Informatics

COVID-19 and Populism in Austrian News User Comments - A Machine Learning Approach

DIPLOMARBEIT

zur Erlangung des akademischen Grades

Diplom-Ingenieur

im Rahmen des Studiums

Data Science

eingereicht von

Ahmadou Wagne, B.A.
Matrikelnummer 12002293

an der Fakultät für Informatik

der Technischen Universität Wien

Betreuung: Univ.Ass.in Mag.a rer.nat. Dr.in techn. Julia Neidhardt
Mitwirkung: Projektass. Dipl.-Ing. Thomas Elmar Kolb, BSc

Wien, 16. März 2023

Ahmadou Wagne

Julia Neidhardt

Technische Universität Wien
A-1040 Wien ▪ Karlsplatz 13 ▪ Tel. +43-1-58801-0 ▪ www.tuwien.at

Informatics

COVID-19 and Populism in Austrian News User Comments - A Machine Learning Approach

DIPLOMA THESIS

submitted in partial fulfillment of the requirements for the degree of

Diplom-Ingenieur

in

Data Science

by

Ahmadou Wagne, B.A.
Registration Number 12002293

to the Faculty of Informatics

at the TU Wien

Advisor: Univ.Ass.in Mag.a rer.nat. Dr.in techn. Julia Neidhardt
Assistance: Projektass. Dipl.-Ing. Thomas Elmar Kolb, BSc

Vienna, 16th March, 2023

Ahmadou Wagne

Julia Neidhardt

Erklärung zur Verfassung der Arbeit

Ahmadou Wagne, B.A.

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst habe, dass ich die verwendeten Quellen und Hilfsmittel vollständig angegeben habe und dass ich die Stellen der Arbeit – einschließlich Tabellen, Karten und Abbildungen –, die anderen Werken oder dem Internet im Wortlaut oder dem Sinn nach entnommen sind, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht habe.

Wien, 16. März 2023

Ahmadou Wagne

Danksagung

Zuallererst möchte ich der Person danken, die mich immer unterstützt hat und es mir ermöglicht hat eine akademische Ausbildung zu verfolgen. Das war für mich zu keinem einzigen Zeitpunkt selbstverständlich. Danke Mama!

Ich möchte zudem meine Dankbarkeit an die Betreuer meiner Arbeit, Julia Neidhardt und Thomas Kolb, für ihre kontinuierliche Unterstützung und ihr Feedback im Zuge dieser Arbeit ausdrücken. Ich habe die Arbeit im RecSys Lab sehr geschätzt und es war eine sehr aufschlussreiche und spannende Erfahrung. Mein Dank gilt auch den verbleibenden Mitgliedern des Labs, für deren Beitrag und Feedback zu dieser Thesis.

Die Ergebnisse dieser Arbeit wären ohne die Unterstützung meiner Familie und Freunde nicht zustande gekommen. Oscar, Anthony und Jakob, viele Dank für eure Zeit und Mühen. Eure Mithilfe hat mir sehr viel bedeutet.

Zu guter Letzt möchte ich dir, Lena, meinen Dank aussprechen. Deine aufbauende Unterstützung und Motivation haben mich immer vorangebracht und dafür gesorgt, dass ich meine Ziele nie aus den Augen verloren habe. Deine Geduld mit mir hat mir sehr dabei geholfen, diese intensive Zeit durchzustehen, ohne jemals daran zu denken, aufzugeben.

Acknowledgements

First and foremost, I want to thank the person who always backed me up and gave me the opportunity to follow an academic path. Not one day passed where I took that for granted. Thank you, Mum!

I also want to express my gratitude to my supervisors, Julia Neidhardt and Thomas Kolb, for their continuous support and feedback throughout the process of this work. I really valued working within the RecSys lab as it was an insightful and exciting opportunity. Furthermore, I want to thank the remaining members of the lab for their input and feedback on my work.

The results of this thesis could not have been obtained without the support of my family and friends. Oscar, Anthony, and Jakob, thank you very much for your time and effort. Your contributions meant a lot to me.

Last but certainly not least, I want to thank you, Lena. Your uplifting support and motivation always kept me going and ensured that I never lost focus on my goals. Your patience with me helped me get through this intense time without ever thinking about giving up.

Kurzfassung

Die COVID-19-Pandemie und die daraus resultierenden Regierungsmaßnahmen zur Eindämmung des Virus haben in Österreich eine Welle von Protesten und Demonstrationen ausgelöst. Einige Demonstranten griffen dabei auf populistische Rhetorik zurück, um ihren Unmut auszudrücken, was in einigen Fällen zu antidemokratischen Tendenzen führte. Populistische Aussagen waren nicht auf die Öffentlichkeit beschränkt, sondern entstanden vielmehr aus sozialen Medien und anderen Online-Plattformen wie Foren von Nachrichtenseiten. Daher besteht ein Bedarf an automatisierten Methoden zur Erkennung populistischer Aussagen in diesen Texten.

Während bisherige Forschung zu Populismus sich auf Politiker konzentrierte, haben jüngste Studien die Rolle von Bürgern als populistische Akteure hervorgehoben. Die meisten aktuellen Methoden zur Erkennung von Populismus in Texten beruhen jedoch auf manueller Datenannotierung oder auf wörterbuchbasierten Ansätzen. Nur wenige Arbeiten haben versucht, maschinelles Lernen für diese Aufgabe zu nutzen, was zu einem Mangel an annotierten Datensätzen führt, insbesondere für die deutsche Sprache.

Um diese Lücke zu schließen, verwendet diese Arbeit eine minimalistische ideologische Definition von Populismus und führt verschiedene Experimente mit BERT-basierten Transformer-Modellen durch, um die Erkennung populistischer Nutzerkommentare zu verbessern. Zusätzlich erstellt diese Arbeit den ersten annotierten Datensatz für populistische Nutzerkommentare unter Nachrichtenartikeln in deutscher Sprache durch die Durchführung einer Annotationsstudie.

Das vorgeschlagene Modell übertrifft den State-of-the-Art in diesem Bereich und wird in einer Fallstudie zur Analyse des Zusammenhangs zwischen COVID-19 und Populismus in österreichischen Nutzerkommentaren angewendet. Eine umfassende Analyse von Kommentaren im Nachrichtenforum der österreichischen Tageszeitung *Der Standard* wird durchgeführt, und die Studie zeigt, dass das Thema COVID-19 in Nachrichtenartikeln während der Pandemie mehr populistische Kommentare anzog als andere Themen. Aus diesen Erkenntnissen lassen sich Schlussfolgerungen für künftiges Krisenmanagement und -kommunikation ziehen.

Insgesamt trägt diese Arbeit zu unserem Verständnis der Prävalenz populistischer Rhetorik im Kontext der COVID-19-Pandemie bei und betont die Bedeutung der Entwicklung automatisierter Methoden zur Erkennung populistischer Aussagen in Texten.

Abstract

The COVID-19 pandemic and the resulting government measures have triggered a wave of protests and demonstrations in Austria. We saw some protestors resorting to populist rhetoric to express their dissatisfaction, which in some cases, led to anti-democratic tendencies. Populist talking points have not been confined to the public sphere but rather emerged from social media and other online platforms, including news comments. Thus, there is a need to develop automated methods to detect populist statements in those texts.

While previous research on populism has focused on politicians, recent studies have emphasized the role of citizens as populist actors. However, most of the current methods to detect populism in text rely on manual coding or dictionary-based approaches. Only a few scholars have attempted to employ machine learning for this task, resulting in a shortage of annotated data, particularly for the German language.

To address this gap, this thesis adopts a minimalistic ideational definition of populism and performs various experiments using BERT-based transformer models to enhance the detection of populist user-generated content. Additionally, the thesis presents the first annotated dataset for populist news user comments in the German language by conducting an annotation study.

The proposed model outperforms the state-of-the-art in this area and is applied in a case study analyzing the correlation between COVID-19 and populism in Austrian news user comments. A large-scale analysis of comments in the news forum of the Austrian daily newspaper *Der Standard* is conducted and the study reveals that the topic of COVID-19 in news articles attracted more populist comments than other topics during the pandemic. From these findings, implications can be drawn for future crisis management and communication.

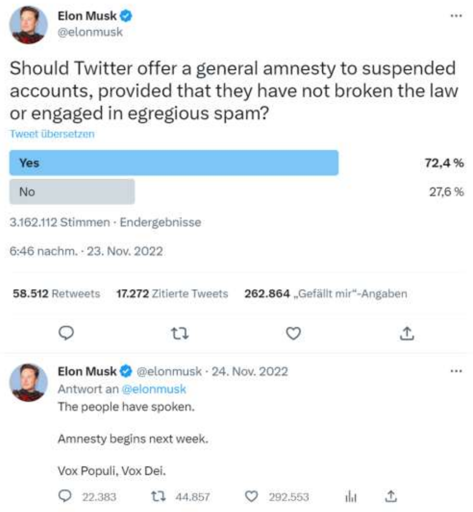
Overall, this thesis contributes to our comprehension of the prevalence of populist rhetoric in the context of the COVID-19 pandemic and underscores the importance of developing automated methods to detect populist statements in text.

Contents

CHAPTER 1

Introduction

"Vox populi, vox dei" - "The voice of the people is the voice of god". This expression, which was first recorded in a letter from Alcuin to Charlemagne in the 8th century [Rat11], has played a significant role in the history of political thought. Its demand for popular sovereignty emphasizes the authority of the common people and puts the power to ultimately govern in their hands. A very recent use of the phrase could be seen in a tweet by current Twitter CEO Elon Musk in Figure 1.1, where he used a Twitter poll to decide on the amnesty of priorly suspended accounts and claimed that "the people have spoken".



Elon Musk
@elonmusk

Should Twitter offer a general amnesty to suspended accounts, provided that they have not broken the law or engaged in egregious spam?

Tweet übersetzen

Yes	72,4 %
No	27,6 %

3.162.112 Stimmen · Endergebnisse

6:46 nachm. · 23. Nov. 2022

58.512 Retweets 17.272 Zitierte Tweets 262.864 „Gefällt mir"-Angaben

Elon Musk @elonmusk · 24. Nov. 2022
Antwort an @elonmusk
The people have spoken.

Amnesty begins next week.

Vox Populi, Vox Dei.

22.383 44.857 292.553

Figure 1.1: Twitter poll by Elon Musk[1]

[1] https://twitter.com/elonmusk/status/1595473875847942146 last accessed 21.01.2023

Here he used a Twitter poll that was active for 24 hours and visible to a very limited amount of people to demonstrate that the decisions of his company, are an expression of the will of "the people". By giving the impression that these polls provide a platform for political engagement, Musk is suggesting that traditional avenues for political participation are insufficient or lacking. This simplistic and non-representative view of popular opinion is an example of the rhetoric utilized in the phenomenon of populism. While the letter from Alcuin suggests that people using such rhetoric should not be heard, "since the riotousness of the crowd is always very close to madness" [Rat11], many populist movements have been successful throughout history by praising the properties of the people and positioning themselves as their voice. Populism is a highly controversial phenomenon, and defining the term populism itself is not without controversy. However, given that populism is a global phenomenon that is reproduced in various political contexts [SG18], there is a growing desire to find ways to detect and compare it on a large scale. In the context of the shown tweet, identifying such texts as populist can help to put these statements into perspective, measure how prominent they are in various forms of discourse and critically reflect on them.

1.1 Motivation and Problem Statement

Populist narratives have become a relevant part of liberal democracies over the last decades. All over the world many political parties that share populist views, like Austrian Freiheitliche Partei Österreichs (FPÖ), German Alternative für Deutschland (AfD), Argentinian Partido Justicialista (PJ) or French Rassemblement National (RN), have gained attraction and electoral success. Therefore it is of interest to define features that are common for populists disregarding their location or political direction. The concept of populism itself still lacks a fully agreed-upon definition. There are several approaches that propose an ideational [Mud04, WEW$^+$16] political-strategic [Bar09, Jan11, Lac05] or stylistic [Kr4, Can99, BM17] definition. However, they share a common set of core motives that are fundamental to the essence of populism. With these motives, it is possible to characterize and compare populist actors such as politicians or parties.

The basis of these motives is another definition that is widely utilized in the majority of the literature and serves as the foundation of this thesis. The so-called *ideational definition* has been set by Mudde in 2004 [Mud04] (and was extended in 2017 [MRK17]), who defines populism as a "thin centered ideology that considers society to be ultimately separated into two homogeneous and antagonistic groups 'the pure people' versus 'the corrupt elite', and which argues that politics should be an expression of the volonté générale (general will) of the people". From this definition, three primary motives can be derived: *people-centrism*, *people-sovereignty*, and *anti-elitism*. There is a widely accepted consensus that these motives are characteristic of populism [Roo19a].

With these motives, it is possible to categorize texts or speeches as populist or determine the presence of populist ideologies. Research in this direction has already been conducted in numerous works, beginning with Jagers' and Walgrave's [JW07] examination of populist content. However most of the work that focuses on text analysis deals with populism expressed by politicians or political parties [RP11, HRK19] and analyses party manifestos or speeches manually or with dictionary methods. There are only a few attempts that classify populism with the use of machine learning (ML) algorithms [UP21, HS16, DCM22], but they also rely on manifestos and speeches as input. Especially through the rise of social media and the means to express oneself to a larger audience, another populist actor has gained importance, who is the citizen that is at the center of the ideology. Some publications already indicated that citizens particularly on social media, also express populist ideas through their messages [EEEB17, Kr7]. Here populist ideas originating from politicians or the media are reproduced, which is why the same definition can be applied. Only a small amount of work is available that deals with the analysis of user comments [GT19, BEEE19, CAFS21]. They are mostly qualitative analyses and only the work of Cabot et al. [CAFS21] uses a ML approach. The latter focuses on a "US vs. Them" mentality that specifically separates an in-group from a defined out-group that is not necessarily elite, but can be a minority as well. This is a motive, which is excluded and not determined to be necessary in most works.

Therefore we can see that there is a clear lack of large-scale content analysis (especially using ML) or annotated data for the detection of populism in the German language specifically. The only validated and promising solution is a dictionary built by Gründl [Gr2] that was tailored to detect populism in social media content of German politicians. While this approach gave highly valid results, it is not flexible to different contexts of populist texts, as it is optimized for high precision. It is good at capturing general features on the word level and could serve as a basis to gather data for supervised ML models. Nonetheless, it suffers from general problems with dictionary-based approaches, as it relies on certain words to be present, and can not deal with negation or incorporate context.

Another perspective that motivated this work is the strong presence of populist talking points in context of the COVID-19 pandemic. Mudde and Rovira Kaltwasser [MRK17] suggest that populism is a threat to democracy. Especially in Germany and Austria we saw many demonstrations related to COVID-19 prevention measures like lockdowns or vaccinations. Originally a medical topic, it quickly turned into a political one. On these demonstrations, people from various political, social and economical background went to the streets. People that attended these demonstrations propagated populist motives to the extent that there have been anti-democratic messages (antisemitism etc.) [NSF20]. These messages were shared in social media and have been brought to the streets. This development makes it interesting to use it as a case study to analyze populism expressed by citizens. Thiele [Thi22] already used the mentioned dictionary-based method to detect populist user comments on social media platform Facebook and suggested that the topic of COVID-19 triggered populist comments under post from news pages.

To extend on that, this thesis works with a dataset of anonymized user comments posted directly on the news forum of the Austrian daily newspaper *Der Standard*[2] from 2019 until November 2021. Opposing to Facebook, users are more anonymous there because they only have to register with an e-mail and set a community nickname to comment. Furthermore, the variety and number of news posts is greater and people do not only post under the snippet of an article but directly view the whole text. The work of Thiele [Thi22] also indicated that populist comments grew over time, but only used data until May 2021. This is interesting because there are implications that critical situations increase populist attitudes from citizens [Ham18]. As the pandemic is still ongoing and developing after the study of Thiele, this needs to be further inspected with our data that contains comments posted during the fourth COVID-19 wave until November 2021.

1.2 Aim of the Work

The first aim of the work is to establish an annotation scheme for populist user comments based on research of the relevant literature. This scheme is used to set up a user study on a chosen platform, where multiple human annotators label a pre-defined set of comments. The outcome is the first German annotated corpus of populist user comments. This dataset is used as a gold standard to train, optimize and evaluate ML models, which is the second aim of this work. As there is a lack of annotated data and methods to detect populist user comments and acquiring a great amount of labelled data is very resource intensive, this work features an attempt to make use of existing resources, like the dictionary approach proposed by Gründl [Gr2] and proposes a model that combines it with ML. The goal is to overcome the limitations of the dictionary alone and improve on its performance, by using it to label training data according to word occurrences. This method can be used to detect populism in different contexts for the German language by feeding it new comments. Although the identified populist comments do not fit every possible definition of populism, they can serve as a means to drastically reduce human work and make it feasible to analyze a large corpus of data and only manually label a few samples for validation. Additionally, other experiments attempt to maximize the performance of a classifier using only a small training sample comprised entirely of human-annotated data.

After choosing the best performing model, the next goal is to apply it to a dataset of user comments provided by *Der Standard* as a case study. The focus here lies on the relationship between COVID-19 and populist user comments. Subsequently, the number of populist comments on COVID-19 articles, non-COVID-19 articles, and pre-COVID-19 articles are compared. This results in an indication of whether articles including the topic

[2] https://www.derstandard.at/

of COVID-19 trigger more populist comments than other topics during the pandemic and whether the COVID-19 crisis increased the number of populist comments.

The last goal is to observe the development of the number of populist comments under COVID-19 related articles to see if the observations of Thiele [Thi22] hold over a longer period of time.

To achieve these goals, this thesis aims to answer the following research questions:

- RQ1: What is an appropriate ML model to improve the detection of populism in Austrian news comments?

- RQ2a: How does the COVID-19 crisis affect the number of populist user comments?

- RQ2b: How does the topic of COVID-19 affect the number of populist user comments posted during the crisis?

- RQ3: How does the amount of populist user comments under COVID-19-related articles evolve over time?

1.3 Methodological Approach

This section describes the scientific methods used and summarizes the steps that were planned, changed and adjusted iteratively while conducting the work. The overall structure of this thesis can be divided into four core parts: a literature review, the annotation study, the model development and the application of the model on the data provided by *Der Standard*. Furthermore, the process of this thesis follows the Cross Industry Standard Process for Data Mining (CRISP-DM) guidelines proposed by Chapman et al.[3] [CCK+00], which can be seen in Figure 1.2 and are still highly used in the industry.

[3] https://s2.smu.edu/~mhd/8331f03/crisp.pdf

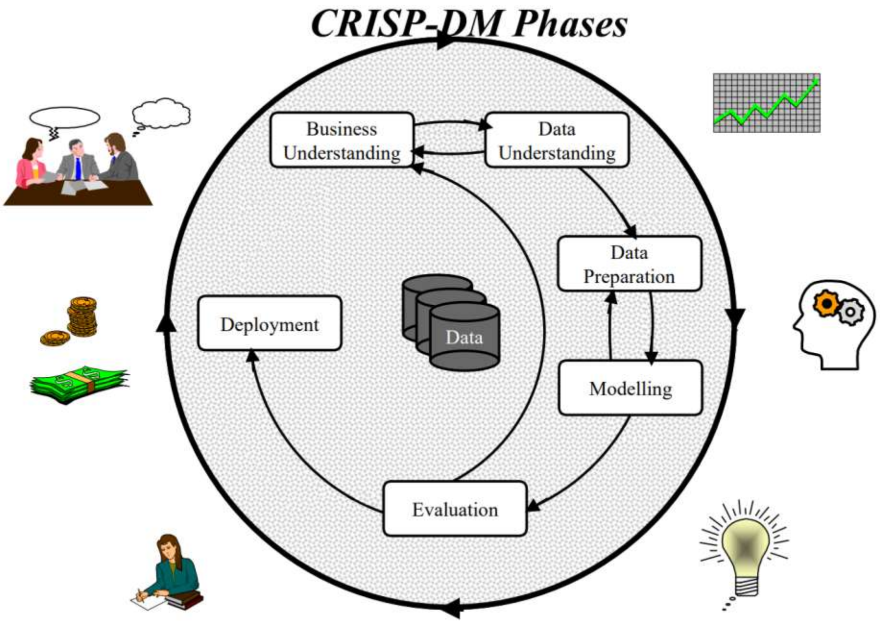


Figure 1.2: Phases of the CRISP-DM model visualized by Pete Chapman

1.3.1 Literature Review

To be able to have a solid foundation for this thesis, the first step is an extensive literature research to get an overview of the state-of-the-art (SOTA) and establish fitting scientific methods to address the identified open problems. The results of this research are summarized in Chapter 2. This references the establishment of business understanding. As populism is a controversial phenomenon, it is important to get an understanding of the literature coming from various different disciplines of social sciences. To be able to classify populism in user comments, the current SOTA of natural language processing (NLP) is reviewed. Finally, this includes research about literature dealing with populism in the context of the COVID-19 pandemic, to then build up on that with the case study.

1.3.2 Data Preparation and Annotation Study

The first crucial step is to establish a gold standard set of populist comments that can be used for model development and evaluation. Therefore we have to get a good understanding of the provided data. Descriptive analysis of the data set is followed by initial cleaning to then generate a sample, whose size is feasible for manual annotation with the resources available. As populism is considered to be rare among all comments, it is important to consider the problem of an unbalanced sample. To mitigate that, the dictionary of Gründl [Gr2] is used to identify potential populist comments and fill the sample with random comments, to reduce the bias of the dictionary. A comment is considered a candidate if it contains at least one dictionary term. Subsequently, those are manually labelled by human coders in an annotation study to obtain ground truth labels in order to overcome the issue that the dictionary only relies on certain words to be present, not real motives.

In the next step, the data is split into three parts. The first part contains only comments from articles that are related to COVID-19, identified by a keyword search. The second part contains comments from articles that were posted during the COVID-19 crisis, but deal with other topics. The last part contains comments that were posted recently before the outbreak of COVID-19 (in the year 2019), which serves as a reference. From each of those, a balanced sample of the same size is drawn. These serve as a gold standard to train and evaluate the algorithm and the baselines on. Blassnig et al. [BEEE19] provide an operationalization of the ideational definition and its motives, which we use as a foundation, to create annotation guidelines that fit the task of labelling news comments. Therefore, we use this operationalization to introduce to the topic of populism and establish a shared concept for the annotators. Furthermore, the participants are provided with selected example comments that showcase how this concept is reflected in news comments. A comment should then be labelled as containing people-centrism, people-sovereignty and/or anti-elitism and is considered populist if at least one of the motives is identified. Based on that, 1,200 comments are annotated by five annotators in total. Each comment is labelled by exactly three annotators and inter-annotator agreement is measured. Concluding, a comment receives a populist gold label by majority vote, meaning that if at least two people vote for a populist motive, the comment is considered populist.

1.3.3 Model Development

In the next step, the outcomes of the data preparation are used to build a model and evaluate it for the final deployment. The goal is to experiment with pre-trained large language models (LLMs), especially Bidirectional Encoder Representation for Transformers (BERT) [DCLT18] based models. Additionally, we build upon the aforementioned dictionary-based method and combine it with supervised ML. This has already been done before for tasks like sentiment analysis [SJ21, MKNMN20, TWC08] and showed success. A problem with purely dictionary-based approaches is that they rely on certain words to be present and are therefore prone to lack in precision, as they can not deal with features like negation or irony, because they do not incorporate context. Although populism is comparable across different countries or systems because of the specified motives, it can still feature context-specific peculiarities [Roo14], which is a problem for dictionary-based methods that are more general and do not adapt to developments in the used language for example.

The goal is to use the dictionary to identify a great amount of training data that is not present in the test set. This results in a self-supervised ML approach, where we initially feed unlabelled data into the dictionary. The advantage and improvement is that ML algorithms understand context and do not just rely on words being present. With this training data as a basis some ML algorithms, like support vector machine [CST00], random forest [Bre01] or logistic regression [Cox58], are implemented (as done before for political manifestos [HS16]) and serve as a baseline. Also, the German version of the first proposed populism dictionary by Rooduijn [RP11] is evaluated on the test set. Finally, the benchmark we want to improve on is the Gründl dictionary [Gr2] itself. The metrics used for comparison are precision, recall, and F1-score. In this setting, it is difficult to beat the dictionary regarding recall, as the test sample is created with its aid, and populism is expected to be rare in the random sample. So the focus lies on increasing precision to the extent that the F1-score is also improved. All experiments are conducted with a training, validation and test split of the gold labels, to try to maximize the performance on a small training set and compare it to the hybrid approach. In order to find the best-performing model two types of transformer models are fine-tuned on our training data and compared. The starting point is the German BERT model [CSM20], which is a deep transformer-based language model tailored for German data based on BERT [DCLT18]. Experiments with the models include different pre-processing steps, different training data and different strategies to deal with the issue of class imbalance. The models are evaluated on the human labelled test data and an improvement on the SOTA answers RQ1.

1.3.4 Case study

In the deployment stage, the validated, best-performing model is then be applied to the data set of all comments. Statistical tests are used to test the assumption that COVID-19 articles trigger more populist comments, which answers research questions RQ2a and RQ2b. This is done for the absolute count of populist comments per article. For RQ3 the correlation between the date and the number of populist comments is measured by introducing an increasing variable that starts with zero on the day of the first COVID-19 article and increases by one each day. All results are visualized for comparability to give further insight into the data. For the validation of these results, they are compared with the existing reports of Thiele [Thi22] in the time from March 2020 until May 2021.

1.4 Structure of the Work

To follow the methods mentioned above, this section gives a quick overview on how the work is structured. In Chapter 2, the results of the literature review are presented to clarify definitions and justify the methods used. Next, in Chapter 3, the used data is described, along with the process of drawing a sample for annotation and the processing steps applied in advance. Chapters 4, 5 and 6 give a detailed description of the data annotation, model building and deployment process and explain the exact steps taken, as well as the measurements used to validate them. Then the results of each core component of this work (besides the literature review) are displayed in Chapter 7. Eventually, the last Chapter 8 reflects on the conclusions drawn, lessons learned and shortcomings, to come up with potentials for future research.

CHAPTER 2

Related Work

This chapter gives a brief overview of the current state of the relevant research fields. We describe and define populism from different viewpoints in Section 2.1 and narrow it down to the phenomenon of citizens as populist actors in Section 2.2. Then we focus on ways to detect populism in Section 2.3 and look at the SOTA of NLP in Section 2.4, to determine methods that we want to use for text classification. To close this chapter, Section 2.5 presents the different publications that already connected the topics of COVID-19 and populism.

2.1 Definitions of Populism

To treat populism as a classification problem in the context of ML, it is important to find a definition that is both widely agreed on and operable to be identified at the text level. A uniform definition of populism remains a subject of ongoing debate to this day. Other then "thick" political ideologies such as socialism, conservatism, liberalism and others, populism primarily focuses on a narrow set of issues and is often characterized by simplistic solutions. In contrast, "thick" ideologies are comprehensive political beliefs that address a wide range of social issues and have well-developed theories and programs. They often have a more complex and nuanced understanding of the world and offer detailed solutions to societal problems. On the other hand, populism can coexist with and borrow elements from other ideologies [Kal12] and is not specific to any kind of political direction. Therefore, for a lot of researchers, it is not even clear that populism can be considered an ideology, which means "a body of normative and normative-related ideas about the nature of man and society as well as the organization and purposes of society" [Sai80]. To give an overview of some of the existing definitions and perspectives and justify the choice of this work to operate with the ideational definition, the following subsections discuss the relevant literature.

2.1.1 Populism as a Political Strategy

In his state of the field study in 2019, Rooduijn [Roo19a] depicts a history of the term populism and states that before a rise of populist studies in the early 2000s, there was no research that describes populism on a level that reaches beyond the area of one nation, as no one could define features that capture it globally. With this rise of studies many scholars aimed for a more comparative research of populism [Mud04, Rob06, Wey01]. In the following years, there was a growing consensus about the essentials of populism and most of the scholars agree on populism including anti-elitism and people-centrism [Roo19a]. However, the question of what populism exactly is continues above this essence.

Some researchers see it as a political strategy expressed through discourse. In his book about Venezuela's Chavism and comparative populism, Hawkins for example understands populism as "a Manichaean worldview linked to a characteristic language or discourse" [Haw10]. Manichaean here means a worldview that divides into good and evil forces, while discourse can be defined as "structured totalities articulating both linguistic and non-linguistic elements" [LM01]. Therefore populism is not seen as the ideology of a populist actor, but rather the expression of a set of ideas through language and communication. Subsequently, this definition fits to analyze for example how populist parties or politicians act or govern, as we can study their communication by observing their public output. This definition builds up on the core of the set of ideas, which sees the struggle of the "common people" against the "corrupt elite".

Ernesto Laclau in his book "On Populist Reason" [Lac05], also states that populism is a strategy to mobilize the masses and no fixed ideology, but rather flexible and manifested differently in different political contexts. Here he says that populism is the way to construct political identities that form a sense of belonging amongst the people, which is an expression of the idea of people-centrism. This strategy can either be used to acquire power or electoral success [Wey01] or retain legitimacy [AR07] by engaging a group of citizens. Barr sums this up by calling populism a "mass movement lead by an outsider [...] to gain or maintain power by using anti-establishment appeals and plebiscitary linkages" [Bar09]. Using this definition, it is possible to talk about the nature, causes and consequences of populism in various instances of it, which helps in understanding why populist strategies are applied, but not how it is expressed in terms of natural language.

2.1.2 Populism as a Communication Style

Next, there is a stream of publications on populism that analyze and define it as a communication style [Kr4, Can99, BM17, Mof16]. Here specific stylistic features are assigned to populist messages. These can include colloquial and emotional language, toxicity/incivility or simplification [BM17]. Moreover, it is said that the populist motives in mind are conveyed by a certain way of expressing oneself. This means that the antagonism between people or elites is described by using emotional words, to cause reactions like indignation amongst the people and engage their feelings. Also, simplification can be used to direct every social issue like criminality, poverty or the lack of employment towards an elite. Populists use this to offer simple solutions to far more complex problems [CG16], which can be appealing to citizens under certain circumstances like during a crisis [Ham18]. Elchardus and Spruyt [ES14] identified two traits of populist voters that are served by these means of communication, namely declinism and relative deprivation. Relative deprivation is a sentiment of helplessness and injustice, triggered by a sense of discrimination by an elite that denies the people what they deserve, while declinism is a general negative sentiment towards the current state and development of society. The essence of these sentiments can again be found in Mudde's ideational definition [Mud04]. The list of stylistic elements can be extended by including dramatization, polarization, moralization, directness or ordinariness, which Bos et al. [BvdBdV11] for example suggest.

While we now have a way to talk about how populism is expressed and we have debatable means to detect for example emotions [ANMC21] or incivility [MCGM18] in data science, we still can not determine what is exactly said in populist messages. We can try to detect those stylistic elements, but the list of attributes is exhaustive and detecting those attributes, will not lead to detecting populist content in text. Mudde [Mud17] raises critique to this perspective, as he states that not every populist message includes those stylistic elements and they are not sufficiently distinct to other forms of discourse or communication.

2.1.3 Ideational Definition

When observing populism on the text level, we want to know what is being said. In cases where this is of interest, the most widely used definition is based on the already presented definition by Mudde, who defines it as a thin ideology:

> An ideology that considers society to be ultimately separated into two homogeneous and antagonistic groups, 'the pure people' versus 'the corrupt elite', and which argues that politics should be an expression of the volonté générale (general will) of the people. [Mud04]

This definition, as minimal as it is, sums up the characteristics that are shared by all populist actors, disregarding their political beliefs, which makes for example right- and left-wing populism comparable. The core notion of this ideology is that in the current system, the people lack power and feel the need to be represented more in shaping society, laws, etc. This representation is suppressed by a certain group of elites or by representatives of the elites like leading politicians. Different from anti-political establishment parties, elites do not necessarily have to be politicians in power in this case. Elites are defined by holding some sort of power and can be political elites, financial elites, cultural elites or media elites [Roo19a]. This creates a Manichaean worldview, where elites are demonized and assigned bad traits, by calling them corrupt, oppressing, lying and patronizing towards the people. The people are seen as a homogeneous group that shares a variety of good common attributes, opinions or beliefs, which are praised by populists. Some definitions propose more out-groups than just elites, like immigrants for example [JW07, CAFS21], to form a “thick” populism, but this often leads to fluent borders to ideologies like nativism and is not included in most definitions. A populist puts themselves into the position of speaking for the people and expressing their needs and desires that are denied by elites. This is a very simplistic view of societies and denies pluralism. However, this view can also empower people that are actually not represented by political elites in an unjust way and serve to correct flaws in democratic systems [Kal12], but is more often seen as a threat to liberal democracy [Can99, MC12, Maz08].

From Mudde’s definition, we can derive three motives that build the main framework for the detection of populism in this work. In Figure 2.1, we can see a conceptual model built based on the content of the ideational definition, which includes anti-elitism, restoring sovereignty (called people-sovereignty in this work) and people-centrism and the interplay of the different actors.

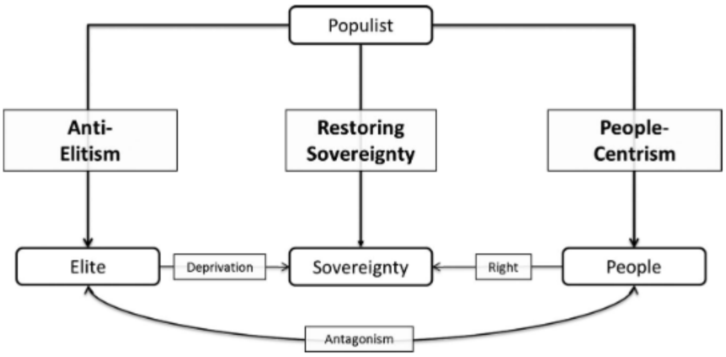


Figure 2.1: Conceptual model of thin populism by Ernst et al.[EEB+17]

Wirth et al. further conceptualized and operationalized the motives and described their features [WEW+16]. Their concepts are broken down into key messages, which are tied to an underlying ideology, to facilitate a more fine-grained distinction and definition of the motives. These key messages are detectable in texts and are therefore considered an operable framework for this work. Anti-elitism refers to the antagonism between the people and the elites, it can be expressed by disparaging the elites and assigning malicious attributes to them. Also, elites are made responsible for the negative development of the country and seen as acting against the people. Moreover, a populist actor tends to stress the distance between the elites and the people and deliberately separates them. Next, there is people-centrism, which can be identified, when the actor sums up the people as a homogeneous group and describes their shared needs and feelings. In addition, the people are celebrated as righteous and their honorable characteristics are stressed. The actor also puts themselves actively in the position of speaking for the people, by stressing that they are "one of them", disregarding their background (as a politician for example). Ultimately the motive of people-sovereignty features two perspectives: claiming more direct influence on decisions and consequently sovereignty for the people and disavowing the same for elites.

These three dimensions of populism are moreover used in a lot of other works on the matter [SMS+18, EBE+19, Wir18, Roo14]. Besides that there is also a critical opposing view that disagrees with the operationalization as an ideology, which can be found in the work of Aslanidis [Asl16]. Nevertheless, this ideological framework gives us the possibility to now answer the question of what is being said in populist communication and enables the analysis of content. Combining the three definitions, it would additionally be possible to talk about who populist actors are, by observing an actor's style, discourse and content of their messages, but this is beyond the scope of this work.

2.2 Citizens as Populist Actors

Now that we have a good understanding of how populism is defined, it is important to translate that definition to our use case. As we want to observe populism in news comments, we focus on the regular citizen as a populist actor. Most of the previous research that deals with analyzing populism in texts, focused on politicians or political parties as populist actors (for example [HRK19, EBE+19, RP11]). However recent work has also dealt with people expressing populism, especially in messages or comments posted online [EEEB17, Kr7]. De Vreese et al. [dVEA+18] define three key actors that convey populist communication, political actors, media and citizens. When talking about the citizen, we can also observe how they are influenced by or exposed to populism, like in the case of declinism and relative deprivation [ES14]. More interesting for this work is how and where citizens express populist messages themselves. In the context of news, Esser et al. describe "populist citizen journalism" [ESH17] and criticized that some, although rather neutral themselves, news outlets open space for reader's populist comments and Hameleers et al. [HBdV16] claim that populism is especially prevalent in the online forums of tabloid newspapers. In general, the online space and social media gave citizens the means to reach a broad audience by sharing their opinions and worldviews, which offers this position as a populist actor. Galpin and Trenz call this phenomenon "participatory populism" [GT19] and find that in the context of the 2014 European parliament election in Germany and the UK, the means to interact with news by writing comments, activates users to unite and express their dissatisfaction with the elites in a form of "low engagement [...] through mainstream media channels". However this selection of users is not representative of the people they pretend to speak for, but they participate in forming public opinion.

When we look at the content of citizens' populist messages, we can apply our established framework, as it is agnostic to the role of the sender of a message. This assumption is reinforced by the little research on said content. Hameleers [Ham19] made the first attempt to study the way Dutch citizens communicate populist messages online and found group dynamics that form a unified identity that empowers the collective, which expresses its discontent towards the system. The people in this collective are likely to surround themselves with like-minded people, which leads to the conviction that they represent public opinion and are not heard by the elites. Subsequently, they demand more power and influence from an antagonized elite. Fernández-Garcia and Salgado discovered similar patterns in an analysis of online comments in Portugal and Spain [FGS20] and added that of the core motives, the most frequently found one was anti-elitism. They additionally found an increased level of hostile, uncivil and hateful language in populist comments than in all others. Finally, they state that there is a great proportion of populist comments that express a worldview that is not related or represented in the source (e.g. news article) that is commented on. This theoretical foundation gives us now the means to capture populism in comments posted by citizens, which is done in the annotation study.

2.3 Populism Detection

Detecting populism on a textual level has emerged as a scientific field beginning with the work of Jagers and Waalgrave [JW07] in 2007. Ever since various methods have been used to code and detect populist content. The methods used can hereby be divided into two subgroups. First, there is manual content analysis, where trained coders or experts analyzed mostly political texts. Yet this is very resource expensive and not scalable on the large amount of data that is available, especially when observing user-generated content. This is why the second approach is to develop automated methods for populism detection. The following subsections give a brief overview of the work that has been done for both approaches.

2.3.1 Manual Analysis

Jagers and Waalgrave opened the field of populist content analysis by qualitatively studying Belgian parties' political broadcast [JW07]. Hawkins followed with an analysis of Hugo Chávez populism in Venezuela [Haw09] and used holistic grading [Whi85] to classify political speeches, which was also applied in other studies [HCS19, HK18]. Other work focused on coding smaller documents like paragraphs of manifestos [Roo14, RP11, PR15], statements and semantic clauses [Asl18, MSW$^+$17, MW17] or social media posts [EEEB17]. While the coding and results are of high qualitative value, the downside is the said expense. In those studies, coders were extensively trained to classify a few documents, which limits the scope of analysis. Especially, when one wants to perform a comparative analysis of different countries, it is not easily feasible to get a sufficiently large collection of coded data. Furthermore, the goal of those studies was mostly to measure the degree of populism of parties determined by their output on a linguistic level. Another approach to this is using (expert) surveys [PRB$^+$17, SAK17], which deals with the same issues and does not incorporate the process of directly classifying text. When processing a large amount of text, one can not rely on manual analysis and other methods had to be developed.

2.3.2 Automated Analysis

The use of computational power to automatically detect and classify populist text aims at quantifying large amounts of documents. There are two methods at the top level that are used for that, dictionary-based and ML methods. The former uses a pre-defined set of terms and identifies them in documents, which means that we are looking for word occurrences. The preparation of such a dictionary is labour-intensive and requires a careful selection of terms that can capture a complex phenomenon and a solid theoretical foundation. In addition, dictionary-based methods need a scoring method to classify text. We could for example label a text as populist if one dictionary term is present, but that would leave us with imprecise results. Other options would be to use relative frequency and thus take document length into account, which would again be problematic when dealing with generally short text. Grimmer and Steward stress that the main pitfall in using dictionaries is validation, especially when they are used in different contexts than they were constructed for [GS13].

In the context of populism, we already mentioned the dictionaries by Rooduijn and Pauwels [RP11] and Gründl [Gr2], which we will focus on, as they provide resources for German text especially. Besides that, the Gründl dictionary showed to be the most valid resource in the context of our application. Other dictionary-based methods can be found in the publications of Bonikowski and Gidron [BG15, BB16]. Rooduijn and Pauwels [RP11] established a dictionary-based method in comparison to their aforementioned qualitative analysis of party manifestos. They found that the dictionary showed less validity in terms of content (measure to evaluate if a method measures all aspects of a phenomenon), face (subjective measure to assess if a method measures what it is intended to) and concurrent (measures the correlation between the scores of two methods measuring the same phenomenon) validity, but they still considered the results sufficiently valid. They also provide a German translation for their dictionary. Gründl criticized the existing approaches as being too narrow (the formerly mentioned dictionary only includes forms of 20 German words), but sees them as a good starting point, upon which he constructed his dictionary. He expanded them in an iterative way, optimized on recall and precision and adapted it to social media content. Additionally, he provides an R-package[1] of his work that captures multi-word expressions and uses regular expressions to detect various forms of the 238 included terms. The validation of his dictionary was two-fold: first of all, he applied it to social media posts of members of German and Austrian political parties and compared his results on the party level to expert surveys. The second validation was done by split-half reliability. This is a way to measure if the terms of the dictionary measure the same phenomenon by splitting the dictionary

[1] https://github.com/jogrue/popdictR

in half and measuring the correlation of the results of both halves. In both cases, he significantly outperformed the existing approaches and is therefore considered SOTA and features as an aid and basis of this work. An additional advantage is that Gründl uses the same definition as this work and opted to identify anti-elitism, people-centrism and people-sovereignty in particular.

When we now look at ML approaches, the first application dates back to 2016, where Hawkins [HS16] and Castanho Silva applied basic ML algorithms to classify party manifestos and political speeches. They found elastic net regression [ZH05] and logit boosting [FHT00] performed best on their data, however, they were only good at detecting non-populist documents and rather poor at detecting populist ones. Despite those unsatisfactory results, they saw potential in using ML to detect populist comments and determined a lack of annotated data as their main problem. In the following years, some resources were published, like *The PopuList* [Roo19b] or the *Global Populism Database* [HRK19], but they either contain coded party manifestos or speeches or populism scores or labels for parties. Subsequently, there is to date a lack of (openly available) annotated data to use for supervised ML. Party speeches and manifestos were also the basis for the other available attempts to classify populist text with ML. Di Cocco and Monechi [DCM22] measured the populism of parties, by splitting manifestos into sentences, and assigning each sentence of populist parties a populist label, which was criticized [JH22]. They used a *Random Forest* classifier and got good results in classifying populist parties. Other resources using *Random Forest* algorithms can be found in the works of Dai [Dai19, DK22]. The only attempt to use a deep learning or BERT-based model is a publication by Ulinskaite and Pukelis, who again classify manifestos and consider the results of BERT as largely successful [UP21].

To the best of our knowledge, there is no available publication to date that uses machine learning to classify populist user comments or social media data according to the minimal, comparable, ideational definition. Cabot et al. manually labelled English Reddit[2]-comments and used a Robustly Optimized BERT Pretraining Approach (RoBERTa)-based model to classify them [CAFS21]. Nonetheless, they focus on the "Us vs Them" mentality and inspect the data for various non-elite out-groups like immigrants, liberals or Muslims. However, we already stated that this is mostly not considered a part of populism in general, but rather seen as an expression of the underlying political attitude [Roo19a]. So we can see that there is an identified gap in the literature that this work will address.

2.4 Natural Language Processing

Summing up the literature presented until now, we can see that most of it was done in various fields of social sciences. In this thesis, we want to contribute to the field

[2] https://www.reddit.com/

by applying the SOTA of data science, ML and for the most part NLP. NLP includes a variety of different tasks that we can use machine learning techniques for, to deal with textual data. Those include question-answering, named-entity recognition, machine translation and the list can be extensively prolonged. The specific task of interest for this work is classification, which means assigning a label, out of a predefined set, to a body of text based on its content. Earlier approaches of NLP use a bag of words (BOW)-assumption, that represents each document as a list of its containing words, disregarding the order in which they appear. Especially with the emergence of deep learning and moreover transformer models, we are now able to derive more meaningful features and especially capture the context within text, by modelling text sequences. Vaswani et al. proposed the method of transformers in their paper "Attention is all you need" [VSP⁺17] to originally use it for machine translation and it can be described as follows:
Transformers use an encoder-decoder structure, which models an input text vector to a continuous representation vector of the same length. The decoder then uses this continuous representation and sequentially generates an output sequence of an arbitrary length, but we will mainly focus on the encoder here. The representations are learned by using a mechanism called attention, which is used to weight the importance of each input token for each output token, based on a learned alignment between them. Here each input token is assigned with three learned vectors: a query vector of dimension d_q, a key vector of dimension d_k and a value vector of dimension d_v. Multiple input tokens are combined to query, key and value matrices Q, K and V. Those matrices are mapped to an output vector by applying a softmax function on the dot product of Q and K-transposed divided by $\sqrt{d_k}$ and use that as weights for V:

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

This is called *Scaled Dot-Product Attention* (see Figure 2.2) and is applied as self-attention here, which means that the relation of each token of a sequence to each other token of the same sequence is computed to get the representation. To attend to multiple parts of the input in parallel, transformers use *Multi-Head Attention*, as seen in Figure 2.2. Therefore the queries, keys and values are projected h times, using different linear projections of the same dimensions. For those, attention is computed in parallel and the h d_v-dimensional output values are concatenated and linearly projected again:

$$\text{MultiHead}(Q, K, V) = \text{Concat}\left(\text{head}_1, \ldots, \text{head}_h\right) W^O$$

$$\text{where head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$$

The said projections are the parameter matrices $W_i^Q \in \mathbb{R}^{d_\text{model} \times d_k}$, $W_i^K \in \mathbb{R}^{d_\text{model} \times d_k}$, $W_i^V \in \mathbb{R}^{d_\text{model} \times d_v}$, and $W^O \in \mathbb{R}^{hd_v \times d_\text{model}}$ and are learned during training. The learned representation of each input is a d_{model} = N-dimensional embedding vector, which is then enriched with positional embeddings of dimension d_{model}, to be able to represent the order of sequences.

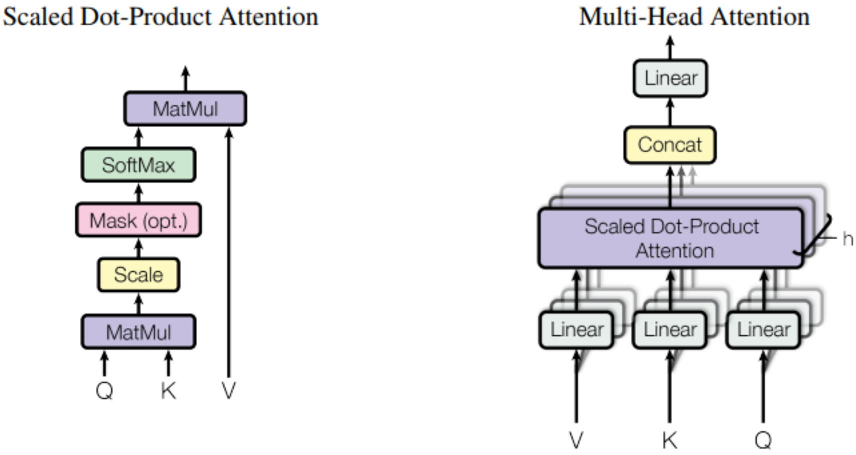


Figure 2.2: (left) *Scaled Dot-Product Attention* (right) *Multi-Head Attention* [VSP+17]

Positional embeddings are calculated as sine and cosine functions of different frequencies:

$$PE_{(pos,2i)} = sin(pos/10000^{2i/d_{\text{model}}})$$

$$PE_{(pos,2i+1)} = cos(pos/10000^{2i/d_{\text{model}}})$$

with pos being the position and i the dimension. The encoder stack then takes the sum of the learned embeddings and the positional embeddings and is built by using M identical layers, where each layer consists of a multi-head self-attention sub-layer and a position-wise fully connected feed-forward network. Both sub-layers use residual connections and layer normalization, which help to alleviate the vanishing gradient problem and improve training stability [HZRS15]. The output of each encoder layer is of dimension d_{model} to enable the residual connections and is passed on to the next layer as input, until the final output of the encoder stack is then used as input to the decoder stack or the output layer of the model, depending on the task [VSP+17].

The transformer architecture was adapted to train many complex language models, which are used in NLP. Especially pre-trained BERT-based models are omnipresent and fine-tuned to reach SOTA results in many tasks. The original BERT model was proposed by Google [DCLT18] and consists of 12 bidirectional Transformer blocks, each using twelve self-attention heads, and uses a dimension of $d_{model} = 768$. The key functionality here is the ability to perform bidirectional training, which means that both the preceding and the following token can be attended during training, which helps to model context as opposed to other left-to-right/right-to-left approaches [PNI+18]. Another strength of BERT lies in the general structure of its architecture, which can be seen in Figure 2.3 and which makes it applicable to many downstream tasks. Therefore, the input has to

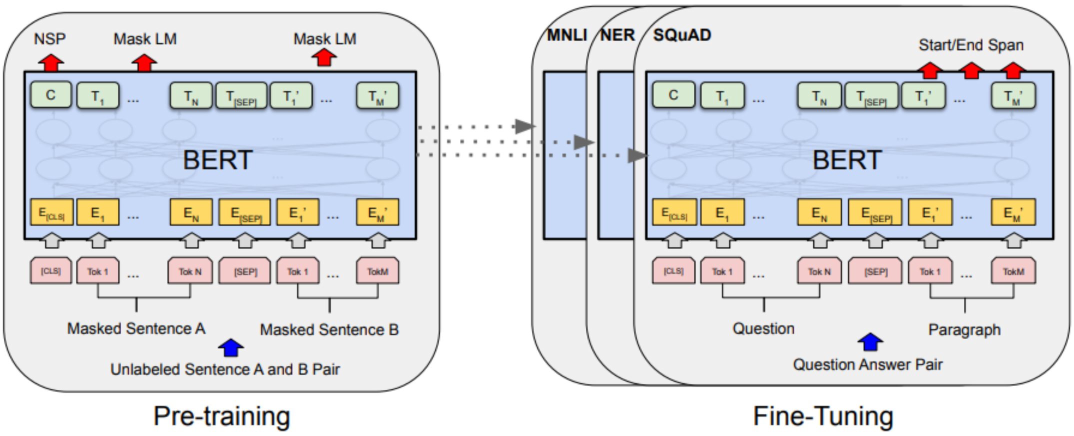


Figure 2.3: BERT architecture for pre-training and fine-tuning [DCLT18]

follow a certain structure. BERT uses an extension of WordPiece embeddings [SSS$^+$20] to create tokenized embeddings as a representation of input sequences, where a *[CLS]* and a *[SEP]* token are added. The *[CLS]* token is added to the beginning of each sequence and in case of paired input sequences (e.g. for question answering), the *[SEP]* token is used between the two concatenated sequences. The input is then the token embedding, a positional embedding and another embedding that indicates the hierarchy of paired sequences if needed. The model is pre-trained in a self-supervised fashion by performing two tasks: masked language modelling (MLM) and next Sentence prediction (NSP). MLM hides some tokens of an input sequence and predicts those masked tokens, while NSP predicts the next sentence after an input sentence, to understand the relationship between sequences. To then apply it to a certain task, the model is fine-tuned, by simply feeding BERT new input-output pairs and tuning the parameters end-to-end. For a classification task, we just have to add an additional fully connected layer including a softmax function, that receives the *[CLS]* representation and outputs class probabilities [DCLT18]. Following the breakthrough of BERT, many researchers tried to build on its success or adapt it to other languages. Through the *Hugging Face*[3] transformers library, many already extensively pre-trained models are publicly available and applicable for people with standard hardware, as fine-tuning and inference is not expensive compared to the pre-training. There is even a BERT adaption specialized on political text called *PoliBERT* [Gup20], which is trained on an English corpus though and is therefore not considered for this work, as manual translation on the whole data set would not be feasible with the given resources. Therefore, we focus on BERT models that were specially trained on German data, like GBERT [CSM20], which was trained on over 160 GB of German text data and reached SOTA results on a lot of German benchmark data sets. Pre-trained models are used to feed them our labelled populist user comments and try to fine-tune them for the populism classification task.

[3] https://huggingface.co/

2.5 COVID-19 and Populism

To come back to the topic of populism and why it is interesting in the context of the COVID-19 pandemic, we now take a look at how the two topics are connected. The pandemic was arguably the most dominant topic in the news after its outbreak. The debate rapidly changed from a medical to a political topic, because of the drastic changes of social life and the guidance of the government, which introduced measures to counteract the spread of the virus. It was already stated that crisis situations nurture populist ideas because people tend to sympathize with the offers of simple solutions and search for a target that they can blame for the circumstances [Ham18]. Especially anti-elitism is a strong motive that could be observed from a group of people. The main antagonists here were political elites, the media and moreover scientists that helped in advising the government by lending their expert knowledge. Virologists like Christian Drosten or Dorothee von Laer became public figures and were subsequently targeted and threatened by frustrated people online and also on “coronasceptic” demos in the respective countries[4]. Nachtwey et al. already showed the strong prevalence of populist stances in the German and Swiss populations that include a high distrust of the government and its measures and conspirative theories that explicitly state that the government wants to deceive the people with their measures. Even clearly anti-democratic views could be measured, like anti-semitism, where people blame Jewish elites for globally influencing politicians behind the curtains. They state that the people expressing those stances come from a great variety of social and ideological backgrounds, but are predominantly members of the educated middle class [NSF20]. In general conspiracy theories tied to the COVID-19 pandemic became very popular and accused for example Bill Gates of controlling global population numbers with vaccination. Eberl et al. [EHG21] argue that populism is at the center of those conspirative beliefs and additionally state that they could not find evidence that the underlying thick ideology (e.g. left or right populism) makes a difference. Scholars and academic experts have already been seen as an opposed elite for populists before, because they claim the power of the “truth”, without being legitimized by “the people” [WEW$^+$16, CAR17]. Because a great portion of public life shifted online during the pandemic [PCF21], we can follow a lot of those hostilities online on social media or in (news) forums, to the extent that Boberg et al. described a “pandemic populism” unfolding online [BQSEF20]. They identified anti-systemic messages that oppose the view of mainstream media, experts and the political establishment, which foster a “contradictory, menacing and distrusting worldview” [BQSEF20], which furthermore reinforces the connection between COVID-19 and populism.

Observing online “pandemic populism” in Germany and Austria is of particular interest, as the “coronasceptic” demonstrations evolved close to the extent of a civic movement [Thi22]. In the context of news comments, Eisele et al. stressed the role of mass media in distributing information on crisis measures [ELB$^+$22]. They analyzed comments from

[4] https://www.dw.com/de/corona-krise-virologen-werden-zur-zielscheibe/a-55909046 last accessed 13.02.2023

krone.at[5] and derStandard.at[6] and measured the emotionality they could find under articles that cover the COVID-19 crisis. They found that people react with more affection towards the mentions of political actors in articles than towards crisis measures and that decision-makers play a central role in sparking emotional responses [ELB+22], which is interesting in the light of anti-elitism in our research. A major reference for this work is a publication by Thiele [Thi22], who examined Facebook comments on the pages of news outlets in Germany and Austria on a large scale using the dictionary of Gründl [Gr2]. His analysis covers the time from 1 January 2020 until 20 May 2021 and two central hypotheses were that posts about COVID-19 attract more populist comments than other posts and that the attraction of COVID-19 posts for populist comments rises over time, which he could both verify. He states that over time the dissatisfaction of the people with the topic grew and they expressed it by showing their discontent by writing comments that contain populist ideas, which is interpreted as a form of reactance [DS05]. However, he already states that using the dictionary for large-scale analysis is a limitation on the depth of his results [Thi22]. This study is taken as a reference and to contribute to the field, by adding another source of user comments, increasing the timeframe of the analysis and trying to improve on the performance of the dictionary. A study of comments under articles from *Der Standard* and their Facebook page showed that incivility is significantly more prevalent on their news forum than on Facebook, which could be caused by the higher use of pseudonyms [SH22]. Anonymity in online discussions increases incivility, which can often be found in populist content and people tend to express their opinion in a more unfiltered way [San14], which is why observing news website comments can help to gain new insights. A further analysis beyond the timeframe of Thiele's work can also help to identify and measure problems of distributing information during crisis situations and communicating the importance of countermeasures in the future, to reduce the appeal of conspiracies originated by populist appeals and strengthen the trust in governmental decisions.

[5] www.krone.at

[6] www.derStandard.at

CHAPTER 3

Data Analysis

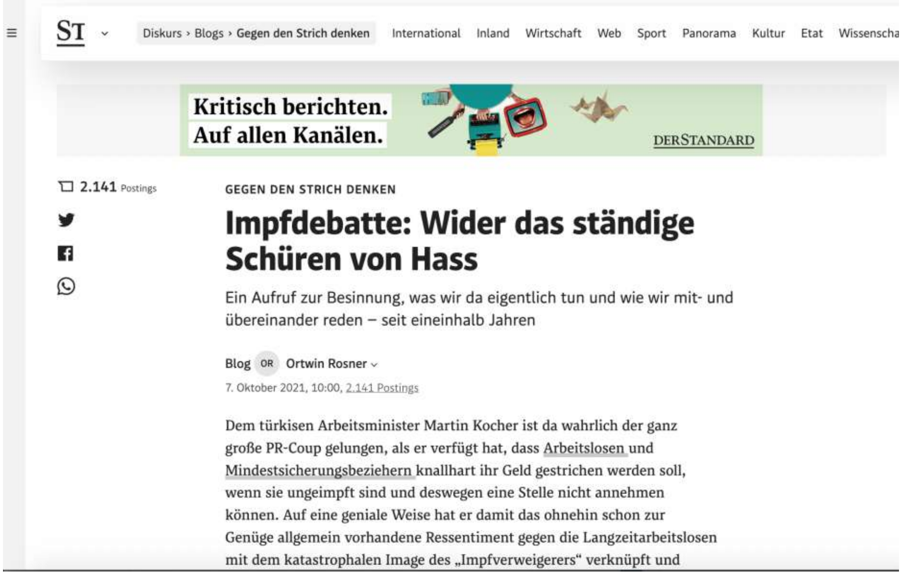
In this chapter, we analyze the data that we work with. First, we give an overview of the available data set and perform descriptive data analysis on the three samples that we want to compare in the case study. Next, it is important to draw a sample for manual annotation and explain why certain decisions are made in the process. Finally, we list the pre-processing steps that are taken before feeding our data into the model.

3.1 *Der Standard* Data

Our data set was kindly provided by Austrian daily newspaper *Der Standard* and features data from 1999 until November 2021 originating from their news forum. *Der Standard* is one of the two most read quality newspapers in Austria and considers itself as liberal and politically neutral, but is classified as left-liberal by press review *eurotopics*[1]. As tabloid news attract more populist comments and content and Blassnig et al. state that populist news articles lead to more populist user comments [BEEE19], the choice of a rather neutral source mitigates this effect. *Der Standard* provides their news forum as a comment section below an article, which is shown in a random COVID-19 article in Figure 3.1 and its comment section underneath in Figure 3.2. This enables the users to directly post their opinion on a specific topic and interact with other users by replying to already existing comments. The replies are organized hierarchically, and users have the option to express their opinion through up- or down-voting a comment without having to respond with textual content. These votes are reflected in the red and green bar attached to a comment. In order to leave a comment, a user must register with an email and a pseudonym, which provides a means to maintain anonymity.

[1] https://www.eurotopics.net/de/148488/der-standard last accessed 13.02.2023

[2] https://www.derstandard.at/story/2000129938584/impfdebatte-wider-das-staendige-schueren-von-hass last accessed 13.02.2023



GEGEN DEN STRICH DENKEN

Impfdebatte: Wider das ständige Schüren von Hass

Ein Aufruf zur Besinnung, was wir da eigentlich tun und wie wir mit- und übereinander reden – seit eineinhalb Jahren

Blog OR Ortwin Rosner

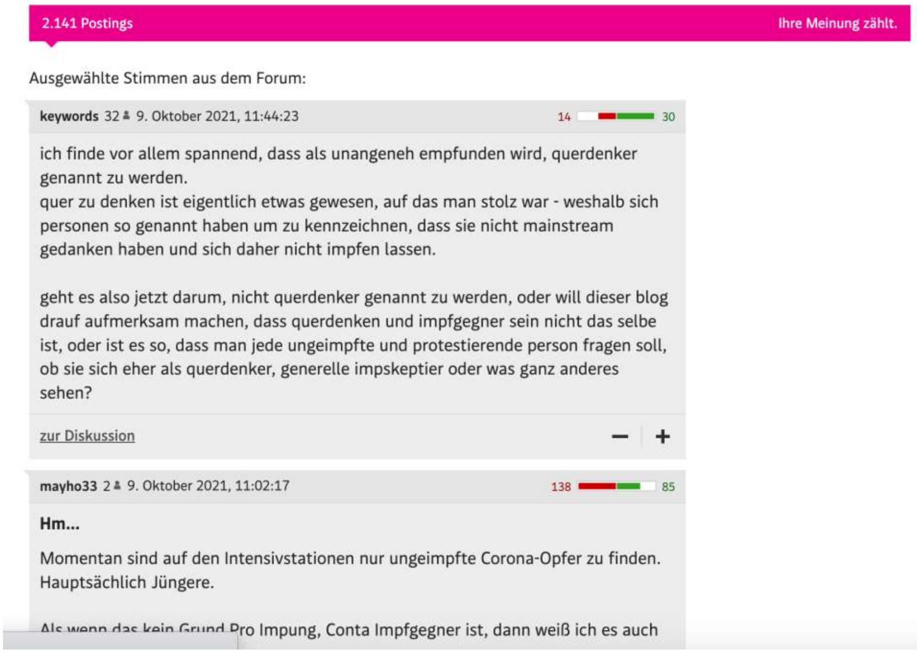
7. Oktober 2021, 10:00, 2.141 Postings

Dem türkisen Arbeitsminister Martin Kocher ist da wahrlich der ganz große PR-Coup gelungen, als er verfügt hat, dass Arbeitslosen und Mindestsicherungsbeziehern knallhart ihr Geld gestrichen werden soll, wenn sie ungeimpft sind und deswegen eine Stelle nicht annehmen können. Auf eine geniale Weise hat er damit das ohnehin schon zur Genüge allgemein vorhandene Ressentiment gegen die Langzeitarbeitslosen mit dem katastrophalen Image des „Impfverweigerers“ verknüpft und

Figure 3.1: Body of a randomly selected COVID-19 related article on the website of *Der Standard*[2]

The available database features a variety of metadata on articles, users, authors, etc. For our study, three tables are of interest and their relevant content is listed here:
Postings: includes a comments unique identifier, the article's identifier it was posted under, the username of the poster, the comment's content, headline, and a timestamp.
Classified Articles: includes a JSON object that contains classifications of the content of the article (e.g. which persons are mentioned), the unique identifier of the article and most important a list of keywords that classifies the relevant topics.
Content Details: includes metadata about all articles, like their headline, a unique identifier and a timestamp.

The original *Postings*-table includes a flag of whether a comment is the parent of another comment, but We choose to ignore the hierarchy of comments and treat parents the same as replies. Then, the first step is to filter the relevant data for the time span we want to observe. As we include reference data from 2019, we extract all comments posted from January 2019 until November 2021, which sums up to a total of 42,559,535 comments posted under 145,189 different articles. We encountered an issue when attempting to merge the three tables and incorporate keyword classifications into the user comments. As a result, 5211 articles remained unclassified following the join. We find an inconsistency in the database, as the identifiers could not be found in the *Content Details*, but all missing classifications are retrieved by just joining the comments with the *Classified Articles*, so we do not lose any data here.


2.141 Postings

Ihre Meinung zählt.

Ausgewählte Stimmen aus dem Forum:

keywords 32 9. Oktober 2021, 11:44:23 14 30

ich finde vor allem spannend, dass als unangeneh empfunden wird, querdenker genannt zu werden.
quer zu denken ist eigentlich etwas gewesen, auf das man stolz war - weshalb sich personen so genannt haben um zu kennzeichnen, dass sie nicht mainstream gedanken haben und sich daher nicht impfen lassen.

geht es also jetzt darum, nicht querdenker genannt zu werden, oder will dieser blog drauf aufmerksam machen, dass querdenken und impfgegner sein nicht das selbe ist, oder ist es so, dass man jede ungeimpfte und protestierende person fragen soll, ob sie sich eher als querdenker, generelle impskeptier oder was ganz anderes sehen?

mayho33 2 9. Oktober 2021, 11:02:17 138 85

Hm...

Momentan sind auf den Intensivstationen nur ungeimpfte Corona-Opfer zu finden. Hauptsächlich Jüngere.

Als wenn das kein Grund Pro Impung, Conta Impfgegner ist, dann weiß ich es auch

Figure 3.2: Comment section underneath the article in Figure 3.1

Before performing a descriptive analysis, we split the data set into three parts. The first one is our reference sample, which is simply created by filtering for all comments posted in 2019. To split the data featuring comments posted during the pandemic, we perform a keyword search to identify COVID-19-related articles. Our goal is to have high precision in finding relevant articles and avoid false positives, which is why the list of search terms should not be too extensive. Finally, a search is conducted for the substrings "Corona", "Covid", and "SARS" (ignoring case sensitivity) within the list of unique keywords for all articles. Of the 64054 different keywords in total, 58 contain at least one of our search terms. Subsequently, we can identify 14,165 different COVID-19-related articles. In the following, we reference our samples as *reference sample* (contains all posts from 2019), *COVID-19 sample* (contains all posts under COVID-19 related articles) and *non-COVID-19 sample* (contains all posts during the pandemic under non-COVID-19 related articles). We find the first COVID-19 article on January 6 and use that as the starting point for the pandemic by cutting off all comments posted before that from the *COVID-19 sample* and the *non-COVID-19 sample*.

3.1.1 Descriptive Analysis

We first examine the sample sizes to get an overview of the data. The *reference sample* contains 9,682,153 comments, the *COVID-19 sample* contains 14,391,704 comments and the *non-COVID-19 sample* contains 16,885,895 comments. There are however many empty comments (431912/375235/733336) among those, which we remove for the further analysis. The number of empty comments could be caused by users just writing their comments in the headlines and leaving the actual comment empty. Table 3.1 lists the user and article stats of all samples. We can immediately see that despite the number of articles published in 2019 exceeding threefold the number of COVID-19 articles, a higher user engagement, in terms of commenting, can be observed for the latter. When looking at the comments posted per article, it is noticeable that there is a substantially higher engagement on COVID-19-related articles, indicated by the mean and median comments per article. This is an interesting finding, as it supports the statement that most of social life has shifted online during the pandemic. Furthermore, this could indicate the controversial nature of the topic. The comments per user feature some outliers for each set, which means we have a group of highly active users that posted up to 101,066 comments during the time captured by our samples. It is remarkable that the top 13 users in the *COVID-19 sample* posted at least 10,000 comments more in total than the top user in the *non-COVID-19 sample*. However, when looking at the quantiles, we can see that it is rather a higher engagement in comments per article than comments per user. This finding suggests that the topic of COVID-19 activated a larger number of users who participated in discussions on the news forum. The equal median of eight and a lower value for the 75% quantile of the COVID-19 comments per user indicates that we have more highly active outlier users that increase the mean in this statistic.

	Reference	COVID-19	Non-COVID-19
Unique articles	49,421	14,135	68,642
Mean comments per article	187.2	991.6	235.3
50% quantile	45	158	63
75% quantile	164	524	216
Maximum article comments	52,135	102,246	95,845
Unique users	55,780	64,401	82,354
Mean comments per user	165.8	217.6	196.1
50% quantile	8	8	8
75% quantile	62	57	68
Maximum user comments	29,664	101,066	35,992

Table 3.1: Article and user stats from all samples

The stats of the comment's content also differ for the *COVID-19 sample*, regarding the length in terms of characters and words (note that comments are bound to a character limit of 1,500). The statistics shown in Table 3.2, indicate that users tend to write shorter comments overall compared to the other samples. It is also noteworthy that the *COVID-19 sample* contains a considerable number of comments comprising only a single word. The general prevalence of short comments could be attributed to various factors, such as the presence of social bots, which is deemed problematic on social media platforms [ZQCL22]. Future research could investigate these reasons to gain a better

	Reference	COVID-19	Non-COVID-19
Maximum characters	1,487	1,492	1,496
Mean characters	180	133.5	181.9
50 % quantile	112	74	114
75 % quantile	241	164	244
Maximum words	572	577	591
Mean words	26.3	19.8	26.8
50 % quantile	17	11	17
75 % quantile	35	24	36
Comments with one word	310,400	803,244	549,895

Table 3.2: Word and character count stats from all samples

understanding of the issue.
This overview of the sample's statistics shows that there are peculiarities in the way users engage with news content related to the topic of COVID-19. In our case study, we concentrate further on the semantic content of the comments and aim to identify potential differences in the utilization of populist rhetoric.

3.2 Annotation Sample

After examining the data, we want to draw a sample that is labelled in the annotation study and later serves as our gold standard for the evaluation of the models. For that, we first filter out short comments that we consider to feature too little semantic information be identifiable as populist concretely. Therefore, all comments containing ten or fewer words are removed, which totals to 3,128,099 posts of the *reference sample*, 6,647,285 posts of the *COVID-19 sample* and 5,370738 posts of the *non-COVID-19 sample*. To remove overtly long outlier comments, we choose to remove all comments outside the 99% quantile of the whole data, which affects posts with more than 110 words (413,270 posts).

A general problem in ML is having an imbalanced sample. This means there is a great difference in numbers between samples of the majority and minority class in our case of binary classification. Out of the whole body of comments, we consider populism to be a rare phenomenon and therefore use the dictionary of Gründl [Gr2] to pre-select potentially populist comments. To achieve this, we apply the dictionary to all samples and assess the number of dictionary terms present in each comment. This count is called score and is listed in Table 3.3. The dictionary analysis confirms the expected rarity of comments containing populism, as across all samples a maximum of 3.7% include at least one term. Within those, there are only a few comments that feature six or more

Score	Reference	COVID-19	Non-COVID-19
0	5,771,093 / 96.3%	7,081,789 / 97.5%	10,175,049 / 96.5%
1	210961 / 3.5%	172356 / 2.4%	355,125 / 3.4%
2	10,133 / <1%	6538 / <1%	17,075 / <1%
3	801 / <1%	481 / <1%	1,302 / <1%
4	104 / <1%	51 / <1%	151 / <1%
5	18 / %	15 / <1%	17 / <1%
6	3 / <1%	1 / <1%	5 / <1%
7	1 / <1%	0 / 0%	2 / <1%
8	0 / 0%	0 / 0%	0 / 0%
9	0 / 0%	0 / 0%	1 / <1%
=>10	2 / <1%	0 / 0%	1 / <1%

Table 3.3: Dictionary score for all samples (absolute/percentage)

populist words, which makes it possible to manually observe them. There is one outlier with a score of 60 that repeats the word “Schande” 60 times, which is a problem for most comments with a high score. Consequently, all posts with excessive repetition are removed, which affects every comment with a score greater than seven for the *reference sample* and nine for the *non-COVID-19 sample.*

Considering our available resources for the annotation study, we decide to draw a sample of the size of 1,200 that entails 400 comments from each of the samples to account for potential differences in populist wordings based on the context. This means that we want to be able to measure populism in the different contexts provided in the samples. To address the issue of class imbalance, we opt to randomly select half of the sample exclusively from comments that were identified as populist by the dictionary. With the assumption that more populist words result in a higher likelihood of populism, we decide to draw the sample for this purpose from all comments with a score greater than or equal to two. The other half is a completely random sample and the composition of the final annotation sample is visualized in Figure 3.3. Here we can see that we draw a sample of 200 random comments and a sample of 200 dictionary-annotated comments (with a score of two or higher) from each of the three samples to end up with the annotation sample of 1,200 comments. The random sample is drawn to account for the bias of the dictionary. Due to the infrequency of populist comments, it is probable that the randomly selected sample will comprise only a limited number of such comments. This must be considered when determining the recall of the dictionary during evaluation.

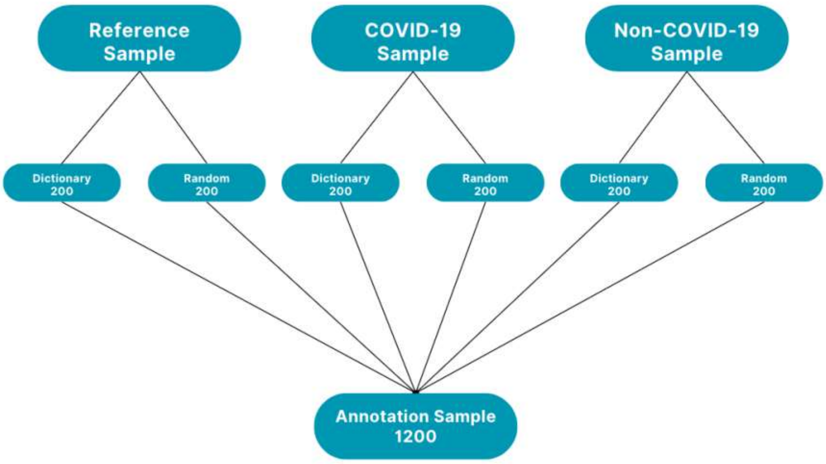


Figure 3.3: Flowchart of the sampling process (numbers indicate the number of comments included in each sub-sample)

Score	Number of comments
0	587
1	12
2	551
3	46
4	4

Table 3.4: Absolute distribution of dictionary scores for the annotation sample

The final sample's populism scores can be found in Table 3.4. There are 587 comments with a score of zero, which indicates that the random sample only includes 13 potentially populist comments with a dictionary score of one or more. The annotation study is performed on a randomly shuffled version of this prepared sample, later serving as a gold standard for the model evaluation.

3.3 Data Cleaning

Prior to conducting the experiments and case study, it is necessary to clean the data to minimize noise and computational costs, given the substantial volume of data involved. In consideration of the case study, our initial endeavor is to reduce the impact of articles with low engagement, as we want to measure the proportion of populist comments relative to the entirety of comments on a given article. Therefore we remove the 10% quantile of each of the three samples. This means removing all articles with less than four comments for the *reference sample*, less than eight for the *COVID-19 sample* and less than four for the *non-COVID-19 sample*. Another source of noise is comments that lack any semantic value, and as a result, comments without any alphabetic characters are excluded. In social media or news comment sections, responses solely comprised of emoticons like ":)" are frequently encountered, but they are irrelevant to our analysis. In our initial cleaning for the case study, we additionally exclude comments with two or fewer words, which might be increased later. After those steps, 9,202,742 posts of the *reference sample*, 12,574,910 posts of the *COVID-19 sample* and 15,187,393 posts of the *non-COVID-19 sample* remain for the case study.

In preparation for our experiments, where we intend to use the dictionary to label training data, we also implement several cleaning steps on our original samples. Those are similar to the pre-selection for the annotation sample, as we simply exclude comments with ten or fewer words and those that are already enclosed within the annotation sample. Further pre-processing steps and their influence on the model performance are subjects of the model development phase.

CHAPTER 4

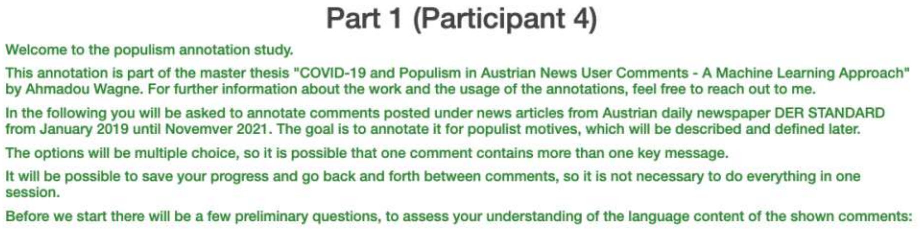
Data Annotation

A major contribution of this work is the first annotated German dataset for populism. The present chapter provides a detailed overview of the data annotation process conducted to create the gold sample. The objective of this chapter is to describe the decisions that are made during the annotation process, as well as the operationalization of the three populist motives anti-elitism, people-centrism, and people sovereignty. We also describe the measures taken to assess the reliability and validity of the annotated data set. This includes a discussion of the measurements and the quality control procedures used to ensure consistency and accuracy in the final outcome. To achieve the goal of creating a reliable and comprehensive annotated data set, a systematic and structured annotation study is conducted. Furthermore, the chapter highlights the methodology used to prepare the study, as well as the tools and criteria used for the annotation process.

4.1 Annotation Study

To collect ground truth data, a labelling study involving multiple coders is conducted to thoroughly examine each comment for any instances of populist content by multiple people. The goal is to annotate the comments for the three populist motives presented earlier. There are some scholars that argue that populism is a non-compensatory multidimensional phenomenon, which means that it requires all motives to be present [WSS20]. However, the vast majority of research reviewed in Chapter 2 sees it as sufficient if at least one of the motives is present, which is also reasonable given the brevity of the texts we are analyzing. Rather than reflecting the ideology as a whole, single statements mostly refer to certain aspects of it [EFL17, EEB^{+}17, ESH17]. Thus, the setting entails a multiple-choice selection from the three motives of anti-elitism, people-centrism, and people-sovereignty, along with a “none” option. Hawkins showed that inter-rater reliability is high for political speeches and party manifestos and often a single coder is sufficient [HS16]. Given the substantially distinct nature of our data, we aim to augment the validity of our annotation by involving external coders. With the available resources, four participants with various different academic backgrounds from our personal environment, are invited to participate in the study. The goal is that each comment gets annotated by exactly three different people. We also label all of the comments ourselves. Due to the very time-consuming nature of a labelling task, the data is split into three parts of 400 comments each. While we and one annotator review all 1,200 comments, the remaining three participants are assigned a subsample of 400 comments for annotation.

The study is carried out using the open-source tool LimeSurvey[1] on a web server. It is a useful tool for this case, as it allows the participants to simultaneously work on the labelling and save their progress with an easy-to-use web interface. Before launching the survey it is very important to establish precise annotation guidelines so that everyone has a shared concept of what populism is and how the motives can be defined. The main information is presented in text form, but the relatively small group of participants also allows it to give a quick personal briefing in advance. Additionally, they receive a few questions to collect demographic data and ask about their understanding of the content, because of the peculiarities in Austrian German compared to Standard German. In the following, the established procedure of the annotation study is explained step by step. The exact formulation and presentation is a refined version, after a test run with an additional participant that is conducted with a sample of 30 comments. The purpose of the test run is to measure the approximate time needed for the labelling and to identify any inconsistencies or pitfalls in the initial version of the guidelines.



Part 1 (Participant 4)

Welcome to the populism annotation study.

This annotation is part of the master thesis "COVID-19 and Populism in Austrian News User Comments - A Machine Learning Approach" by Ahmadou Wagne. For further information about the work and the usage of the annotations, feel free to reach out to me.

In the following you will be asked to annotate comments posted under news articles from Austrian daily newspaper DER STANDARD from January 2019 until Novemver 2021. The goal is to annotate it for populist motives, which will be described and defined later.

The options will be multiple choice, so it is possible that one comment contains more than one key message.

It will be possible to save your progress and go back and forth between comments, so it is not necessary to do everything in one session.

Before we start there will be a few preliminary questions, to assess your understanding of the language content of the shown comments:

Figure 4.1: Welcome page of the annotation study

The first page the participants see during the personal briefing is the welcome page in Figure 4.1. Here they are presented with the most important information about the study. The instructor stresses that the goal is not to focus on detecting populism in the comments but to rather detect comments that actually contain populist messages. This is done to emphasize the fact that a considerable proportion of comments are evidently not populist, and the objective is not to force any identification of populism in them. During the test run the problem occurred that the participant did not recognize it as a binary task to label populist and non-populist comments, because he was presented with four answers and therefore assumed the task is rather about finding the

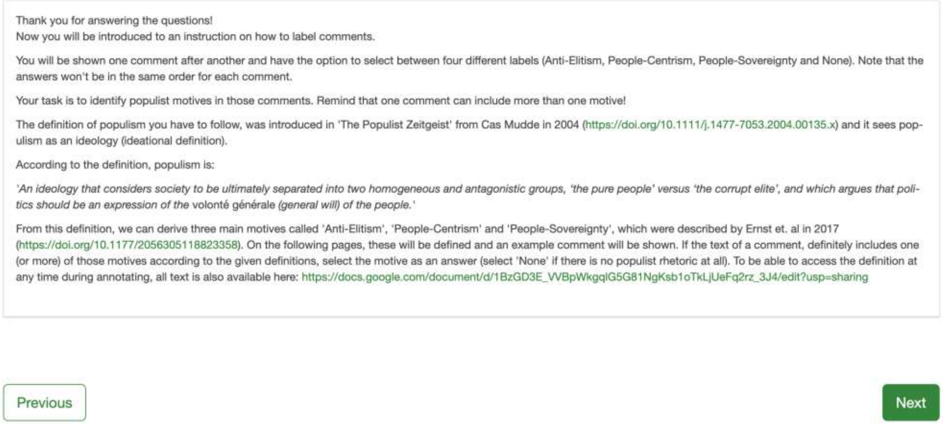
[1] https://www.limesurvey.org/de/

populist key messages in the examples. It is furthermore explicitly mentioned that it is a multiple-choice task and comments can and will often include multiple motives. The last step of the briefing includes technical details about the functionalities to save the progress and the possibility to go back and forth between questions. Another note here is that the database limitations of LimeSurvey require that the surveys are split into parts of 200 comments each, which is also explained. After the participants can start to work on the survey and have the possibility to reach out for questions at any time.

Before the demographic questions, the participants get asked about their own definition of populism, to see how they construct the phenomenon with their previous knowledge about it and to furthermore indicate that they should not incorporate their own view on populism in their labelling decisions. The following questions ask about their knowledge of the language and demographic data:

- What is your native language? (options: Austrian German, German, Other)
- How long have you been living in Austria (answer in years)? (open question)
- In which part of Austria did you live for the longest time until now? (options: list of all Austrian federal states and “I’ve never lived in Austria”)
- What is your highest degree of education (if you are currently studying, state your aspired degree)? (options: no degree, secondary school, completed vocational training, bachelor, master/diploma, higher academic degree)
- What is your industry/field of study? (open question)

This information is collected in order to be able to assess the quality of the annotations in terms of the understanding of the content and the context of Austrian news comments. Next, the actual guidelines on how to classify the content are displayed, which follow the guidelines of the work Ernst et al. [EBE$^+$19], who already conducted an annotation study on social media content, using the same definitions. In Figure 4.2 the explanations are shown. The ideational definition of Mudde [Mud04] is given as the central reference, along with a link to the publication. According to that, the three motives are introduced and the reference to the relevant publication is included so that the participants can get further information if desired. The following pages include the explanation of the motives, which can be seen in Figures 4.3 4.4 and 4.5. The participants are provided with a link to a publicly accessible Google Doc that contains identical information. This intention is to facilitate the process of labelling by enabling participants to consult the information in the Google Doc while conducting the survey, without having to revisit the introduction if any uncertainties arise. In addition, we showcase a pre-selected example comment from the *Der Standard* data for each motive, providing an impression of how content related to that motive could look like. The presentation of the motives comes with explanations in three dimensions: the populist key message that should be identified,



Thank you for answering the questions!
Now you will be introduced to an instruction on how to label comments.

You will be shown one comment after another and have the option to select between four different labels (Anti-Elitism, People-Centrism, People-Sovereignty and None). Note that the answers won't be in the same order for each comment.

Your task is to identify populist motives in those comments. Remind that one comment can include more than one motive!

The definition of populism you have to follow, was introduced in 'The Populist Zeitgeist' from Cas Mudde in 2004 (https://doi.org/10.1111/j.1477-7053.2004.00135.x) and it sees populism as an ideology (ideational definition).

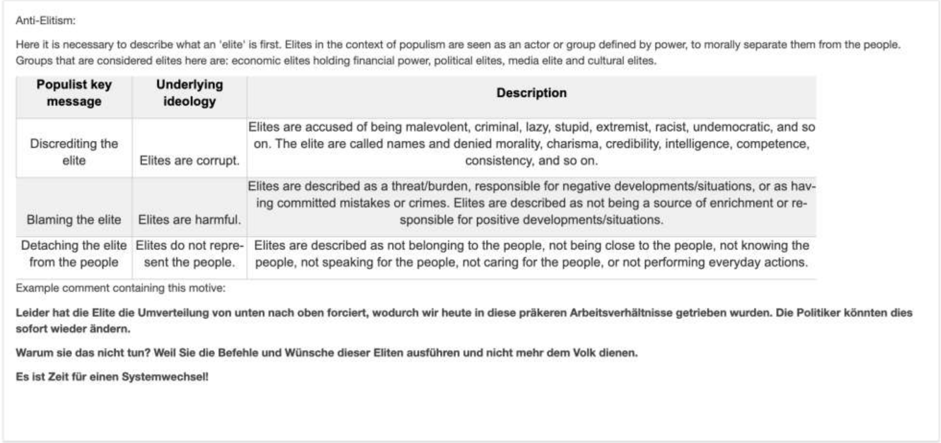
According to the definition, populism is:

'An ideology that considers society to be ultimately separated into two homogeneous and antagonistic groups, 'the pure people' versus 'the corrupt elite', and which argues that politics should be an expression of the volonté générale *(general will) of the people.'*

From this definition, we can derive three main motives called 'Anti-Elitism', 'People-Centrism' and 'People-Sovereignty', which were described by Ernst et. al in 2017 (https://doi.org/10.1177/2056305118823358). On the following pages, these will be defined and an example comment will be shown. If the text of a comment, definitely includes one (or more) of those motives according to the given definitions, select the motive as an answer (select 'None' if there is no populist rhetoric at all). To be able to access the definition at any time during annotating, all text is also available here: https://docs.google.com/document/d/1BzGD3E_VVBpWkgqlG5G81NgKsb1oTkLjUeFq2rz_3J4/edit?usp=sharing

Figure 4.2: Introduction on how to label comments in the annotation study

along with the underlying ideology of it and a description of how it can manifest in a populist text. For anti-elitism, we added further information on who or what can be seen as an elite and for people-centrism we added a note that the German translation for people is “Volk” in this context, to make sure that people are seen as a group that includes the countries population.



Anti-Elitism:

Here it is necessary to describe what an 'elite' is first. Elites in the context of populism are seen as an actor or group defined by power, to morally separate them from the people. Groups that are considered elites here are: economic elites holding financial power, political elites, media elite and cultural elites.

Populist key message	Underlying ideology	Description
Discrediting the elite	Elites are corrupt.	Elites are accused of being malevolent, criminal, lazy, stupid, extremist, racist, undemocratic, and so on. The elite are called names and denied morality, charisma, credibility, intelligence, competence, consistency, and so on.
Blaming the elite	Elites are harmful.	Elites are described as a threat/burden, responsible for negative developments/situations, or as having committed mistakes or crimes. Elites are described as not being a source of enrichment or responsible for positive developments/situations.
Detaching the elite from the people	Elites do not represent the people.	Elites are described as not belonging to the people, not being close to the people, not knowing the people, not speaking for the people, not caring for the people, or not performing everyday actions.

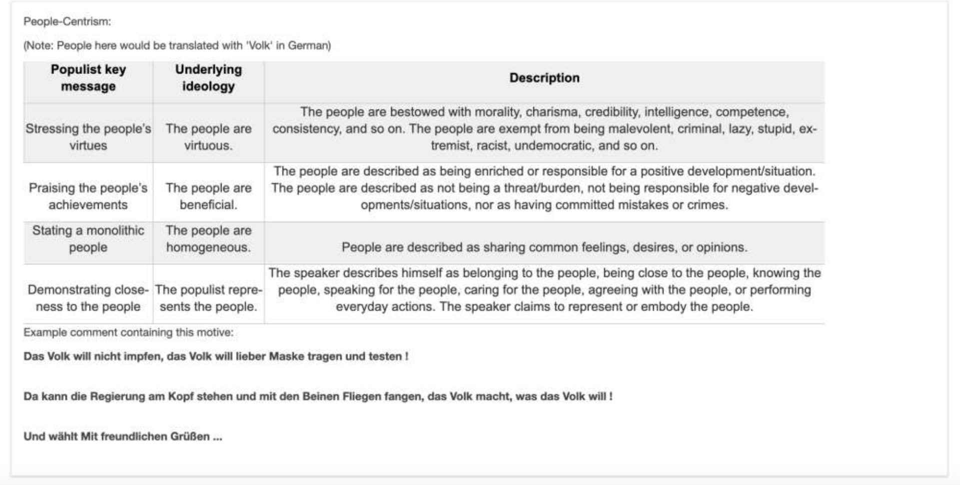
Example comment containing this motive:

Leider hat die Elite die Umverteilung von unten nach oben forciert, wodurch wir heute in diese präkeren Arbeitsverhältnisse getrieben wurden. Die Politiker könnten dies sofort wieder ändern.

Warum sie das nicht tun? Weil Sie die Befehle und Wünsche dieser Eliten ausführen und nicht mehr dem Volk dienen.

Es ist Zeit für einen Systemwechsel!

Figure 4.3: Description of the motive anti-elitism in the annotation study, including an example comment



People-Centrism:

(Note: People here would be translated with 'Volk' in German)

Populist key message	Underlying ideology	Description
Stressing the people's virtues	The people are virtuous.	The people are bestowed with morality, charisma, credibility, intelligence, competence, consistency, and so on. The people are exempt from being malevolent, criminal, lazy, stupid, extremist, racist, undemocratic, and so on.
Praising the people's achievements	The people are beneficial.	The people are described as being enriched or responsible for a positive development/situation. The people are described as not being a threat/burden, not being responsible for negative developments/situations, nor as having committed mistakes or crimes.
Stating a monolithic people	The people are homogeneous.	People are described as sharing common feelings, desires, or opinions.
Demonstrating closeness to the people	The populist represents the people.	The speaker describes himself as belonging to the people, being close to the people, knowing the people, speaking for the people, caring for the people, agreeing with the people, or performing everyday actions. The speaker claims to represent or embody the people.

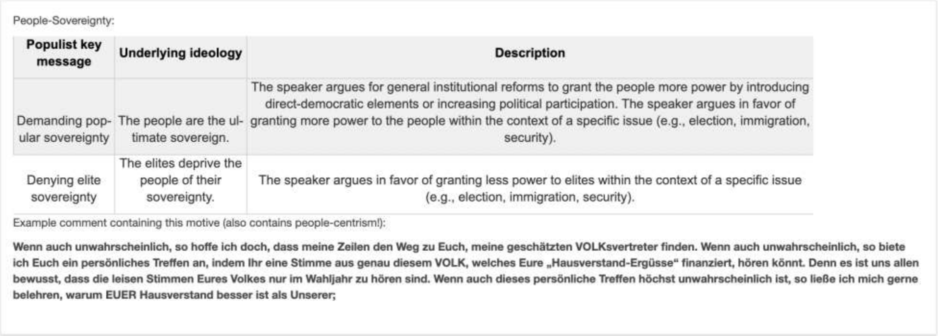
Example comment containing this motive:

Das Volk will nicht impfen, das Volk will lieber Maske tragen und testen !

Da kann die Regierung am Kopf stehen und mit den Beinen Fliegen fangen, das Volk macht, was das Volk will !

Und wählt Mit freundlichen Grüßen ...

Figure 4.4: Description of the motive people-centrism in the annotation study, including an example comment



People-Sovereignty:

Populist key message	Underlying ideology	Description
Demanding popular sovereignty	The people are the ultimate sovereign.	The speaker argues for general institutional reforms to grant the people more power by introducing direct-democratic elements or increasing political participation. The speaker argues in favor of granting more power to the people within the context of a specific issue (e.g., election, immigration, security).
Denying elite sovereignty	The elites deprive the people of their sovereignty.	The speaker argues in favor of granting less power to elites within the context of a specific issue (e.g., election, immigration, security).

Example comment containing this motive (also contains people-centrism!):

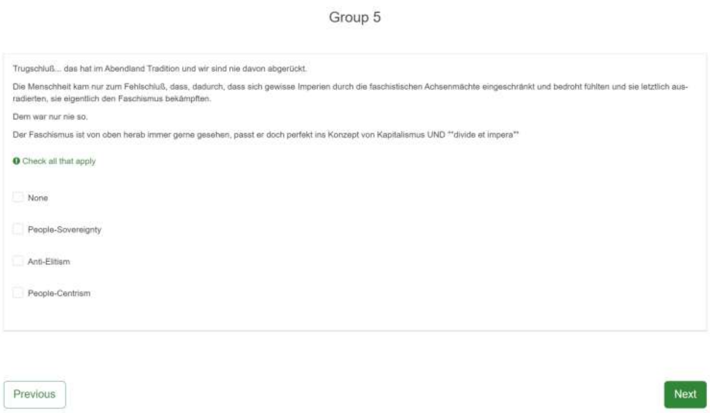
Wenn auch unwahrscheinlich, so hoffe ich doch, dass meine Zeilen den Weg zu Euch, meine geschätzten VOLKsvertreter finden. Wenn auch unwahrscheinlich, so biete ich Euch ein persönliches Treffen an, indem Ihr eine Stimme aus genau diesem VOLK, welches Eure „Hausverstand-Ergüsse“ finanziert, hören könnt. Denn es ist uns allen bewusst, dass die leisen Stimmen Eures Volkes nur im Wahljahr zu hören sind. Wenn auch dieses persönliche Treffen höchst unwahrscheinlich ist, so ließe ich mich gerne belehren, warum EUER Hausverstand besser ist als Unserer;

Figure 4.5: Description of the motive people-sovereignty in the annotation study, including an example comment

After this introduction, the participants are presented with one comment after another and can track their progress on the top (see Figure 4.6). To add further validation of the annotation guidelines and to check if the participants have a shared understanding of the concept, gold label comments are added. This means that for every batch of 400 comments, three comments are manually created following the definitions of the motives. The participants' annotations for those comments are extracted later and a 100% agreement is expected.

Each batch features one comment of every motive, which means that the following nine golden samples are inserted at a random position:

- “Die da oben mit ihrer Coronapolitik sind doch an allem Schuld!” (anti-elitism)
- “Das ist doch eh alles nur noch gesteuert von den finanziellen Eliten, die mit den Problemen von uns Normalos nichts zu tun haben.” (anti-elitism)
- “Auf das was die Medien berichten, kann man sich als Bürger nicht mehr verlassen, die plappern doch nurnoch die Politiker nach” (anti-elitism)
- “Die letzten Wochen haben doch ganz klar die Meinung des Volkes gezeigt. Ich gebe hier nur die allgemeine Stimmung wieder.” (people-centrism)
- “Diese Entwicklung ist definitiv der Bevölkerung zuzuschreiben, ich bin stolz auf alle Österreicher” (people-centrism)
- “Ich denke, ich spreche hier für alle Österreicher, wenn ich sagen, dass dieses Gesetz keiner braucht” (people-centrism)
- “Nach den letzten gescheiterten Versuchen, sollte die Bevölkerung nun doch selbst entscheiden dürfen, was sie will. Abstimmung über Lockdowns jetzt!” (people-sovereignty)
- “Das ist doch alles keine Demokratie mehr, wenn der gemeine Bürger nichts mehr zu sagen hat.” (people-sovereignty)
- “Also in Zukunft sehe ich schwarz für dieses Land, wenn die Meinung der Bürger nicht stärker ins Gewicht fällt..” (people-sovereignty)



Group 5

Trugschluß... das hat im Abendland Tradition und wir sind nie davon abgerückt.

Die Menschheit kam nur zum Fehlschluß, dass, dadurch, dass sich gewisse Imperien durch die faschistischen Achsenmächte eingeschränkt und bedroht fühlten und sie letztlich ausradierten, sie eigentlich den Faschismus bekämpften.

Dem war nur nie so.

Der Faschismus ist von oben herab immer gerne gesehen, passt er doch perfekt ins Konzept von Kapitalismus UND **divide et impera**

Check all that apply

□ None

□ People-Sovereignty

□ Anti-Elitism

□ People-Centrism

Figure 4.6: Example of a comment shown in the annotation study

4.2 Measurements

After conducting the annotation study, all results are collected and processed. Each participant's outcome is a LimeSurvey export ".csv"-file, which is converted into a data frame containing a boolean value for every populist motive indicating whether it was found or not (none is also included). From that, a new value is created, which is our final label. This label is called *populism* and is assigned if a participant found any of these motives. We report the distribution and the agreement of all motives separately because those could be of interest for further work, but the criterion, which is important for our analysis is the agreement on the *populism* label. We measure the agreement of two or more coders by using Krippendorff's α [HK07]. α is an inter-rater reliability measure used to assess the agreement of multiple coders on different types of data (nominal in our case). Its basic assumptions rely on the concept of observed and expected agreement. Observed agreement is the proportion of actual agreement of the coders on the same class, while expected agreement is the proportion of agreement by chance. With observed disagreement as D_o and expected disagreement as D_e, Krippendorff's α in our binary case is calculated as:

$$\alpha = 1 - \frac{D_o}{D_e}$$

With observed disagreement D_o:

$$D_o = \frac{1}{N} \sum_{i=1}^{N} \sum_{j=i+1}^{N} [x_i \neq x_j]$$

Where N is the number of all coded comments and x_i is the rating assigned by the participant i. $[x_i \neq x_j]$ equals 1 if the ratings by participant i and j are different, and 0 if they are the same. It is basically the distance between the annotations i and j, which is simply 0 (agreement) or 1 (disagreement) in the binary classification scenario. The expected disagreement D_e is calculated as follows:

$$D_e = p(1 - p)$$

where p is the proportion of populist comments in the whole sample. This is applied to incorporate the probability of random disagreement based on the positive class. With this approach, α has a number of advantages over other inter-rater reliability measures. It is able to handle any number of coders, deals with different sizes of data sets and uses nominal weights to measure the importance of different categories, which is beneficial for our imbalanced sample. Hawkins and Castanho Silva [HS16] observed an agreement of $\alpha = 0.87$, which is high for such a complex phenomenon. When dealing with short, user-generated text, we do not expect such a high value. Thiele measured an agreement of $\alpha = 0.81$ for populist Facebook comments, which is very satisfactory. We additionally calculate the agreement on our self-created gold label comments separately and report those numbers to assess the accuracy of our annotation guidelines. After this, we create the final labels by majority vote, which means that a comment is populist if at least two participants detect any populist motive and subsequently agree that the comment features populist content, as the boundaries of the motives are sometimes fluent and can still be subjective.

CHAPTER 5

Model

This chapter aims to address the research questions *RQ1* (see Section 1.2) by developing a model that outperforms the SOTA for automated classification of populism in German news comments. The objective of this chapter is to describe the process of building and evaluating the model, including the approaches used to train it and the settings for the experiments. First, the two different settings of the training data are examined. Then we choose baselines that we want to improve on and illustrate their specifications. In order to maximize the performance of our model, we set up an experiment that tests different model types and manipulations on the input data. We also discuss the challenges encountered during the model-building process and the decisions made to address these challenges. Finally, we explain the measurements chosen to evaluate performance and explain the optimization process. All experiments and data manipulations are implemented in Python scripts and Jupyter Notebooks.

5.1 Training Data

We already presented the approach of using a dictionary to establish a great amount of training data for sentiment analysis [SJ21, TWC08, MKNMN20]. The aim here is to adapt this for populism detection, to reduce the number of resources necessary to label a large training set. Furthermore, the expectation is that this outperforms the dictionary alone, as the transformer model learns the context of the comments. There are two things to take into consideration when using the dictionary to label data: defining a criterion that decides on the label and choosing the size of the training set. The approach to assign labels used for sentiment analysis is to count the positive and negative words for every input sample and assign a weighted score that is relative to the length of the text. Subsequently, a comment is labelled positive, neutral or negative, if it is above, between or below a pre-defined threshold of this score. When deciding on the size of the training set, the goal is to find a size where performance is high and computational cost is acceptable. There is no general rule-of-thumb of optimal sample size in ML tasks, but for this work, the decision is made to conduct the experiments with a training size of 12,000 samples, as this allows to conduct a great variety of experiments with the available computational resources. In order to account for sample size, the most effective setting is applied to a training sample of 24,000 to get an impression of the model's performance with a larger amount of data.

As the amount of available data is huge, a strategy to pre-select the training data is needed. An initial approach could be to assign a weighted score to each comment, similar to sentiment analysis, and then select the top N comments with the highest score as the most likely to contain populist language. This is problematic when dealing with a rare phenomenon and a small dictionary, as the majority of texts may contain only a single dictionary term, resulting in a bias towards shorter texts when using a ranking. Consequently, the classification strategy of Thiele [Thi22] is used, which simply labels a comment populist, if one dictionary term is present. The next problem for the pre-selection is the class imbalance. A stratified sample of size 12,000, which reflects the distribution of the original population, would only include 387 populist comments. Therefore the decision is made to oversample to an equal class distribution. To ensure that data from all three samples is included, a sample of equal size is drawn from each of them. Subsequently, there are 2,000 populist and 2,000 non-populist comments each from the *reference sample*, the *COVID-19 sample* and the *non-COVID-19 sample* in the final training data.

However, fine-tuning for a downstream task with a BERT model can also achieve good results with only a few training samples, considering the extensive pre-training. This is why we additionally use a sub-sample of 800 comments from our human-labelled data for fine-tuning. The advantage is that we have actual gold labels in our training data and exclude the false positives found by the dictionary, because of its limitations. By using a training sample of 800 comments, a split of the gold standard data into 2/3 training data, 1/6 (200 comments) validation data and 1/6 (200 comments) test data is performed. In this case, we do not use a standard 80%/10%/10% split so that we have a greater test set size, because all models are evaluated on the same set. Additionally, the final test set is desired to represent our golden sample's label distribution, which requires drawing stratified training, validation and test samples. Eventually, all experiments are conducted in the same fashion for both training sets.

5.2 Baselines

To have a comparison of the model performance, a set of baselines are used as a reference. The first baseline is the Gründl dictionary, which is applied to the test data. The model's performance is evaluated in two cases: labeling a comment as populist if it contains at least one dictionary term and labeling a comment as populist if it contains at least two dictionary terms. This dictionary has an implementation as an R-package[1], so the data is exported and loaded into R-studio[2] to count the populist terms. The second baseline is the German version of the dictionary of Rooduijn and Pauwels [RP11], which is however only tailored towards anti-elitist content. Next, the best-performing model setting of Hawkins and Castanho Silva [HS16] is used, as well as their pre-processing steps. They suggest using elastic net (EN) regression [ZH05] with the following pre-processing steps:

- lower case all text
- remove punctuation
- remove numerical characters
- remove stop words
- perform word stemming

For the stop word removal, the stop word list of the nltk Python library[3] is used and Cistem[4] is used for stemming, as it achieved SOTA performance for German data [WF17]. A few adjustments are made as we deal with user-generated content and the original approach was made for political texts. As a first step, URLs are removed before the pre-processing. Further, in the original paper, they create a document-term-matrix with a cut-off of high and low-frequency terms, which is not done here because of the small document size. The model tuning is also reproduced by tuning for the “alpha” (controls the strength of the penalty term) and the “l1_ratio” (parameter that controls the combination of L1 and L2 penalty) parameters, using 5-fold cross-validation.
For further reference, a logistic regression (LR), a support vector machine (SVM) and a random forrest (RF) classifier are implemented using the standard parameters given by the sklearn Python library[5] . These algorithms were also used as baselines for Hawkins’ and Castanho Silva’s work.

5.3 Experiment Design

In Section 2.4, we highlighted the strengths and advantages of using large pre-trained LLMs. Therefore we select existing models that are publicly available in the Hugging Face transformers library[6] and trained on German data. For those, we set up systematic experiments to address the challenges of using them on user-generated content. We

[1] https://github.com/jogrue/popdictR

[2] https://posit.co/

[3] https://www.nltk.org/search.html?q=stopwords&check_keywords=yes&area=default last accessed on 18.02.2023

[4] https://github.com/LeonieWeissweiler/CISTEM

[5] https://scikit-learn.org

[6] https://huggingface.co/

generally deal with very noisy data because user comments can not be compared with reviewed and stylistically consistent political texts. Users often quickly type their comments to express their opinion without checking for errors. Another issue is that we often do not deal with Standard German text, as Austrian German has its own peculiarities and vocabulary, which is most likely not reflected in the training data used for the BERT models. As we consider the noisy nature of the data to be the primary challenge for our model, we opt to experiment with various techniques for manipulating the input data to address this issue instead of focusing on hyper-parameter tuning. Another reason for this is that using a further grid search to find the optimal performance would blow up the number of experiments and the computational cost. The framework used for training and inference is PyTorch [PGM+19], a powerful Python library that features the implementation and customization of various deep learning architectures and enables us to speed up training by running our model on a GPU. The pre-trained model and its tokenizer are loaded using the transformers library [WDS+19], which includes tools that make it easy to use the vast amount of available resources on Hugging Face out of the box.

The general training process is the same for all experiment settings. The first step is to encode our data for the BERT model. The tokenizer is directly taken from the selected model and loaded with *BertTokenizer.from_pretrained()*. For our BERT model, this is usually an adaption of the WordPiece tokenizer [SSS+20]. The following code is used to create an encoding dictionary that features the IDs of our input sequences and the attention masks:

```
tokenizer.encode_plus(
    input_text,
    add_special_tokens=True,
    max_length=260,
    truncation=True,
    pad_to_max_length=True,
    return_attention_mask=True,
    return_tensors='pt'
)
```

The variable “input_text” is a single comment, “add_special_tokens” ensures that the special tokens (e.g. [CLS]) of the specific model are encoded, “max_length” is hard-coded here, as it is the pre-computed maximum length of our manually labelled training data, which is used as a cut-off for long sequences. The code furthermore pads shorter sequences to the maximum length by adding [PAD] tokens to have equally long sequences. Equally, longer sequences are truncated to the maximum length. The encoder returns the encoded sequences and the attention masks as PyTorch-tensors. Attention masks are used to indicate to the model which tokens in the input sequence are actual input tokens and which ones are used for padding.

For all experiments, we fix the training hyper-parameters to the recommendations of the original BERT paper [DCLT18]:

- **batch size**: the number of training examples that are processed together in each iteration of the training process. Value: **16**
- **optimizer**: algorithm that calculates the gradient of the loss function and updates parameters into the gradient's direction. Value: we use **AdamW** here instead of the regular Adam optimizer, as it adds an additional weight decay term, which prevents over-fitting [LH17].
- **learning rate**: controls the step size that the optimizer takes towards the gradient in each iteration. Value: **5e-5**
- **epsilon**: hyper-parameter for the optimizer that controls numerical stability by preventing the denominator of the adaptive learning rate update from becoming too small. Value: **1e-08**

The next step is to load the selected model as a starting point for fine-tuning. The architecture of the BERT model follows the structure shown in Section 2.4, which means that we have twelve transformer blocks ($d_{model} = 768$), followed by a pooling layer and an extra linear layer with two-dimensional output for the classification task. As our data most likely differs from the learned BERT embeddings, we choose not to freeze the parameters of the loaded model, but rather train all parameters of the pre-trained model along with our classification layer. In this fashion, each model setting is trained for a number of five epochs, which means that the model sees the entire training data five times. At the end of every epoch, the performance metrics are calculated on the 200 validation samples created in Section 4.1. We implement early stopping in the fashion that a new best-performing model on the validation set is saved if it outperforms a previous epoch. After training, the best-performing model is loaded and utilized to predict the independent test data. Because training can be unstable, especially for small sample sizes [ZWK^{+}20], we repeat each experiment five times with different random seeds (12, 23, 42, 99, 1337), calculate the mean performance for each indicator on the test set and use that to compare the different settings.

Each experiment consists of a selection of four different parameters so that each parameter combination is executed once. The chosen parameters are training data (TD), model name (MN), initial cleaning (IC) and spelling correction (SC) and are further described in the following. TD has two modes corresponding to the training set used for fine-tuning. Either the dictionary-labelled set of 12,000 comments or the human-labelled set of 800 comments is used here. MN corresponds to the name of the pre-trained model, which we fine-tune. The German adaption of the base BERT ("bert-base-german-cased", further referenced as *model 1*) is the starting point here. We further add another German transformer model called "deepset/gbert-base" (further referenced as *model 2*), which is also provided by Chan et al. [CSM20]. Utilizing the WordPiece tokenizer that draws from the pre-trained model's vast vocabulary does not require extensive pre-processing beforehand. WordPiece splits a sequence on white space and punctuation and then uses a greedy maximum-matching strategy, first checking for a perfect match of a word in the vocabulary. If there is no perfect match, the word is broken down into sub-words, which are marked with a "##", until each subpart of the word is matched, which we can see in Figure 5.1. With this strategy, BERT can handle out-of-vocabulary words, by breaking them down into smaller parts it already knows. If it is impossible to find any match, the word gets represented as an unknown ([UNK]) token.

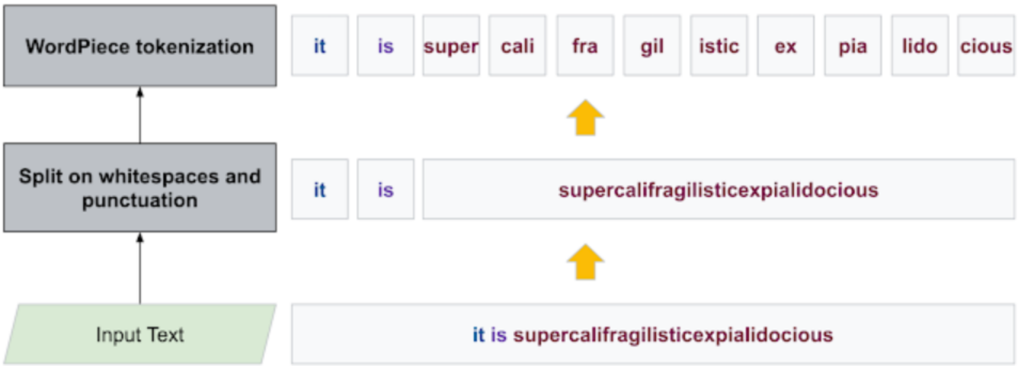

WordPiece tokenization
it
is
super
cali
fra
gil
istic
ex
pia
lido
cious
Split on whitespaces and punctuation
it
is
supercalifragilisticexpialidocious
Input Text
it is supercalifragilisticexpialidocious

Figure 5.1: Illustration of an example sentence tokenized with WordPiece from Google Research[7]

IC consequently only includes a number of basic cleaning steps and serves to reduce noise in the data. The process involves removing various elements, including HTML tags, non-ASCII characters, digits, single-letter words and multiple white spaces. Another source of noise in comments are typos and other errors on the text level. The underlying notion is to decrease the number of tokens that are not recognized (i.e. tokens resulting in an [UNK]-token) or perfectly matched (i.e. tokens that get split with ##) by BERT by rectifying possible mistakes. To address this issue, we experiment with the use of automated spell-checking. To this end, we employ a Python implementation[8] of one of the most widely used spell-checking softwares, Hunspell[9], which is used by several applications such as OpenOffice, LibreOffice and Google Chrome. Hunspell utilizes a pre-existing dictionary as a reference and generates a list of recommended spelling corrections for words that do not have an exact match, which is sorted by the probability of correctness. A great strength of the algorithm of Hunspell is that it can handle compound words, which is important for the German language in particular. The algorithm takes

[7] https://ai.googleblog.com/2021/12/a-fast-wordpiece-tokenization-system.html last accessed 17.02.2023

[8] https://pypi.org/project/hunspell/

[9] https://hunspell.github.io/

	S1	S2	S3	S4	S5	S6	S7	S8
IC	X	X	X	X				
MN	1	1	2	2	1	2	1	2
SC	X			X	X	X		

Table 5.1: Experiment setups for the model selection ("X" marks that a process is applied)

into account several factors to determine its recommendations, such as the frequency of similar words in the language, the similarity of the misspelt word to dictionary terms, and the length of the word. To deal with peculiarities of the Austrian German language, we choose the latest version of the dictionary that LibreOffice uses for this language[10]. We implement this by applying Hunnspell word by word for every sample and automatically replacing it with the most probable correction if one is suggested. A limitation of this is that the dictionary tries to match every unknown word to a known term, which might influence some actually correctly spelt neologisms, names or slang words. This could lead to semantic changes in the input sentences, which is not desired.

In total there are $4 * 4 = 16$ experiments that are compared. Every experiment has four properties, each having two possible settings (small or large training set, IC applied or not, SC applied or not and model 1 or model 2). In order to reference the setups later, Table 5.1 shows an overview of every possible experiment that is conducted for each training set and names them from setup 1 (S1) to setup 8(S8).

5.4 Measurements

During both the training and evaluation phases of a model, various metrics are reported to assess its performance. When considering the populist class as the positive class, a sample can be categorized as true positive (TP), false positive (FP), true negative (TN), or false negative (FN) depending on its true label and the assigned prediction by the model. These categories allow us to calculate the following performance indicators for a binary classification task:

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

[10] https://extensions.libreoffice.org/en/extensions/show/german-de-at-frami-dictionaries last accessed 30.01.2023

$$\text{F1-score} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

Accuracy is a widely used performance metric in classification tasks, which quantifies the percentage of correctly predicted samples for both classes. However, this metric can be misinterpreted or less meaningful when dealing with imbalanced data sets. In such cases, even if the model predicts only the majority class, the accuracy could still be quite high. Because we are more interested in the minority class, precision, recall, and F1-score are introduced, which are commonly used evaluation metrics that originated from information retrieval. In our study, we specifically measure these scores for the minority class. Recall is a metric that measures how many of the actual populist samples are correctly predicted by our model. However, in our study, the recall score needs to be interpreted carefully. This is because we use a dictionary to create parts of the golden sample, and we expect the random half of the sample to only include a few populist comments. Therefore, recall scores obtained from the dictionary and the models trained on the dictionary-labelled data are expected to generally have a high recall. On the other hand, precision is used to measure the accuracy of our prediction for only the populist comments. This metric quantifies how many of the predicted populist comments actually have a populist ground truth label. Our primary objective is to outperform the dictionary, particularly regarding precision. This would indicate that we can reduce the number of false positives resulting from the limitations of the dictionary regarding for example negation or context. As we aim to maintain a high recall while improving precision, we use the F1-score as the primary performance metric to answer *RQ1*. The F1-score is a harmonic mean of precision and recall and provides a balanced measure of both metrics. We optimize our models to achieve the highest F1-score, which serves as the decisive performance criterion.

CHAPTER 6

Case Study

This chapter gives a brief overview of the model application and the procedure of the case study to answer the remaining research questions *RQ2a*, *RQ2b* and *RQ3* (see Section 1.2). Section 6.1 describes how the results of the model are processed to create the dependent and explanatory variables of our tests. This is followed by an introduction of the methods and measures used to answer the research questions in Section 6.2.

6.1 Study Setup

The outcome of the model phase is an additional value that stores the populism prediction for each comment as a binary variable. Our dependent variable for the statistical tests is the number of populist comments per article. This accounts for the different lengths of the covered time spans and the difference in the number of articles between our samples. Therefore, the number of all populist comments under an article is summed up to an article's populism score. The explanatory variable for *RQ2a* and *RQ2b* is the sample the comments originate from, as we want to examine the difference in the number of populist comments between 2019 and the pandemic, as well as between COVID-19-related and other articles posted during the pandemic. For *RQ3*, our explanatory variable is a running count on the day an article is posted, starting with the value zero for March 11 2020, the day the WHO declared the global pandemic[1]. Here, we only use the *COVID-19 sample*, as we want to investigate if the number of populist comments under COVID-19-related articles rise over the course of the pandemic. For *RQ2a*, the *reference sample* is compared to the combined sample consisting of the *COVID-19 sample* and the *non-COVID-19 sample*, to represent the overall time during the pandemic. *RQ2b* specifically focuses on comments made during the pandemic. Therefore, the comparison is made between the

[1] https://www.who.int/director-general/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19—11-march-2020 last accessed 23.02.2023

COVID-19 sample and the *non-COVID-19 sample*. The relevant samples and comparisons for each research questions are visualized in Figure 6.1. The explanatory variables for the comparisons of the samples are turned into binary class labels to make them operable. This means that the positive class "1" indicates a comment was posted during the pandemic for *RQ2a* and that it was posted under a COVID-19-related article for *RQ2b*.

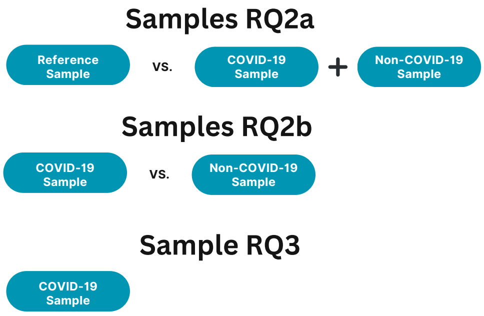


Figure 6.1: Visualization of the samples used and compared in the research questions

Before answering the research questions, we perform a descriptive analysis of the number of populist comments and observe their development over time. Moreover, we utilize the Wikipedia entry outlining the chronology of the pandemic in Austria[2] as a reference to identify a series of critical events that occurred during the crisis, such as the declaration of anti-COVID-19 measures or implementation of lockdowns. To investigate the prevalence of populist comments in relation to these events, we examine the day of the event, as well as the five preceding and following days.

6.2 Measurements

In order to assess the impact of our explanatory variables on the number of populist comments, we conduct the same statistical tests used in Thiele's [Thi22] work, which serves as a point of comparison for our findings. However, we extend Thiele's studies by examining a more extensive timeframe and introducing reference data from before the pandemic. Thiele uses negative binomial regression models to test his hypotheses [Hil11]. Negative binomial regression is a statistical model used especially for count data and is widely used in social sciences to model counts of behaviors or events. It extends the Poisson regression model and accounts for overdispersion, which occurs when

[2] https://de.wikipedia.org/wiki/COVID-19-Pandemie_in_Österreich last accessed 23.02.2023

the variance of the count is greater than the mean, which means that there might be more variation in the data than a Poisson distribution would expect. The assumption of the model is that the populism count follows a negative binomial distribution, which is described by two parameters: the mean and the dispersion, where dispersion measures the degree of overdispersion in the data. Therefore the model estimates the effect of our explanatory variable (sample or day count) on the number of populist comments while accounting for overdispersion in the data. The resulting coefficient can be either positive or negative, depending on the influence of the chosen explanatory variable and serves to answer our research questions. The implementation of the regression model is done with the statsmodels Python package[3]. The exact model is a generalized linear model[4] with "statsmodel.api.families.NegativeBinomial()" as a parameter for the family of the model. We additionally add a constant intercept term to the model and include the logarithm of the number of total comments of an article as an offset to control for the difference in the comment count between the samples [Hil11]. The latter decision goes with the assumption that more total comments lead to more populist comments and is done so that the results of our model are not caused by the higher amount of comments per article, we observed for the *COVID-19 sample* in Section 3.1.

[3] https://www.statsmodels.org/stable/index.html

[4] https://www.statsmodels.org/dev/glm.html

CHAPTER 7

Results and Discussion

This chapter presents the primary findings of the research and provides an analysis of the results. Here, we collect and present the results of the studies and experiments and discuss them to draw implications of our work. Section 7.1 provides an overview of the results of the annotation study and the resulting golden sample for populist news user comments. In the following Section 7.2, we address research question *RQ1* by presenting the results of the model selection phase of our work. We showcase the best-performing model and apply it to the *Der Standard* data. Section 7.3 starts with an overview and a descriptive analysis of the detected populist comments. We then delve into addressing research questions *RQ2a*, *RQ2b*, and *RQ3* by presenting the results of the case study. Each section concludes with a discussion of the findings, which includes a comparison to previous work for the case study in order to interpret the results.

7.1 Annotation Study

This section presents the results of the annotation study and gives insight into the established gold label dataset. Moreover, the results are discussed to draw implications and highlight potential shortcomings.

7.1.1 Annotation Evaluation

In the annotation phase, we collect the labels assigned by all participants to form the golden sample. In total, we report the results of five different coders. In the following, the author is called *participant 0*, the first participant, who also annotated all 1,200 comments, is called *participant 1* and the other participants, who annotated 400 comments each are called *participants 2-4*. We report demographic data collectively and not separately for every participant for privacy reasons. The first question about their personal perception

	Participant 0	Participant 1	Participant 2
Anti-elitism	90	102	76
People-centrism	18	13	32
People-sovereignty	6	10	6
None	302	288	308
Populism	98	112	92

Table 7.1: Distribution of labels for the first batch of 400 comments in the annotation study

	Participant 0	Participant 1	Participant 3
Anti-elitism	95	100	109
People-centrism	19	16	22
People-sovereignty	8	11	7
None	296	290	284
Populism	104	110	116

Table 7.2: Distribution of labels for the second batch of 400 comments in the annotation study

of what populism is gives very interesting insights. The participants mention the conflict between in- and out-groups and describe populism as a simplified worldview with no political content. Furthermore, they highlight the people to be at the centre of the ideology and describe populists as opportunists that put themselves in the position of speaking for the people. Some of the participants explicitly separate left- and right-wing populism. In general, we can find each of our populist motives in the answers and no participant had a completely different definition in advance. Concerning their native language, three participants indicate that they speak German as their mother tongue, whereas one participant reports Austrian German as their native language. One of the participants has never lived in Austria, while the others state that they have lived there for two, six and 25 years and they lived in Vienna or Upper Austria for the most part. The participants across a range of academic domains, including education, veterinary medicine, and international business administration, all report a Master's or Diploma degree as their highest (intended) educational level. The demographic questions reveal that the participants have a strong academic background and exhibit proficiency in the German language, along with a familiarity with the cultural context of Austria for most of them.

Due to the fact that three participants only labelled a batch of 400 comments each, we report the distribution of the given labels separately for those batches in Tables 7.1, 7.2 and 7.3.

	Participant 0	Participant 1	Participant 4
Anti-elitism	87	97	104
People-centrism	14	10	14
People-sovereignty	9	8	9
None	304	299	286
Populism	96	101	114

Table 7.3: Distribution of labels for the third batch of 400 comments in the annotation study

	Agreement
Anti-elitism	0.79
People-centrism	0.54
People-sovereignty	0.72
Populism	0.79

Table 7.4: Krippendorff's α for all participants of the annotation study across all labels

The tables provide an overview of all potential responses and the corresponding populism label assigned upon the detection of at least one motive. The prevalence of *anti-elitism* is prominent among comments categorized as populist by all annotators. Additionally, the results indicate a strong class imbalance, which is noticeable due to the pre-selection of 50% of the sample with the populism dictionary. The agreement of all participants in Table 7.4 shows satisfactory results for our populist label. An α value of 0.79 comes close to the results of Thiele [Thi22] (0.81), which indicates that one annotator may be sufficient for this labelling task. Notably, the highest agreement can be observed for anti-elitism. In contrast, agreement on people-centrism is lower, potentially due to the subjective boundaries between people-centrism and people-sovereignty, as well as the multiple-choice format of the task. The high agreement on anti-elitism suggests that participants share a common understanding of who qualifies as an elite. Based on the high agreement, we use a majority vote and report the resulting label distributions for the golden sample and sub-samples used in the modelling phase in Table 7.5. As expected, the results confirm a high class imbalance, with only approximately 24.8% of comments in the golden sample classified as populist. Additionally, the assumption that the dictionary would have a high recall based on the sampling process is supported by the fact that there are only 30 comments marked as populist in the random half of the sample, which were not detected by the dictionary.

Lastly, we evaluate the agreement on the manually created gold standard samples placed in the annotation study. Here we can observe perfect agreement on each sample (i.e. $\alpha = 1$), further validating the annotation guidelines' quality. An interesting addition

	Full	Training	Validation	Test
Populist	297	198	50	49
None	903	602	350	351

Table 7.5: Gold label distribution of the whole sample and the training, validation and test split

here is that for the gold standard comment "Das ist doch eh alles nur noch gesteuert von den finanziellen Eliten, die mit den Problemen von uns Normalos nichts zu tun haben." which was an example of *anti-elitism*, both annotators added a *people-centrism* label, which is reasonable when revisiting the sample.

7.1.2 Discussion

With the established annotation guidelines, the annotation study reaches a high agreement and we can therefore confidently use the sample to evaluate the ML models. After the annotation study, the participants are asked about inconsistencies or problems they encountered, to consider this for future work. One factor that can lead to a lack of confidence in assigning labels to comments is the challenge of determining whether the mentioned individual or group qualifies as "elite" in power. It is important to note that the political context at the time a comment was posted, or the context of the time to which the comment refers, holds significant relevance to making accurate decisions. The participants state that they sometimes had to perform an external search if specific people were mentioned. Another "issue" they mention is that they sometimes could not distinguish clearly between the three motives, which is not a problem, because they are not at all mutually exclusive. However, regarding the high level of inter-rater agreement and unanimous consensus on the golden samples, the guidelines and resulting dataset hold potential to improve future research endeavors.

Regarding the distribution of the motives in the annotation study it is noticeable that anti-elitism is a lot more prevalent then all other motives. We furthermore observe the highest agreement on anti-elitism. This could be an indicator that the "pandemic populism" does rather focus on blaming the experts and politicians that are held responsible for the countermeasures that drastically changed social life. Investigating the identities of the elites referenced in the comments and integrating this information into the prediction of populist comments may present a potential improvement for the future. Additionally, upon qualitative examination of the data, a notable prevalence of in-group versus out-group dynamics emerges. Users frequently classify individuals into those who adhere to and trust governmental measures versus those who do not, as well as vaccinated versus unvaccinated individuals. This motive is not represented in the minimalistic definition and might be a characteristic of the "pandemic populism".

7.2 Model Performance

In this section, we compare the different models to our baselines to address *RQ1* and choose the best model for the case study. The report is based on the measurements outlined in Section 5.4 and compares the small human-annotated and the larger dictionary-annotated training samples. Furthermore, all results are reported for the performance on the same test set of 200 comments.

7.2.1 Baseline Performance

Our baselines consist of two different kinds of approaches: dictionary-based methods and ML methods. The scores of the dictionary-based methods are listed in Table 7.6. The number behind the dictionary name, indicates how many dictionary terms are required to assign a populist label. The Gründl-dictionary has an expected high recall of 0.88 and only accounts for seven false negatives, which is highlighted in the confusion matrix in Figure 7.1. There is only a slight difference in the thresholds, with only two test samples having a different classification outcome when assigning labels to comments with a dictionary score of two or more, as opposed to using a threshold of one. This is the case because half of the annotated sample is drawn using comments that feature at least two dictionary terms. The low precision of 0.44 and 0.45 is not surprising, as the results of the annotation study already showed that there are only less than 300 samples where the participants actually agreed on the populist label of the dictionary. R&P in Table 7.6 references the dictionary by Rooduijn and Pauwels. Its performance declines when using a higher threshold, which is most likely caused by the small size of the dictionary (20 words). Furthermore, it only has a very low recall of 0.27 as it is solely designed to capture anti-elitism. It only outperforms the Gründl-dictionary in regards to accuracy and precision. The higher precision is noticeable because this means that the R&P-dictionary is slightly more often correct when it assigns a populist label. Nevertheless, the Gründl-dictionary is the best-performing dictionary regarding the F1-score.

	Gründl 1	Gründl 2	R&P 1	R&P 2
Accuracy	0.69	0.7	**0.77**	**0.77**
Precision	0.44	0.45	**0.53**	0.52
Recall	**0.88**	**0.88**	0.51	0.27
F1	**0.59**	**0.59**	0.52	0.36

Table 7.6: Accuracy, precision, recall and F1-score of the dictionary-based methods (best performance for each score is written in bold font)

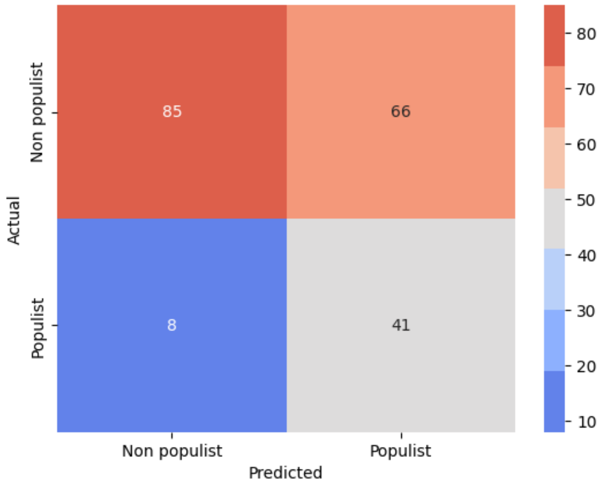


Figure 7.1: Confusion matrix for the Gründl-dictionary with a threshold of two

The remaining baselines are the ML-based approaches introduced in Section 5.2. For the elastic net, which was the best-performing model in the study of Hawkins and Castanho Silva [HS16], the parameter-tuning using five-fold cross-validation results in “alpha” = 0.0001 and “l1_ratio” = 0.4 for the model using a training size of 12,000 and “alpha” = 0.001 and “l1_ratio” = 0.2 for a training size of 800. Table 7.7 shows the scores of the EN, LR, SVM and RF classifiers on the large training set. We can generally see a similar performance of all models, with the RF performing best overall. The precision of each algorithm is lower than the precision of the Gründl dictionary, while the RF surprisingly outperforms in recall. When looking at Table 7.8, which reports the performance with the small training set, we can see that the training size is insufficient to perform satisfactorily for those algorithms because they only find a small number of actual populist comments. However, they tend to be more precise in the few populist predictions they make. As F1 is our decisive performance criterion, we report the Gründl dictionary as our best-performing baseline, which we want to improve on.

	EN	LR	SVM	RF
Accuracy	**0.64**	0.61	0.63	0.63
Precision	0.38	0.37	0.38	**0.39**
Recall	0.82	0.8	0.82	**0.9**
F1	0.52	0.5	0.52	**0.54**

Table 7.7: Accuracy, precision, recall and F1-score of the ML models using a training size of 12,000 (best performance for each score is written in bold font)

	EN	LR	SVM	RF
Accuracy	0.77	**0.81**	0.74	0.77
Precision	0.54	**0.92**	0.46	0.58
Recall	**0.39**	0.22	0.33	0.22
F1	**0.45**	0.36	0.38	0.32

Table 7.8: Accuracy, precision, recall and F1-score of the ML models using a training size of 800 (best performance for each score is written in bold font)

7.2.2 Model Experiments

To outperform the baseline models, we conduct the experiments outlined in Section 5.3 on a NVIDIA GTX 1080 TI GPU. The reported values are the mean across five runs for every experiment setup. The main goal of the experiments is to observe the influence of the different training sets and pre-processing steps on the model performance. We introduced IC and SC as pre-processing steps to reduce the amount of unknown and separated tokens. Tables 7.9 and 7.10 show the effects of applying those steps. IC can remove a substantial number of unknown tokens for both sets, while SC has a greater impact on the number of separated tokens. Performing cleaning, followed by correcting spelling errors, eliminates all unknown tokens and leads to a substantially higher number of perfect matches with the BERT dictionary than using either of these techniques individually.

	No	IC	SC	Both
UNK	165	6	93	0
##	174,224	171,755	169,058	158,896

Table 7.9: Effects of pre-processing on the number of (exact) dictionary matches of BERT for the large training sample

	No	IC	SC	Both
UNK	61	9	46	0
##	16,484	16,505	16,059	15,356

Table 7.10: Effects of pre-processing on the number of (exact) dictionary matches of BERT for the small training sample

The models converged within the five training epochs for both datasets in all cases. The results regarding the chosen performance measurements can be found in Table 7.11 and Table 7.12 for the large and small datasets separately. The proposed approach of using the dictionary to annotate a larger training sample for transformer fine-tuning gives surprisingly bad results when evaluating it on the human-annotated sample. The assumption that we can overcome the limitations of the purely dictionary-based method is not reflected in the results of the experiments. All models reach relatively similar results, meaning that the pre-processing steps have no observable effect on the model performance with this sample size. Table 7.11 further shows no increased performance, when doubling the sample size to 24,000, which meets our assumptions. The Gründl-dictionary outperforms the best experiment run in every reported measurement, besides recall where they are equal. The tendency of the results with a high recall and a low precision resembles the values achieved by the dictionary.

	S1	S2	S3	S4	S5	S6	S7	S8	24,000
Accuracy	**0.63**	**0.63**	**0.63**	**0.63**	**0.63**	**0.63**	**0.63**	**0.63**	0.62
Precision	0.38	**0.39**	**0.39**	**0.39**	**0.39**	**0.39**	**0.39**	0.38	0.38
Recall	0.87	0.86	**0.88**	0.86	0.87	0.87	0.87	0.87	0.87
F1	0.53	0.53	**0.54**	0.53	**0.54**	**0.54**	**0.54**	0.53	0.53

Table 7.11: Results of all experiment setups for the large training sample, including a run with S3 with an increased training size of 24,000 (best performance for each score is written in bold font)

	S1	S2	S3	S4	S5	S6	S7	S8
Accuracy	0.81	0.8	0.79	0.76	0.81	**0.82**	0.79	**0.82**
Precision	0.62	0.57	0.56	0.55	0.61	**0.63**	0.56	0.61
Recall	0.64	**0.79**	**0.79**	0.68	0.74	0.61	0.73	0.64
F1	0.61	**0.66**	0.65	0.56	**0.66**	0.62	0.62	0.62

Table 7.12: Results of all experiment setups for the small training sample (best performance for each score is written in bold font)

Using only human-annotated data, we could improve our results with the small sample. Opposed to the ML baseline models, BERT manages to reach a considerably high recall with a small training set. The experiments with the small sample are furthermore sensitive to the manipulation of the input comments, which is visualized in Figure 7.2 showing a comparison to the stable performance values of the larger set. The F1-scores range from 0.56 to 0.66, with the setup using model 2, IC and SC performing the worst and model 1 with IC and no SC performing best. Concerning IC (included in setup S1, S2, S3 and S4), we observe an increase in average recall of 5% compared to the setups, where it is not applied, while average precision declines by 2%. SC (included in S1, S4, S5, S6) lowers the average recall by 7% and increases precision by 2%. The model choice results in an increased average recall of 5% when choosing model 1 (included in setup S1, S2, S5 and S7) over model 2, with equal average precision.

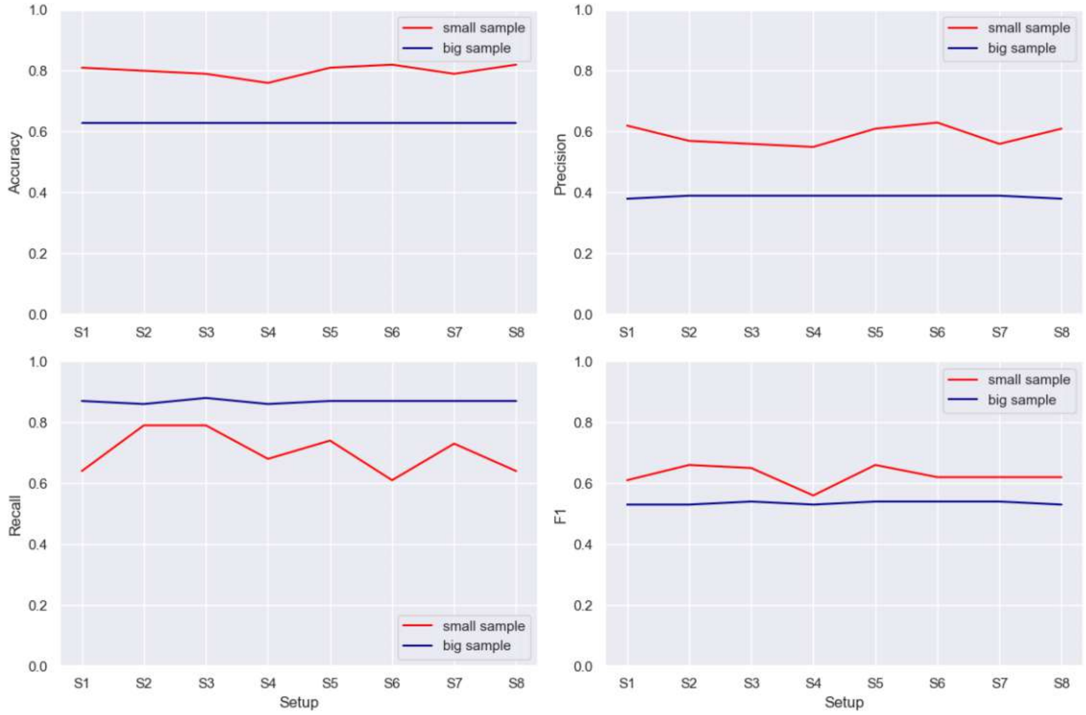


Figure 7.2: Comparison of performance metrics for both training sets and all experimental setups

	Baseline	Best large	Best small	Single best
Accuracy	0.7	0.63 (-0.07)	0.8 (+0.1)	**0.84 (+0.14)**
Precision	0.45	0.39 (-0.06)	0.57 (+0.12)	**0.62 (+0.17)**
Recall	**0.88**	**0.88 (+/-0)**	0.79 (-0.09)	0.86 (-0.02)
F1	0.59	0.54 (-0.05)	0.66 (+0.07)	**0.72 (+0.13)**

Table 7.13: Comparison of the average best-performing setups for the BERT models and the single best-performing model to the baseline (difference compared to the baseline in brackets; best performance for each score written in bold font)

Although our decisive criterion, the F1-score, is equal for S2 and S5, we opt to select S2 as the superior model due to two factors: first, S2 reaches a maximum recall score across all models and second, S5 includes SC, which is a resource-intensive operation when applying it on the full samples in the case study.

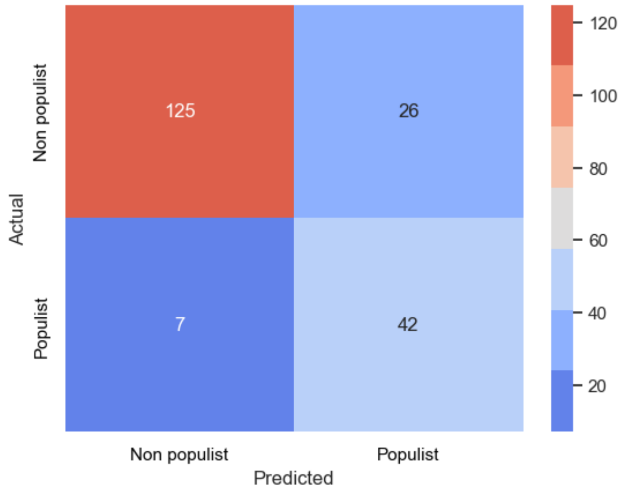


Figure 7.3: Confusion matrix for the single best-performing model

7.2.3 Discussion

Upon revisiting our first research question,

> **RQ1**: What is an appropriate ML model to improve the detection of populism in Austrian news comments?

we can say that our best-performing setup using human-annotated training data, model 1 and IC can be considered an appropriate model to improve the SOTA of populism detection in Austrian news comments. We reach the goal of improving on the precision of the dictionary-based method by 17%, while almost matching its recall, which results in an improved F1-score by 13% that we can see in the comparison in Table 7.13. As early stopping is used to save the best model of all experiment runs, we can utilize the single best-performing model of S2 for the case study. Figure 7.3 shows the confusion-matrix of the selected model, which confirms that the predictions of the BERT-based model are remarkably more precise, as we can lower the false positives of the dictionary from 66 to only 26.

The results of our experiments give interesting insights into the operationalization of the research objectives. Concerning the baseline algorithms, we can confirm the assumption that the dictionary based SOTA method has its limitations in regard to precision and produces a high number of false positives when evaluated on human-annotated data. The dictionary only measures word occurrences and does not weight them, which can cause inconsistencies when applying it to short texts. The dictionary includes terms like “arrogant” or “Irrsinn” (German for insanity), which might increase the probability of a text featuring populist content, but are often used in very different contexts. The baseline results already indicated that ML algorithms using human-annotated might be more precise in the populist predictions they make, but they only predicted a small number of comments as populist, which is why they perform poorly in terms of recall. The results confirm the assumption that the Gründl-dictionary is the baseline we want to improve on.

Our experimental findings indicate that the proposed approach of combining the dictionary and machine learning to decrease the expense of manually annotating a large amount of training data is not viable for populism detection in our particular case, unlike the outcomes obtained for sentiment analysis. This is surprising because even SOTA LLMs that are capable of understanding the context of sequences and relations between words can not derive rich enough features to outperform the dictionary's precision. This suggests that the models could still concentrate on the similarities among the comments with respect to the occurrence of the dictionary terms. Moreover, it has the disadvantage of learning from the false positives in the automatically labelled training data. The experiments involving BERT have demonstrated that the recall performance of the model reaches its threshold at the dictionary level and is even more imprecise.

The experiments with the small sample however could meet our expectations of outperforming the dictionary in terms of precision and subsequently F1-score. This is an interesting finding, as it was achieved with only 800 samples for BERT fine-tuning. The transformer architecture strongly outperforms the other ML methods that served as a baseline and is able to better learn the underlying patterns of the complex phenomenon that populism is. As expected, the method can not match the recall of the dictionary, as the dictionary supported the sample drawing process, but we still retain a high level of recall, which is seen as a success. The improvement in precision can be seen as an improvement in detecting non-populist comments and subsequently the ability to distinguish them better from actual populist comments. especially when compared to the performance of human annotators that reach an agreement of 0.79 on the task.

Populism detection is still a research field that is at an early stage, which is why data is lacking for example, but our findings suggest high potential in the use of the methods data science offers. Although our experiments did not identify concrete patterns in increasing the performance by manipulating the input samples to augment the number of BERT dictionary matches, they support the approach of performing only minimal pre-processing of the data prior to feeding it into a large pre-trained LLM. This is reflected in the observation that for both datasets, the setup of only applying IC reached the most favorable results.

A qualitative analysis of a fraction of the predictions on the *Der Standard* dataset, backs up our decision to use the best-performing BERT model for the case study. The analysis demonstrates that the model is able to detect comments that we would classify as populist based on our operationalization, for all three motives, even though our gold labels included only a limited number of comments containing *people-centrism* or *people-sovereignty*. Concluding this discussion, we therefore show an example comment detected by the model for each motive:

- **Anti-elitism**: "Sie sprechen mir aus der Seele. Allerdings wurde das schon lange vorher abgeschafft. Schon beim Klimawandel wurde nur einer bestimmten Lobbyisten Gruppe hinterhergehirscht. Das Klimathema ist jetzt durch, dafür werden die Leute in Bälde die wirtschaftlichen Folgen ihrer Forderungen genießen dürfen. Ein gewisser Herr Drosten hat schon vor Jahren eine Panik verbreitet, die sich als völlig haltlos erwiesen hat. Jetzt fordert er mehr oder weniger solche Dinge wie Abschaffung der Zulassung von Impfstoffen wurde ja auch völlig überbewertet. Und solchen Leuten rennt man blind hinterher, jegliche wissenschaftliche Diskussion unmöglich ."
- **People-centrism**: "Ich denke ich spreche für alle hier wenn ich sage wir wünschen uns eine Veränderung in der politischen Regierungslandschaft in Österreich. Weg von systematischer Korruption und Freunderlwirtschaft, hin zu einem Neubeginn!! Ich will einen NEUBEGINN!! Scorpions Wind Of Change."
- **People-sovereignty**: "Covid ist verhältnismäßig ungefährlich. maximal sterben daran, selbst Personen, die mit einer erkrankten Person im Haushalt leben, erkranken oft nicht und die Entwicklung eines Impfstoffs war relativ einfach, wa man an der Vielzahl entwickelter Impfstoffe innerhalb eines Jahres erkennen kann. Die Infrastruktur, die Funktionsfähigkeit des Staates war zu keinem Zeitpunkt ernsthaft gefährdet. Einzig die Intensivstationen waren gut gefüllt, für einen kurzen Zeitraum. Umfragen sind kein demokratisches Gremium. Eine Volksabstimmung wäre eine demokratische Entscheidung."

7.3 Model Application

To address the remaining research questions, this section presents the results obtained by applying the best-performing model to the *Der Standard* data. First, we perform descriptive analysis to observe the populism statistics of all our samples and highlight the development of populist comments over time. Following, the amount of populist comments per article are compared across our samples to draw implications about the connection between populism and the COVID-19 pandemic and investigate the development of populism over the course of the pandemic. Eventually, we discuss the results, compare it to previous studies and draw conclusions of our findings.

7.3.1 Descriptive Analysis Populist Comments

Following our model selection, we use a PyTorch script to perform inference on the *reference sample*, *COVID-19 sample* and the *non-COVID-19 sample* on the GPU. We process the samples in batches of 100,00 comments to reduce memory consumption and create save states of our predictions after every batch. After a runtime of approximately five days, each comment is assigned with a binary populism label. Table 7.14 shows the total amount of comments in each sample, along with all comments that the model predicts as populist. In our data analysis in Chapter 3, we already showed that the dictionary detects the lowest ratio between total comments and populist comments in the *COVID-19 sample*, which was surprising but is supported by the model's decision. However, the numbers that are relevant for our research questions are the number of populist comments per article, which we can see in Table 7.15. Here the statistic is dominated by the articles related to the topic of COVID-19, with a mean populist comment count per article that is almost three times higher than in the reference. Nevertheless, these numbers do not yet take the total amount of comments of an article into account. Noticeably, the top three articles that attracted the highest number of populist comments (8,862, 7,520, 4,377) are all articles that report on the recent COVID-19 cases and were released in April 2021, September 2021 and September 2020. The reported standard deviation highlights the overdispersion of the data, which is the reason for the choice of a negative binomial regression model.

	Reference	COVID-19	Non-COVID-19
Populist	1,352,281	959,022	2,056,663
Total	9,201,455	12,572,785	15,186,329

Table 7.14: Number of populist comments and total number of comments for each sample

	Reference	COVID-19	Non-COVID-19
Populist comments per article	27.2	76.2	33.7
Standard deviation	87.7	174.9	91.2

Table 7.15: Mean number and standard deviation of populist comments per article for each sample

To get an impression of the events that nurtured the number of populist comments, Figure 7.4 shows the development of this number aggregated per week for 2019, 2020 and 2021 (the 2021 data ends with November, as there is no further data available). We can moreover follow the four different waves of the pandemic in Austria, with the fourth still ongoing after the cut-off of our data. In 2019 we can see a major peak in the number of populist comments in May, which most likely corresponds to the reaction to the events following the publication of the video of former Austrian vice-chancellor Heinz-Christian Strache and deputy leader Johann Gudenus[1] that sparked the so-called “Ibiza affair”[2]. This was a highly relevant and controversial topic throughout the course of the year and is furthermore reflected in the peak reached at the end of September and beginning of October, where Sebastian Kurz was re-elected as chancellor[3] after the preceding dissolution of the Austrian National Council. Those events must be considered when using the 2019 data as a reference.

The following year shows the start of the pandemic. Before March, the topic started to emerge but was not highly covered in the news. When COVID-19 reached Austria at the beginning of March and people were directly influenced by the virus and the first countermeasures, we can observe a peak in the number of populist comments under COVID-19-related articles and an all-time low for the populist comments under all other articles. This indicates that the general reporting and the interactions of the users focused on the topic of COVID-19, as the majority of the total amount of comments during this time were posted under COVID-19 articles as well. By the peak’s prominence, we can say that the start of the pandemic saw a rise in populist comments posted by the users of the *Der Standard* news forum. During the months of March and April, there was a rise in the number of articles and user comments related to COVID-19, as compared to other topics. Out of a total of 7,086 articles, a significant proportion of 2,758 were related to COVID-19, which represents approximately 39% of all articles. Remarkably, COVID-19-related articles drew 2,115,926 user comments, while all other articles combined only attracted 620,841 comments in total during this timespan. After that, the distribution of the engagement shifted and the weekly aggregated total number of comments under non-COVID-19 was almost constantly higher. Another peak in the number of populist comments that matches the magnitude of the peak observed for the “Ibiza affair” appears in the reference data during October 2019. This corresponds to another major political event in Austria, because then chancellor Sebastian Kurz declared the withdrawal from his position, after investigations put him under suspicions of corruption[4].

To connect the topics of COVID-19 and populism, we pre-select some of the most important events related to COVID-19 countermeasures in Austria with the information found on the Wikipedia page covering the pandemic in Austria[6] and present the number of populist comments in a higher resolution, by aggregating them per day and showing

[1] https://www.derstandard.at/story/2000103364196/strache-soll-staatsauftraege-fuer-wahlkampfspenden-in-aussicht-gestellt-haben last accessed 27.02.2023

[2] https://en.wikipedia.org/wiki/Ibiza_affair last accessed 27.02.2023

[3] https://orf.at/stories/3139981/ last accessed 27.02.2023

[4] https://www.derstandard.at/story/2000130311035/nach-juengsten-enthuellungen-sebastian-kurz-vor-rueckzug-als-kanzler last acceessed 27.02.2023

[5] https://viecer.univie.ac.at/corona-blog/

[6] https://de.wikipedia.org/wiki/COVID-19-Pandemie_in_Österreich last accessed 27.02.2023

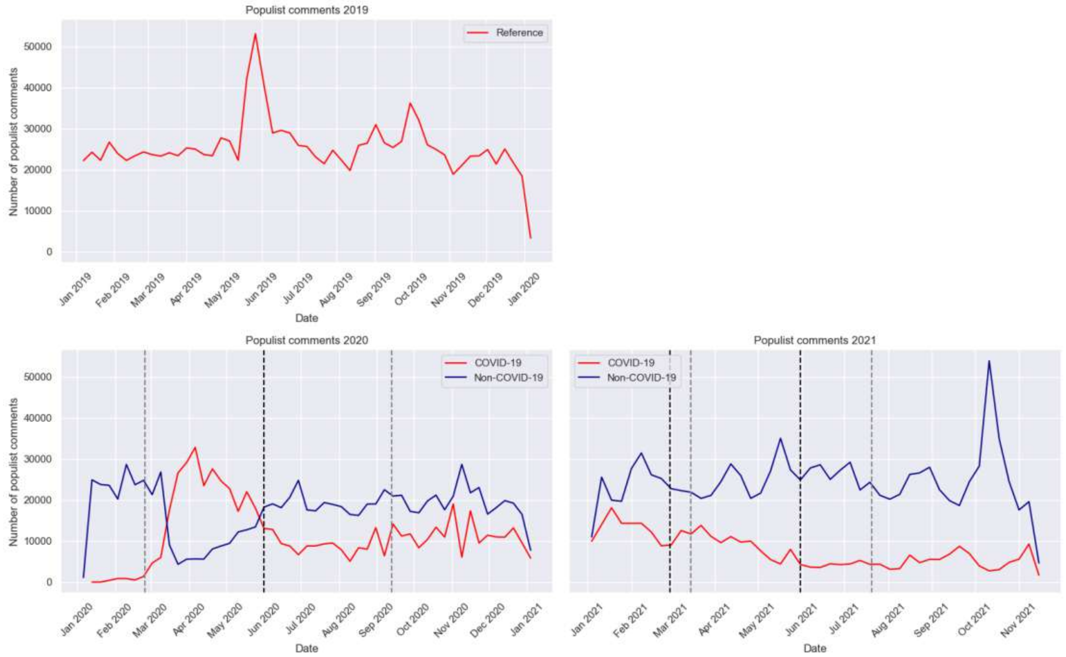


Figure 7.4: Plots of the number of populist comments aggregated per week for each sample and subdivided in the years covered by the samples (the grey dotted lines indicate the approximate starts and the black line the approximate ends of the four COVID-19 waves in Austria following information from the corona-blog of the Vienna Center for Electoral Research[5])

the development of the number during the five preceding and following days of an event. Figure 7.5 shows the development of populist comments around the following events:

- 16.03.2020: Beginning of the first lockdown in Austria
- 30.03.2020: Decision to make it mandatory to wear masks in shops
- 14.09.2020: Reinstatement of mandatory masks indoor
- 17.11.2020: Beginning of the second lockdown in Austria
- 26.12.2020: Beginning of the third lockdown in Austria
- 15.09.2021: Reinstatement of mandatory FFP-2 masks and tightening of the 3G-rules

For the first lockdown, we can see the start of the major rise in populist comments at the beginning of the pandemic that increases further towards the announcement of compulsory masks in supermarkets, after which we can see a peak at the beginning of April. Those were events that were very restrictive for the people, who were still

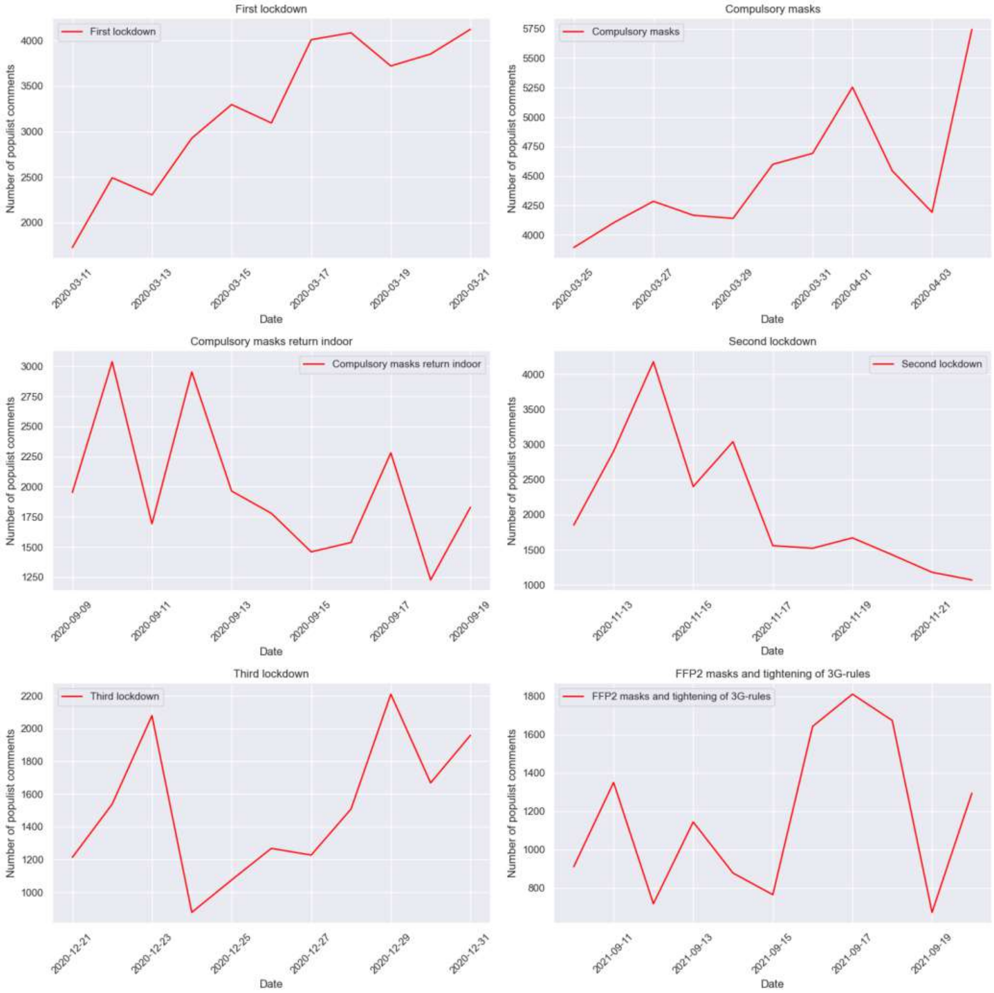


Figure 7.5: Plots showing the number of populists comments aggregated by day around major COVID-19 events

unfamiliar with the situation, but seemingly express their discontent with the government and experts that they hold responsible for the decisions made. After a more relaxed summer in terms of restrictions, we see the return of mandatory masks and subsequently the second lockdown, where we observe more populist comments before the measurements come into effect and a decline afterward. This could be the case because the preceding days show the days after the announcement of the measures, where media coverage of the topics is most likely higher than on the day the measures actually come into effect. For the third lockdown, we can observe two spikes before and after the start. The last plot shows an event that led to more restrictions, especially for unvaccinated people, including a majority of the "coronasceptic" people. This is represented in another relative peak during that time, which could refer to the tension and discontent of "coronasceptic" people who felt unfairly treated by the state.

7.3.2 Populism Comparison

Our major research goal in the model application is the comparison of populist comments per article between our different samples. Therefore, we fit negative binomial regression models to our data to answer the research questions:

- RQ2a: How does the COVID-19 crisis affect the number of populist user comments?
- RQ2b: How does the topic of COVID-19 affect the number of populist user comments posted during the crisis?

To observe the effect of the pandemic on the number of populist user comments, we compare all articles posted during 2019 to every article posted during the pandemic disregarding the article's topic. The outcome of our model gives us insight into whether the situation of the pandemic as a crisis led to more populist user comments in general. A rise would, for example, indicate an increased anti-elitist stance and discontent with the government. In Table 7.18, we can however see that the model returns a negative coefficient for our explanatory variable. This means that populist comments were less prevalent during the pandemic than in the reference year before. With the reported p-value, our results are highly significant. In this case, we can directly interpret the coefficient as the average difference in populist comments for an article posted before and during the pandemic. Subsequently, an article posted during the pandemic attracted approximately 7.8% less populist comments.

Coefficient	Standard error	z	p-value	[0.025	0.975]
-0.0778	0.007	-11.783	<0.001	-0.091	-0.065

Table 7.16: Test statistics of the negative binomial regression model to measure the relation between the populist comments per article and the pandemic in reference to 2019

Coefficient	Standard error	z	p-value	[0.025	0.975]
0.17	0.01	16.449	<0.001	0.15	0.19

Table 7.17: Test statistics of the negative binomial regression model to measure the relation between the populist comments per article and the topic of COVID-19

For *RQ2b*, we want to test whether the observations of Thiele [Thi22] hold over a longer period of the pandemic. Here, we directly measure the influence of the topic of COVID-19 on the number of populist comments by comparing the *COVID-19 sample* and the *non-COVID-19 sample* and report the results in Table 7.17. The regression model returns a significant positive coefficient in this case, indicating that the topic of COVID-19 indeed attracts more populist comments than other topics. This is furthermore an interesting finding, as it still holds despite the peaks of populist comments under other articles we observed in 2021, including the withdrawal of Sebastian Kurz. Subsequently, we have evidence that COVID-19 nurtured populist talking points and stances, as a COVID-19 article on average attracts 17% more populist comments than other articles.

7.3.3 Populism over Time

In this section, we deal with the last remaining research question:

- RQ3: How does the amount of populist user comments under COVID-19-related articles evolve over time?

The goal is to measure whether the number of populist comments under COVID-19-related articles rises or declines with the ongoing pandemic. Therefore, our negative binomial regression model measures the effect of the difference in days to the start of the pandemic on the number of populist comments for an article. The results in Table 7.18 suggest that there is no noticeable influence of the duration of the pandemic on the observed populism. The small negative coefficient of -0.0006 is only a minor effect that can be neglected. We can therefore not detect a trend in the development of populist comments under COVID-19-related articles during the pandemic.

Coefficient	Standard error	z	p-value	[0.025	0.975]
-0.0006	$< 0{,}0001$	-10.16	<0.001	-0.001	0

Table 7.18: Test statistics of the negative binomial regression model to measure the relation between the populist comments per article and the days passed since the start of the pandemic

7.3.4 Discussion

To sum up the insights of the model application on our research questions, we want to elaborate on the findings and compare them to the work of Thiele [Thi22] in a final discussion. The first research question in this context:

- RQ2a: How does the COVID-19 crisis affect the number of populist user comments?

has not been covered in the existing literature, but added new insights to the prevalence of populism in Austria news comments. The significant answer is that, after accounting for the rise in total comments, articles published during the pandemic receive a 7.8% lower quantity of populist comments than those published during the reference period. The finding that the amount of populist comments actually declined with the start of the pandemic makes it interesting for future work to investigate the timeframe of 2019. We suggested possible topics, especially the "Ibiza affair", which could cause this effect.

However when looking at the next research question:

- RQ2b: How does the topic of COVID-19 affect the number of populist user comments posted during the crisis?

we can confirm the findings of Thiele, who also noticed an increase of populist comments under COVID-19-related articles by 14%. Our study shows that this still holds with the ongoing pandemic until November 2021. As we approached the autumn of 2021, tensions between "coronasceptics" and other citizens and experts increased due to the tightening of 3G rules and the introduction of 2G rules, which placed more restrictions on unvaccinated individuals. These additional factors provided further grounds for expressing discontent with the situation and blaming the elites for what was perceived as unfair treatment of citizens.

The results of our tests for the final research question:

- RQ3: How does the amount of populist user comments under COVID-19-related articles evolve over time?

are surprising in the light of the previous study by Thiele [Thi22], who could confirm in his work that populist comments under Facebook posts of news outlets rose over the time of the pandemic, which could be interpreted with growing discontent with the governmental decisions and the way they communicated them. Furhtermore, he suggested this could be signs of "reactance" [DS05] and consequences of failed "fear appeals" [Wit92]. We could not confirm that for our observed timeframe, where we could not find a connection between the length of the pandemic and COVID-19. With these results it is interesting to test, if Thiele's hypothesis holds for the comments directly posted on the news forum of *Der Standard* for his observed timeframe. Therefore, we cut our data off on 30 May and run our model again. With the adjusted timeframe, we obtain a coefficient of -0.0004 which is again no indicator of an observable trend. This may be caused by a discrepancy between the data sources, as Thiele utilized a diverse range of news outlets. However, the more important difference could lie in the nature of Facebook comments, where users only engage with a snippet and the headline of an article, as opposed to being exposed to the entire content of the article on the news website prior to accessing the comment section, which would require further investigation.

CHAPTER 8

Conclusion

This chapter concludes this thesis by briefly summing up the most important steps and findings of the work and furthermore highlights the contribution made to the research field. After that we discuss the limitations of the conducted research and draw implications about potential future work.

8.1 Summary

COVID-19 is a topic that started as a medical one and quickly turned into a political topic with the pandemic affecting the lives of people all over the world and governments having to make decisions to contain the spread of the virus. This uncertain situation introduced a very complex problem to governments in finding a balance between preventing the health care system from collapsing and restricting the social life of millions of people. This opened a gateway for populist rhetoric that seeks to provide simplistic solutions to complex issues by constructing an adversary in the form of an elite group, which is held responsible for any unfavorable developments that affect the general public, thereby distancing them from said elites. This work showed that these rhetorics can be identified by an operationalization of the minimalistic ideational definition of Mudde [Mud04]. In an extensive literature review, we demonstrated the importance of the definition in the research field and that it is the best fit for our task. Through an annotation study, we demonstrated that, during the observed period, the most dominant populist motive in news user comments was anti-elitism. Additionally, we found that human annotators displayed a high level of agreement in identifying instances of populism in short texts. The strong prevalence of anti-elitism is congruent with the “anti-systemic messages” and distrust with mainstream media and political establishment that Boberg et al. determine as main characteristic of the “pandemic populism” they observed [BQSEF20].

After the annotation study, we demonstrated that we can outperform existing methods for automated populism detection with he use of pre-trained BERT models and human-annotated data. We could especially reduce the amount of false positives, which gives us more precise results and helps to capture the phenomenon more accurately. Nevertheless, we rejected the approach of using the dictionary to label a large amount of training data, as the observed performance could not improve the dictionary alone.

The application of the model on the *Der Standard* data could confirm the influence of the topic of COVID-19 on the amount of populist news user comments, even during a time were other events that dealt with corruption investigations of politicians in Austria provided a breeding ground for populist rhetoric. Focusing on citizens as the primary propagators of populist messages in our analysis provides a perspective that is not as widely represented in the literature as that of politicians' or media's populism. Detecting populism in user comments could help for further crisis management, by observing the effects of certain strategies and the communication of countermeasures on the reactions of a group of people. The inclusion of a reference timeframe uncovered additional insights, enabling us to contextualize "pandemic populism" and demonstrate that the pandemic did not result in a general increase in populist user comments. With the results obtained by answering the final research question, we could also find no significant evidence of a stronger prevalence of populist rhetoric over the course of the pandemic, which opposes previous findings.

Overall this thesis provides the following contributions to the research field:

- We provide the first annotated dataset of Austrian news user comments for populism.
- We propose a model that outperforms the SOTA for populism detection by 0.17 regarding F1-score.
- We add a reference timeframe for the analysis of populism during the COVID-19 pandemic.
- We perform analysis on a dataset of more than 30,000,000 comments, which increases the volume of data used in previous studies and furthermore covers a longer time of the pandemic.

To support the findings of this work, supplementary material is publicly available in a GitHub repository[1].

[1] https://github.com/ahmadouw/COV-Populism-Standard

8.2 Limitations and Future Work

Upon reflecting on the limitations of our work, we can identify the potential for further research. Considering the high prevalence of anti-elitism in our observations, incorporating information about the elites mentioned in the comments and introduce it as a feature for ML models could improve the results. By identifying popular figures that had a strong medial presence or were targeted by conspiracy theories, we could enrich the purely textual input of our model with more semantic information.

As we only investigated populism in the dimension of single comments and articles, it can be interesting to further concentrate on the user level, to for example classify certain users as populist or not or measure a degree of populism based on their textual output. With this information, we could furthermore investigate the networks that populist users form, as it is an ideology that forms groups that reinforce themselves in the conviction that they actually represent the people and claim to express the majority's opinion [GT19]. After inspecting citizens as populist actors, another connection that has to be observed is the behaviour of populist parties and the success of populist parties within the pandemic. Future work should investigate, whether the emerging "pandemic populism" is reflected in the electoral decision of citizens and whether populist politicians can profit and gain influence during such a crisis.

Considering the scale of this work, the choice of *Der Standard* as a media source that is not considered to have a populist orientation gives a good overview, but incorporating other news sources might give a more diverse representation of public opinion. Geographically the work is limited to Austria, which is a good general example given the magnitude of "coronasceptic" demonstrations, but comparing the findings to the response in other countries can solidify the connection of populism and COVID-19 further, given the global relevance of the crisis. As the pandemic is still ongoing after the cut-off of our data, analysis of temporal development beyond November 2021 is of interest to expand the scale. The controversy about the introduction of compulsory vaccination in Austria is for example not fully included in the current work.

This work aims at making populism comparable, as it adds a reference timeframe. Therefore the choice of incorporating data from all three samples and using the minimalistic ideational definition of populism in the data annotation is fitting for the use case. However, future research could be directed towards investigating the unique features of "pandemic populism". This could involve defining the phenomenon and developing a specialized model that might be able to make more accurate predictions and enable isolated analysis.

Regarding the examination of *Der Standard* data, there are two aspects that were disregarded in this work: first the inclusion of headlines and second the hierarchy of the comments. During the annotation study, some comments seemed unnatural for the participants, as they start in the middle of a sentence. This is caused by a misuse of the headline feature, where people put actual parts of their messages in the headline instead of a short summary statement. However, the influence of this might be neglectable. The hierarchy of comments means the information of whether a comment was posted as a parent comment directly to the article or a reply under an already existing comments. We chose to include all hierarchical levels of comments, because of the reinforcing and group building nature previously found in populist online communication [GT19], which we expect to be reflected in users replying to each other. Incorporating the hierarchy of the comments might however be a beneficial aspect for analyzing populist user networks.

Eventually, a final problem that can generally limit the performance of ML is the lack of annotated data. The class imbalance made it necessary to involve the dictionary in the sample drawing process, which is why we can not really assess the performance of our model on a lot of populist comments that the dictionary does not detect (only 30 of those were included in the random sample). With the availability of more resources for labeling larger amounts of data and a growing interest for populism detection within the scientific community, there is potential for enhancing model performance and validation of new approaches in a still emerging research field. Future endeavours could then aim to go beyond the measurement of common performance metrics and try to add explainability to model predictions, given the complexity of the phenomenon.

List of Figures

List of Tables

Acronyms

AfD Alternative für Deutschland. 2

BERT Bidirectional Encoder Representation for Transformers. 7, 8, 18, 20, 21, 42, 44–46, 59–63, 74, 79

BOW bag of words. 19

CRISP-DM Cross Industry Standard Process for Data Mining. 5

EN elastic net. 43, 58

FN false negative. 47

FP false positive. 47

FPÖ Freiheitliche Partei Österreichs. 2

IC initial cleaning. 45–47, 59, 60, 62, 63

LLM large language model. 7, 43, 63

LR logistic regression. 43, 58

ML machine learning. 3, 4, 7, 8, 11, 17–19, 29, 42, 56–60, 63, 75, 76, 79

MLM masked language modelling. 21

MN model name. 45

NLP natural language processing. 6, 11, 19, 20

NSP next Sentence prediction. 21

PJ Partido Justicialista. 2

RF random forrest. 43, 58

RN Rassemblement National. 2

RoBERTa Robustly Optimized BERT Pretraining Approach. 18

SC spelling correction. 45, 47, 59–61

SOTA state-of-the-art. 6, 8, 11, 18–21, 41, 43, 62, 63, 74

SVM support vector machine. 43, 58

TD training data. 45

TN true negative. 47

TP true positive. 47

Bibliography

[ANMC21] Francisca Adoma Acheampong, Henry Nunoo-Mensah, and Wenyu Chen. Transformer models for text-based emotion detection: a review of bert-based approaches. *Artificial Intelligence Review*, pages 1–41, 2021.

[AR07] Koen Abts and Stefan Rummens. Populism versus democracy. *Political Studies*, 55(2):405–424, 2007.

[Asl16] Paris Aslanidis. Is populism an ideology? a refutation and a new perspective. *Political Studies*, 64(1):88–104, 2016.

[Asl18] Paris Aslanidis. Measuring populist discourse with semantic text analysis: an application on grassroots populist mobilization [quality quantity]. *Quality and Quantity*, 53:1241–1263, 05 2018.

[Bar09] Robert R. Barr. Populists, outsiders and anti-establishment politics. *Party Politics*, 15(1):29–48, 2009.

[BB16] Noam Gidron Bart Bonikowski. Populism in legislative discourse : Evidence from the european parliament , 1999-2004. 2016.

[BEEE19] Sina Blassnig, Sven Engesser, Nicole Ernst, and Frank Esser. Hitting a nerve: Populist news articles lead to more frequent and more populist reader comments. *Political Communication*, 36(4):629–651, 2019.

[BG15] Bart Bonikowski and Noam Gidron. The Populist Style in American Politics: Presidential Campaign Discourse, 1952–1996. *Social Forces*, 94(4):1593–1621, 12 2015.

[BM17] Roberta Bracciale and Antonio Martella. Define the populist political communication style: the case of italian political leaders on twitter. *Information, Communication & Society*, 20(9):1310–1329, 2017.

[BQSEF20] Svenja Boberg, Thorsten Quandt, Tim Schatto-Eckrodt, and Lena Frischlich. Pandemic populism: Facebook pages of alternative news media and the corona crisis – a computational content analysis, 2020.

[Bre01] L Breiman. Random forests. *Machine Learning*, 45:5–32, 10 2001.

[BvdBdV11] Linda Bos, Wouter van der Brug, and Claes de Vreese. How the media shape perceptions of right-wing populist leaders. *Political Communication*, 28(2):182–206, 2011.

[CAFS21] Pere-Lluís Huguet Cabot, David Abadi, Agneta H. Fischer, and Ekaterina Shutova. Us vs. them: A dataset of populist attitudes, news bias and emotions. In *EACL*, 2021.

[Can99] Margaret Canovan. Trust the people! populism and the two faces of democracy. *Political Studies*, 47(1):2–16, 1999.

[CAR17] DANIELE CARAMANI. Will vs. reason: The populist and technocratic forms of political representation and their critique to party government. *American Political Science Review*, 111:54–67, 02 2017.

[CCK$^+$00] Peter Chapman, Janet Clinton, Randy Kerber, Tom Khabaza, Thomas P. Reinartz, Colin Shearer, and Richard Wirth. Crisp-dm 1.0: Step-by-step data mining guide. 2000.

[CG16] Manuela Caiani and Paolo R. Graziano. Varieties of populism: insights from the italian case. *Italian Political Science Review / Rivista Italiana di Scienza Politica*, 46(2):243–267, 2016.

[Cox58] D. R. Cox. The regression analysis of binary sequences. *Journal of the Royal Statistical Society. Series B (Methodological)*, 20(2):215–242, 1958.

[CSM20] Branden Chan, Stefan Schweter, and Timo Möller. German's next language model. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6788–6796, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics.

[CST00] Nello Cristianini and John Shawe-Taylor. *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press, 2000.

[Dai19] Yaoyao Dai. Measuring populism in context: A supervised approach with word embedding models. *Manifesto Corpus Conference, Berlin*, 2019.

[DCLT18] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2018.

[DCM22] Jessica Di Cocco and Bernardo Monechi. How populist are parties? measuring degrees of populism in party manifestos using supervised machine learning. *Political Analysis*, 30(3):311–327, 2022.

[DK22] Yaoyao Dai and Alexander Kustov. When do politicians use populist rhetoric? populism as a campaign gamble. *Political Communication*, 39(3):383–404, 2022.

[DS05] James Dillard and Lijiang Shen. On the nature of reactance and its role in persuasive health communication. *Communication Monographs*, 72:144–168, 06 2005.

[dVEA⁺18] Claes H. de Vreese, Frank Esser, Toril Aalberg, Carsten Reinemann, and James Stanyer. Populism as an expression of political communication content and style: A new perspective. *The International Journal of Press/Politics*, 23(4):423–438, 2018. PMID: 30886670.

[EBE⁺19] Nicole Ernst, Sina Blassnig, Sven Engesser, Florin Büchel, and Frank Esser. Populists prefer social media over talk shows: An analysis of populist messages and stylistic elements across six countries. *Social Media + Society*, 5(1):2056305118823358, 2019.

[EEB⁺17] Nicole Ernst, Sven Engesser, Florin Büchel, Sina Blassnig, and Frank Esser. Extreme parties and populism: an analysis of facebook and twitter across six countries. *Information, Communication Society*, 20:1–18, 05 2017.

[EEEB17] Sven Engesser, Nicole Ernst, Frank Esser, and Florin Büchel. Populism and social media: how politicians spread a fragmented ideology. *Information, Communication & Society*, 20(8):1109–1126, 2017.

[EFL17] Sven Engesser, Nayla Fawzi, and Anders Olof Larsson. Populist online communication: introduction to the special issue. *Information, Communication & Society*, 20(9):1279–1292, 2017.

[EHG21] Jakob-Moritz Eberl, Robert Huber, and Esther Greussing. From populism to the "plandemic": Why populists believe in covid-19 conspiracies. *Journal of Elections Public Opinion and Parties*, 31:272–284, 06 2021.

[ELB⁺22] Olga Eisele, Olga Litvyak, Verena K. Brändle, Paul Balluff, Andreas Fischeneder, Catherine Sotirakou, Pamina Syed Ali, and Hajo G. Boomgaarden. An emotional rally: Exploring commenters' responses to online news coverage of the covid-19 crisis in austria. *Digital Journalism*, 10(6):952–975, 2022.

[ES14] Mark Elchardus and Bram Spruyt. Populism, persistent republicanism and declinism: An empirical analysis of populism as a thin ideology. *Government and Opposition*, -1:1–23, 09 2014.

[ESH17] Frank Esser, Agnieszka Stępińska, and David Nicolas Hopmann. *Populism and the media: cross-national findings and perspectives. In T. Aalberg, F. Esser, C. Reinemann, J. Strömbäck C. d. Vreese (Eds.), Populist Political Communication in Europe. Routledge, 365-380.* 01 2017.

[FGS20] Belén Fernández-García and Susana Salgado. Populism by the people: An analysis of online comments in portugal and spain*. pages 210–219, 07 2020.

[FHT00] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Additive logistic regression: a statistical view of boosting (With discussion and a rejoinder by the authors). *The Annals of Statistics*, 28(2):337 – 407, 2000.

[Gr2] Johann Gründl. Populist ideas on social media: A dictionary-based measurement of populist communication. *New Media & Society*, 24(6):1481–1499, 2022.

[GS13] Justin Grimmer and Brandon M. Stewart. Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, 21(3):267–297, 2013.

[GT19] Charlotte Galpin and Hans-Jörg Trenz. Participatory populism: Online discussion forums on mainstream news sites during the 2014 european parliament election. *Journalism Practice*, 13(7):781–798, 2019.

[Gup20] Bolden S.E. Kachhadia J. Korsunska A. Stromer-Galley J. Gupta, S. Polibert: Classifying political social media messages with bert., 2020.

[Ham18] Michael Hameleers. A typology of populism: Toward a revised theoretical framework on the sender side and receiver side of communication. *International Journal of Communication*, 12(0), 2018.

[Ham19] Michael Hameleers. The Populism of Online Communities: Constructing the Boundary Between "Blameless" People and "Culpable" Others. *Communication, Culture and Critique*, 12(1):147–165, 03 2019.

[Haw09] Kirk A. Hawkins. Is chávez populist?: Measuring populist discourse in comparative perspective. *Comparative Political Studies*, 42(8):1040–1067, 2009.

[Haw10] Kirk A. Hawkins. *Venezuela's Chavismo and Populism in Comparative Perspective*. Cambridge University Press, 2010.

[HBdV16] Michael Hameleers, Linda Bos, and Claes de Vreese. *The Netherlands: A Heartland Full of Insights into Populist Communication*. 07 2016.

[HCS19] K.A. Hawkins and B. Castanho Silva. *Textual Analysis: Big Data Approaches*. Routledge, London, 2019.

[Hil11] Joseph M. Hilbe. *Negative Binomial Regression*. Cambridge University Press, 2 edition, 2011.

[HK07] Andrew F. Hayes and Klaus Krippendorff. Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures*, 1(1):77–89, 2007.

[HK18] Kirk A. Hawkins and Cristóbal Rovira Kaltwasser. Measuring populist discourse in the United States and beyond. *Nature Human Behaviour*, 2(4):241–242, April 2018.

[HRK19] Aguilar R. Castanho Silva B. Jenne-E. K. Kocijan B. Hawkins, K. A. and C. Rovira Kaltwasser. Measuring populist discourse: The global populism database. *EPSA*, 2019.

[HS16] Kirk A. Hawkins and Bruno Castanho Silva. A head-to-head comparison of human-based and automated text analysis for measuring populism in 27 countries. 2016.

[HZRS15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

[Jan11] Robert S. Jansen. Populist mobilization: A new theoretical approach to populism. *Sociological Theory*, 29(2):75–96, 2011.

[JH22] Michael Jankowski and Robert A Huber. When correlation is not enough: Validating populism scores from supervised machine-learning models, Jan 2022.

[JW07] Jan Jagers and Stefaan Walgrave. Populism as political communication style: An empirical study of political parties' discourse in belgium. *European Journal of Political Research*, 46(3):319–345, 2007.

[Kal12] Cristóbal Rovira Kaltwasser. The ambivalence of populism: threat and corrective for democracy. *Democratization*, 19(2):184–208, 2012.

[Kr4] Benjamin Krämer. Media Populism: A Conceptual Clarification and Some Theses on its Effects. *Communication Theory*, 24(1):42–60, 01 2014.

[Kr7] Benjamin Krämer. Populist online practices: the function of the internet in right-wing populism. *Information, Communication & Society*, 20(9):1293–1309, 2017.

[Lac05] Ernesto. Laclau. *On populist reason / Ernesto Laclau*. Verso New York, 2005.

[LH17] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2017.

[LM01] Ernesto Laclau and Chantal Mouffe. *Hegemony and Socialist Strategy: Towards a Radical Democratic Politics*. Verso, paperback edition, 2001.

[Maz08] Gianpietro Mazzoleni. *Populism and the Media*, pages 49–64. Palgrave Macmillan UK, London, 2008.

[MC12] Cas Mudde and Kaltwasser Cristóbal, Rovira, editors. *Populism in Europe and the Americas: Threat or Corrective for Democracy?* Cambridge University Press, hardcover edition, 2012.

[MCGM18] Suman Kalyan Maity, Aishik Chakraborty, Pawan Goyal, and Animesh Mukherjee. Opinion conflicts: An effective route to detect incivility in twitter. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):1–27, 2018.

[MKNMN20] Alaa Mahmood, Siti Kamaruddin, Raed Naser, and Maslinda Mohd Nadzir. A combination of lexicon and machine learning approaches for sentiment analysis on facebook. *Journal of System and Management Sciences*, 10, 11 2020.

[Mof16] Benjamin Moffitt. *The Global Rise of Populism: Performance, Political Style, and Representation*. Stanford University Press, hardcover edition, 2016.

[MRK17] Cas Mudde and Cristóbal Rovira Kaltwasser. *Populism: A Very Short Introduction*. Oxford University Press, 02 2017.

[MSW+17] Philipp Müller, Christian Schemer, Martin Wettstein, Anne Schulz, Dominique S. Wirz, Sven Engesser, and Werner Wirth. The Polarizing Impact of News Coverage on Populist Attitudes in the Public: Evidence From a Panel Study in Four European Democracies. *Journal of Communication*, 67(6):968–992, 2017.

[Mud04] Cas Mudde. The populist zeitgeist. *Government and Opposition*, 39(4):541–563, 2004.

[Mud17] Cas Mudde. Populism: An Ideational Approach. In *The Oxford Handbook of Populism*. Oxford University Press, 2017.

[MW17] Luca Manucci and Edward Weber. Why the big picture matters: Political and media populism in western europe since the 1970s. *Swiss Political Science Review*, 23, 08 2017.

[NSF20] Oliver Nachtwey, Robert Schäfer, and Nadien Frei. Politische soziologie der corona-proteste, Dec 2020.

[PCF21] Jeremy Pressman and Austin Choi-Fitzpatrick. Covid19 and protest repertoires in the united states: an initial description of limited change. *Social Movement Studies*, 20(6):766–773, 2021.

[PGM+19] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library, 2019.

[PNI+18] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations, 2018.

[PR15] Teun Pauwels and Matthijs Rooduijn. *Populism in Belgium in times of crisis: Intensification of Discourse, Decline in Electoral Support*. ECPR Press, Colchester, UK, 2015.

[PRB+17] Jonathan Polk, Jan Rovny, Ryan Bakker, Erica Edwards, Liesbet Hooghe, Seth Jolly, Jelle Koedam, Filip Kostelka, Gary Marks, Gijs Schumacher, Marco Steenbergen, Milada Vachudova, and Marko Zilovic. Explaining the salience of anti-elitism and reducing political corruption for political parties in europe with the 2014 chapel hill expert survey data. *Research & Politics*, 4(1):2053168016686915, 2017.

[Rat11] Susan Ratcliffe. *Concise Oxford Dictionary of Quotations, 5th ed.* Oxford paperback reference. Oxford University Press, 6. ed edition, 2011.

[Rob06] Kenneth M. Roberts. Populism, political conflict, and grass-roots organization in latin america. *Comparative Politics*, 38(2):127–148, 2006.

[Roo14] Matthijs Rooduijn. The nucleus of populism: In search of the lowest common denominator. *Government and Opposition*, 49(4):573–599, 2014.

[Roo19a] Matthijs Rooduijn. State of the field: How to study populism and adjacent topics? a plea for both more and less focus. *European Journal of Political Research*, 2019.

[Roo19b] Van Kessel S. Froio C. Pirro A.-De Lange S. Halikiopoulou D. Lewis P. Mudde C. Taggart P. Rooduijn, M. The populist: An overview of populist, far right, far left and eurosceptic parties in europe., 2019.

[RP11] Matthijs Rooduijn and Teun Pauwels. Measuring populism: Comparing two methods of content analysis. *West European Politics*, 34(6):1272–1283, 2011.

[Sai80] R. M. Sainsbury. Russell on constructions and fictions. *Theoria*, 46(1):19–36, 1980.

[SAK17] Yannis Stavrakakis, Ioannis Andreadis, and Giorgos Katsambekis. A new populism index at work: identifying populist candidates and parties in the contemporary greek context. *European Politics and Society*, 18(4):446–464, 2017.

[San14] Arthur D. Santana. Virtuous or vitriolic. *Journalism Practice*, 8(1):18–33, 2014.

[SG18] Maria Sousa Galito. Populism as a political phenomenon. *JANUS.NET, e-journal of International Relation*, 9:53–69, 05 2018.

[SH22] Cornelia Schroll and Brigitte Huber. Assessing levels and forms of incivility and deliberative quality in online discussions on covid-19: A cross-platform analysis. *Frontiers in Political Science*, 4, 02 2022.

[SJ21] Salim Sazzed and Sampath Jayarathna. Ssentia: A self-supervised sentiment analyzer for classification from unlabeled data. *Machine Learning with Applications*, 4:100026, 2021.

[SMS$^+$18] Anne Schulz, Philipp Müller, Christian Schemer, Dominique Wirz, Martin Wettstein, and Werner Wirth. Measuring populist attitudes on three dimensions. *International Journal of Public Opinion Research*, 30:316–326, 02 2018.

[SSS$^+$20] Xinying Song, Alex Salcianu, Yang Song, Dave Dopson, and Denny Zhou. Fast wordpiece tokenization, 2020.

[Thi22] Daniel Thiele. Pandemic populism? how covid-19 triggered populist facebook user comments in germany and austria. *Politics and Governance*, 10(1):185–196, 2022.

[TWC08] Songbo Tan, Yuefen Wang, and Xueqi Cheng. Combining learn-based and lexicon-based techniques for sentiment detection without using labeled examples. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '08, page 743–744, New York, NY, USA, 2008. Association for Computing Machinery.

[UP21] Jogile Ulinskaite and Lukas Pukelis. Identifying populist paragraphs in text: A machine-learning approach. *CoRR*, abs/2106.03161, 2021.

[VSP$^+$17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017.

[WDS$^+$19] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan

Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Huggingface's transformers: State-of-the-art natural language processing, 2019.

[WEW+16] Werner Wirth, Frank Esser, Martin Wettstein, Sven Engesser, Dominique S. Wirz, Anne Schulz, Nicole Ernst, Florin Büchel, Daniele Caramani, Luca Manucci, Marco R. Steenbergen, Laurent Bernhard, Edward Weber, Regula Hänggli, Caroline Dalmus, Christian Schemer, and Philipp Müller. The appeal of populist ideas, strategies and styles: A theoretical model and research design for analyzing populist political communication. (88), 2016.

[Wey01] Kurt Weyland. Clarifying a contested concept: Populism in the study of latin american politics. *Comparative Politics*, 34(1):1–22, 2001.

[WF17] Leonie Weissweiler and Alexander Fraser. Developing a stemmer for german based on a comparative analysis of publicly available stemmers. In *German Society for Computational Linguistics*, 2017.

[Whi85] Edward M. White. *Teaching and Assessing Writing (Jossey Bass Higher Adult Education Series)*. Jossey-Bass Inc Pub, 1985.

[Wir18] Dominique Stefanie Wirz. Persuasion through emotion? an experimental test of the emotion-eliciting nature of populist communication. *International Journal of Communication*, 12:25, 2018.

[Wit92] Kim Witte. Putting the fear back into fear appeals: The extended parallel process model. *Communication Monographs*, 59(4):329–349, 1992.

[WSS20] Alexander Wuttke, CHRISTIAN SCHIMPF, and Harald Schoen. When the whole is greater than the sum of its parts: On the conceptualization and measurement of populist attitudes and other multidimensional constructs. *American Political Science Review*, 114:1–19, 02 2020.

[ZH05] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 67(2):301–320, 2005.

[ZQCL22] Menghan Zhang, Xue Qi, Ze Chen, and Jun Liu. Social bots' involvement in the covid-19 vaccine discussions on twitter. *International Journal of Environmental Research and Public Health*, 19:1651, 01 2022.

[ZWK+20] Tianyi Zhang, Felix Wu, Arzoo Katiyar, Kilian Q. Weinberger, and Yoav Artzi. Revisiting few-sample bert fine-tuning, 2020.