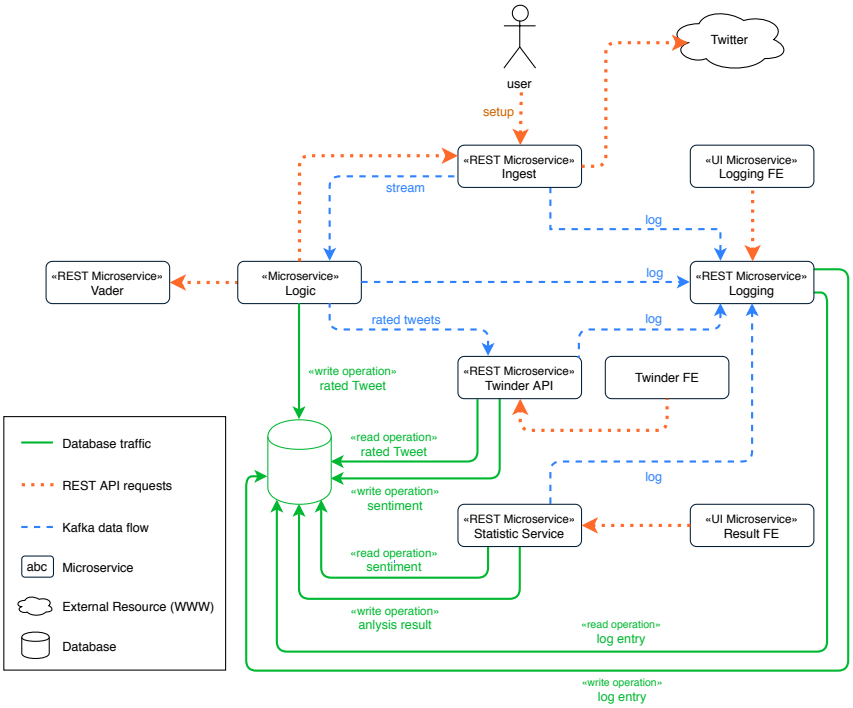

Fig. 2: The architecture of the test environment

this call, the subject of the analysis is set. When the rating process of a tweet is finished, Kafka transports the rated tweet to the so called *Twinder API*, which saves it to a Database. They are then manually rated by (human) test subjects, through a simple frontend in order to be able to evaluate the performance of our approach. After this processing step is completed, the tweets are evaluated in the *statistics service*, that assesses the sentiment rating quality of the different approaches.

4.4 Data Acquisition

The test data with the subject *Captain Marvel* was acquired on the 12th of March 2019. Over the timespan from 8:13pm to 8:55pm CET, 429 tweets with a textual mention of *Captain Marvel* were captured and analyzed with *deepSEA*. The corresponding thresholds of the SEA step were set as follows:

- init SEA: 0.6
- AUS SEA: 0.5
- Like SEA: 0.5

The procedure of finding the init SEA threshold consisted of a series of experiments roughly resembling the gradient descent method [Sc17]. Through iteratively adjusting the threshold, we found, that a value of 0.6 provided the best results.
The confidence of the *AUS* resp. *Like* sentiment is calculated by deviding the count of tweets that show the dominant sentiment by the count of all relevant tweets that were taken into account. Since the overall probability of occurrence of relevant tweets is relatively small, we decided that an absolute majority (i.e. a ratio >50%) is the most sensible threshold to choose, since it *always* provides a usable result, if a dominant sentiment exists.

4.5 Manual Rating Process

Additionally, each tweet was rated manually by four test subjects. The subjects were asked to identify the attitude of the original author towards the topic (in this case the movie *Captain Marvel*) rather than the general tonality of the tweet. The subjects were presented each tweet exactly once and were supposed to rate it as positive, negative or neutral. The results of this manual rating served as reference sentiment for the automatic SEA methods.

5 Discussion

5.1 Data Collection and Evaluation

We chose the topic of *Captain Marvel*'s cinema release, because we expected controversial sentiments, which - as one can derive from the results presented in Table 1 - turned out to be mostly true. The true sentiment of the tweets was found via our rating frontend *Twinder*. We specifically did not want test subjects to rate the tonality of the language, like a sentiment analysis software would, but rather the author's true stance on the topic. This meant that they had to consider irony and disagreements with people of polar opinions. If three test subjects agreed on a sentiment, we took it as ground truth and compared our analysis to this value. Non-english tweets and spam was discarded. After this data cleansing, we ended up with a total of 429 tweets. Their ratings are summarized in table 1.

	True Sentiment	**deepSEA**	**VADER**
positive	303	388	367
neutral	92	6	18
negative	34	35	44

Tab. 1: Overview of sentiment values

Our algorithm rated 191 of these 429 tweets, since they did not match our specified certainty threshold for the initial sentiment analysis. Of these 191 tweets, 101 had a correctly detected final sentiment. This means deepsea could not manage to outperform VADER, which detected the sentiment of 106 tweets correctly (but - as explained - all with insufficient certainty). However, the vast majority of tweets was still classified correctly only by analysing the original posts context. This is an indicant, that our foundational assumptions were correct.
Nevertheless, the data shows that deepSEA failed to improve VADER ratings in some cases. It is noticcable, that many tweets were falsely rated to be positive and hardly any neutral tweets were detected.

5.2 Further Thoughts

The deepSEA approach appears to be well suited in the case of a person with a positive opinion on the topic expressed in negative language, like in a discussion between two users. VADER can not detect these sentiments, whereas our context based approach takes a user's previous comments and interactions on the subject into account to find their true opinion.
The opposite scenario however, causes a lot of false classifications: It seems appears more unlikely, that a user would post a negative opinion using positive words. This could explain why our system classifies too many tweets as positive. Some edge cases come to mind that could further explain the false ratings:

- A user has expressed a negative opinion even though their past statements were positive.
- They did not use a *like* as an agreement.
- They generally use negative language in all of their tweets.
- Etc.

Since we rely on VADER in our submethods, the original problem of not being able to detect context or irony persists even in the lower stages of our analysis. VADER also failed to classify a lot of neutral tweets, since the numeric sentiment range for a neutral rating is quite small. This means the software prefers to give positive or negative ratings, even though the manual classification determined a higher amount of neutral tweets for the analyzed topic.

5.3 Possible Improvements

Our sample data set was quite small and four participants are most likely not enough to construct a valid true sentiment. Hence our first goal in further research is to optimize our test data not only in regards to size, but also tweet content. In order to correctly identify the shortcomings of our algorithm, we need to test with an equal amount of handpicked positive, neutral and negative tweets. Our current data set was quite imbalanced, the most common sentiment was *neutral*. Increasing the threshold for neutral ratings could also improve results.
When we incorporated the context into our rating, we only looked at user context, i.e. other tweets from the same author. Another possibility for improvement could be to construct a rating from conversational context. By looking at replies in a comment thread, it might be easier to figure out a user's agreement or disagreement. We also did not take temporal changes into account: a user's opinion on a subject could change over time. This causes uncertain ratings in our algorithm.
Our sentiment analysis did not run, when VADER's initial rating had a high certainty. In that case we did not consider the possibility of the VADER sentiment being wrong. In future tests we will start our analysis on every tweet and compare it with the initial sentiment.

6 Conclusion

In this paper we have pointed out weaknesses of modern sentiment analysis algorithms. In order to overcome them, we came up a new analysis strategy that takes the context of a social media posting into account: Whenever the underlying lexicon approach failed to return a sufficiently certain sentiment, the algorithm considered older posts and liked posts by other users to derive the sentiment of the original one. We implemented a system to test

our approach with real time test data and statistically evaluated the results.
We found out that, despite being seemingly successful in enhancing the detection of tweets with a negative sentiment, our algorithm yet has issues in improving it for other sentiments as well. We discussed the most probable reasons for those findings and possible solutions, that might lead to better results in future tests.

References

[AG17] Allcott, H.; Gentzkow, M.: Social Media and Fake News in the 2016 Election. In: Journal of Economic Perspectives. 2017.

[Ar12] Arsmeteo, A. C.: Sentiment analysis in the planet art, http://www.di.unito.it/~patti/arsemotica.htm, 2012, visited on: 03/15/2019.

[Co15] Costin, D.: Banking Business and Social Media – A Strategic Partnership. In: Theoretical and Applied Economics. 2015.

[GBH09] Go, A.; Bhayani, R.; Huang, L.: Twitter sentiment classification using distant supervision. Processing 150/, Jan. 2009.

[HG14] Hutto, C. J.; Gilbert, E.: Vader: A parsimonious rule-based model for sentiment analysis of social media text. Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media/May, pp. 216–225, 2014, URL: http://comp.social.gatech.edu/papers/icwsm14.vader.hutto.pdf.

[Ho13] Hogenboom, A.; Bal, D.; Frasincar, F.; Bal, M.; de Jong, F.; Kaymak, U.: Exploiting emoticons in sentiment analysis. In: Proceedings of the ACM Symposium on Applied Computing. Pp. 703–710, Mar. 2013.

[Ho19] Hootsuite: 28 Twitter Statistics All Marketers Need to Know in 2019, https://blog.hootsuite.com/twitter-statistics/, 2019, visited on: 03/15/2019.

[HPH14] Holsapple, C.; Pakath, R.; Hsiao, S.-H.: Business Social Media Analytics: Definition, Benefits, and Challenges. In: Twentieth Americas Conference on Information Systems, Savannah. 2014.

[HR17] Hernandez Farias, D.; Rosso, P.: Irony, Sarcasm, and Sentiment Analysis. 2017.

[KL18] Kowald, D.; Lex, E.: Studying Confirmation Bias in Hashtag Usage on Twitter./, pp. 2–4, 2018, URL: http://arxiv.org/abs/1809.03203.

[Le19] Leetaru, K.: A Reminder That Machine Learning Is About Correlations Not Causation, 2019, URL: https://www.forbes.com/sites/kalevleetaru/2019/01/15/a-reminder-that-machine-learning-is-about-correlations-not-causation, visited on: 03/15/2019.

[Li12] Liu, B.: Sentiment Analysis and Opinion Mining. Morgan & Claypool Publishers, 2012.

[Mä18] Mäntylä Graziotin, K.: The evolution of sentiment analysis. In: A review of research topics, venues, and top cited papers. 2018.

[MKP02] Mehra, N.; Khandelwal, S.; Patel, P.: Sentiment Identification Using Maximum Entropy Analysis of Movie Reviews./, Jan. 2002.

[Pa18] ParallelDots, I.: Breakthrough Research Papers and Models for Sentiment Analysis, https://blog.paralleldots.com/data-science/breakthrough-research-papers-and-models-for-sentiment-analysis/, 2018, visited on: 03/15/2019.

[PP10] Pak, A.; Paroubek, P.: Twitter as a Corpus for Sentiment Analysis and Opinion Mining. In: Proceedings of the Seventh conference on International Language Resources and Evaluation. European Languages Resources Association (ELRA), Valletta, Malta, May 2010.

[Sc17] Schöni, M.: Was ist das Gradientenabstiegsverfahren?, https://zkmablog.com/2017/02/11/was-ist-das-gradientenabstiegsverfahren/, July 2017, visited on: 07/04/2019.

[SHA12] Saif, H.; He, Y.; Alani, H.: Semantic Sentiment Analysis of Twitter. In: The Semantic Web–ISWC 2012. Vol. 7649, pp. 508–524, Nov. 2012.

[Tw] Twitter: Glossary, URL: https://help.twitter.com/en/glossary, visited on: 07/05/2019.

[VCB14] Vanzo, A.; Croce, D.; Basili, R.: A context-based model for Sentiment Analysis in Twitter./, Aug. 2014.