PotTS: The Potsdam Twitter Sentiment Corpus

Uladzimir Sidarenka
Applied Computational Linguistics
FSP Cognitive Science
University of Potsdam
Karl-Liebknecht Straße 24-25
14476 Potsdam
sidarenk@uni-potsdam.de

Abstract

In this paper, we introduce a novel comprehensive dataset of 7,992 German tweets, which were manually annotated by two human experts with fine-grained opinion relations. A rich annotation scheme used for this corpus includes such sentiment-relevant elements as opinion spans, their respective sources and targets, emotionally laden terms with their possible contextual negations and modifiers. Various inter-annotator agreement studies, which were carried out at different stages of work on these data (at the initial training phase, upon an adjudication step, and after the final annotation run), reveal that labeling evaluative judgements in microblogs is an inherently difficult task even for professional coders. These difficulties, however, can be alleviated by letting the annotators revise each other's decisions. Once rechecked, the experts can proceed with the annotation of further messages, staying at a fairly high level of agreement.

Keywords: language resources, sentiment analysis, social media, Twitter

1. Introduction

As people share more and more personal opinions via social media services, rapidly analyzing these subjective statements in an automatic way becomes a vital necessity for the success of modern social and commercial endeavors. This analysis, however, presupposes the existence of sufficiently big manually annotated corpora, since these are inevitably required for training new systems and testing existing applications. Although several attempts have already been made to create such datasets for English Twitter (Go et al., 2009; Pak and Paroubek, 2010; Nakov et al., 2013), the number of opinion corpora for this service in less-resourced languages still remains low.

In this paper, we try to overcome this limitation by presenting a novel collection of 7,992 German tweets, which were annotated with fine-grained sentiment relations by two human experts. With this resource, we not only aim at mitigating the data scarceness problem for German but also attempt to improve the current state of the art of opinion mining corpora in general by offering a dataset, which, in contrast to previous distantly supervised and therefore fundamentally noisy works (Go et al., 2009; Barbosa and Feng, 2010; Davidov et al., 2010), was created fully manually with a high level of inter-rater agreement; which, unlike many other hand-labeled sentiment data, is big enough to train and validate various machine-learning techniques; and, finally, which not only covers just one aspect of subjective judgements – as it was done, for instance, in the SemEval training set (Nakov et al., 2013), where only expression- and message-level polarities were annotated – but covers sentiments as a whole, providing precise information about their textual spans as well as the spans of their targets, sources, corresponding polar terms with their possible modifiers.

We begin our introduction by summarizing related work on opinion corpora for Twitter done so far. After describing the tracking procedure, which was used to initially collect and pre-select the message data, we present the annotation scheme that our experts relied on while annotating this project. Section 5. provides further details about the markup tool and format used for storing the annotations. Afterwards, in Section 6., we explain the evaluation metrics that we applied to estimate the inter-rater reliability at different stages of work on the corpus. We perform a brief analysis of the remaining disagreements in Section 8. before drawing conclusions and making suggestions for future research in the final part of this paper.

2. Related Work

Despite their relatively short history, opinion mining in general and sentiment corpora in particular have already attracted much attention of researchers from both computational and linguistic perspectives. A particularly important role in this regard has been played by datasets created for social media, such as Livejournal, Facebook, or Twitter, due to the crucial role that these services play in our society.

One of the first attempts to create a sentiment corpus of microblogs was made by Go et al. (2009). In their experiment, the authors gathered a collection of 1,6 M messages containing emoticons and automatically assigned polarity classes to the collected tweets using these smileys. A similar approach was also taken by Pak and Paroubek (2010), who applied distant supervision in order to obtain positive, negative, and neutral posts, subsequently training a Naïve Bayes classifier on these data.

Further works in this direction include those of Davidov et al. (2010), who retrieved 65 K microblogs featuring one of 50 emotionally laden hashtags and 15 common smileys, considering these entities as gold opinion categories for the downloaded tweets; Barbosa and Feng (2010), who used three publicly available automatic sentiment services to annotate a set of 200 K messages; and Kouloumpis et al. (2011), who adopted the hashtag approach of Davidov et al. (2010) to annotate the Edinburgh Twitter corpus (Petrović et al., 2010).

All of these resources, however, were created either fully automatically or in a semi-supervised way. The presumably biggest manually annotated public sentiment dataset for Twitter today is the SemEval corpus of Nakov et al. (2013). This set comprises 15 K messages, which were labeled with their overall polarities and polarities of their opinion expressions by five human experts on the crowdsourcing platform Amazon Mechanical Turk.

Unfortunately, much less work in this regard has been done for the non-English segment of Twitter so far. Notable exceptions to this are the labeled subset of the TWITA corpus (Basile and Nissim, 2013) and the Senti-TUT dataset of Bosco et al. (2013) created for Italian, as well as the TASS shared task data (Villena-Román et al., 2013) developed for Spanish.

The TWITA collection was used to create a smaller subcorpus of 2 K tweets, which were later manually labeled with their message-level polarities. The Senti-TUT tweebank comprises 3,288 messages pertaining to the election of Mario Monti and 1,159 microblogs obtained from the Twitter section of a popular Italian web portal,[1] which were annotated by five human coders with the following gold sentiment categories: positive, negative, ironic, mixed, or none. A different set of polarity classes (viz., strong negative, negative, neutral, positive, and strong positive) was distinguished in the TASS corpus, which provides 70 K Spanish messages.

With this work, we aim at filling the gap of such resources for German, a language, for which, to the best of our knowledge, only few automatic Twitter datasets exist to date (Tumasjan et al., 2010; Narr et al., 2012).

[1] http://www.spinoza.it

3. Data

In order to collect the initial data for our corpus, we were tracking German microblogs between March and September 2013 on the basis of extensive keyword lists (with several dozens entries each) pertaining to the following four topics:

- the federal elections in Germany in 2013,
- the papal conclave 2013,
- discussions of general political issues,
- and casual everyday conversations.

We obtained messages for the last part by taking the German Twitter snapshot of Scheffler (2014). This collection comprises $\approx$ 24 M tweets posted in April 2013, which were gathered by querying common German stop words from the Twitter Streaming API (Twitter, 2016). According to Scheffler (2014), this method allows to retrieve up to 95% of all Twitter posts written in German.

Our choice of topics was motivated by the wish to reduce the scarceness of sentiments in the resulting corpus by including political subjects, which a priori incite people to express more subjective judgements. However, in order to mitigate the bias introduced by this steered selection, we also added messages not prefiltered by any topical criteria to the resulting sampling set.

With this procedure, we were able to obtain a total of 27 M microblogs. To get a representative excerpt from this collection, we grouped tweets obtained for each topic into three disjunctive bins based on the following formal features:

- We put messages that contained at least one polar term from the sentiment lexicon SentiWS (Remus et al., 2010) into the first bin;
- Microblogs which did not satisfy the first criterion but had at least one emoticon or an exclamation mark were put into the second group;
- Finally, all remaining tweets were allocated to the third set of their respective topic.

Using such stratification, we, again, hoped to increase the recall of sentiments by separately analyzing messages with already known polar terms, which were indirectly more likely to contain subjective opinions as well.

In order to find such terms, we considered three major German polarity lists: SentiWS (Remus et al., 2010), German Polarity Clues (Waltinger, 2010), and the Zurich Sentiment Lexicon of Clematide and Klenner (2010), choosing in the end the first one due to its moderate size, acceptably high precision, and the availability of inflection forms of its entries.

Since no polar lexicon, however, is guaranteed to provide for the full coverage of opinionated expressions and, moreover, because Twitter users are renowned for their creativity in instantly inventing new language forms (Eisenstein, 2013), we also applied a bail-out approach by separately collecting messages which did not have any lexicon terms but did contain a smiley or exclamation point, assuming that either these elements alone would suffice to express subjective opinions or that they would reinforce the meaning of some accidentally missed polar words.

Finally, as we did not make any hypotheses about the distribution of sentiments in the rest of the tweets, we allocated all remaining microblogs to the same group, hoping that a uniform sampling from these data would provide us with further positive and negative opinion examples.

To get the final dataset, we eventually chose 666 random messages from each of the three bins of each of the four topics, getting a total of 7,992 microblogs: 666 tweets $\times$ 3 formal criteria $\times$ 4 topics.

4. Annotation Scheme

In the next step, we defined an annotation scheme for our corpus.[2] Since our goal was to get a maximally full coverage of all sentiment-relevant aspects, we devised an extensive list of elements that had to be annotated by the experts. This list included:

- **emotional expressions**, which we defined as words or phrases that unequivocally possessed some evaluative lexical meaning (e.g., *gut* "good", *schlecht* "bad", or *lieben* "to love");
- **intensifiers**, which were specified as elements that increased the expressivity or the polar sense of emotional items (for instance, *sehr* "very" or *außergewöhnlich* "extraordinarily");
- **diminishers**, which we described as lexical instances that decreased the polar sense of an emotional expression (for example, *wenig* "little" or *kaum* "hardly");
- and **negations**, which were linguistic means that turned the polarity of an emotional expression to the complete opposite.

As noted by one of the reviewers and as also confirmed in our later experiments, this definition of emotional expressions turned out to be not ideal though. Because we outlined these elements as *evaluative* lexical items, our experts annotated cases like "Held" (*hero*), "mögen" (*to like*), or "toll" (*awesome*) with this tag, but occasionally omitted emotionally connoted entities, such as "Erfolg" (*success*), "misslingen" (*to fail*), or "grimmig" (*grim*), since these did not appraise anything in particular but rather described general states of feeling.[3]

Additionally, in order to capture the interrelationship between the polar terms (emotional expressions in our case) and the entities they characterized, we also introduced the following elements in our scheme:

- **targets**, which we described as objects or events that were evaluated by sentiment expressions;
- **sources**, which were immediate author(s) or holder(s) of evaluative opinions;
- and actual **sentiments**, which we specified as minimal complete syntactic or discourse-level units in which both target and evaluative expression appeared together.

A sample tweet annotated according to our definitions is provided below:

Example 4.1

[[Diese Milliardeneinnahmen]$_{target}$ sind selbst [Schäuble]$_{source}$ [peinlich]$_{emo\text{-}expression}$]$_{sentiment}$

[[*These billions of revenues*]$_{target}$ *are* [*embarrassing*]$_{emo\text{-}expression}$ *even for* [*Schäuble*]$_{source}$]$_{sentiment}$

Beside annotating text spans of opinion-relevant items, our experts also specified the values of the attributes associated with these elements. For emotional expressions and sentiments, they determined the *polarity* and *intensity* of the respective judgements, distinguishing between positive, negative, and comparative cases for the former attribute and using a three point scale (weak, medium, and strong) for assessing the intensity.

Furthermore, we explicitly addressed the cases of *sarcasm* by providing a separate attribute for ironically

[2] Our annotation guidelines and corpus are available online at https://github.com/WladimirSidorenko/PotTS

[3] We are currently revising these cases, planning to release the first updated version of the corpus by June 2016.

meant polar terms and opinions. In total, our experts were able to find 145 cases of sarcastic sentiments, also detecting 82 examples of mocking lexical phrases.

For diminishers and intensifiers, the annotators were also asked to set the *degree* of these elements, which showed by how much they were changing the intensity of their respective emotional expressions. Finally, in cases when sources and targets were expressed by pronouns, our coders had to specify the respective *antecedents* of these pro-forms.

5. Annotation Tool and Format

All annotations were done using MMAX2 – a freely available text markup tool.[4] The primary reasons for choosing this program were *a)* its non-commercial license, *b)* its portability to a wide variety of platforms, and *c)* a mature set of annotation features, such as possibility to create link attributes (used for coreference), mark overlapping elements, and assign multiple annotations of the same class to one token (which was heavily used by our experts for the cases when one sentiment statement was included into another opinion, e.g., *She loves this ugly jacket*).

Since MMAX2 relies on a token-oriented stand-off XML format (where all annotations are stored separately from the original text and only refer to the ids of the tokens they are spanning), we first had to split the downloaded tweets into tokens in order to create an annotation project[5] for our dataset. For this purpose, we applied a minimally modified version of Christopher Potts' social media tokenizer,[6] which we slightly adjusted to the peculiarities of German spelling – we accounted for the capitalized form of German nouns and the dot at the end of ordinal numbers.

To ease the annotation process and minimize possible data loss during labeling, we split our complete dataset into 80 smaller project files with 99 – 109 tweets each. In each such file, we put microblogs pertaining to the same topic, making sure that the formal groups of that topic were represented in equal proportions.

In the last preparation step, we created the corresponding scheme and customization files, which specified what kinds of elements with which attributes were to be annotated by the human coders, and how these elements had to look like.

[4] http://mmax2.sourceforge.net/

[5] In MMAX2, an annotation project refers to a collection of all XML files pertaining to one corpus, including text data, annotation files, scheme definition etc.

[6] http://sentiment.christopherpotts.net/code-data/happyfuntokenizing.py

6. Inter-annotator Agreement

For estimating the inter-annotator agreement (IAA), we adopted the Cohen's κ metric (Cohen, 1960). Following the standard practice for computing this term, we calculated the observed agreement p_o as the ratio of tokens with matching annotations to the total number of tokens:

$$p_o = \frac{T - A_1 + M_1 - A_2 + M_2}{T},$$

where T denotes the total number of tokens, A_1 and A_2 are the numbers of tokens annotated with the given class by the first and second annotators respectively, and M_1 and M_2 represent the number of tokens with matching annotations for that class.

We also estimated the chance agreement p_c in the usual way as:

$$p_c = c_1 \times c_2 + (1.0 - c_1) \times (1.0 - c_2),$$

where c_1 and c_2 are the proportions of tokens annotated with the given class in the first and second annotations respectively, i.e., $c_1 = \frac{A_1}{T}$ and $c_2 = \frac{A_2}{T}$.

Two questions that arose during this computation, however, were *a)* whether tokens belonging to several overlapping annotation spans of the same class in one annotation had to be counted multiple times when computing the A scores (for instance, if we had to count the words *this*, *nice*, and *book* in Example 6.1 twice as sentiments when computing A_1 and A_2), and *b)* whether we had to assume that two annotated spans from different experts agreed on all of their tokens when these spans had at least one word in common (e.g., if we had to consider the annotation of the token *My* in Example 6.1 as matching, regarding that the rest of the corresponding sentiment spans agreed).

Example 6.1
Annotation 1:
[*My father hates* [*this nice book*]$_{sentiment}$.]$_{sentiment}$

Annotation 2:
My[*father hates* [*this nice book*]$_{sentiment}$.]$_{sentiment}$

To address these issues, we introduced two separate agreement metrics: *binary* and *proportional* kappa. With the former variant, we counted tokens belonging to multiple eponymous annotation spans multiple times and considered all tokens belonging to the given annotation instance as matching if this span agreed on at least one token with the annotation from the other expert. With the latter metric, every labeled token was counted only once, and we only calculated the actual number of tokens with matching annotations when computing the M scores.

Element	Initial annotation					Adjudication step					Final corpus				
	M_1	A_1	M_2	A_2	κ	M_1	A_1	M_2	A_2	κ	M_1	A_1	M_2	A_2	κ
Binary Kappa															
Sentiment	4,215	7,070	3,484	9,827	**38.05**	8,198	8,530	8,260	14,034	**67.92**	14,748	15,929	14,969	26,047	**65.03**
Target	1,103	1,943	1,217	4,162	**35.48**	3,088	3,407	2,814	5,303	**65.66**	5,765	6,629	5,292	9,852	**64.76**
Source	159	445	156	456	**34.53**	573	690	545	837	**72.91**	966	1,207	910	1,619	**65.99**
EExpression	1,951	2,854	2,029	3,188	**64.29**	3,164	3,298	3,261	4,134	**85.68**	5,574	5,989	5,659	7,419	**82.83**
Intensifier	57	101	59	123	**51.71**	111	219	113	180	**56.01**	192	432	194	338	**49.97**
Diminisher	3	10	3	8	**33.32**	9	16	10	16	**59.37**	16	30	17	34	**51.55**
Negation	21	63	21	83	**28.69**	68	84	67	140	**60.21**	111	132	110	243	**58.87**
Proportional Kappa															
Sentiment	3,269	6,812	3,269	9,796	**31.21**	7,435	8,243	7,435	13,714	**61.94**	13,316	15,375	13,316	25,352	**58.82**
Target	898	1,905	898	4,148	**26.85**	2,554	3,326	2,554	5,212	**57.27**	4,789	6,462	4,789	9,659	**56.61**
Source	153	439	153	456	**33.75**	539	676	539	833	**71.12**	898	1,180	898	1,604	**64.1**
EExpression	1,902	2,851	1,902	3,180	**61.36**	3,097	3,290	3,097	4,121	**82.64**	5,441	5,977	5,441	7,395	**80.29**
Intensifier	57	101	57	123	**50.81**	111	219	111	180	**55.51**	192	432	192	338	**49.71**
Diminisher	3	10	3	8	**33.32**	9	16	9	15	**58.05**	16	30	16	33	**50.78**
Negation	21	63	21	83	**28.69**	67	83	67	140	**60.03**	110	131	110	242	**58.92**

Table 1: Inter-coder agreement at different annotation stages.
($M1$ – number of tokens with matching labels in the first annotation, $A1$ – total number of labeled tokens in the first annotation, $M2$ – number of tokens with matching labels in the second annotation, $A2$ – total number of labeled tokens in the second annotation)

7. Annotation Procedure

After setting up the agreement metrics, we finally let our experts annotate the data. The annotation procedure was carried out in three steps:

- First, both annotators labeled one half of the corpus after only minimal training. Unfortunately, their mutual agreement at this stage was relatively low, reaching only 38.05% for sentiments (measured with binary κ) and being consequentially even lower for their corresponding targets and sources (amounting to 35.48 and 34.53% respectively). A notable fact, however, is that the consensus about emotional expressions at this time was notably higher, attaining 64.29%, which, according to Landis and Koch (1977), already suggests a substantial result;

- In the second step, in order to improve the inter-rater reliability, we automatically determined the differences between the two annotations, adding and highlighting unmatched elements as a separate class of labelings. We subsequently let our experts resolve these discrepancies by either correcting their own decisions or rejecting the alternative annotations of the other coder. As in the previous stage, we allowed the annotators to consult the supervisor (the author of this paper) about dealing with ambiguous cases, but did not let our assistants communicate with each other directly. This adjudication has lead to significant improvements on all annotation levels: The agreement on sentiments has improved by 29.87 percentage points, reaching 67.92%. Similar effects were also observed for targets, sources, emotional expressions, and their modifiers, resulting in an average IAA increase of 25.96 percent;

- Finally, after the adjudication was complete, our assistants proceeded with the annotation of the remaining files. Working completely independently, one of the experts has annotated 78.8 percent of the full corpus, whereas another coder has labeled the complete dataset.

8. Evaluation

The agreement results for each annotation stage computed with the two adjusted κ-metrics are shown in Table 1. As can be seen from the table, the inter-rater reliability of sentiments strongly correlates with the inter-annotator agreement on sources and targets, where higher sentiment figures inevitably lead to a better consensus about the holders and subjects of the opinions and vice versa. The same applies to the correlation between the emotional expressions and their modifiers. The latter elements, however, show generally lower scores apparently due to their smaller number in the corpus.

Regarding the annotation stages, one can observe that the peak of the annotators' agreement is reached upon the adjudication. Even though it decreases afterwards in the final step, this drop is not dramatic (typically amounting to only a couple of percentage points), and the average reliability is still almost two times better than the results obtained in the initial run.

As to the different variants of the κ-measure, we clearly can see that the binary metric has a direct relation to the proportional κ. The biggest difference be-

tween the two scores is noticed for sentiments and targets, reaching an average delta of $7.18^{\pm 1.37}$ in the final annotation, whereas, for the rest of the elements, it only runs up to $1.1^{\pm 1.07}$ percentage points. We explain this divergence by the fact that opinions and their targets are typically expressed by syntactic or discourse units (nominal phrases or clauses), whose boundaries are difficult to determine exactly because of their possible adjuncts. Sources, on the other hand, are most commonly represented by pronouns, which are much easier to spot as they usually lack any syntactic attributes. The same claim is true for emotional expressions and their supplementary elements, which are explicitly defined in our guidelines as *lexical* items, i.e., single words or clearly discernable idioms.

A sample case of diverging sentiment annotations is provided in Example 8.1. As can be seen from the labels, the second coder correctly interpreted the emoticon *:)* at the end of the tweet as an evaluative expression with respect to the complete content of the sentence. The first expert, on the other hand, stayed more conservative and did not regard this message as a sentiment case.

Example 8.1
Annotator 1:
@TinaPannes immerhin ist die #afd nicht dabei :)

Annotator 2:
@TinaPannes [immerhin ist die #afd nicht dabei :)]$_{sentiment}$

@TinaPannes [*anyway the #afd is not there :)*]$_{sentiment}$

As confirmed by our statistics (cf. Table 1) and as also shown in Example 8.2, the second annotator generally preferred annotating more opinions and emotional expressions than the first one. This also involved cases of emotionally connoted but non-evaluative terms mentioned previously. In the following example, for instance, the second expert considered three polar facts (*Angriff* "attack", *Bombe* "bomb", and *Frieden* "peace") as emotional expressions, whereas the first assistant did not include these words in his annotation.

Example 8.2
Annotator 1:
Syrien vor der Angriff – bringen diese Bomben den Frieden?

Annotator 2:
Syrien vor der [Angriff]$_{emo\text{-}expression}$ – bringen diese [Bomben]$_{emo\text{-}expression}$ den [Frieden]$_{emo\text{-}expression}$?

Syria facing an [*attack*]$_{emo\text{-}expression}$ *– will these* [*bombs*]$_{emo\text{-}expression}$ *bring* [*peace*]$_{emo\text{-}expression}$*?*

This difference of interpretations is partially due to the adjudication procedure that we applied, because a closer analysis of these cases revealed that, at the initial stage, our experts had had opposite preferences regarding the non-evaluative polar terms (viz., the first annotator had typically annotated them, whereas the second coder had usually skipped these entities). During the adjudication, both assistants interpreted their decisions as false, and both changed their minds. Even though the rest of their changes made in that step has still lead to significant improvements, the possibility of mutual concessions during adjudication needs to be kept in mind when applying this method in future.

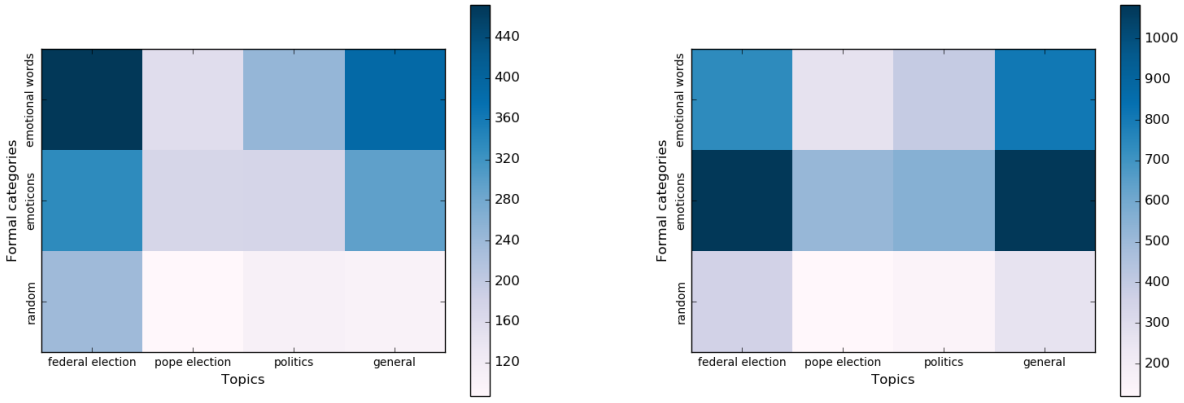
Nevertheless, even despite these deviating annotation cases, polar terms are still the most reliably labeled entities, having an agreement of more than 80%, which means an almost perfect score according to the scale of Landis and Koch (1977). The second best result (65.99%) is attained by sources. Intensifiers, on the other hand, show the lowest reliability (totaling 49.97 and 49.71% when measured with the binary and proportional κ respectively). However, after manually inspecting these disagreements, we came to the conclusion that the prevailing majority of these differences could be explained by the simple fact that the second annotator had considered exclamation marks as intensifying elements, whereas the first expert had only marked pure lexical items with this tag.

Element	Polarity κ	Intensity α
Sentiment	58.8	73.54
EExpression	87.12	78.79

Table 2: Inter-annotator agreement on polarity and intensity of sentiments and emotional expressions.

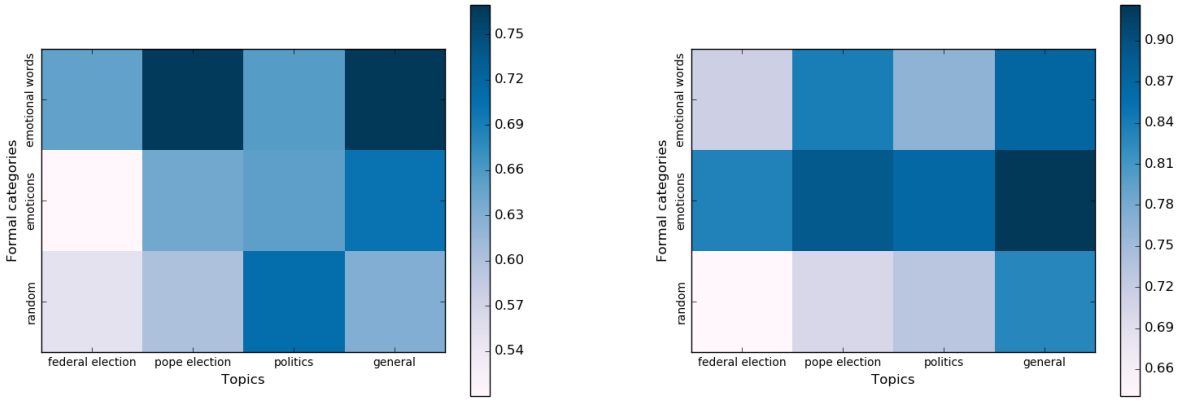
In order to see whether our experts also agreed on the attributes once they were at one about the elements, we additionally computed the Cohen's κ and Krippendorff's α (Krippendorff, 2007) for the polarities and intensities of agreeing opinions and polar expressions. For estimating the latter term, we used the ordinal distance measure, interpreting weak, medium, and strong intensities as zero, one, and two respectively. The results of these computations are shown in Table 2.

As can be seen from the scores, reaching a consensus about the polarities of emotional terms was much easier than agreeing on the value of this attribute for complete opinions. Similar to Example 8.1, one of the main reasons for these disagreements were subjective opinions containing smileys, especially in the



(a) `Sentiments` (b) `Emotional expressions`

Figure 1: Distribution of sentiments and emotional expressions across topics and formal categories.



(a) `Sentiments` (b) `Emotional expressions`

Figure 2: Inter-annotators agreement on sentiments and emotional expressions across topics and formal categories.

cases when the polarity of the emoticon contradicted the polarity of its preceding sentence, e.g., "Ich hasse die Piratenpartei ☺" (*I hate the Pirate Party* ☺).

Finally, to check how the selection criteria that we applied initially for sampling our corpus affected the final distribution of sentiments and polar expressions, we generated statistics plots on the frequencies and agreement level of these elements in the annotated dataset. As can be seen from the figures, the stratification according to topics and formal features has notably influenced both the number of these elements and the difficulty of their interpretation.

According to Figure 2, federal elections and topically unfiltered tweets are the ones that contain the major part of the opinions. A similar tendency is also observed for emotional expressions, though, in this case, the formal grouping appears to play a more important role than topics. Interestingly enough, the higher number of polar terms does not necessarily imply a higher number of targeted sentiments. We can recognize that from the fact that, even though most of the polar terms show up in the second row of the plot (i.e., in microblogs with smileys), the biggest number of opinions appear in row one (i.e., in tweets containing terms from the SentiWS lexicon).

Regarding the inter-annotator agreement, we can see that the highest reliability of annotated opinions is achieved on general tweets taken from casual everyday conversations. This group is also the one with the highest IAA scores for emotional expressions. A different situation, however, is observed for these two element types as to the formal groups of the tweets.

In this case, the first formal category (i.e., tweets with lexicon terms) appears to comprise messages with the most reliably annotated sentiments. For emotional expressions, however, the emoticons category, again, is the one with the highest achieved result, whereas, for opinions, the agreement scores in this row are among the lowest. This finding suggests that, even though, smileys are typically recognized as polar entities, the question whether they relate to something particular in the tweet or rather express the general mood of the author might often be difficult to answer.

9. Summary and Conclusions

Based on the descriptions outlined in the previous parts and summarizing our observations made above, we can formulate the main contributions and conclusions of this paper as follows:

- Our dataset notably contributes to the existing language resources by providing authentic German tweets with high-quality manual sentiment annotations;
- The rich annotation scheme used for this corpus touches on virtually every conceivable aspect of contemporary sentiment analysis, not only offering information about the textual spans of the opinions and the spans of their targets and holders but also showing details about polar terms and their modifiers;
- Additional attributes associated with these elements, such as intensities and polarities of emotional expressions and opinions, open up new possibilities for exploring the vast variety of ways, in which semantic compositionality works in sentiments;
- Beside providing the data, we also address theoretical agreement issues by specifically adopting the established Cohen's κ-metric (Cohen, 1960) to the peculiarities of the sentiment analysis task;
- Furthermore, after showing how difficult the opinion annotation might be even for pre-trained coders, we provide an efficient way of dealing with these difficulties using adjudication;
- Finally, a detailed inter-rater reliability study carried out at the end of the annotation process not only proves our claims about the high quality of the offered annotations but also reveals the specific influence that the applied topical and formal sampling criteria had on the resulting sentiment distribution.

Even though our study is based on a substantial amount of work, more things still need to be done. In particular, we are currently testing our annotation guidelines on a new group of undergraduate students to see whether the provided descriptions generalize to other coders as well. In addition to that, another experienced expert is currently revising the existing annotations of emotional expressions, resolving the cases of non-evaluative emotionally connoted terms. In this connection, we are also expanding the set of possible attributes for opinionated terms by providing a special flag for polar facts (objective words with inherent emotional associations, e.g., *disease*, *medicine*, *explosion*) and giving our annotator the possibility to specify how certain he is about his decisions.

Despite these remaining steps, however, we strongly believe that our corpus is already prepared for being used in further data exploration and classification experiments. To that end, we offer our dataset for free at the following URI address:

`https://github.com/WladimirSidorenko/PotTS`

10. Bibliographical References

Barbosa, L. and Feng, J. (2010). Robust sentiment detection on Twitter from biased and noisy data. In Huang and Jurafsky (Huang and Jurafsky, 2010), pages 36–44.

Basile, V. and Nissim, M. (2013). Sentiment analysis on Italian tweets. In *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 100–107, Atlanta, Georgia, June. Association for Computational Linguistics.

Bosco, C., Patti, V., and Bolioli, A. (2013). Developing corpora for sentiment analysis: The case of irony and senti-tut. *IEEE Intelligent Systems*, 28(2):55–63.

Nicoletta Calzolari, et al., editors. (2010). *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010, 17-23 May 2010, Valletta, Malta*. European Language Resources Association.

Clematide, S. and Klenner, M. (2010). Evaluation and extension of a polarity lexicon for German. In *Proceedings of the First Workshop on Computational Approaches to Subjectivity and Sentiment Analysis,*, pages 7–13.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Davidov, D., Tsur, O., and Rappoport, A. (2010). Enhanced sentiment learning using twitter hashtags

and smileys. In Huang and Jurafsky (Huang and Jurafsky, 2010), pages 241–249.

Eisenstein, J. (2013). What to do about bad language on the internet. In Lucy Vanderwende, et al., editors, *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 9-14, 2013, Westin Peachtree Plaza Hotel, Atlanta, Georgia, USA*, pages 359–369. The Association for Computational Linguistics.

Go, A., Bhayani, R., and Huang, L. (2009). Twitter sentiment classification using distant supervision. *Technical report*, pages 1–6.

Chu-Ren Huang et al., editors. (2010). *COLING 2010, 23rd International Conference on Computational Linguistics, Proceedings of the Conference, 23-27 August 2010, Beijing, China*. Tsinghua University Press.

Kouloumpis, E., Wilson, T., and Moore, J. D. (2011). Twitter sentiment analysis: The good the bad and the omg! In Lada A. Adamic, et al., editors, *Proceedings of the Fifth International Conference on Weblogs and Social Media, Barcelona, Catalonia, Spain, July 17-21, 2011*. The AAAI Press.

Krippendorff, K. (2007). Computing Krippendorff's Alpha Reliability. Technical report, University of Pennsylvania, Annenberg School for Communication, June.

Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, (33):159–174.

Nakov, P., Rosenthal, S., Kozareva, Z., Stoyanov, V., Ritter, A., and Wilson, T. (2013). Semeval-2013 task 2: Sentiment analysis in twitter. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 312–320, Atlanta, Georgia, USA, June. Association for Computational Linguistics.

Narr, S., Hülfenhaus, M., and Albayrak, S. (2012). Language-independent twitter sentiment analysis. In *Proceedings of the Workshop on Knowledge Discovery, Data Mining, and Machine Learning (KDML-2012)*.

Pak, A. and Paroubek, P. (2010). Twitter as a corpus for sentiment analysis and opinion mining. In Calzolari et al. (Calzolari et al., 2010).

Petrović, S., Osborne, M., and Lavrenko, V. (2010). The Edinburgh Twitter Corpus. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Linguistics in a World of Social Media*, WSA '10, pages 25–26, Stroudsburg, PA, USA. Association for Computational Linguistics.

Remus, R., Quasthoff, U., and Heyer, G. (2010). SentiWS - A publicly available German-language resource for sentiment analysis. In Calzolari et al. (Calzolari et al., 2010).

Scheffler, T. (2014). A German Twitter snapshot. In Nicoletta Calzolari, et al., editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014), Reykjavik, Iceland, May 26-31, 2014.*, pages 2284–2289. European Language Resources Association (ELRA).

Tumasjan, A., Sprenger, T. O., Sandner, P. G., and Welpe, I. M. (2010). Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment. In William W. Cohen et al., editors, *Proceedings of the Fourth International Conference on Weblogs and Social Media, ICWSM 2010, Washington, DC, USA, May 23-26, 2010*. The AAAI Press.

Twitter. (2016). Twitter Streaming API.

Villena-Román, J., Lana-Serrano, S., Martínez-Cámara, E., and Cristóbal, J. C. G. (2013). TASS - workshop on sentiment analysis at SEPLN. *Procesamiento del Lenguaje Natural*, 50:37–44.

Waltinger, U. (2010). GermanPolarityClues: A Lexical Resource for German Sentiment Analysis. In Calzolari et al. (Calzolari et al., 2010).