SB-CH: A Swiss German Corpus with Sentiment Annotations

Ralf Grubenmann, Don Tuggener, Pius von Däniken, Jan Deriu, Mark Cieliebak
Zurich University of Applied Sciences, Winterthur, Switzerland
SpinningBytes AG, Küsnacht, Switzerland
rg@spinningbytes.com, tuge@zhaw.ch, piusvd@gmail.com, deri@zhaw.ch, mc@spinningbytes.com

Abstract

We present the SB-CH corpus, a novel Swiss German corpus with annotations for sentiment analysis. It consists of more than 200,000 phrases (approx. 1 Mio tokens) from Facebook comments and online chats. Additionally, we provide sentiment annotations for almost 2000 Swiss German phrases. We describe the methodologies used in the collection and annotation of the data, and provide the first baseline results for Swiss German sentiment analysis.

Keywords: Swiss German, Sentiment Analysis, Emotion Corpus, Language Resources

1. Introduction

Swiss German denotes a collection of Alemannic dialects spoken in the German-speaking part of Switzerland. Although not one of the official languages, it is spoken on a daily basis by an estimated 4,5 million speakers, i.e. by more than three fifths of the Swiss population (Coray and Bartels, 2017). Swiss German is almost exclusively used for private communications between native speakers, while Swiss Standard German, which closely resembles the written German in Germany, is used for official and public communications. Due to its non-official status and the phonetic differences between the dialects, there is no standardized spelling for Swiss German, and it has been hardly used in written form (Baumgartner, 2003). Through the advent of social media and messaging systems, Swiss German has increasingly expanded to written form (Stark et al., 2015, e.g.). However, resources for written Swiss German are still sparse, and it can be considered a low-resourced language. Combined with the fluent nature of the (mostly) oral language, it is non-trivial to find a large collection of written dialect. Furthermore, there exists no Swiss German corpus for Sentiment Analysis.

Contributions In this paper, we present the following[1]:

- A new corpus *SB-CH* composed of more than 200,000 Swiss German phrases[2] and approx. 1 Mio tokens.
- *Sentiment annotations* for 1843 phrases with labels positive, negative, or neutral.
- *Baselines* for Swiss German sentiment classification.

2. Related Work

Language Resources for Swiss German At present, we are aware of the following resources for Swiss German:

- Swiss SMS Corpus (Stark et al., 2015): a collection of SMS which were sent by the Swiss public in 2009/2010. It contains 10,708 SMS in Swiss German, together with demographic information about the author.
- NOAH's Corpus of Swiss German Dialects (Hollenstein and Aepli, 2015): a compilation of 7,453 Swiss German texts from various genres, collected between 2010-2014. Text sources include the Alemannic Wikipedia, blog posts, novels by Viktor Schobinger, newspaper articles from "Blick am Abend", and the Swatch Annual Business Report. The texts were manually annotated with part-of-speech tags for 106,987 tokens.
- Sprachatlas der Schweiz (Baumgartner, 2003): a documentation of regional differences in Swiss German dialects, based on data collected between 1939-1957. It contains more than 1500 maps, depicting dialect varieties in Switzerland.
- Kleiner Sprachatlas der Deutschen Schweiz (Christen and Renn, 2010): an excerpt of the "Sprachatlas der Schweiz" with 120 maps and explanations.
- ArchiMob corpus (Samardžić et al., 2016): this corpus contains 300 interviews in Swiss German about the Second World War. Of these, 34 recordings were manually transcribed into "Schwyzertütschi Dialäktschrift" (Dieth, 1986), resulting in 528,381 tokens.

To complement these resources with a large-scale corpus, we have compiled a new corpus of written Swiss German, called SB-CH. This corpus is composed of 203,242 Swiss German phrases and 981,247 tokens. The texts were retrieved from Facebook comments and chat messages.

Sentiment Analysis Automatic sentiment analysis is a fundamental research area in natural language processing (NLP) and serves as a flagship task for other classification problems. Generally, the goal of sentiment analysis is to classify a text into one of the classes positive, negative, or neutral. Sentiment analysis has gained the interest of both academia and industry due to its important applications in e.g. social media monitoring or customer care. Various initiatives exist in the scientific community, such as shared

[1]The corpus including the annotations is available here: `https://www.spinningbytes.com/resources/swissgermansentiment`

[2]The term *phrases* is used to describe a unit of the dataset, which is typically a sentence or a short paragraph of text.

tasks at SemEval (Nakov et al., 2013) or TREC (Ounis et al., 2008). At present, most solutions for sentiment analysis incorporate supervised machine learning algorithms such as Support Vector Machines (Jaggi et al., 2014) or Convolutional Neural Networks (Deriu et al., 2016). In order to develop these systems, manually annotated training data is required. Various sentiment corpora exist for e.g. English (Nakov et al., 2013), German (Cieliebak et al., 2017), French (Bosco et al., 2016), or Italian (Barbieri et al., 2016). However, we are not aware of any sentiment corpora for Swiss German. For this reason, we enriched the SB-CH corpus with sentiment annotations for 1843 distinct phrases. In this paper, we describe the annotation process and measure the agreement among annotators along with key metrics of the corpus. Furthermore, we present results of baseline sentiment classifiers trained on our dataset. To the best of our knowledge, SB-CH is the largest collection of Swiss German documents at present, and the first one that contains sentiment annotations.

3. SB-CH Corpus Overview

The corpus consists of two parts: the full corpus of collected Swiss German texts and the subsection containing the phrases annotated with sentiment. This section details the properties of the full corpus, whereas the next section focuses on the annotated phrases.

3.1. Sources

To compile the corpus, we crawled two messaging platforms.

Facebook page "Schwiizerdütsch" The Facebook page "Schwiizerdütsch" (Swiss German)[3] posts about current events, news, and tradition in Switzerland. The page was crawled for comments on posts, most of which are written in Swiss German by native speakers. The comments were extracted using the official Facebook API, and no private posts or profiles were included. The comments were stored without user information and sanitized of user mentions. In total, 70,904 comments were crawled covering a time period from 2010 to 2017.

Chatmania The online chat platform "Chatmania"[4] was crawled for chat messages in public chatrooms. Since the platform is based on the Internet Relay Chat (IRC) protocol, messages could be obtained by simply joining the IRC channels and logging all messages. The messages were then filtered from the logs and cleaned of usernames and other possibly identifying information.

3.2. Corpus Statistics

SB-CH contains 203,242 Swiss German phrases with 981,247 tokens. Figure 1 shows the distribution of phrase lengths over the entire corpus, and Table 1 provides statistical data per source type.

The average phrase length in SB-CH is 28.6 characters, which is a lot shorter compared to other Swiss German corpora (e.g. 88.6 for NOAH corpus and 117.5 for Swiss SMS Corpus). This is probably due to the fact that SB-CH is built from social media posts and from turns in online chats, which can be quite short.

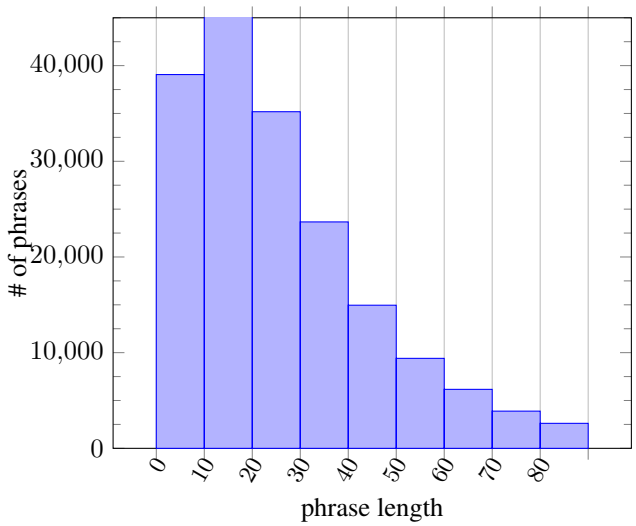

Figure 1: Histogram of phrase lengths in SB-CH by number of characters

Source	Phrases	Tokens	Characters per Phrase
Facebook	87,892	424,185	31.8
Chatmania	115,350	557,062	28.1
Total	**203,242**	**981,247**	
Corpora included for sentiment annotation			
NOAH corpus	7,453	106,987	88.6
Swiss SMS Corpus	10,708	217,940	117.5

Table 1: Corpus composition

4. Sentiment Annotation

In this section, we focus on the ongoing sentiment annotation of SB-CH. We describe the annotation scheme, the annotation guidelines, the annotation process, and provide statistics of the resulting annotations.

4.1. Sources

Two existing corpora were included alongside our newly created corpus SB-CH to get a more varied pool of phrases for annotation with sentiment polarity. These corpora are the NOAH Corpus (Hollenstein and Aepli, 2015) and the Swiss SMS Corpus (Stark et al., 2015). Random samples were then drawn from the pool and presented to annotators.

4.2. Annotation Scheme

The corpus sampled for sentiment annotation mainly consists of social media messages (chats, micro-blogging posts, SMS) and is thus similar to the corpus used in the SemEval tasks on sentiment detection (Nakov et al., 2013)

[3] https://www.facebook.com/Schwiizerduetsch/

[4] http://www.chatmania.ch/

which is compiled from Twitter and SMS messages in English. Following the SemEval scheme, we asked annotators to annotate message level polarity.[5] In addition to the standard labels for positive, negative and neutral (POS, NEG, NEUT), we used the label NA (i.e. *not applicable*) for phrases that were unintelligible due to e.g. errors in sentence splitting, not Swiss German, or too short ($<$ 3 words). Furthermore, the UNSURE label was introduced for phrases that contain both positive and negative sentiments, or phrases that depend on context for disambiguation w.r.t. polarity. The UNSURE label was also used for messages containing irony and sarcasm. The distinction between NA and UNSURE was introduced to separate noisy data from messages that might be interesting for future work (e.g. irony detection).

4.3. Annotation Process

Annotations were performed by 5 annotators using a custom web-based annotation platform. The group of annotators was composed of scientific assistants employed by the Zurich University of Applied Sciences (ZHAW) and employees of SpinningBytes AG. The annotators were from the cantons[6] of Aargau, Schaffhausen, Bern, St. Gallen and Zurich, giving a good distributions of speakers of different dialect groups.

A small set of phrases, picked following a uniform random distribution, was first annotated and studied to help develop the annotation guidelines. This was followed by two iterations of annotations, with annotators labeling randomly picked samples of the corpus in the first iteration, followed by the second iteration where the annotators labelled phrases previously annotated by a different annotator to ensure multiple annotations per phrase.

4.4. Annotation Guidelines

The texts in the corpus were randomly sampled from the heterogeneous sources of the corpus and presented to the annotators without context. Thus, we applied a guideline that focuses on text-level polarity, abstracting away from the (opaque) writer's intent. That is, annotators were asked to judge the sentiment of the phrases mainly by the polarity of the vocabulary encountered in them and regardless of pragmatics, where possible. For example:

- *Guete Morge!* is a common Swiss German greeting, *Good Morning*, and contains the word for *good*, which has a clearly positive polarity, so the text is judged as positive.
- *Lüg nid!* is an imperative, meaning *Don't lie!*, which contains the clearly negative word *lying*, so it is judged as negative.

The latter example could also appear as a flirtatious response in an online chat and be intended to carry an enticing, positive polarity. However, without the accompanying dialogue context, this cannot be inferred, and thus the phrase is labelled as negative based on the word *lying*.

[5]However, we did not ask annotators to annotate individual words or word spans within the messages with polarity.

[6]Member states of the Swiss Confederation

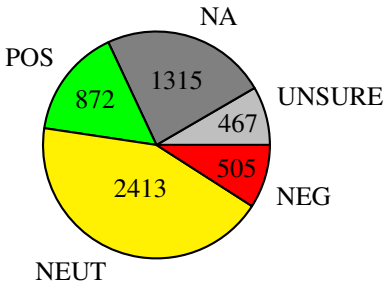

Figure 2: Distribution of labels in annotations

Questions were judged as neutral by default, since question polarity depends highly on pragmatics and intentions of the questioner cannot be determined reliably without it.
The distinction between subjectively or objectively positive/negative was of lesser importance. *"Ich mues is gfängnis"* (*"I have to go to prison"*) is negative for the writer, but can also be judged to be objectively positive (*a criminal has been caught*). But the word *prison* has a negative polarity, so the phrase is judged as negative.
While including salutations and greetings in the annotations is less common, our motivation to annotate them was to be able to capture polar information of important Swiss German adjectives (e.g. *Guet*) that are helpful in classification. This was also the driving idea behind putting a general focus on the lexical items in the annotation guidelines. Ideally, our annotations can be used to bootstrap a classifier for other domains, such as customer support messages in Swiss German in the future.

4.5. Annotation Statistics and Inter-Annotator Agreement

The statistics of the annotations are summarized in figure 2. The chart shows the number of phrases annotated with each label so far. The annotations show a distribution skewed towards the *neutral* label, similar to a related corpus for sentiment analysis comprised of Tweets in Standard German (Cieliebak et al., 2017), but also encountered in an English Twitter corpus for sentiment (Nakov et al., 2013).
On average, each phrase has been annotated by 2.45 annotators, and more than 80% of the phrases have been labelled by more than one annotator.
To calculate inter-annotator agreement, we applied Krippendorff's α metric (Krippendorff, 2007). We obtained an α of 0.42 over all annotations, and an α of 0.75 if only POS, NEG and NEUT annotations were regarded. We attribute this comparably high agreement to the relatively straightforward annotation guidelines and the fact that annotators did not have to select spans in the phrases, but only had to label the phrase-level polarity. Since UNSURE and NA describe inherent uncertainty regarding the annotations, the low agreement when including these classes is to be expected.

4.6. Sentiment Classification Baselines

In the following, we provide baseline sentiment polarity results for SB-CH using a traditional Support Vector Machine model (Cortes and Vapnik, 1995) and fastText (Joulin et al.,

2016), a model based on a shallow neural network. As features for the SVM approach, we use the TF-IDF vectors of the unigrams in the phrases to predict their polarity. Since our data is highly skewed, we apply the class imbalance to weight the samples by inverse frequency of their labels. For fastText, we tried different parameter settings and settled for the best performing one, i.e. reducing the number of dimensions of the embeddings and increasing the number of epochs.[7]

For our experiments, we harmonized the annotations per phrase by majority vote (and removed the phrases with a tie) to create a gold standard. We then sampled 20% of the phrases using stratified sampling to create a test set and used the remaining 80% of the phrases as training data. We trained and evaluated the models using two sets of labels. The first run made use of only the POS/NEG/NEUT labels as is common in the sentiment analysis literature. Additionally, we used all labelled phrases to simulate a more realistic setting where a system has to classify all phrases encountered in a stream of messages, as in e.g. social media streams.

Common labels

	Precision	Recall	F1-score
NEG	0.35	0.28	0.31
NEUT	0.75	0.86	0.80
POS	0.61	0.39	0.48
avg	0.68	0.70	0.68

All labels

	Precision	Recall	F1-score
NA	0.78	0.79	0.79
UNSURE	0.00	0.00	0.00
NEG	0.28	0.23	0.25
NEUT	0.63	0.66	0.64
POS	0.45	0.48	0.47
avg	0.63	0.64	0.63

Table 2: Results of SVM baseline

The results of the baseline models are presented in Tables 2 and 3. We report Precision, Recall, and F1-Score for each of the labels and weighted average F1. We see that both the SVM and fastText achieve similar performance. For the common labels, both models achieve rather low F1-Scores for the POS and NEG labels. This is not surprising due to the small size of our data at the current stage, and we believe results will improve once more phrases have been annotated. Unsurprisingly, including all labels in the classification task leads to a drop in performance for both models regarding the POS, NEG, and NEUT labels. The models perform quite well in labeling the NA class, but struggle with the UNSURE class. This is due to the lower amount of examples for the UNSURE class and the fact that the NA class subsumes noisy data that is often not Swiss German and thus consists of a different lexical items than the other classes.

The performance of the baselines at this stage of the annotation process meets our expectations and indicates that the annotation effort is going in the right direction.

[7] flags: -minCount 0 -dim 50 -epoch 50

Common labels

	Precision	Recall	F1-score
NEG	0.39	0.23	0.29
NEUT	0.74	0.90	0.81
POS	0.59	0.34	0.43
avg	0.67	0.70	0.67

All labels

	Precision	Recall	F1-score
NA	0.73	0.79	0.76
UNSURE	1.00	0.10	0.18
NEG	0.25	0.10	0.14
NEUT	0.62	0.75	0.68
POS	0.47	0.30	0.37
avg	0.63	0.65	0.62

Table 3: Results of fastText baseline

5. Conclusion

We presented, to the best of our knowledge, the largest collection of Swiss German texts, and the first annotated corpus of sentiment polarity for Swiss German. We achieved a high inter-annotator agreement of 0.75 (Krippendorff α) for our ongoing manual sentiment annotation, and we created baselines with reasonable F1-Scores for automatic sentiment prediction.

We expect that these resources will enable other researchers to address new and interesting research questions for Swiss German. One such question that we would like to tackle next is how resources and technologies for Standard German, which are available, can be utilized to improve solutions for Swiss German. If successful, such technology transfers could be applied to other low-resourced languages.

6. Acknowledgements

Zurich University of Applied Sciences (ZHAW) and SpinningBytes AG are collaborating in a joint research project to develop state-of-the-art solutions for text analytics tasks in several European languages. This research has been funded by Commission for Technology and Innovation (CTI) project no. 18832.1 PFES-ES.

7. References

Barbieri, F., Basile, V., Croce, D., Nissim, M., Novielli, N., and Patti, V. (2016). Overview of the EVALITA 2016 sentiment polarity classification task. In *Proceedings of Third Italian Conference on Computational Linguistics (CLiC-it 2016) & Fifth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2016)*.

Baumgartner, H. (2003). *Sprachatlas der deutschen Schweiz*, volume 9. Francke.

Bosco, C., Lai, M., Patti, V., and Virone, D. (2016). Tweeting and being ironic in the debate about a po-

litical reform: the French annotated corpus Twitter-Mariagepourtous. In *LREC*.

Christen, H. and Renn, M. (2010). *Kleiner Sprachatlas der deutschen Schweiz*. Huber.

Cieliebak, M., Deriu, J., Egger, D., and Uzdilli, F. (2017). A Twitter corpus and benchmark resources for German sentiment analysis. In *Proceedings of the 4th International Workshop on Natural Language Processing for Social Media (SocialNLP 2017)*.

Coray, R. and Bartels, L. (2017). *Schweizerdeutsch und Hochdeutsch in der Schweiz*, volume BFS-Nummer 1762-1700. Bundesamt für Statistik, Schweizerische Eidgenossenschaft (BFS).

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3):273–297, Sep.

Deriu, J., Gonzenbach, M., Uzdilli, F., Lucchi, A., De Luca, V., and Jaggi, M. (2016). SwissCheese at SemEval-2016 task 4: Sentiment classification using an ensemble of convolutional neural networks with distant supervision. In *SemEval@ NAACL-HLT*, pages 1124–1128.

Dieth, E. (1986). *Schwyzertütschi Dialäktschrift. Dieth-Schreibung*. Verlag Sauerländer.

Hollenstein, N. and Aepli, N. (2015). A resource for natural language processing of Swiss German dialects. In *GSCL*, pages 108–109.

Jaggi, M., Uzdilli, F., and Cieliebak, M. (2014). Swiss-Chocolate: Sentiment detection using sparse SVMs and part-of-speech n-grams. In *SemEval@ COLING*, pages 601–604.

Joulin, A., Grave, E., Bojanowski, P., and Mikolov, T. (2016). Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Krippendorff, K. (2007). Computing Krippendorff's alpha reliability. *Departmental papers (ASC)*, page 43.

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)*, volume 2, pages 312–320.

Ounis, I., Macdonald, C., and Soboroff, I. (2008). On the TREC blog track. In *ICWSM*.

Samardžić, T., Scherrer, Y., and Glaser, E. (2016). Archimob - a corpus of spoken Swiss German. In Nicoletta Calzolari, et al., editors, *Language Resources and Evaluation (LREC 2016)*, Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), pages 4061–4066. s.n., May.

Stark, E., Ueberwasser, S., and Ruef, B. (2015). Swiss SMS corpus. www.sms4science.ch.