Overview of the GermEval 2020 Shared Task on Swiss German Language Identification

Pius von Däniken **Manuela Hürlimann** **Mark Cieliebak**

Institute of Applied Information Technology
Zurich University of Applied Sciences
{vode, hueu, ciel}@zhaw.ch

Abstract

In this paper, we present the findings of the Shared Task on Swiss German Language Identification organised as part of the 7th edition of GermEval, co-located with SwissText and KONVENS 2020.

1 Introduction

Language Identification is the task of determining which language(s) a given piece of text is written in. It is an important step in many modern language processing pipelines, especially when working with online data sources as well as for tasks where downstream processing is language-dependent. While it has previously been proclaimed ”a solved problem” (McNamee, 2005), there are still several open challenges: handling short, noisy, user-generated text from social media is much harder than working with carefully composed and edited documents, such as news articles. Similarly, while some languages are easy to distinguish from each other, the more fine-grained the distinction we want to make, the harder it is to train systems to do so automatically. For instance, while it may be relatively easy to distinguish Arabic from English, it is difficult to distinguish different variations of Arabic from each other (Zampieri et al., 2018).

In this shared task, we are specifically interested in identifying Swiss German. While Standard German is one of the official languages of Switzerland (the others are French, Italian and Romansh), people in the German-speaking part of Switzerland speak a variety called Swiss German. It is composed of a range of local dialects, none of which have a standardized writing system. Nonetheless, the advent of the internet and social media has led to an increase in the written usage of Swiss German (Siebenhaar, 2003).

Since its written usage has only picked up in recent years, and there are only few native speakers to begin with, Swiss German can be considered a low-resource language. As such, it is not supported by most modern language identification tools.

In this task we are interested in identifying Swiss German as it is written on social media. We propose a binary classification task of distinguishing Swiss German from any other language. To that end we create a new data set from messages from the social media platform *Twitter*[1].

2 Related Work

Jauhiainen et al. (2018) have recently summarized the long history of language identification and the various approaches that have been explored over the years.

Recent editions of the VarDial workshop included many different language identification tasks (Zampieri et al., 2019, 2018, 2017). The tasks usually revolve around distinguishing similar languages, such as dialects of Arabic. Most importantly, it also included tasks on German Dialect Identification, which challenged participants to distinguish four regional dialects of Swiss German. The task data was taken from the ArchiMob corpus of Spoken Swiss German (Scherrer et al., 2019), which consists of interviews transcribed following the ”Schwyzertütschi Dialäktschrift” by Dieth (1986).

Linder et al. (2019) gathered a corpus of Swiss German from web resources. To build their corpus, they developed a language identification system based on the Leipzig text corpora (Goldhahn

[1] https://twitter.com

et al., 2012), reporting an accuracy of 99.58% using a fine-tuned BERT model (Devlin et al., 2019). Previously, von Däniken and Cieliebak (2018) built a simple binary SVM classifier based on character n-grams and trained it on data from the SB-CH corpus (Grubenmann et al., 2018).

2.1 Corpora

NOAH NOAH's corpus of Swiss German Dialects (Aepli et al., 2018) is a compilation of Swiss German texts from various sources and domains. It contains newspaper articles, blog posts, articles from the Alemannic Wikipedia, novels by Viktor Schobinger, and the Swatch Annual Business Report. Its 115'000 tokens have been annotated with Part-of-Speech tags.

Swiss SMS Corpus The Swiss SMS Corpus (Stark et al., 2009-2014) contains 25'947 SMS sent by the Swiss public in 2009 and 2010, of which around 41% are written in Swiss German.

ArchiMob The previously mentioned ArchiMob corpus (Scherrer et al., 2019) contains interview transcriptions. The latest release includes 43 transcripts with an average length of 15'000 tokens. The transcription script (Dieth, 1986) aims at a close phonetic representation of the pronunciation and is unfortunately not representative of how Swiss German is written on social media. For this reason, the corpus is not as useful for our purposes.

SB-CH Grubenmann et al. (2018) extended NOAH and the Swiss SMS Corpus with two new sources. The first is 87'892 comments crawled from a Facebook page dedicated to Swiss German, and the second are 115'350 messages gathered from the online chat platform "Chatmania". They provide sentiment annotations for parts of their corpus.

SwissCrawl Recently, Linder et al. (2019) built a large corpus of 562'521 Swiss German sentences from web resources.

3 Task Description

We propose a binary classification task of deciding whether a given Tweet is written in Swiss German (GSW) or any other language (NOT_GSW). The provided data comes from Twitter, which is a notoriously noisy data source. For training we only provided Tweets from the positive class (GSW), forcing participants to seek out a diverse set of additional resources to build robust systems, as the goal is to build a system that can generalize beyond Twitter.

Evaluation Participants were asked to submit predicted labels, as well as classifier scores, such as confidences, distances to decision boundary, or similar. We evaluate *Precision*, *Recall*, and *F1-score* of the predicted labels and rank systems according to their *F1-score* for the GSW class. Additionally we use the classifier scores to plot the Receiver Operating Characteristic (ROC) curve and Precision-Recall curves. We compute the Area Under the ROC curve (AUROC) and Average-Precision (AP) as secondary criteria to rank the submissions. While it is standard practice to use *F1-score* to evaluate text classification systems, we were also interested in the specific precision-recall trade-offs of the different submissions. We are particularly interested in applying insights of the submitted systems to collect further Swiss German samples, and for that it is useful to be able to adapt the classification threshold to limit false positives in practice.

4 Data

Instead of sampling data from Twitter directly, we chose to rely on the Swiss Twitter Corpus (Nalmpantis et al., 2018). It contains Tweets from 2017 and 2018 related to Switzerland, based on geolocation data, keywords related to Switzerland, and other criteria. The corpus contains a substantial subset of Tweets written in Swiss German, as well as a variety of other languages.

To build our data set, we sampled one million entries from the Swiss Twitter Corpus, and ranked them according to the SVM scores of von Däniken and Cieliebak (2018). We selected the top 10000 Tweets according to this score for manual annotation.

Every Tweet was annotated by one native speaker of Swiss German into one of four categories: The labels GSW and NOT_GSW were used for Tweets that are unambiguously written in Swiss German (GSW) or any other language (NOT_GSW). The label INDIST (short for *indistinguishable*) was used for Tweets where a distinction between GSW and NOT_GSW is not possible. This is for instance the case for short utterances consisting entirely of loanwords (*Merci!*, *Hallo*) or utterances where all tokens have the same surface form as another language but slightly different pronunciation in Swiss German (e.g. *Viel Spass!*). Finally, the label OTHER was used for Tweets that seemed to be nonsensical or spammy. A summary of the raw annotations is shown in Table 1.

Class	Count
GSW	5994
NOT_GSW	3908
INDIST	39
OTHER	59

Table 1: Overview of the number of raw annotations

For the released shared task data we excluded the categories INDIST and OTHER, since we deemed them not useful to evaluate language identification due to their nature and low occurrence rate (see Table 1). Since we only published Tweet IDs and their labels, in accordance with Twitter's Terms of Service, we also excluded Tweets which were not available anymore at the time of publication. We also manually removed a few duplicate entries before publication. The composition of the final released data set[2] can be seen in Table 2.

	Train		Test	
	freq	%	freq	%
GSW	2001	100	2592	48.2
NOT_GSW	0	0	2782	51.8
Total	2001	100	5374	100

Table 2: Class distribution in training and test data

5 Participants and Approaches

We had three groups participating in our shared task.

Models All three teams employed very different models and input representations. Team *jj-cl-uzh* trained a bi-directional GRU on character sequences (Goldzycher and Schaber, 2020). Team *IDIAP* applied an auto-encoder architecture to character n-gram BoW representations (Parida et al., 2020). Finally, team *Mohammadreza Banaei (MB)* employed a fine-tuned BERT model followed by a FastText classifier (Banaei, 2020).

Additional Corpora Used Table 3 shows additional corpora that the participants used. The following sources of Swiss German data were used: SwissCrawl, NOAH, the chatmania subcorpus from SB-CH, and the Swiss SMS Corpus. Similarly, the following corpora were used for NOT_GSW data: the Leipzig Corpora collection (Goldhahn et al., 2012), the Hamburg Dependency Treebank (Foth et al., 2014), the data for the second DSL shared task (DSLCCv2) (Zampieri et al., 2015), and the Ling10 corpus (Olafenwa and Olafenwa, 2018).

Fine Grained Classification The two leading teams (see Section 6) noticed that they get an improvement in performance when splitting the NOT_GSW class into sub-classes and training their classifiers on the fine-grained labels.

Data Augmentation Since the provided Tweets are substantially noisier than most of the other data sets, Team *jj-cl-uzh* chose to inject character- and token level noise into samples during training.

6 Results and Discussion

System	Precision	Recall	F1
MB	**0.984**	0.979	**0.982**
jj-cl-uzh	0.945	0.993	0.968
IDIAP	0.775	**0.998**	0.872

Table 4: Precision, Recall, and F1 scores for the positive class (GSW) of all submissions

The full evaluation results can be seen in Table 4 and Figure 1. Overall all teams achieve good scores with the two top teams ranking closely together and solving the task almost perfectly. Especially notable are the PR- and ROC-Curves, showing that one can achieve near perfect precision (recall) without sacrificing too much recall (precision).

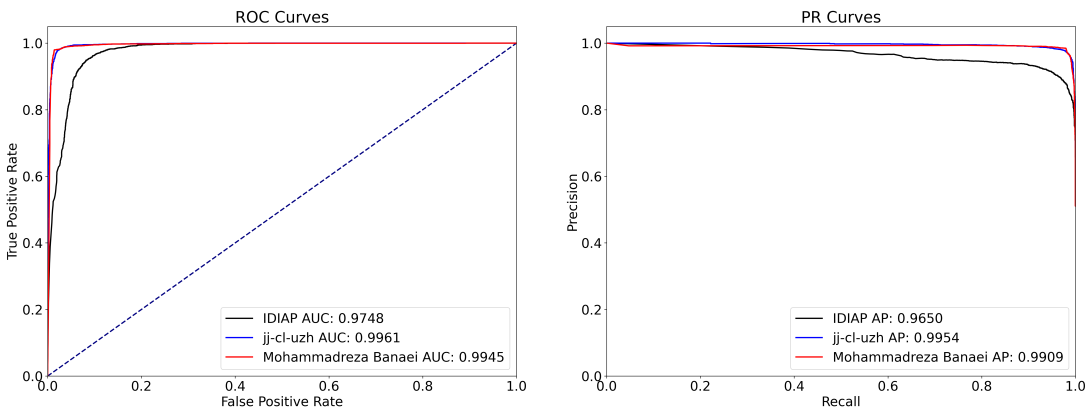
System Design Given that there were only three participating systems, it is hard to draw any general conclusions about the effectiveness of different systems and features. Nevertheless, given that both top performing systems applied fine-grained classification by sub-dividing the NOT_GSW class, this seems a good principle for other one-versus-all style language identification tasks.

Task and Data Overall we can conclude that the task of identifying Swiss German is indeed solv-

[2]The task data is available at: `https://github.zhaw.ch/vode/gswid2020/`

	IDIAP	jj-cl-uzh	MB
SwissCrawl	✓	✓	✓
NOAH	✓	✓	
SB-CH		✓	
Swiss SMS Corpus			✓
Leipzig Corpora Collection		✓	✓
Hamburg Dependency Treebank		✓	
DSLCC v2	✓		
Ling10	✓		

Table 3: Overview of the additional corpora used by participants



(a) Receiver Operating Characteristic Curve for all submissions and their respective Area Under Curve

(b) Precision Recall Curve for all submissions and their respective Average Precision

Figure 1: Evaluation Results based on classifier scores of all submissions

able to a high degree of fidelity, even when facing short and noisy user-generated utterances.

Future Work We see several important directions for future work. First of all we have to show that the results of this evaluation hold up to bigger data sets from a bigger range of domains. One source of noise in this task's data set is the propensity of users to code-switch to English and other languages. Therefore it would be interesting to generalize the current task to token-level language identification. Finally, good language identification enables us to gather larger high-quality corpora of Swiss German texts. This has already been achieved to an extent by Linder et al. (2019). Once enough Swiss German texts are available, the community can shift its efforts to extending the annotations of these corpora (cf. Section 2) and building up a collection of standard Natural Language Processing tools for Swiss German.

7 Conclusion

We described the findings of the Shared Task on Swiss German Language Identification which was part of GermEval 2020. The three participating teams achieved high evaluation scores, with the best system reaching an *F1-score* of 0.982 on the Swiss German class (evaluated on 5374 Tweets). This indicates that Swiss German language identification is feasible with high fidelity even for short, noisy, user-generated text.

References

Noëmi Aepli, Nora Hollenstein, and Simon Clematide. 2018. NOAH 3.0: Recent Improvements in a Part-of-Speech Tagged Corpus for Swiss German Dialects. In *Proceedings of the 3rd Swiss Text Analytics Conference - SwissText 2018*.

Mohammadreza Banaei. 2020. Spoken dialect identification in Twitter using a multi-filter architecture. In *Proceedings of the 5th Swiss Text Analytics Con-*

ference (SwissText) & 16th Conference on Natural Language Processing (KONVENS).

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Eugen Dieth. 1986. *Schwyzertütschi Dialäktschrift, 2nd Edition*. Sauerländer.

Pius von Däniken and Mark Cieliebak. 2018. Swiss German Language Detection in Online Resources. In *Proceedings of the 3rd Swiss Text Analytics Conference - SwissText 2018*.

Kilian A. Foth, Arne Köhn, Niels Beuck, and Wolfgang Menzel. 2014. Because Size Does Matter: The Hamburg Dependency Treebank. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland. European Language Resources Association (ELRA).

Dirk Goldhahn, Thomas Eckart, and Uwe Quasthoff. 2012. Building Large Monolingual Dictionaries at the Leipzig Corpora Collection: From 100 to 200 Languages. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*.

Janis Goldzycher and Jonathan Schaber. 2020. "Hold up, was zur höu isch ds?" Detecting Noisy Swiss German Web Text Using RNN- and Rule-Based Techniques. In *Proceedings of the 5th Swiss Text Analytics Conference (SwissText) & 16th Conference on Natural Language Processing (KONVENS)*.

Ralf Grubenmann, Don Tuggener, Pius von Däniken, Jan Deriu, and Mark Cieliebak. 2018. SB-CH: A Swiss German Corpus with Sentiment Annotations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Tommi Jauhiainen, Marco Lui, Marcos Zampieri, Timothy Baldwin, and Krister Lindén. 2018. Automatic Language Identification in Texts: A Survey. *Journal of Artificial Intelligence Research*, 65.

Lucy Linder, Michael Jungo, Jean Hennebert, Claudiu Musat, and Andreas Fischer. 2019. Automatic Creation of Text Corpora for Low-Resource Languages from the Internet: The Case of Swiss German.

Paul McNamee. 2005. Language identification: A solved problem suitable for undergraduate instruction. *Journal of Computing Sciences in Colleges*, 20:94–101.

Christoforos Nalmpantis, Fernando Benites, Michaela Hnizda, Daniel Kriech, Pius von Däniken, Ralf Grubenmann, and Mark Cieliebak. 2018. Swiss Twitter Corpus. In *Proceedings of the 3rd Swiss Text Analytics Conference - SwissText 2018*.

John Olafenwa and Moses Olafenwa. 2018. Ling10.

Shantipriya Parida, Esaú Villatoro-Tello, Qingran Zhan, Petr Motlicek, and Sajit Kumar. 2020. Idiap Submission to Swiss German Language Detection Shared Task. In *Proceedings of the 5th Swiss Text Analytics Conference (SwissText) & 16th Conference on Natural Language Processing (KONVENS)*.

Yves Scherrer, Tanja Samardžić, and Elvira Glaser. 2019. ArchiMob: Ein multidialektales Korpus schweizerdeutscher Spontansprache. *Linguistik Online*, 98(5):425–454.

Beat Siebenhaar. 2003. Sprachgeographische Aspekte der Morphologie und Verschriftung in schweizerdeutschen Chats. *Linguistik online*, 15.

Elisabeth Stark, Simone Ueberwasser, and Beni Ruef. 2009-2014. Swiss SMS Corpus. University of Zurich. https://sms.linguistik.uzh.ch.

Marcos Zampieri, Shervin Malmasi, Nikola Ljubešić, Preslav Nakov, Ahmed Ali, Jörg Tiedemann, Yves Scherrer, and Noëmi Aepli. 2017. Findings of the VarDial evaluation campaign 2017. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, pages 1–15, Valencia, Spain. Association for Computational Linguistics.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Ahmed Ali, Suwon Shon, James Glass, Yves Scherrer, Tanja Samardžić, Nikola Ljubešić, Jörg Tiedemann, Chris van der Lee, Stefan Grondelaers, Nelleke Oostdijk, Antal van den Bosch, Ritesh Kumar, Bornini Lahiri, and Mayank Jain. 2018. Language Identification and Morphosyntactic Tagging: The Second VarDial Evaluation campaign. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, Santa Fe, USA.

Marcos Zampieri, Shervin Malmasi, Yves Scherrer, Tanja Samardžić, Francis Tyers, Miikka Silfverberg, Natalia Klyueva, Tung-Le Pan, Chu-Ren Huang, Radu Tudor Ionescu, Andrei M. Butnaru, and Tommi Jauhiainen. 2019. A report on the third VarDial evaluation campaign. In *Proceedings of the Sixth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 1–16, Ann Arbor, Michigan. Association for Computational Linguistics.

Marcos Zampieri, Liling Tan, Nikola Ljubešić, Jörg Tiedemann, and Preslav Nakov. 2015. Overview of the DSL Shared Task 2015. In *Proceedings of the Joint Workshop on Language Technology for Closely Related Languages, Varieties and Dialects*, pages 1–9, Hissar, Bulgaria. Association for Computational Linguistics.