A German Corpus on Topic Classification and Success of Social Media Posts from Facebook

Max-Emanuel Keller, Johannes Forster, Peter Mandl
Munich University of Applied Sciences
Munich, Germany
{max-emanuel.keller, johannes.forster}@hm.edu, mandl@cs.hm.edu

Frederic Aich, Jacqueline Althaller
ALTHALLER communication GbR
Munich, Germany
{fa, ja}@althallercommunication.de

***Abstract*—We provide a corpus consisting of 6,000 posts of German food delivery services from five brand pages on the online social network Facebook. The brand pages include Call a Pizza, Deliveroo, Domino's, Lieferando, Mundfein and Smiley's. A group of social media marketing experts annotated each post with one or more topic labels from eleven marketing related categories describing its content. Additionally an assessment on the success of the social media post is provided as a binary label. The inter-rater reliability over all annotators according to Fleiss' Kappa is 0.4835 for the topics and 0.6674 for success. Furthermore, baseline measurements with machine learning based text classification with an F_1-score up to 0.7173 are presented as a first experiment on this new corpus. The data set of the corpus on German topic classification and success (GTCS6k) is publicly available here: https://ccwi.github.io/corpus-gtcs6k**

I. Introduction

Online social networks are a medium that has experienced enormous popularity in recent years and is now used by a tremendous number of users. As a result, many companies have adopted it as an additional marketing channel, they use to operate brand pages to inform their customers about their company, products and services. The campaigns created for this purpose consist of several posts, where each of them is causing considerable expense in terms of costs and working time.

As a consequence, these companies want to determine whether their investments lead to the desired output. For this purpose, they would like to know for each single post they publish whether it will be successful or not. The success of a post can be measured by how it is perceived by the readers of the page. A strong indicator of this is the way the readers of the page interact with the post. Facebook gives users the ability to interact with posts in several ways. Users can leave different reactions, create a comment or share the post. Successful posts can be identified by the fact that they are often shared, receiving almost only comments with positive sentiment and predominantly positive reactions (like, love, wow, haha) while bad ones (angry, sad) are very rare. The goal of every company is to create as many successful posts as possible and avoid less successful ones. But in order to be able to create successful posts, the reasons for success must be determined. Success depends on many factors, whereby the topic plays an important role. However, the topics and contents that lead to successful posts can vary from company to company. Important characteristics are the company's industry, the demographics of the target group and the language of communication.

In order to evaluate the success of posts in relation to their topics, annotated posts are needed that classify the content of a post according to its topic and evaluate its success. There are many corpora with posts from social networks that have been annotated according to aspects like sentiment of comments [1, 2] or signs of hate speech [3]. However, to the best of our knowledge, there is no publicly available corpus with posts annotated according to their topic and success.

We have therefore created a new corpus consisting of Facebook posts from the food delivery services sector in Germany. The posts belong to six of the industry's most important brands and are in German language. The annotation was done by a group of social media marketing experts who annotated each post according to topic and success. To annotate the topic, the experts worked out a list of 11 topical classes of corporate communication. Each post was assigned one or more of these topical classes, which best describe its content. Success, on the other hand, was classified into *successful* and *not successful*, taking into account the formal markers such as user reactions but also the content, sentiment and context of the comments.

The corpus offers the possibility to examine the success of posts depending on their topic. Our goal is to develop a framework that can evaluate the success of a post and even predict the potential success of a new post before it is published [4]. However, we believe that the corpus can also be useful for other applications in natural language processing (NLP) and classification. This is why we have decided to make the corpus publicly available to the scientific community.

In this paper, we make the following contributions:

- We propose a methodology to annotate social media posts (section II) and describe the actual annotation process (section III).
- We provide a first baseline for a text classification of the posts by the topical categories, including natural language processing and different algorithms (section IV).
- We make the annotated corpus publicly available for research purpose (section VI).

II. Data set

The corpus presented in this work consists of 6,000 posts from Facebook. The posts come from six of the most important brand pages of the food delivery services industry in Germany and are written in German. Among the pages are Call a Pizza, Deliveroo, Domino's, Lieferando, Mundfein and Smiley's. The focus of the corpus lies on the content of the posts published by the companies on their pages. Each post has an ID, the post text, a publication date and other media associated with it, such as images, videos and links. In addition, there are interactions created by users, such as comments, reactions (like, wow, etc.) and shared posts.

The posts were annotated by a group of five experts who rated the success of a post as either *successful* or *not successful* and assigned one or more out of a total of 11 topical classes that best describe the content. All of the data mentioned above are part of the corpus that we make available to the scientific community. The details and requirements for the use of the corpus are described in section VI.

In the following sections we describe the procedure for the creation of the corpus. First, we give an overview of the annotation process, which is divided into several phases (section II-A). We then describe the classes used in the annotation process, which include success and the 11 topical classes (section II-B). Finally, in section II-C, we discuss the calculation of the inter-rater reliability, which measures the consistency of the experts in the annotation of posts.

A. Annotation process

The annotation process consists of three consecutive phases. During Phase 1, the experts were prepared to create a common understanding of the annotation process. The experts reviewed several posts and familiarized themselves with the classes that had to be annotated. Subsequently, the goal of Phase 2 was to ensure that the experts actually annotate according to the same rules. For this purpose, the agreement of the experts in the annotation was determined by calculating an inter-rater reliability. Phases 1 and 2 thus represented two training phases, hence their results were rejected. Finally, in Phase 3 further posts were annotated, that represent the productive part of the corpus.

B. Classes to annotate

During the annotation, the experts assessed two aspects of a post, its success and the topic. Success depends on the perception of the post by the users and can be rated as *successful* or *not successful*. The topic, on the other hand, describes the content of the post. Since it can have several topics at the same time, one or more topics can be chosen that best describe the content of the post. The following eleven topical classes, developed by the experts, were available for selection:

1) Product / Service: Product launch, preview, review
2) Event / Fair
3) Interactions: Contest, survey, question
4) News: News from the environment
5) Entertainment: Memes, jokes, virals, contests
6) Knowledge: Tip, expertise, insight, case study, FAQ, research
7) Recruiting / HR: Employee feature, interview, testimonial, job advertisement
8) Corporate social responsibility (CSR)
9) Advertising / Campaign: Testimonial, discounts, lead generation
10) Sponsoring
11) Other: None of the above categories

C. Inter-rater reliability

A common method to calculate the agreement of annotators during annotation is to let the same documents be evaluated by all annotators and then compare the results. The quality of the agreement can then be calculated by an inter-rater reliability.

A well known measure for the inter-rater reliability is Cohens Kappa [5], shown in eq. (1). It computes the observed match between two annotators $\bar{P}$ as well as the probability for an agreement based on chance $\bar{P}_e$ and from this it calculates the agreement κ. The height of the kappa value is thereby a measure for the quality of the annotation. An interpretation of the kappa value by Landis and Koch [6] is given in Table I.

$$\kappa = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e} \tag{1}$$

TABLE I. INTERPRETATION OF KAPPA [6]

Kappa Statistic	Strength of Agreement
< 0.00	Poor
$0.00 - 0.20$	Slight
$0.21 - 0.40$	Fair
$0.41 - 0.60$	Moderate
$0.61 - 0.80$	Substantial
$0.81 - 1.00$	Almost Perfect

However, Cohens Kappa has a major disadvantage. If there are three annotators instead of two, three kappa values for the three different combinations of the annotators can be calculated, however, a simple aggregation of these values to a single value is not possible. This drawback is solved by Fleiss Kappa [7], which represents a single kappa value that can be calculated with two or more annotators.

As Cohens Kappa, Fleiss kappa is calculated with eq. (1), whereby the probabilities are defined differently as illustrated by $\bar{P}$ in eq. (2) and eq. (3) as well as $\bar{P}_e$ in eq. (4) and eq. (5). Let N represent the number of subjects, which in our case is the number of posts, while n is the number of annotations per subject, and k is the number of categories. The subjects are indexed by $i = 1, 2, \ldots N$ and the categories by $j = 1, 2, \ldots k$. Hence, n_{ij} is the number of annotators who assigned the i-th post to the j-th class.

The number of categories is $k = 2$ for success, but not $k = 11$ for the topical classes. The second would require that the topical classes are mutually exclusive, which is not the case, as each post is assigned to one or more classes. We therefore calculate each class individually and binary with $k = 2$ and then make an average over the agreements.

$$\bar{P} = \frac{1}{N} \sum_{i=1}^{N} P_i \tag{2}$$

$$P_i = \frac{1}{n(n-1)} \sum_{j=1}^{k} n_{ij}(n_{ij} - 1) \tag{3}$$

$$\bar{P_e} = \sum_{j=1}^{k} {p_j}^2 \tag{4}$$

$$p_j = \frac{1}{Nn} \sum_{i=1}^{N} n_{ij} \tag{5}$$

III. Details

The annotation of the corpus was performed according to the procedure previously outlined in section II. In the following section III-A we first describe the procedure in greater detail. Then, in section III-B, we present the results of the annotation.

A. Annotation

The annotation was conducted by the communication agency *ALTHALLER communication GbR*, which advises enterprises in social media communication. Five experts from the agency, who regularly produce social media posts on behalf of clients, took over the work of the evaluation over a period of 2 months (May to June 2019).

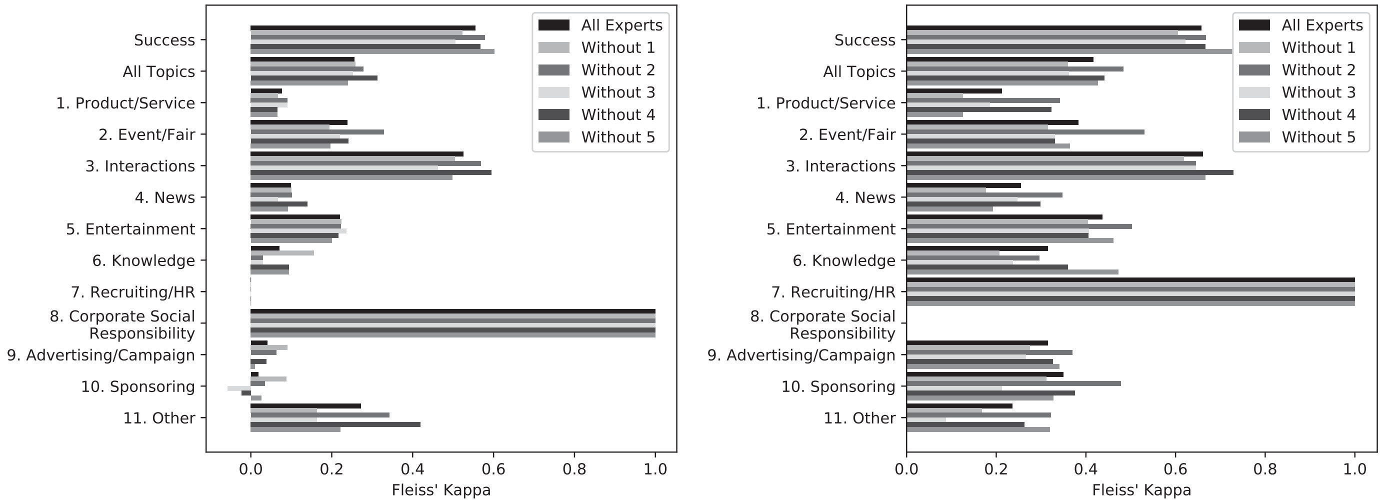
During Phase 1, the experts met for a discussion and reviewed 10 selected posts, which were manually selected to ensure that each of the posts belongs to a different one of the first ten topical classes. For each post they read the text and examined other media associated with it, such as pictures, videos and links. They also reviewed the number of interactions in the form of comments, reactions and shared posts that the post had received from the users. Regarding the assessment of the reactions they distinguished between the positive ones (like, love, wow, haha) and the negative ones (angry, sad). They also took a closer look at the comments and analyzed the content and sentiment. The positive and negative amount of reactions on their own are not sufficient indicators of the success of a post, as the context of written comments have to be taken into account. For example, there could be comments which are actually not related to the post at all but purely to other activities of the posting company or personal experiences of the users to the company which are not related to the actual post. Therefor also the success is rated manually as *successful* or *not successful* and at least one of the eleven classes was selected to describe the content. If the experts came to different views at first on how to assess one of these two points, this point was discussed until a consensus was reached that everyone approved. This approach was intended to create a common understanding among the experts on how to assess these two aspects.

The following Phase 2 should determine whether the experts actually evaluate with the same standards. To validate this, each of the experts received the same randomly selected 50 posts, which they then annotated on their own according to the rules previously established during Phase 1. Afterwards the results of the individual experts were used to determine the uniformity of the annotation. Therefore an inter-rater reliability with Fleiss' Kappa was calculated. The agreement for all experts was $\kappa = 0.5551$ for the success, which corresponds to a moderate agreement, and $\kappa = 0.2562$ for the topical classes, which in contrast only represents a fair agreement according to interpretation of Landis and Koch [6] mentioned above in Table I. Due to this second rather poor result for the topical classes, we examined the underlying causes. In order to determine the influence of every single expert, we calculated additional values for all combinations of $n - 1$ experts. The results are given in Table II in column *Phase 2a*. As the interpretation for the topics shows, only expert 4 has a slightly higher negative influence, since without him a $\kappa = 0.3123$ could be achieved. Overall, however, the values differed only slightly. To exclude further uncertainties regarding individual topics, we calculated an additional kappa value for each single category. Again, we computed a value for all experts as well as further values for all combinations of $n-1$ experts. Fig. 1a shows the results. As we see, the values for most topics barely differ. Only the topic *7. Recruiting/HR* has the anomaly that all values are $\kappa = 0$, which is due to the fact that none of the experts assigned this topic to even one of the posts. However, there were no obvious indicators for a single expert or a particular topic as a cause for the low agreement. From this result, we conclude that the experts had not formed a sufficient common understanding of annotation yet to be able to annotate uniformly.

TABLE II. Inter-rater reliability with Fleiss' Kappa

	Phase 2a		Phase 2b	
	Success	Topics	Success	Topics
All	0.5551	0.2562	0.6574	0.4163
Without 1	0.5224	0.2589	0.6054	0.3602
Without 2	0.5782	0.2782	0.6674	**0.4835**
Without 3	0.5053	0.2522	0.6221	0.3618
Without 4	0.5676	**0.3123**	0.6661	0.4412
Without 5	**0.6020**	0.2406	**0.7351**	0.4270

In order to improve the uniformity of annotation regarding the topical classes, the experts reviewed the 50 posts annotated during the first round of Phase 2 and discussed the rules of annotation again. Then a second round of Phase 2 was conducted, in which the experts received another 50 posts for annotation. To prevent that once again none of the posts belongs to the topic *7. Recruiting/HR*, three posts from this topic were selected manually, while the remaining 47 posts were randomly chosen as before. Table II presents the kappa values of the second round in column *Phase 2b*. As the results show, the agreement over all experts regarding success increased by 0.1023 to a substantial $\kappa = 0.6574$. But also the agreement concerning the topical classes had increased significantly by 0.1601 to $\kappa = 0.4163$, and now corresponded to a moderate agreement. At the same time, there are no significant differences for the individual topics, as shown by Fig. 1b. However, as in Phase 2a, there is again a topic, *8. Corporate Social Responsibility*, with all values at $\kappa = 0$, due to the fact that it were not selected at all. To further improve the agreement, we examined the kappa values in Table II in column *Phase 2b*. As we can see, expert 5 has the strongest negative influence on success, while expert 2 has it for the topical classes. In the balance between a solid agreement for the topical classes and an even higher one for success, we decided in favour of the topical classes for the compromise to exclude expert 2 from any further participation. This corresponds to the best agreement for the topical classes with $\kappa = 0.4835$ and the second best for the success with $\kappa = 0.6674$. A comparable corpus of Schabus *et al.* [8], we described in greater detail with related work in section V, comes to similar kappa values between 0.3 and 0.6.



(a) Phase 2a

(b) Phase 2b

Fig. 1. Agreement of Phase 2 - Inter-rater reliability with Fleiss' Kappa per class for each combination of $n - 1$ annotators

During the final Phase 3, the four remaining experts processed a further 6,000 randomly selected posts. Due to the solid inter-rater reliability, each post was annotated by only one of the experts. Table III summarizes the details of the annotation for the phases, including the number of annotations, number of unique posts and the experts involved.

TABLE III. PHASES OF THE ANNOTATION PROCESS

		Number of annotations		Experts	
Phase	Posts	Total	Per expert	Included	Excluded
1	10	10	-	1-5	
2a	50	250	50	1-5	
2b	50	250	50	1-5	
3	6000	6000	average 1500	1, 3-4	2

B. Corpus Statistics

During the three phases of the annotation, a total of 6,510 posts were annotated. Of these, 510 were processed during Phases 1 and 2, which only served to train the experts. The annotations of these posts were created before or for the purpose of calculating the inter-rater reliability and therefore have no guaranteed quality. Phase 3 was the first phase in which 6,000 further posts were annotated, whose quality is assured, which is why they form the core of the corpus.

Table IV provides the quota of the 6,000 posts assigned to each of the topical classes. As shown, the three most common topics are *9. Advertising/Campaing*, *3. Interactions* and *5. Entertainment*. In contrast, the topics *7. Recruiting/HR* and *8. Corporate Social Responsibility*, which were not included in Phases 2a and 2b respectively, were also very rarely found in the rest of the corpus. Finally, the relatively high proportion of the topic *11. Other*, with 8.93 %, indicates that the topical classes should possibly be extended by further ones. Table V shows the distribution of the second criterion success, which was assessed as either *successful* or *not successful*, with approximately three quarters of the posts being rated as *not successful*.

TABLE IV. NUMBER OF POSTS PER CLASS IN 6,000 POSTS (MULTI-LABEL)

Category	Posts	Percent
1. Product/Service	316	5.22 %
2. Event/Fair	368	6.07 %
3. Interactions	2370	39.12 %
4. News	547	9.03 %
5. Entertainment	978	16.14 %
6. Knowledge	390	6.44 %
7. Recruiting/HR	65	1.07 %
8. Corporate Social Responsibility (CSR)	40	0.66 %
9. Advertising/Campaign	4098	67.63 %
10. Sponsoring	322	5.31 %
11. Other	541	8.93 %

TABLE V. NUMBER OF POSTS BY SUCCESS IN 6,000 POSTS

	Posts	Percent
Not successful	4578	76.3 %
Successful	1422	23.7 %

IV. EXPERIMENTS

On the basis of the 6,000 annotated posts, we created a first baseline classification that predicts the association of the posts with the topical classes.

This classification task is a multi-label classification characterized by the fact that each post is assigned to a set of one or more classes, which in our case represent the eleven topical classes. A multi-label classifier would predict for a post whether it belongs to each one of the classes or not. We decided to simplify the problem by breaking it down into several binary classification problems. Therefore, we trained a separate binary classifier for each of the eleven topical classes, which predicts whether a post belongs to a certain topic or not.

As features of the classification, only the post text was used, while other attributes of the post, such as the page and date of publication or the number of interactions by users, were not taken into account. The problem thus represents a text classification for which there are a number of established algorithms. Two of these algorithms that we used are SVM (Support-vector machine) and ANN (Artificial neural network). The representation of the texts was implemented with BOW (Bag-of-words), where the texts are divided into tokens, which allows to count the number of occurrences of every single token in the text. Each text is then represented by a vector that indicates the number of occurrences of each token.

For the evaluation of the algorithms, the corpus of 6,000 posts was split 2/3 into a training set and 1/3 into a test set. The training set was then used to train each of the classifiers in a 5-fold cross-validation. Finally the classifiers were validated with the test set. To evaluate the predictions, the metrics Precision, Recall and F_1 were used.

The results of the baseline classification are given in Table VI. The best values of each row are depicted in bold, while the worst are slanted respectively. As the data show, the classes are predicted with varying quality by the different algorithms. However, the tendency can be seen that the height of the class distribution, shown in Table IV, has a major impact on the classification quality. Classes with balanced representation, like *3. Interactions* and *9. Sponsoring* could be predicted by all algorithms with good quality. In contrast, the classes *7. Recruiting/HR* and *8. Croporate Social Responsibility (CSR)*, which were only annotated at 1.07 % and 0.66 % respectively and thus very unevenly distributed, were not correctly classified by any of the algorithms, why we had to omit them. In order to improve the results, the classification would have to be optimized to unevenly distributed classes, which, however, would go beyond the focus of this work. Nevertheless, these classes may serve as a basis for research towards classification of imbalanced or skewed data [9, 10, 11].

TABLE VI. RESULTS OF THE BASELINE CLASSIFICATION

		Bag of words	
Cat.	Meas.	SVM	ANN
1.	Prec.	***0.1000***	***0.1000***
	Rec.	*0.0505*	**0.0909**
	F_1	*0.0671*	**0.0952**
2.	Prec.	**0.1186**	*0.1010*
	Rec.	*0.0588*	**0.0840**
	F_1	*0.0787*	**0.0917**
3.	Prec.	**0.4357**	*0.3991*
	Rec.	**0.3495**	*0.3420*
	F_1	**0.3879**	*0.3684*
4.	Prec.	**0.2105**	*0.1758*
	Rec.	*0.1117*	**0.1788**
	F_1	*0.1460*	**0.1773**
5.	Prec.	**0.1940**	*0.1487*
	Rec.	*0.1193*	**0.1437**
	F_1	**0.1477**	*0.1462*
6.	Prec.	*0.0192*	**0.0571**
	Rec.	*0.0078*	**0.0465**
	F_1	*0.0110*	**0.0513**
9.	Prec.	**0.6927**	*0.6924*
	Rec.	**0.7437**	*0.6991*
	F_1	**0.7173**	*0.6957*
10.	Prec.	**0.0278**	*0.0247*
	Rec.	*0.0094*	**0.0189**
	F_1	*0.0141*	**0.0214**
11.	Prec.	**0.1237**	*0.0941*
	Rec.	*0.0750*	**0.1000**
	F_1	*0.0934*	**0.0970**
Worst	Prec.	2	8
	Rec.	7	2
	F_1	6	3
Best	Prec.	8	2
	Rec.	2	7
	F_1	3	6

V. RELATED WORK

The creation of corpora is and has been a major task for computational linguistics. Corpora are used as evaluation base line and as training data for machine learning models in natural language processing. Hence there are numerous related corpora in the field of topic classification and in social media. In this section we focus on well known topic classification and social media corpora or specifically related corpora to our provided data set.

One of the most common corpora for topic classification is the 20 News Group data set of Lang [12]. It is a collection of nearly 20,000 newsgroup documents evenly split among 20 different newsgroups. These newsgroups labels act as the topical classes. However it is focused on news related language and only available in English. Therefore it is of limited use for German social media topic classification.

Larger news related data sets are provided by Reuters Ltd. There have been multiple different volumes of this data set. The well know Reuters-21578 data set has been replaced by the RCV1 dataset containing 810,000 news articles in English language. Related to our research work is the RCV2 data set with 487,000 multilingual documents in thirteen languages [13]. A subset of the Reuters corpus also represents the basis of the MLDoc corpus of Schwenk and Li [14], that is focused on cross-lingual document classification. Unfortunately, the stories are not parallel translated so only the German fraction is of interest. Additionally it is also not related to social media data.

Nobata *et al.* [15] provide a data set on abusive language with 1.2 million and 2.1 million comments from Yahoo! Finance and News, that were annotated after whether they were abusive or not. Another corpus on abusive language is the one of Waseem [3] that contains 6,909 posts from Twitter, that were annotated by amateurs and experts with the labels racist, sexist, both or neither. Although related to social media, the corpora are not suitable for the classification of topics or success.

Verhoeven *et al.* offer a multilingual social media corpus extracted from Twitter. It consists of the posts of 18,168 authors in six different languages and can be used for the classification of the gender of the authors and their personality. It includes posts in the languages German, Dutch, French, Spanish, Italian and Portuguese. The personality categories are split into four opposing binary categories: introvert vs. extrovert, sensing vs. intuition, thinking vs. feeling and judging vs. perceiving. As an additional category, the authors are split into male and female. Hence it is not suitable for a classical topic classification.

As there are only limited data sets in German, we draw the inspiration for our research work and our methodology from the One Million Post Corpus of Schabus *et al.* [8]. This corpus is a collection of one million comments from an Austrian online newspaper site, of which 11,773 were classified in seven categories. Since recently it is the baseline of the 10k German News Article data set [17] which can be used for topic classification tasks and challenges.

Closer to our domain of social media posts, is the work of Bretschneider and Peters [18]. This corpus also uses Facebook as a source. It consists of three data sets based on the comments of three different Facebook pages. It can be used to detect offending statements and hate speech against foreigners and therefore is still not a classical source for topic classification in an commercial related way.

Similar to the above is the multilingual corpus of Narr *et al.* [19, 2], which they use to train a language independent sentiment analysis. The data set consists 10,000 Twitter posts in English, German, French and Portuguese, which have been manually annotated into the sentiment categories of positive, neutral or negative.

VI. How to use the corpus?

We provide the corpus presented in this paper, consisting of the posts and the annotations, to the scientific community through a website (https://ccwi.github.io/corpus-gtcs6k). However, for legal reasons, we are not allowed to share the entire data of the posts directly. In order to still publish the corpus while respecting the rights of third parties, we instead provide the annotations along with the IDs of the posts and a script that allows interested readers to retrieve the posts of the corpus on their own. Besides the corpus, we also include the implementation of the experiments presented in section IV, which allows to reproduce the results.

VII. Conclusion and future work

In this work a corpus was presented which consists of 6,000 posts in German language that belong to six brand pages of German food delivery services on Facebook. The posts were annotated by experts according to topic and success, achieving a solid degree of agreement, as an inter-rater reliability according to Fleiss' Kappa revealed, which was 0.4835 for the topics and 0.6674 for success. To evaluate the corpus, a first baseline text classification was presented, where the text of each post was used to predict its association with each of the 11 topical classes, while comparing different algorithms. To make the corpus also usable for other applications in natural language processing and classification, we provide it as data set on German topic classification and success (GTCS6k) to the scientific community. In a future work, we will examine the annotations of the success for the posts depending on their topic. Our goal is to develop a framework that can evaluate the success of a post and even predict the potential success of a new post before it is published.

Acknowledgment

Our special thanks goes to the experts of *ALTHALLER communication GbR* who contributed to the annotation of the corpus. The presented work was conducted as part of a project funded by *Forschungs- und Entwicklungsprogramm Informations- und Kommunikationstechnik des Freistaates Bayern*. Funding reference number: IUK482/002.

References

[1] M. Cieliebak, J. M. Deriu, D. Egger, and F. Uzdilli, “A twitter corpus and benchmark resources for german sentiment analysis,” in *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, L.-W. Ku and C.-T. Li, Eds., event-place: Stroudsburg, PA, USA, Association for Computational Linguistics, 2017, pp. 45–51. DOI: 10.18653/v1/W17-1106.

[2] S. Narr, M. Hülfenhaus, and S. Albayrak, *Language-Independent Twitter Sentiment Analysis*. DAI-Labor, Technical University Berlin, Germany, 2012. [Online]. Available: http://www.dai-labor.de/fileadmin/Files/Publikationen/Buchdatei/narr-twittersentiment-KDML-LWA-2012.pdf.

[3] Z. Waseem, “Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter,” in *Proceedings of the First Workshop on NLP and Computational Social Science*, D. Bamman, A. S. Doğruöz, J. Eisenstein, D. Hovy, D. Jurgens, B. O'Connor, A. Oh, O. Tsur, and S. Volkova, Eds., event-place: Stroudsburg, PA, USA, Association for Computational Linguistics, 2016, pp. 138–142. DOI: 10.18653/v1/W16-5618.

[4] M.-E. Keller, B. Stoffelen, D. Kailer, P. Mandl, and J. Althaller, “Predicting the success of posts for brand pages on facebook,” in *Proceedings of the 17th International Conference WWW/Internet 2018, Budapest, Hungary: IADIS*, 2018.

[5] J. Cohen, “A coefficient of agreement for nominal scales,” *Educational and Psychological Measurement*, vol. 20, no. 1, pp. 37–46, Apr. 1960, ISSN: 0013-1644, 1552-3888. DOI: 10.1177/001316446002000104.

[6] J. R. Landis and G. G. Koch, “The measurement of observer agreement for categorical data,” *Biometrics*, vol. 33, no. 1, p. 159, 1977, ISSN: 0006341X. DOI: 10.2307/2529310.

[7] J. L. Fleiss, “Measuring nominal scale agreement among many raters.,” *Psychological Bulletin*, vol. 76, no. 5, pp. 378–382, 1971, ISSN: 0033-2909. DOI: 10.1037/h0031619.

[8] D. Schabus, M. Skowron, and M. Trapp, “One million posts,” in *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR '17*, N. Kando, T. Sakai, H. Joho, H. Li, A. P. de Vries, and R. W. White, Eds., event-place: New York, New York, USA, ACM Press, 2017, pp. 1241–1244, ISBN: 978-1-4503-5022-8. DOI: 10.1145/3077136.3080711.

[9] Y. Sun, A. K. C. Wong, and M. S. Kamel, “CLASSIFICATION OF IMBALANCED DATA: A REVIEW,” *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 23, no. 4, pp. 687–719, Jun. 2009, ISSN: 0218-0014, 1793-6381. DOI: 10 . 1142 / S0218001409007326.

[10] M. Galar, A. Fernandez, E. Barrenechea, H. Bustince, and F. Herrera, “A review on ensembles for the class imbalance problem: Bagging-, boosting-, and hybrid-based approaches,” *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 42, no. 4, pp. 463–484, Jul. 2012, ISSN: 1094-6977, 1558-2442. DOI: 10.1109/TSMCC.2011.2161285.

[11] V. López, A. Fernández, S. García, V. Palade, and F. Herrera, “An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics,” *Information Sciences*, vol. 250, pp. 113–141, Nov. 2013, ISSN: 00200255. DOI: 10.1016/j.ins.2013.07.007.

[12] K. Lang, “Newsweeder: Learning to filter netnews,” in *Machine Learning Proceedings 1995*, Elsevier, 1995, pp. 331–339.

[13] D. D. Lewis, Y. Yang, T. G. Rose, and F. Li, “Rcv1: A new benchmark collection for text categorization research,” *Journal of machine learning research*, vol. 5, pp. 361–397, Apr 2004.

[14] H. Schwenk and X. Li, “A corpus for multilingual document classification in eight languages,” in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, N. C. (chair), K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis, and T. Tokunaga, Eds., Miyazaki, Japan: European Language Resources Association (ELRA), May 7, 2018, ISBN: 979-10-95546-00-9.

[15] C. Nobata, J. Tetreault, A. Thomas, Y. Mehdad, and Y. Chang, “Abusive language detection in online user content,” in *Proceedings of the 25th International Conference on World Wide Web - WWW '16*, J. Bourdeau, J. A. Hendler, R. N. Nkambou, I. Horrocks, and B. Y. Zhao, Eds., event-place: New York, New York, USA, ACM Press, 2016, pp. 145–153, ISBN: 978-1-4503-4143-1. DOI: 10.1145/2872427.2883062.

[16] B. Verhoeven, W. Daelemans, and B. Plank, “Twisty: A multilingual twitter stylometry corpus for gender and personality profiling,” in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, 2016, pp. 1632–1637.

[17] T. Block. (2019). Ten thousand german news articles dataset, 10kgnad - a german topic classification dataset., [Online]. Available: https://tblock.github.io/10kGNAD/.

[18] U. Bretschneider and R. Peters, “Detecting offensive statements towards foreigners in social media,” in *Proceedings of the 50th Hawaii International Conference on System Sciences (2017)*, ser. Proceedings of the Annual Hawaii International Conference on System Sciences, Hawaii International Conference on System Sciences, 2017. DOI: 10.24251/HICSS.2017.268.

[19] S. Narr, M. Hulfenhaus, and S. Albayrak, “Language-independent twitter sentiment analysis,” *Knowledge discovery and machine learning (KDML), LWA*, pp. 12–14, 2012.