Universität Zürich[UZH]

Bachelor thesis
for obtaining the academic degree
Bachelor of Arts
from the Philosophical Faculty of the University of Zurich

Analysing Social Media comments with Zero-Shot and Few-Shot Learning Techniques

Author: Moritz Preisig
Immatriculation Nr.: 19-739-283

Referent: Gerold Schneider, PD Dr.

Institute of Computational Linguistics

Date of submission: 01.12.2022

Abstract

In recent years social media platforms like Facebook, Twitter, etc., have gained lots of attention, attracting millions of users who contribute daily on these platforms. The massive amount of data generated by users has made it more and more challenging to deal with all the information present at hand. Therefore, the need for automatic classification of social media texts has risen as it allows us to work against the significant unclassifiable stream of information. This thesis tries to show a solution by working on three different tasks proposed by the Shared Task GermEval 2021 [Risch et al., 2021] on the Identification of Toxic, Engaging, and Fact-Claiming Comments. Two approaches are proposed to solve the three tasks of the competition. Firstly, the task is reformulated as a textual entailment problem and then solved by a Zero-Shot Text Classification which tries to classify the text without any labeled training data by relying on the power of pre-trained language models. Secondly, a Few-Shot Text Classification approach is used where small amounts of training data (8, 16, 32, 126, 256) are used. These data-efficient approaches try to open up the field for many other tasks where only limited training data is available, as state-of-the-art systems rely on a vast number of training data. My results show that using the method of Hypothesis Engineering can improve performance in some of the sub-tasks, which was also the case when adding small amounts of training data. Reaching an f1-score of 61.0%, 64.4%, and 63.3% on Toxic, Engaging, and Fact Claiming detection does not compete with the top submissions of the shared task but delivers a solid data-efficient approach. It also is the first submission on GermEval21 using Zero-Shot and Few-Shot techniques which makes it novel work in this discipline.

Zusammenfassung

Nutzergenerierte Inhalte im Internet, insbesondere in den sozialen Medien, sind zu einem festen Bestandteil unseres Alltags geworden. Um mit der dabei entstehenden, stetig wachsenden Menge von Text umgehen zu können, sind automatische Textklassifizierungen gefragter denn je. In dieser Arbeit erläutere ich einen Ansatz, im Rahmen des Shared Task GermEval 2021, zur Klassifizierung von toxischen (toxic), dem Diskurs positiv beitragenden (engaging) und Tatsachen behauptenden (fact-claiming) Texten. Hierfür schlage ich zwei verschiedene Methoden der automatischen Textklassifikation an Hand von maschinellem Lernen, nämlich Zero-Shot (Null-Schuss) und Few-Shot Learning (Wenig-Schuss), vor und vergleiche sie mit dem neuesten Stand der Forschung. Meine Ergebnisse zeigen, dass die Verwendung der Methode des Hypothesis Engineering die Leistung in einigen der Teilaufgaben verbessern kann, ebenso wie das Hinzufügen einer kleinen Menge von Trainingsdaten. Das Erreichen eines F1-Scores von 61,0%, 64,4% und 63,3% bei der Erkennung von Toxizität, Engagement und Tatsachenbehauptungen konkurriert nicht mit den besten Resultaten des Shared Tasks, liefert aber einen soliden ressourcenschonenden Ansatz.

Acknowledgement

I want to thank Dr. Gerold Schneider and Janis Goldzycher for introducing me to the topic of Zero-Shot learning and assisting me in setting up my experiments and evaluating the results. Many thanks also go to my parents and Gian Radler for proofreading this thesis. Lastly, I thank Marco Leder for helping me set up the cloud computing platform, saving me quite a few headaches.

Contents

List of Figures

List of Tables

List of Acronyms

BERT	Bidirectional Encoder Representation from Transformers
CNN	Convolutional Neural Network
GRU	Gated Recurrent Unit
LM	Language Model
LSTM	Long-Short Term Memory
ML	Machine Learning
NLI	Natural Language Inference
NLP	Natural Language Processing
RNN	Recurrent Neural Network
SVM	Support Vector Machines

1 Introduction

1.1 Motivation

The rise of social media platforms has led to millions of users being active on platforms like Twitter, Facebook, and others, creating vast amounts of textual data. Social media allows you to gather and share information on a scale that is hard to compare with any other phenomenon. Using it in a meaningful and beneficial way can help us solve problems we were unable to solve previously. To maximize the use of social media, the approach of automatic text classification to help moderate the platforms is crucial as it can dissect the immense stream of information into manageable parts and work against the improper use of social media. Of the same opinion were the organizers of the Shared Task GermEval21 [Risch et al., 2021] competition, which created three sub-tasks in the realm of Social Media Text Classification: Toxic, Engaging, and Fact Claiming Detection. Fifteen different teams took part in finding the state-of-the-art approach for each of the three sub-tasks. In this thesis, I present my attempt at the competition using two different machine learning approaches: Zero-Shot and Few-Shot Text Classification. Zero-Shot learning is a relatively new approach to text classification. The attractive characteristic is that it does not use any labeled data compared to most state-of-the-art systems that use an abundance of training data. It applies to many different tasks and, in the field of NLP, is an exciting step towards Artificial General Intelligence as it uses general language models to predict the classifications. To my knowledge, the approach that I present in this thesis is also the first submission in GermEval21 that uses Zero-Shot techniques. Being the first to tackle this problem is very motivating, as it can be seen as novel work in this field. On the other hand, this thesis is one of the few Zero-Shot and Few-Shot experiments in German that does not use a multilingual model to make its predictions.

1.2 Research Questions

The thesis describes my approach to solving the GermEval21 tasks. The research questions that shall be answered in this thesis are:

1. How well does the Zero-Shot method work compared to the models developed for the Shared Task?
2. How much can the approach of Hypothesis Engineering, the combination of different hypotheses, help improve the experiments?
3. How much does the introduction of a small amount of training data, so-called Few-Shot learning, help improve the performance?

1.3 Thesis Structure

In this first chapter, I start this thesis by showing overall motivation and listing the research questions.
Chapter 2 introduces the previous work on three sub-tasks, including some submissions of competitors of GermEval21 but also shows some related work on Zero- and Few-Shot work.
Chapter 3 presents the dataset at hand and its more fine-grained subcategories.
Chapter 4 presents the Zero-Shot experiments, firstly in their base form and later combined with Hypothesis Engineering and its corresponding results.
Chapter 5 presents the Few-Shot experiments and results.
Chapters 6 and 7 contain a short discussion and conclusion and what I have learned in the process of writing this thesis.
In the Appendix, you can find the detailed results of the experiments that I conducted.
For all of the code that is used to produce the results, please refer to:
`https://github.com/moprei21/Bachelor_Thesis`

2 Related Work

2.1 GermEval

To give an idea of how the shared task is built up it is helpful to learn about the organization and previous occurrences of the shared task. GermEval 2021 is the seventh edition as part of a series of workshops on German NLP tasks, with its first edition in 2014. The shared task is self-organized by different special interest groups of the German Society of Computational Linguistics (GSCL). The workshops are held within the scope of the Conference of Natural Language Processing (KONVENS), which was held in Düsseldorf. Since 2014 many topics have been treated in the shared task ranging from Sentiment Analysis, Named Entity Recognition, over lexical substitutions, and hierarchical classification of blurbs to identifying offensive language. The latter topic was part of previous renditions in 2018 and 2019 and was part of the motivation of the organizers to create the Toxic sub-task in GermEval21.

2.2 Toxicity

Detecting Toxic language, also known as offensive language, hate speech, abusive language, or incivility, is currently one of the most researched fields in NLP. Due to the availability of annotated datasets, the majority of study on this topic is based on English data. To get an idea of the datasets, please refer to [Zampieri et al., 2019] and [Rosenthal et al., 2020]. For other languages, the research examines offensive content over various social media platforms in Greek, Italian, and Portuguese, to name a few. When looking at German, there are the two datasets from previous editions of GermEval, GermEval2018 [Wiegand et al., 2018], and GermEval2019 [Struß et al., 2019], which contained comments of the social media platform Twitter or DeTox [Demus et al., 2022] which includes different datasets, including all available GermEval datasets but also other datasets like for instance the HASOC 2019 dataset [Mandl et al., 2019].
The first approaches to solve the problem of classifying offensive content on so-

cial media used different approaches from traditional machine learning, such as SVMs and Logistic Regression [Malmasi and Zampieri, 2018]. The introduction of transformer-based architecture like BERT [Devlin et al., 2018] caused a stir in the field of NLP, also affecting the detection of hate speech and toxic language. It does not sound surprising that in GermEval21, all teams except one considered such contextualized embeddings like BERT. This is also the case for the best-performing submission that used two different large pre-trained neural networks: BERT and ELECTRA [Clark et al., 2020]. They fine-tuned the German versions of said neural networks, created ensembles, and analyzed how the performance reacted on the different ensemble members [Bornheim et al., 2021]. With this approach, they reached a f1-score of 71.75 % on the test set. Other submissions used different types of classifiers ranging from SVMs over Logistic Regressions to Random Forests or other deep-learning architectures like CNNs, GRUs, or LSTMs. As an honorable mention, one team used a rule-based approach but could not reach comparable performance.

2.3 Engagement

The motivation behind the task of detecting engaging comments stems from the idea of highlighting comments that encourage and foster reasoned and civil discussions [Ziegele et al., 2018]. The first contribution to the task was by Napoles et al. [2017], who created an annotated dataset of engaging, respectful, and informative conversations. In their groundwork, they analyzed the characteristics of these conversations. They found that characteristics such as being on-topic in context and persuasive but not sarcastic and mean lead to an engaging environment. Their follow-up work used a CNN with word embeddings to classify engaging conversations. Regarding deep learning [Risch and Krestel, 2020] applied some methods to classify engaging comments by analyzing the replies specific comments received and their respective upvotes. They use a trained RNN and CNN model for the classification and achieve accuracies ranging from 68 to 72 percent. With the introduction of GermEval21, the task was solved by transformer-based approaches for the first time. The best performing approach focused on feature-engineering with conventional classification methods [Hildebrandt et al., 2021]. They combine a pre-trained BERT embedding with a writing style embedding in the framework of ADHOMINEM [Boenninghoff et al., 2019] with additional numerical features like Average Emoji Representation, Number of References, and Spelling Mistakes. A majority voting approach is applied to generate predictions for each sub-task using ensembles of Logistic Regression classifiers and SVMs. With this, they reach a f1-score of 69.9%.

2.4 Fact Claiming

Fake news is deliberately created to encourage readers to believe incorrect information, making it challenging and nontrivial to identify based on news content. To open up this field, the task is to find check-worthy factual claims that could be analyzed manually or semi-automatic in a second step to find out if they are factually correct. Note that detecting fact claiming is not the same as detecting fake news, as these two tasks require different approaches. The first approach to solve this was proposed by Hassan et al. [2017], providing a semi-automated approach for fact-checking. They created a knowledge base inspired by the e U.S. general election debate transcripts. As a result, they also presented the *ClaimBuster* dataset, which contains check-worthy factual claims. There has also been a series of shared tasks on the verification and automatic classification of social media texts called *CLEF - CheckThat! Lab* on different languages and tasks like check-worthiness, retrieving previously fact-checked claims, evidence retrieval, and claim verification [Barrón-Cedeño et al., 2020]. The best performing model in GermEval21 was again transformer-based using an ensemble strategy by majority (hard) voting [Tran and Kruschwitz, 2021]. They used three different BERT models, namely: a Twitter-based, a German-based, and a multilingual model, of which the Twitter-based model performed the best when evaluating the models individually.

2.5 Zero-Shot and Few-Shot Text Classification

The idea of Zero-Shot Text Classification was first introduced by the paradigm of *Dataless Classification* by Chang et al. [2008]. In the first approach, the method of Explicit Semantic Analysis Gabrilovich et al. [2007] is used, which maps the labels and content into a single high-dimensional space of concepts. The label with the highest matching score out of all the labels is picked. Chang et al. [2008] found that the crucial part of this process is the representation of the labels and the representation of learning the text. Both label and text need to be processed and assessed with equal importance, as shown in further work down the line. With the introduction of semantic embeddings like word embeddings, [Mikolov et al., 2010] a new form of representation of text got introduced and was adapted to Zero-Shot Learning [Sappadla et al., 2016]. With the introduction of transformer-based models, Yin et al. [2019] proposed Zero-Shot Learning as an Entailment approach by reformulating the text classification into a Natural Language Inference (NLI) task and thus converting it into a fine-tuning task. In an NLI task, a model has to

predict if a given premise either entails, contradicts, or acts neutral to a proposed hypothesis. Yin et al. [2019] used this structure to classify topics by using the text to be classified as the premise and creating hypotheses that contain information about the different topics to be classified. A hypothesis is formulated in the following way: *"This text is about {topic}"*. For each topic, the model now generates a probability of the hypothesis corresponding to the premise. They assume that the given text is about the subject of the hypothesis if the model successfully predicts entailment for a particular hypothesis and vice versa if a model predicts non-entailment.
When looking at Few-Shot Learning, the goal behind the approach is to determine what improvements can be achieved by only training a model on a limited amount of training data. This idea was first introduced to the field of image processing, where models tried to learn simple visual concepts through one example, also called One-Shot learning [Lake et al., 2011]. This idea was transferred into the field of NLP processing where huge language models like BERT [Devlin et al., 2018] are fine-tuned with a Masked Language Model (MLM) objective, see Chapter 4, which is based on the idea of a Cloze task [Taylor, 1953]. Several works have reformulated Few-Shot Learning tasks as cloze questions, reusing previously trained LMs. Schick and Schütze [2020] show that presenting their algorithm PET, even giving a pre-trained model a few annotated training examples, can lead to nontrivial performance for several tasks and languages. [Wang et al., 2021] also found that the reformulation into an entailment task increases the performance over many different NLP tasks when applying Few-Shot Learning. They systematically evaluated 18 different tasks in which offensive language detection was included. Other notable approaches of Few-Shot Learning in text classification are prompting [Liu et al., 2021], and task descriptions [Raffel et al., 2020].

3 Data and Task Descriptions

3.1 Data Collection

The dataset provided by the GermEval21 [Risch et al., 2021] consists of more than 4000 Facebook comments manually annotated by different annotators, mainly student assistants from the Heinrich Heine University in Düsseldorf. The comments used in the dataset are from the Facebook page of a German political talk show of a national public television broadcaster. They contain political topics discussed on the talk show, the comments of TV presenters, and political standpoints. The comments are split into a training and a test set, leading to more than 3000 comments in the training set and a test set of about 1000 comments. Another significant difference was the time frame these comments were collected in. The training set was collected during a span from January to July 2019 and the test set from September to December 2020 as this constitutes a realistic use-case according to [Risch et al., 2021]. They further state that they consider the phenomenon of topic bias and person bias Wiegand et al. [2019] as unlikely as 157 active users are commenting on 141 discussion threads. As for the annotation guidelines, they used a theory-based annotation scheme which allows them to define fine-grained forms of all the three sub-tasks (Toxic, Engaging, Fact Claiming). These detailed forms are explained in the upcoming subsections regarding each sub-task. The annotation for these three coarse-grained categories is the only part that has been made public of the dataset. It is also important to note that some comments are labeled as belonging to more than one category, for example, a comment can be engaging and fact-claiming simultaneously.

3.1.1 Toxic Comments

As one of the arguably most culturally significant technological developments, online platforms such as Facebook, Instagram, and Twitter have reached immense popularity in the last decade. The platforms allow thousands of users to communicate and distribute content but also have become a place of disturbing toxic communi-

Figure 1: Comment Examples

cation, such as cyberbullying and harassment [Sheth et al., 2022]. The use of such comments decreases the quality perception of participants and observers and may trigger stereotypical thinking, hateful commenting behavior, or withdrawal from the debate. [Hsueh et al., 2015], [Prochazka et al., 2018], [Ziegele et al., 2018]. Toxic comments can be further specified into different subcategories defined by Wilms [2021] as listed in Table 1. If a comment falls into the scope of proposed categories, it will be labeled as TOXIC by the annotators and NOT TOXIC otherwise. An example of a toxic and non-toxic comment can be found in Figure 1.

	train		test	
subcategory	frequency	percentage	frequency	percentage
screaming	163	5.0	101	9.2
vulgar language	190	5.8	37	3.4
insults	25	6.3	79	7.2
sarcasm	419	12.9	295	27.0
discrimination	104	3.2	145	13.3
discrediting	360	11.0	26	2.4
accusation of lying	136	4.1	75	6.9
total	1122	34.5	504	46.2

Table 1: Overview Toxic Comments shows the number of comments and percentages of each subcategory in both train an test set

3.1.2 Engaging Comments

Approaches such as the Online Deliberation Theory [Friess and Eilders, 2015] assume that engaging comments, with subcategories like rational, respectful, and reciprocal comments lead to a balanced, constructive and peaceful environment [Stroud et al., 2015]. This has also sparked the attention of moderators and community managers who are interested in knowing which users are producing valuable content such that these users get more publicity by highlighting said comments [Risch and Krestel, 2020]. As the definition of engagement has quite a wide scope which comments can fall into, there are also quite some subcategories to inspect further. Table 2 gives an overview of all categories and to get a more practical insight one can refer to Figure 1.

	train		test	
subcategory	frequency	percentage	frequency	percentage
argument	506	15.5	197	18.0
additional information	184	5.6	37	3.4
personal experience	125	3.8	25	2.3
solution proposal	89	2.7	58	5.3
empathy	31	0.9	10	0.9
mutual respect	59	1.7	24	2.2
polite salutation	30	0.9	11	1.0
total	865	26.6	293	26.8

Table 2: Overview Engaging Comments shows the number of comments and percentages of each subcategory in both train an test set

3.1.3 Fact Claiming Comments

As several indicators suggest, the number of social network users keeps rising, with its members using social networks as a significant source of information and news. This leads to much information being distributed rapidly but barely regulated regarding misinformation. A report published by the World Economic Forum, as quoted by Tong et al. [2018], even states that spreading misinformation is one of the top global economic risks. The growing spread of misinformation results in demand for systems that allow the identification of social media contributions that should be fact-checked manually. In the dataset, all comments containing any assertion of facts are considered fact-claiming comments [Risch et al., 2021]. Into this category also falls the use of external sources cited in a comment. To see the exact distribution

of subcategories refer to Table 3 and so find a example of a fact-claiming comment refer to Figure 1. Here it is also essential to distinguish fact-claiming from fake news detection or fact-checking, as the latter part is done manually or semi-automatically after the classification.

	train		test	
subcategory	frequency	percentage	frequency	percentage
assertion of facts	1013	31.2	343	31.4
provision of evidence	184	5.6	37	3.4
total	1103	34.0	353	32.3

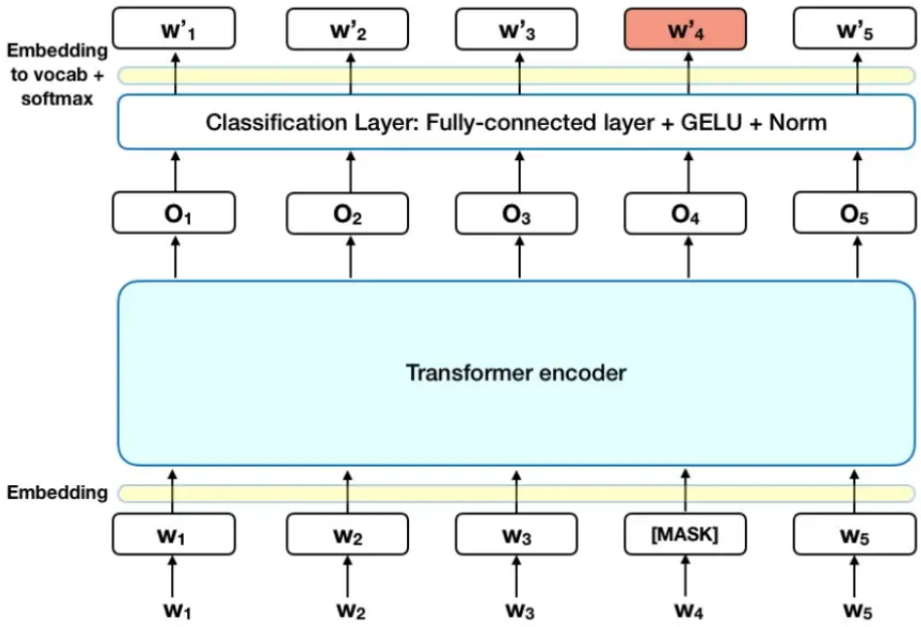
Table 3: Overview Fact Claiming Comment shows the number of comments and percentages of each subcategory in both train an test set

4 Zero-Shot

4.1 Evaluating Zero-Shot Prediction

In this thesis, three goals are formulated to define the evaluation process of the Zero-Shot method. For each task, the goal is (1) to analyze the performance of Zero-Shot Learning in a most basic setting of a single hypothesis and (2) to compare the performances of different hypothesis templates and candidate labels. Lastly, a further goal is (3) to develop evaluation strategies based on the Codebook of the annotation for the Shared Task GermEval21 [Wilms, 2021]. This process of proposing strategies is known as Hypothesis Engineering [Goldzycher and Schneider, 2022].

BERT and NLI In the last couple of years, the field of Machine Learning has experienced a considerable increase in attention, especially in the field of NLP. A significant impact on this increase has been the introduction of the Transformer architecture [Vaswani et al., 2017] and its usage as Bidirectional Encoder Representation for Transformers [Devlin et al., 2018]. It has achieved state-of-the-art performance on various NLP tasks such as Natural Language Inference, Question Answering, and many more. This paragraph should work as a short introduction to better understand the experiment setup and give insight into the models used in this thesis. On a high level, BERT is a model which allows predictions of the following word/sentence in a sequence based on a mechanism called attention which tries to learn contextual relations between words from a large amount of text data. It achieves this by reading an entire sequence of text as a whole, compared to other approaches, which read text sequentially. With this approach, the model can learn the context of a word based on the words surrounding it and thus tries to get a general understanding of how natural language is built up. BERT gets a more profound sense of language context by processing the text bi-directionally and being trained in this fashion. The model is built up of two main parts. On one side is an encoder (see Figure 2) that reads the input text and a decoder that allows the prediction of upcoming words in a sequence. The input is text, a series of tokens



w_1 - w_5 represent the input that gets fed into the endcoder, while o_1 - o_5 represents the output of the encoder which gets translated into textual form to create the predictions w'_1- w'_5

Figure 2: Bidirectional Encoder Representation for Transformers

embedded into vectors before being processed by the neural network. Here, the mechanism of attention comes into play. By creating the embedding of the input word, the model looks at the important parts based on relevance to the other words in the sequence. This information about the crucial parts is stored in multiple different attention layers. These layers try to learn more about the input, such as discovering syntactic or semantic information. For example, in the sentence "She is eating a green apple", the model learns to get high attention between the words "eating" and "apple" but low attention between "eating" and "green". For a more detailed description of attention, refer to Vaswani et al. [2017]. The decoder takes the context vector created by the encoder and turns the vector back into textual form. Two different approaches are used for training these models. The technique of Masked Language Modelling(MLM) feeds a unique token into the input sequence, which is [MASK] instead of the original word. When processing this token, the model predicts the masked word by its context, the unmasked words. The probability of a word in the vocabulary is calculated, and the embedding matrix is updated based on a classification layer's predictions. This is a relatively slow process that needs lots of training data. In a larger text predicting coherence between sentences

is a problem many language models struggle with. BERT works against this by predicting a further token for the Next Sentence Prediction. A token [CLS] is inserted at the beginning and a different unique token [SEP] at the end of each sentence. The model then predicts the connectedness of these two sentences with these tokens under the assumption that random sentences will be disconnected from the original sentence in contrast to the following one. These two mechanisms allow the model to create an understanding of natural language and its specific context.

The task of Natural Language Inference was proposed to give an example to test if a model can infer information from a text and use this information to make a statement if a different text also contains it. In the case of BERT, an already pre-trained model predicts, when given a premise, if the following hypothesis contradicts, entails, or is neutral towards said premise. For example: If a given premise: *"Ich bin äh, Chief Master Sergeant, im Ruhestand, wie Rick sagte."* and the corresponding hypothesis: *"Rick sagte dir, dass ich im Ruhestand war."* would be labeled as entailment, as the hypothesis entails information mentioned in the premise. In contrast to this, using the same premise, the hypothesis: *"Ich arbeite noch immer, bis heute."* would be labeled as a contradiction.

4.2 Experiment Setup

The experiments in this thesis build on the same core idea and structure present when doing an NLI task. The method proposed to classify hate speech by testing different hypotheses on the premise that represents the text that has to be classified [Wang et al., 2021]. This process of generating a hypothesis has been taken further by Goldzycher and Schneider [2022] by providing an extensive evaluation of different hypotheses. In the experiments, the same method is used and translated into German. The hypotheses are reformulated to predict in the context of each of the three GermEval21 tasks. The hypotheses are built up in the following fashion: *hypothesis = hypothesis template + candidate label"* where the candidate label is a word that was used by the Codebook [Wilms, 2021] to annotate a sentence to a corresponding sub-task. Four different hypothesis templates are used to build grammatically correct sentences and used as hypotheses to be evaluated. A base evaluation is conducted using all these hypotheses to get an overview of how the model evaluates the different options. Section 4.3 explains this process in more detail. All the experiments are conducted with a German Zero-Shot Model, which is a GBERT [Chan et al., 2020] which was fine-tuned on the German part of the XNLI dataset [Conneau et al., 2018] which is a translated version of the Multi-Genre Nat-

ural Language Inference dataset (MNLI) [Williams et al., 2017]. The Huggingface transformers library [Wolf et al., 2020] and its user *Sahajtomar* provide this model called *German_Zeroshot* [1]. The model predicts a probability of entailing information from the premise for each of the hypotheses. To decide if a test is classified as *entailment*, a threshold is specified, and if the probability exceeds this threshold, the text will be classified as such. This procdure is depicted in Figure 3 .This optimal threshold is tested by using a range of different thresholds from 0.1 to 1. To find the best threshold, threshold testing is done on a smaller eval set containing a smaller sample of the training set (size = 0.3 of the test set). Once an optimal threshold is found, the model predicts the probabilities on the test set and is evaluated on different metrics such as accuracy, and f1-score, of which the latter is used by GermEval21 to compare your results.

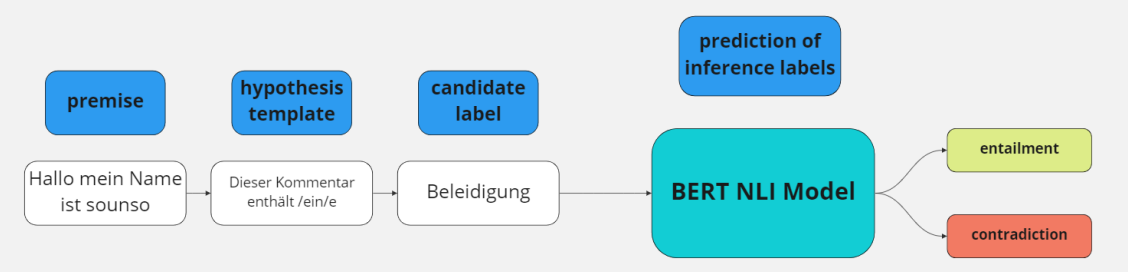


Figure 3: Prediction Pipeline Zero-Shot

4.3 Base Evaluation

As seen in Chapter 3, there are many ways to claim a comment as toxic, engaging, or fact-claiming. Goldzycher and Schneider [2022] have shown that choosing a suboptimal way to express this claim can lead to lower accuracy. To understand the best-performing hypotheses, different candidate labels are used for each sub-task according to the Codebook of the annotation of the dataset. Four hypothesis templates are used for each candidate label, which can be found in Table 4. A hypothesis that is evaluated would look like the following: *"Dieser Kommentar entspricht einer Beleidigung"*. The following paragraphs will show the results of the base evaluation for each sub-task. The full evaluation results are given in Appendix A.

[1] You can find the huggingface model card and playground here: `https://huggingface.co/Sahajtomar/German_Zeroshot`

hypothesis template
Dieser Kommentar ist /ein/e {}
Dieser Kommentar entspricht /eine/r {}
Dieser Kommentar enthält /ein/e {}
In diesem Kommentar findet man /ein/e {}

Table 4: Hypothesis templates

Toxic For this sub-task, the seven different candidate labels of the Codebook are the following: *"Schreien", "Vulgäre Sprache", "Beleidigung", "Zynismus/Sarkasmus", "Diskriminierung", "Diskreditierung" and "Lügenvorwurf"*. As the proposed annotation of "Zynismus/Sarkasmus" proposed two alternatives, the better-performing candidate label is chosen. In Table 5, you can find the results of the best-performing hypothesis for each candidate label. Note that these values are based on our generated eval set and thus cannot be directly compared to the results on the test set. The best performing hypothesis is *"In diesem Kommentar findet man eine Diskriminierung"* achieving an accuracy of 64 % and a f1-Score of 61.2 %. Two things are interesting to note here. Firstly, the differences in performance between the best hypotheses only differ by approximately 3 pp. when comparing this to Goldzycher and Schneider [2022]. However, looking at the full results, the highest difference between hypotheses is found to be 12pp. when looking at the candidate label of *"Schreien"*.
Secondly, when looking at the hypothesis templates, there is not a single one that does not work apart from *"Dieser Kommentar enthält /ein/e* {}*"* which is clearly underperforming.

Engaging Regarding the engagement classification sub-tasks, the Codebook provides us with the most detailed subcategories, namely *"Begründung", "Externer Beleg", "Interner Beleg", "Lösungsvorschlag, "Höfliche Anrede", "Respektbekundung", "Schlichtung" and "Empathie"*. For two of the subcategories, the category's name was changed from a more abstract term to a word that was part of the definition's inspection ("Prüfung"). *"Externer Beleg"* was changed into *"externe Quelle"* and *"Interner Beleg"* was changed to *"persönliche Erfahrung"*. Table 5 shows the results regarding this sub-task. The best performing hypothesis is *"Dieser Kommentar ist eine persönliche Erfahrung"* which reaches a similar f1-score as the Toxic sub-task of 61.7 % but reaches an accuracy of 74.2 %. An increase in accuracy is mainly due to the distribution of the labels in the dataset, which is more skewed compared

task	hypothesis	acc	f1
toxic	Dieser Kommentar ist Zynismus	0.622	0.587
	Dieser Kommentar enspricht einer Beleidigung	0.608	0.589
	Dieser Kommentar ist vulgäre Sprache	0.608	0.581
	Dieser Kommentar entspricht Schreien	0.629	0.560
	Dieser Kommentar ist eine Diskreditierung	0.682	0.518
	In diesem Kommentar findet man Diskriminierung	0.640	0.612
	In diesem Kommentar findet man ein Lügenvorwurf	0.632	0.581
engaging	Dieser Kommentar ist eine Begründung	0.689	0.540
	Dieser Kommentar ist eine Wertanschauung	0.675	0.536
	Dieser Kommentar ist eine persönliche Erfahrung	0.717	0.617
	Dieser Kommentar ist ein Lösunsvorschlag	0.742	0.546
	Dieser Kommentar enthält Empathie	0.650	0.518
	Dieser Kommentar ist eine Respektbekundung	0.750	0.499
	In diesem Kommentar findet man eine höfliche Anrede	0.689	0.474
	Dieser Kommentar enthält eine Schlichtung	0.682	0.500
fact claiming	Dieser Kommentar ist eine externe Quelle	0.632	0.568
	Dieser Kommentar ist ein Wahrheitsanspruch	0.548	0.481
	Dieser Kommentar ist eine Tatsachenbehauptung	0.629	0.537

Table 5: Base Evaluation Results Overview

to the other sub-tasks. When looking at the differences in performance between the different candidate labels, it becomes apparent that a few options do not work either. A difference of 14 pp. shows that choosing the right candidate labels can significantly influence performance. One also has to consider that some of the candidate labels represent part of the dataset and thus are not that well fitted to make a general prediction over the whole dataset. When looking at the hypothesis templates, the template that leads to the most performing results is, in most cases, *"Dieser Kommentar ist /ein/e {}"*.

Fact Claiming As the Codebook provided in-depth annotation schemes for the first two sub-tasks when assessing fact claiming, however there are only two subcategories *"Tatsachenbehauptung" and "Externer Beleg"*. Like in the Engaging subtask, *"Externer Beleg"* was changed into *"externe Quelle"*. To get a broader evaluation of fact claiming, the candidate label of *"Wahrheitsanspruch"* was added as it is mentioned in the inspection of the definition that only assertions of facts which also

have a claim to some truth should be classified as such. When looking at the results in Table 5, we can see that the hypothesis: *"Dieser Kommentar ist eine externe Quelle"* achieves the highest performance with a f1-score of 56.8 % and accuracy of 63.2 %. This is the lowest performance over the three sub-tasks, with a difference of about 5 pp. Interestingly, however, for all of the three best-performing hypotheses, the same hypothesis template is being used. Here we can also see that the difference in performance between hypothesis templates can vary up to 14 pp. in accuracy, which is quite drastic.

Overall Results As mentioned, the base evaluation is based on evaluating the hypotheses on a smaller dataset formed from the training set. To get comparable results of the base evaluation to the submissions of GermEval21, the best-performing hypotheses from each sub-task are evaluated against the whole test set. The event organizers mentioned that the official results test set was a filtered version with a similar distribution to the training and the original test set. In Table 6, you can see the best-performing hypotheses for each sub-task. This approach ranks in the lower quarter of submissions when comparing macro f1-score, but there is still a considerable difference to the median/mean of all submissions.

task	hypothesis	acc	f1
toxic	Dieser Kommentar ist eine Beleidigung	**0.609**	**0.552**
engaging	Dieser Kommentar ist eine persönliche Erfahrung	**0.644**	**0.574**
fact claiming	Dieser Kommentar ist eine externe Quelle	**0.599**	**0.531**

Table 6: Base Evaluation Results

4.4 Hypothesis Engineering

To further improve the base evaluation, the method of Hypothesis Engineering [Goldzycher and Schneider, 2022] is applied. This method proposes that a combination of different hypotheses exploit particular patterns of classifications. They used the definitions of different cases of hate speech to create filtering mechanisms that help detect said cases. An example of this is, for instance, not to classify a sentence which predicted entailment for the hypothesis: *"This contains hate speech"* but a contradiction for *"This text is about black people"* because for a text to be classified as hate-speech a target group has to be mentioned. In the case of this thesis, different strategies are proposed based on the Codebook of Wilms [2021]

and their definition of specific cases of toxicity, engagement, or fact-claiming. For each annotation of a subcategory, the Codebook provides a definition of the variable and a description for the annotators if doubts about different annotations come up. In the following sections, the strategies based on this additional information and corresponding results are presented. If none of the strategies apply to a comment to be classified, the approach of a base strategy that contains the best-performing hypothesis from the base evaluation is used.

4.4.1 Toxic

Screaming [1] To achieve the effect of screaming through text, there are many ways to achieve volume or emphasis. The annotation definition classifies this as the appearance of two or more words where all letters are written in uppercase but excludes standard abbreviations like countries or political parties. Secondly, the use of three or more question marks or exclamation marks is also filtered out and directly classified as TOXIC.

Humorous insults [2] In the subcategory of insults, the Codebook highlights that insults, in contrast to sarcasm, do not have a humorous evaluation. This ensures that only insults and disparaging statements towards persons and groups of persons, subjects, and facts apply. To catch humorous comments the hypothesis: *"Dieser Kommentar ist humoristisch veranlagt"* is used. For this strategy, if the combination of *"Dieser Kommentar ist eine Beleidigung"* and the before mentioned one predicting humor predict entailment, the comment is labeled as NOT TOXIC.

Non humorous sarcasm/cynicism [3] As a counterexample to the strategy before, this strategy tackles sarcasm/cynicism, which does not contain a humoristic motive. Here if the model predicts engagement for sarcasm/cynicism, with the hypothesis: *"Dieser Kommentar ist Zynismus"* but contradiction for humor, the comment is labeled as TOXIC.

Discrimination against target group [4] The last strategy, closely related to one of the strategies also proposed by Goldzycher and Schneider [2022], is described in the introduction of this section called filtering by target. The strategy of only classifying a comment if there is a mention of discrimination against a specific target group was translated into German. To filter for discrimination, the hypothesis *"In diesem Kommentar findet man Diskriminierung"* is used and for comments against

specific target groups: *"Dieser Text ist gegen eine Gruppe gerichtet"*. If both predict entailment, the comment is labeled as TOXIC.

Results In the bottom row of Table 7, you can find the overall performance of the experiment when applying the different strategies above. When comparing this result of the base evaluation, an increase of 6 pp. in the f1-score is noted. This clear result shows that hypothesis engineering can have a positive effect when combining different strategies. When looking at the efficiencies of the different strategies, it becomes apparent that not all strategies work as well as others, especially the two strategies [2], [4] that achieve a lower accuracy than the overall base hypothesis. On a positive note, it has to be mentioned that the introduction of strategies improved the performance of the base hypothesis, which indicates that some examples that belong to a specific subcategory of toxicity have been successfully filtered out. This allows the base strategy to classify the more general examples more accurately than before. One aspect of what this approach seems to be struggling with is that not every strategy filters out all the examples that belong to the corresponding subcategory. As an example of the strategy of screaming, only 28 comments were detected of the 101 annotated. However, this is also hard to control as the fine-grained classification according to subcategories is not published. Thus a kind of error analysis in more depth is not possible.

	detected	true	false	accurracy
Screaming	28	26	2	**0.93**
Non humorous sarcasm/cycnisim	156	70	86	0.44
Humorous insults	25	18	7	**0.72**
Discrimination against target group	157	75	82	0.48
Base Strategy	578	401	177	**0.69**
Without strategies			**f1-score:**	**0.55**
With strategies			**f1-score:**	**0.61**

Table 7: Toxic Strategies Overview shows the number of filtered-out comments of each strategy and its accuracy in predicting it. The bottom two rows state the model's overall performance with or without using strategies.

4.4.2 Engagement

External sources with supported claim [1] To get further information about a topic, external sources allow you to open your scope about a certain topic. As there

is a ton of information out there, to find out if an external source added value to the discussion, it has to come in combination with a supported claim. To be able to filter out meaningless external sources the hypothesis:*"Dieser Kommentar enthält eine Tatsachenbehauptung"* is tested. If the hypothesis predicts entailment on top of the detection of an external source, the comment is labeled as ENGAGING.

Empathy [2] To be able to classify empathy, the Codebook provides signal words like *"verstehen", "nachfühlen", "hineinversetzen", "empfinden"* that indicate empathy being present. This approach is implemented by filtering out comments which contain these signal words.

Ironic Respect [3] Respectful behavior on social media platforms helps create a peaceful environment and gives credit to the authors of the articles. The problem is that many messages seem respectful at first sight but are ironic, thus taking away respectfulness. This strategy tries to capture these false positive expressions of respect by adding the hypothesis of *"Dieser Kommentar enthält Ironie"*. If the hypothesis predicts contradiction, the comment is labeled as NOT ENGAGING.

Polite Salutations [4] The last strategy focuses on the topic of whether polite phrases are used in the commentary. This includes, in particular, formulations of the polite form of address ("Sehr geehrter Herr/Frau; Liebe Herr/Frau") and the delivery of greetings ("Bleiben Sie gesund!; Mit freundlichen Grüßen; Schönes Wochenende"). These formulations are captured in a list of signal words like *"Lieber", "Liebe", "grüßen"* and are labeled as ENGAGING if present.

	detected	true	false	accurracy
External Sources with supported claim	3	2	1	0.67
Empathy	17	6	11	0.35
Ironic Respect	108	89	19	**0.82**
Polite Salutations	6	5	1	**0.83**
Base Strategy	827	552	275	**0.67**
Without strategies			**f1-score:**	**0.57**
With strategies			**f1-score**	**0.58**

Table 8: Engaging Strategies Overview shows the number of filtered-out comments of each strategy and its accuracy in predicting it. The bottom two rows state the model's overall performance with or without using strategies.

Results When looking at the results on the bottom row of Table 8, you can see that the overall improvement on this sub-task is only 1pp., which is less than in the sub-task analyzed before. This can be explained, firstly, by the empathy strategy not working as expected and also that the number of comments filtered out is relatively low so that only 134 out of 994 go through the additional process. This also does not help the base hypothesis, which achieves a lower accuracy when evaluated over the whole test set. On a positive note, the strategy of ironic respect is highly efficient and opens up the question of whether filtering for false positives is a good strategy in hypothesis engineering.

4.4.3 Fact Claiming

Links [1] A widespread way of supporting facts on social media is by posting a link on which a user can access further information about a specific topic. This strategy filters for the appearance of links and classifies them as FACT CLAIMING.

External sources with supported claim [2] As there are not only links regarding external sources, the same strategy as in the previous sub-task of engagement is used. As discussed above, the strategy filters out meaningless external sources with the hypothesis *"Dieser Kommentar enthält eine Tatsachenbehauptung"*. If this hypothesis predicts entailment on top of the detection of an external source, the comment is labeled as FACT CLAIMING.

Statement of fact with truth claim [3] The last strategy is constructed to filter out factual statements that also contain a truth claim. The Codebook highlights that a truth claim has to be present for a factual statement to be annotated as fact claiming. This characteristic ensures that only factual claims are labeled as such, and that simple factual statements do not fall into this category. The hypothesis *"Dieser Kommentar ist ein Wahrheitsanspruch"* is used to filter out truth claims, and *"Dieser Kommentar ist eine Tatsachenbehauptung"* filters out statements containing a fact. If both hypotheses predict entailment, the comment is labeled as FACT CLAIMING.

Results In Table 9, you can find the impact of each of the three strategies on the experiment's performance. Just like in the engagement sub-task, the introduction of combinations of hypotheses brings an increase of f1-score by only 1 pp. Similarly, the three strategies cannot capture a considerable amount of comments, such that

	detected	true	false	accurracy
Links	22	20	2	**0.91**
External Sources with supported claim	44	32	12	**0.73**
Statement of fact with truth claim	42	33	9	**0.79**
Base Strategy	836	499	337	**0.60**
Without strategies			**f1-score:**	**0.53**
With strategies			**f1-score**	**0.54**

Table 9: Fact Claiming Overview shows the number of filtered-out comments of each strategy and its accuracy in predicting it. The bottom two rows state the model's overall performance with or without using strategies.

the strategies are filtering out only 104 comments. A promising aspect of this experiment is that all strategies achieve an accuracy that exceeds the accuracy of the base evaluation. All of the strategies achieve an accuracy of over 70 pp., which opens the path of trying out other different hypotheses that can potentially reach similar performance. The evaluation of the base strategy stayed around the same.

5 Few-Shot

5.1 Methodology

5.1.1 Training Setup

In the previous section, we have seen how we can classify texts without labeled training data using Zero-Shot techniques. This approach goes against the stream of ML models specifically fine-tuned to achieve state-of-the-art on a specific task. In this chapter, I use a technique called Few-Shot learning, where we explore how a small amount of training data affects the proposed structures of hypotheses. I use the same pre-trained GBERT Zero-Shot model but train it on a random subset of 8,16,32,64,128, and 256 examples from the training set of GermEval 21 [Risch et al., 2021]. Training these models is done using the well-known technique of fine-tuning. In this process, a classification layer is added, just like in the Next-Sentence-Prediction of BERT, and a token [CLS] to the sequence to encode the premise and hypothesis information. This token is then evaluated into the output, consisting of three classes: entailment, contradiction, and neutral. Through taking into account training examples, the output of the prediction of this token is learned by adjusting the BERT encoder and the added classification layer based on the output of the pre-trained model.

As we want to stay as close to the Zero-Shot experiments as possible, the training setup is an NLI task formulation based on binary training examples for each subtask. The training examples are used as an NLI task, which is always built on a premise and a hypothesis. For the corresponding hypothesis, we choose the hypotheses that make up the different strategies in the Zero-Shot setting. This allows that all strategies are also incorporated into the training process. NLI tasks are traditionally a ternary text classification with a neutral dimension. This third category is not available in the training data such that positive examples get mapped to entailment, and negative examples get mapped to a contradiction. The distribution of contradictions and entailments in the smaller generated training sets is based on

the distribution of the whole training set.

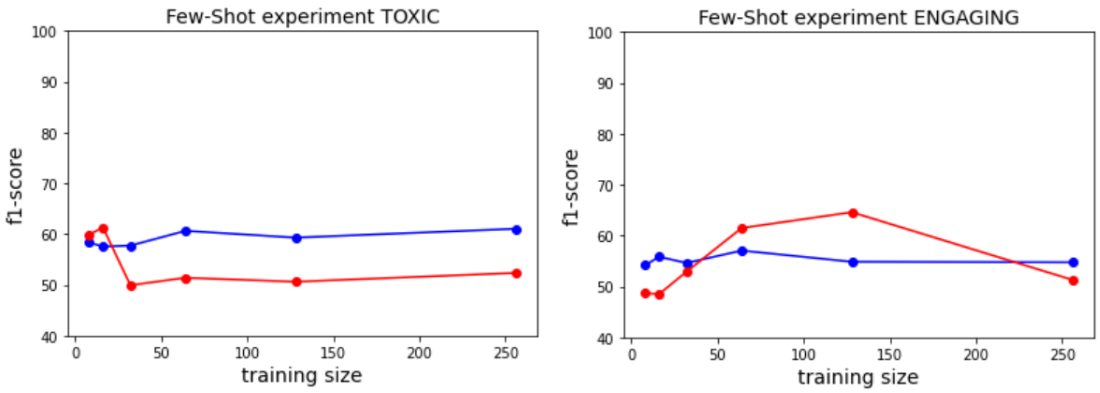
5.1.2 Experiments

Training the models is highly dependent on their components and parameters, so I tried out two different hyper-parameter settings to see how the models respond. In the first setting, I used the default parameters that the huggingface library provides. For each number of examples in the created training sets, we train for three epochs and keep the batch size at 8; hence this is our minimum number of examples in training sets. For optimization, the default optimizer from huggingface AdamW [Loshchilov and Hutter, 2019] is used with a learning rate of $5e^{-5}$ and the standard betas of 0.9 and 0.999. In a second setting, we use the setting proposed in the paper of Goldzycher and Schneider [2022]. Here the learning rate is lowered to $1e^{-5}$, and the number of epochs increases to ten. As fine-tuning the models also depends highly on the training data you provide, the model was trained on three randomly sampled training sets per setting of the number of examples. In the results, I present a mean of the three runs to account for the randomness of the sampling. The evaluation process for the newly created models is the same as for the Zero-Shot models described in the previous chapter. This allows us to compare the two settings and use the insights gained previously. We evaluate the trained models by taking the best-performing hypothesis of the Zero-Shot as the base hypothesis evaluation, displayed in Table 6 and on the proposed strategies. As we have already conducted threshold testing in the Zero-Shot setting, the thresholds to decide if a hypothesis predicts contradiction or entailment are adopted for the base evaluation and the strategies setting. The macro f1-score is used to compare the results to the other submissions of GermEval21.

5.2 Results

In this section, the results of the Few-Shot setting are presented. As for each subtask, the results are different, I will discuss them in separate paragraphs. For an interpretation of the results, please refer to Chapter 6.

Toxic The results are displayed in Figure 4, where we can see that the best result is achieved using 256 examples and a f1-score of 61.03 % averaged over the three runs. This is only a 0.5pp increase over the Zero-Shot setting, which is relatively low. It is interesting to see that the results of the adjusted setting stay quite constant over the

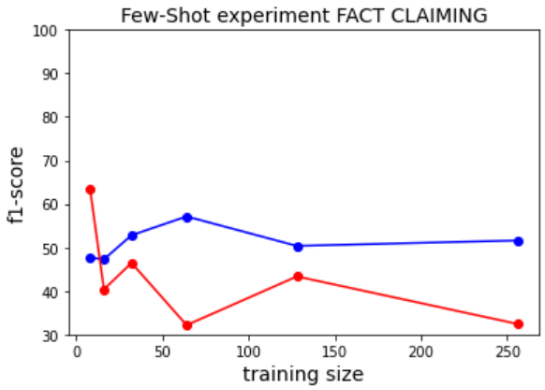


Red line represents the default huggingface setting, while the blue line represents the adjusted setting with a lower learning rate and higher epochs.

Figure 4: Few-Shot Results Toxic & Engaging

different training sizes, and the increase between the lowest and highest performing setting is at 4 pp. Comparing the adjusted setting to the default setting, a few differences can be detected. Firstly the default setting using 16 examples works just as well as the best experiment from the other setting but adding more examples to the model decreases the models' performance massively. After this drop, the performance stays relatively stable, showing similar developments as in the adjusted setting.

Engaging The results for the engagement sub-task are also presented in Figure 4. Here the best-performing experiment reaches a f1-score of 64.6 %, which is a 6 pp. increase over the Zero-Shot setting, which shows that the additional training data improved the model's performance quite strongly. Unlike the toxicity task, the default setting achieves this performance, showing the influence the hyperparameters can have on the performance of the models. Interestingly, within the default setting, the performance increases at the start when seeing up to 128 examples but then drops when given more training examples. This behavior is very different from the toxicity task, which will be part of the discussion section in Chapter 6. When comparing the two training settings, one notices that the adjusted setting is much more stable but achieves a lower performance also when comparing it to the Zero-Shot results. On the other hand, the adjusted setting shows a similar performance development over the different training sizes to the toxicity experiments, which is interesting, as the results of the default setting vary over the two tasks.



Red line represents the default huggingface setting, while the blue line represents the adjusted setting with a lower learning rate and higher epochs.

Figure 5: Few-Shot Results Fact Claiming

Fact Claiming When looking at the results of the Fact-Claiming experiments in Figure 5, the best-performing experiment reaches a f1-score of 63.26 %, which is an increase of 9 pp. over the Zero-Shot setting. This is the highest performance increase over all the sub-tasks and is interesting because it was achieved by using only **eight** examples in training. Interestingly, it is achieved in the default setting, which does not work well for all the other experiments. The lowest performance is nearly half the best-performing model's percentage. This decrease is very unusual as, in the other two sub-tasks, some fluctuation in the default setting up to 15 percent occurred. However, in this experiment, the difference in performance has almost doubled. In the adjusted setting, the best-performing model also improves on the Zero-Shot setting but shows the most significant difference between the different experiments compared to the other sub-tasks. The fluctuation of both settings shows that this is a more challenging task for the models.

6 Discussion

6.1 Result Elaboration

This section compares the results generated by both Zero-Shot and Few-Shot learning. The first step compares them to the results of the Shared Task participants and further assesses the limitations of the model's predictions.

Comparsion to GermEval21 As a first disclaimer, I want to remind the reader again that the participants' submissions are evaluated against a smaller sampled dataset of the test set, which contained a similar distribution of positive and negative examples as in the whole test set. This means that the results are not completely comparable, but as the smaller dataset is not publicly available, it is the best shot at comparing the results. It must also be noted that the results of the submissions in GermEval21 are assumed to represent the state-of-the-art in all three sub-tasks. Firstly when looking at the Zero-Shot results, it becomes apparent that the results presented in this thesis cannot compete with the state-of-the-art of current NLP approaches by reaching 10 pp. less f1-score in the toxicity detection, 11 pp. on engagement detection, and over 18 pp., in fact claiming detection. This overall loss of performance, I assume, is mainly due to the usage of a larger amount of data, as all other submissions used the whole training set to their advantage. Some submissions even used data augmentation methods to increase the number of training examples. When looking at performance from task to task, it is interesting that the results vary enormously. Interestingly, in the GermEval21 competition, the fact-claiming sub-task reached the highest scores, but in my experiment, this sub-task achieved the worst scores. One possible reason for this is that for fact-claiming, there was not a very detailed annotation into different subclasses available, which firstly narrowed down the amount of candidate labels to check and limited the use of different strategies when applying hypothesis engineering. On a positive note, if I compare the result to the only proposed baseline, a majority-class classifier, all models reach more than 16 pp. increase and even up to 23 pp. in the best case. Comparing the Few-Shot results, we can see that in all sub-tasks, an increase in f1-score is reached,

and the Toxic sub-task is only two pp. away from the mean of all submissions of GermEval21. There is still a remarkable difference between the state-of-the-art compared to our models, but in the Few-Shot setting, it has decreased, especially in the Fact-Claiming task, where it has made up 9 pp. compared to the Zero-Shot setting. The difference, in fact claiming, is still the biggest over all tasks, but it stays caught up like before.

Hypothesis Engineering and Error Checking As Goldzycher and Schneider [2022] have found in their work, the performance of NLI strategies largely depends on the accuracy of the supporting hypotheses. The problem with this is that checking the accuracy of each supporting hypothesis is not possible, as the organizers of GermEval21 do not provide fine-grained annotations of the dataset into subcategories. The only way to get an estimate is by evaluating the comments which did get classified by a supporting hypothesis, but this does not necessarily mean that this subcategory has been annotated as such. This process works as a usable estimator, but it is also visible that this process only sometimes works as planned, as seen in the Tables of Section 4.4. I also assume that the disability to make error checks in an automated way also leads to a decrease in performance. Another problem is that only looking at accuracy does not tell everything about the effectiveness of a supporting hypothesis. A hypothesis only works well if it filters out the corresponding comments in the dataset. Proposing different metrics to rate the effectiveness could be an exciting task for future work.

Fine-tuning models When looking at the results of the Few-Shot setting, it is hard to identify a trend along the different experiments conducted. This leads to the conclusion that changing the hyper-parameters, like the learning rate or the number of trained epochs, can considerably influence the predictions of the models. For each of the trained models, the probabilities of the predictions change, but the threshold, if an example gets classified as positive/negative, stays the same. In the base evaluation of the Zero-Shot setting, the testing for different thresholds was only factored by the number of computationally feasible hypotheses. In the Few-shot setting, the factors of finding an optimal threshold are not only the hypotheses but also the strategy setup, number of training examples, and hyper-parameters of the trained model. These changes can lead to different performances, making the optimization process computationally costly. The good performances overall sub-tasks suggest that if an optimal setup is found, a high performance can be achieved. Here it also has to be mentioned that this process is susceptible to the problem of overfitting, which has to be taken into account. One trend that has to be noted is

that in the default setting of huggingface, the model performs worse when confronted with a higher number of training data. It gets to a situation where it is stuck as it predicts similar probabilities for entailment and contradiction, most likely due to reaching a local minimum in optimizing the loss during training.

7 Conclusion

In this thesis, I have shown the capabilities of Zero-Shot and Few-Shot Text Classification on German social media comments. The comments to be classified were part of the Shared Task GermEval21 [Risch et al., 2021], which focused on three sub-tasks Toxic, Engaging, and Fact Claiming Comment detection. In my work, I use a transformer-based approach that reformulates the Shared Task into a Natural Language Inference setting and uses different hypotheses based on the annotation scheme, which has been presented by the organizers, to get my predictions. When looking at my first research question, the performance of the Zero-Shot model is not comparable to the models developed for the Shared Task reaching from 10 to 18 pp. less in f1-score. To answer the second research question, I develop simple strategies by combining hypotheses to catch different variants of toxicity, engagement, and fact-claiming and thus improve performance. Using strategies leads to an increase in performance of up to 6pp. in f1-score on the toxicity detection sub-task but only achieves minor improvements on the other two tasks. As Zero-Shot Learning works without additional labeled data, I wanted to find out the influence of adding a limited amount of training data, so-called Few-Shot learning. Here an increase in performance over all three sub-tasks can be noted, achieving an f1-score of 61.0%, 64.4%, and 63.3% on Toxic, Engaging, and Fact Claiming detection, which answers the third research question. In future work, to maximize the effectiveness of Hypothesis Engineering, I would want to implement an effective way to conduct error checking, as this would allow us to get more information about the supporting hypotheses and Goldzycher and Schneider [2022] have shown benefits of a process like this in their work. Further, it would be interesting to see how these models would prevail in a multilingual setting on different sentiment analysis tasks. This thesis has allowed me to familiarize myself with transformer-based architectures. I got to understand popular frameworks like huggingface and PyTorch, which will help me in my future work as a computational linguist. As the Few-Shot setting was run on the Google Cloud Console, I also learned how to manage the setup of cloud services which took up much of my time. Lastly, I learned the importance of tracking and visualizing my results on the platform *Weights and Biases*, which helped a lot in overviewing my models' results.

References

A. Barrón-Cedeño, T. Elsayed, P. Nakov, G. D. S. Martino, M. Hasanain, R. Suwaileh, F. Haouari, N. Babulkov, B. Hamdan, A. Nikolov, S. Shaar, and Z. S. Ali. Overview of checkthat 2020: Automatic identification and verification of claims in social media. In *Conference and Labs of the Evaluation Forum*, 2020.

B. Boenninghoff, S. Hessler, D. Kolossa, and R. M. Nickel. Explainable authorship verification in social media via attention-based similarity learning. In *2019 IEEE International Conference on Big Data (Big Data)*, pages 36–45. IEEE, 2019.

T. Bornheim, N. Grieger, and S. Bialonski. Fhac at germeval 2021: Identifying german toxic, engaging, and fact-claiming comments with ensemble learning. *arXiv preprint arXiv:2109.03094*, 2021.

B. Chan, S. Schweter, and T. Möller. German's next language model. *arXiv preprint arXiv:2010.10906*, 2020.

M.-W. Chang, L.-A. Ratinov, D. Roth, and V. Srikumar. Importance of semantic representation: Dataless classification. In *Aaai*, volume 2, pages 830–835, 2008.

K. Clark, M.-T. Luong, Q. V. Le, and C. D. Manning. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*, 2020.

A. Conneau, G. Lample, R. Rinott, A. Williams, S. R. Bowman, H. Schwenk, and V. Stoyanov. Xnli: Evaluating cross-lingual sentence representations. *arXiv preprint arXiv:1809.05053*, 2018.

C. Demus, J. Pitz, M. Schütz, N. Probol, M. Siegel, and D. Labudde. Detox: A comprehensive dataset for german offensive language and conversation analysis. In *Proceedings of the 6th Workshop on Online Abuse and Harms (WOAH 2022), Association for Computational Linguistics, Online*, pages 54–61, 2022.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

D. Friess and C. Eilders. A systematic review of online deliberation research. *Policy & Internet*, 7(3):319–339, 2015.

E. Gabrilovich, S. Markovitch, et al. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *IJcAI*, volume 7, pages 1606–1611, 2007.

J. Goldzycher and G. Schneider. Hypothesis engineering for zero-shot hate speech detection. *arXiv preprint arXiv:2210.00910*, 2022.

N. Hassan, F. Arslan, C. Li, and M. Tremayne. Toward automated fact-checking: Detecting check-worthy factual claims by claimbuster. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1803–1812, 2017.

N. Hildebrandt, B. Boenninghoff, D. Orth, and C. Schymura. Data science kitchen at germeval 2021: A fine selection of hand-picked features, delivered fresh from the oven. *arXiv preprint arXiv:2109.02383*, 2021.

M. Hsueh, K. Yogeeswaran, and S. Malinen. "leave your comment below": Can biased online comments influence our own prejudicial attitudes and behaviors? *Human communication research*, 41(4):557–576, 2015.

B. Lake, R. Salakhutdinov, J. Gross, and J. Tenenbaum. One shot learning of simple visual concepts. In *Proceedings of the annual meeting of the cognitive science society*, volume 33, 2011.

P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *arXiv preprint arXiv:2107.13586*, 2021.

I. Loshchilov and F. Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL `https://openreview.net/forum?id=Bkg6RiCqY7`.

S. Malmasi and M. Zampieri. Challenges in discriminating profanity from hate speech. *Journal of Experimental & Theoretical Artificial Intelligence*, 30(2): 187–202, 2018.

T. Mandl, S. Modha, P. Majumder, D. Patel, M. Dave, C. Mandlia, and A. Patel. Overview of the hasoc track at fire 2019: Hate speech and offensive content identification in indo-european languages. In *Proceedings of the 11th Forum for Information Retrieval Evaluation*, FIRE '19, page 14–17, New York, NY, USA,

2019. Association for Computing Machinery. ISBN 9781450377508. doi: 10.1145/3368567.3368584. URL https://doi.org/10.1145/3368567.3368584.

T. Mikolov, M. Karafiát, L. Burget, J. Cernockỳ, and S. Khudanpur. Recurrent neural network based language model. In *Interspeech*, volume 2, pages 1045–1048. Makuhari, 2010.

C. Napoles, J. Tetreault, A. Pappu, E. Rosato, and B. Provenzale. Finding good conversations online: The yahoo news annotated comments corpus. In *Proceedings of the 11th linguistic annotation workshop*, pages 13–23, 2017.

F. Prochazka, P. Weber, and W. Schweiger. Effects of civility and reasoning in user comments on perceived journalistic quality. *Journalism studies*, 19(1): 62–78, 2018.

C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67, 2020.

J. Risch and R. Krestel. Top comment or flop comment? predicting and explaining user engagement in online news discussions. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 579–589, 2020.

J. Risch, A. Stoll, L. Wilms, and M. Wiegand, editors. *Proceedings of the GermEval 2021 Shared Task on the Identification of Toxic, Engaging, and Fact-Claiming Comments*, Duesseldorf, Germany, Sept. 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.germeval-1.0.

S. Rosenthal, P. Atanasova, G. Karadzhov, M. Zampieri, and P. Nakov. A large-scale semi-supervised dataset for offensive language identification. *arXiv preprint arXiv:2004.14454*, 2020.

P. V. Sappadla, J. Nam, E. L. Mencía, and J. Fürnkranz. Using semantic similarity for multi-label zero-shot classification of text documents. In *ESANN*, 2016.

T. Schick and H. Schütze. Exploiting cloze questions for few shot text classification and natural language inference, 2020. URL https://arxiv.org/abs/2001.07676.

A. Sheth, V. L. Shalin, and U. Kursuncu. Defining and detecting toxicity on social media: context and knowledge are key. *Neurocomputing*, 490:312–318, 2022. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2021.11.095. URL

https:
//www.sciencedirect.com/science/article/pii/S0925231221018087.

N. J. Stroud, J. M. Scacco, A. Muddiman, and A. L. Curry. Changing deliberative norms on news organizations' facebook sites. *Journal of Computer-Mediated Communication*, 20(2):188–203, 2015.

J. M. Struß, M. Siegel, J. Ruppenhofer, M. Wiegand, M. Klenner, et al. Overview of germeval task 2, 2019 shared task on the identification of offensive language. 2019.

W. L. Taylor. "cloze procedure": A new tool for measuring readability. *Journalism quarterly*, 30(4):415–433, 1953.

A. Tong, D.-Z. Du, and W. Wu. On misinformation containment in online social networks. *Advances in neural information processing systems*, 31, 2018.

H. N. Tran and U. Kruschwitz. ur-iw-hnt at germeval 2021: An ensembling strategy with multiple bert models. *arXiv preprint arXiv:2110.02042*, 2021.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

S. Wang, H. Fang, M. Khabsa, H. Mao, and H. Ma. Entailment as few-shot learner. *arXiv preprint arXiv:2104.14690*, 2021.

M. Wiegand, M. Siegel, and J. Ruppenhofer. Overview of the germeval 2018 shared task on the identification of offensive language. 2018.

M. Wiegand, J. Ruppenhofer, and T. Kleinbauer. Detection of Abusive Language: the Problem of Biased Datasets. 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 602–608, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1060. URL https://aclanthology.org/N19-1060.

A. Williams, N. Nangia, and S. R. Bowman. A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv:1704.05426*, 2017.

L. Wilms. Annotation guidelines for germeval 2021 shared task on the identification of toxic, engaging, and fact-claiming comments. excerpt of an

unpublished codebook of the dedis research group at heinrich-heine-university düsseldorf (full version available on request). 2021.

T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, and A. Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, Oct. 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-demos.6. URL `https://aclanthology.org/2020.emnlp-demos.6`.

W. Yin, J. Hay, and D. Roth. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. *arXiv preprint arXiv:1909.00161*, 2019.

M. Zampieri, S. Malmasi, P. Nakov, S. Rosenthal, N. Farra, and R. Kumar. Predicting the type and target of offensive posts in social media. *arXiv preprint arXiv:1902.09666*, 2019.

M. Ziegele, M. Weber, O. Quiring, and T. Breiner. The dynamics of online news discussions: Effects of news articles and reader comments on users' involvement, willingness to participate, and the civility of their contributions. *Information, Communication & Society*, 21(10):1419–1435, 2018.

Lebenslauf

Persönliche Angaben

Moritz Preisig
Schaffhauserstrasse 226
8057 Zürich
moritz.preisig@uzh.ch

Schulbildung

2019 - jetzt	Bachelor-Studium Computerlinguistik und Sprachtechnologie an der Universität Zürich

Berufliche und nebenberufliche Tätigkeiten

2020	Tutorat PfL
2020 - jetzt	Vorstand Fachverein Computerlinguistik

A Fine-grained Zero-Shot Results

hypothesis	f1-score	accuracy
Dieser Kommentar enthält Zynismus	0.548	0.572
Dieser Kommentar ist Zynismus	0.587	0.622
Dieser Kommentar enspricht Zynismus	0.562	0.601
In diesem Komentar findet man Zynismus	0.528	0.530
Dieser Kommentar enthält eine Beleidigung gegen eine Person	0.563	0.590
Dieser Kommentar ist eine Beleidiung gegen eine Person	0.573	0.594
Dieser Kommentar entspricht einer Beleidigung gegen eine Person	0.589	0.608
In diesem Kommentar findet man eine Beleidigung gegen eine Person	0.563	0.587
Dieser Kommentar enthält vulgäre Sprache	0.563	0.576
Dieser Kommentar ist vulgäre Sprache	0.581	0.608
Dieser Kommentar entspricht vulgärer Sprache	0.550	0.580
In diesem Kommentar finder man vulgärer Sprache	0.560	0.583
Dieser Kommentar enthält Schreien	0.547	0.587
Dieser Kommentar ist Schreien	0.545	0.625
Dieser Kommentar entspricht Schreien	0.560	0.629
In diesem Kommentar findet man Schreien	0.502	0.509
Dieser Kommentar enthält Diskriminierung	0.595	0.647
Dieser Kommentar ist Diskriminierung	0.584	0.587
Dieser Kommentar entspricht Diskriminierung	0.570	0.580
In diesem Kommentar findet man Diskriminierung	0.612	0.640
Dieser Kommentar enthält einen Lügenvorwurf	0.586	0.597
Dieser Kommentar ist ein Lügernvorwurf	0.572	0.594
Dieser Kommentar entspricht einem Lügenvorwurf	0.570	0.608
In diesem Kommentar findet man ein Lügenvorwurf	0.581	0.633

Table 10: Single Hypothesis Toxic Evaulation

hypothesis	f1-score	accuracy
Dieser Kommentar enthält eine Begründung	0.495	0.597
Dieser Kommentar ist eine Begründung	0.540	0.689
Dieser Kommentar enspricht einer Begründung	0.501	0.686
In diesem Komentar findet man eine Begründung	0.513	0.657
Dieser Kommentar enthält eine Wertanschauung	0.413	0.420
Dieser Kommentar ist eine Wertanschauung	0.536	0.675
Dieser Kommentar entspricht einer Wertanschauung	0.507	0.675
In diesem Kommentar findet man eine Wertanschauung	0.400	0.403
Dieser Kommentar enthält eine persönliche Erfahrung	0.547	0.633
Dieser Kommentar ist eine persönliche Erfahrung	0.617	0.717
Dieser Kommentar entspricht einer persönliche Erfahrung	0.596	0.707
In diesem Kommentar finder man eine persönliche Erfahrung	0.551	0.749
Dieser Kommentar enthält ein Lösungsvorschlag	0.548	0.686
Dieser Kommentar ist ein Lösungsvorschlag	0.546	0.742
Dieser Kommentar entspricht einem Lösungsvorschlag	0.540	0.696
In diesem Kommentar findet man ein Lösungsvorschlag	0.500	0.735
Dieser Kommentar enthält Empathie	0.518	0.650
Dieser Kommentar ist Empathie	0.541	0.604
Dieser Kommentar entspricht Empathie	0.537	0.615
In diesem Kommentar findet man Empathie	0.510	0.600
Dieser Kommentar enthält eine Respektbekundung	0.496	0.714
Dieser Kommentar ist eine Respektbekundung	0.499	0.749
Dieser Kommentar entspricht einer Respektbekundung	0.491	0.693
In diesem Kommentar findet man eine Respektbekundung	0.479	0.671
Dieser Kommentar enthält eine höfliche Anrede	0.466	0.675
Dieser Kommentar ist eine höfliche Anrede	0.467	0.650
Dieser Kommentar entspricht einer höflichen Anrede	0.472	0.686
In diesem Kommentar findet man eine höflche Anrede	0.474	0.689
Dieser Kommentar enthält eine Schlichtung	0.500	0.682
Dieser Kommentar ist eine Schlichtung	0.512	0.590
Dieser Kommentar entspricht einer Schlichtung	0.499	0.629
In diesem Kommentar findet man eine Schlichtung	0.516	0.654

Table 11: Single Hypothesis Engaging Evaluation

hypothesis	f1-score	accuracy
Dieser Kommentar enthält eine externe Quelle	0.563	0.622
Dieser Kommentar ist eine externe Quelle	0.568	0.633
Dieser Kommentar enspricht einer externer Quelle	0.510	0.583
In diesem Komentar findet man eine externe Quelle	0.530	0.654
Dieser Kommentar enthält einen Wahrheitsansspruch	0.469	0.583
Dieser Kommentar ist eine Wahrheitsanspruch	0.482	0.548
Dieser Kommentar entspricht einem Wahrheitsanspruch	0.461	0.484
In diesem Kommentar findet man ein Wahrheitsanspruch	0.458	0.463
Dieser Kommentar enthält eine Tatsachenbehauptung	0.511	0.572
Dieser Kommentar ist eine Tatsachenbehauptung	0.483	0.590
Dieser Kommentar entspricht einer Tatsachenbehauptung	0.469	0.484
In diesem Kommentar findet man eine Tatsachenbehauptung	0.537	0.629

Table 12: Single Hypothesis Fact Claiming Evaluation

B List of packages and framework

This appendix contains a list of things I used for my work.

- packages
  - `transformers`
  - `datasets`
  - `evaluate`
  - `torch`
  - `numpy`
  - `pandas`
- frameworks
  - `huggingface :)`
  - `google colab`
  - `google cloud compute`
  - `weights and biases`