Institut für Maschinelle Sprachverarbeitung
Universität Stuttgart
Pfaffenwaldring 5B
D-70569 Stuttgart

Master Thesis

Enhancing a German Dialect Corpus with Neural Methods

Wolfgang Tessadri

Study program: M.Sc. Computational Linguistics

Examiners: Prof. Dr. Eleonore Brandner
Prof. Dr. Sabine Schulte im Walde

Supervisors: Prof. Dr. Eleonore Brandner
Prof. Dr. Jonas Kuhn

Start of thesis: 1 December 2022

End of thesis: 30 June 2023

Statement of Authorship

This thesis is the result of my own independent work, and any material from work of others which is used either verbatim or indirectly in the text is credited to the author including details about the exact source in the text. This work has not been part of any other previous examination, neither completely nor in parts. It has neither completely nor partially been published before. The submitted electronic version is identical to this print version.

Wolfgang Tessadri

Abstract

With the advent of modern chat applications, an increasing number of German dialect speakers use their dialects for written communication. The DiDi Facebook corpus (Frey et al. 2016) captures this phenomenon for South Tyrolean dialects. While the authors included a dialect/standard variety tag on the posting level, a third of these tags was undefined. By training DeBERTa and XLM-RoBERTa for dialect/standard classification we reduce these undefined instances by over 75%. We also use XLM-RoBERTa to add explicit variety labels to individual tokens. By performing a linear regression analysis of socio-linguistic variables and a label-derived dialectality metric we show that the generated labels are highly meaningful. Finally, we describe how the implemented Transformer models can be applied to gather geo-referenced dialect samples on Twitter and we discuss how this data can enrich future dialectometric research.

Table of Contents

PART I – Theoretical background

1.1 Introduction

With approximately 134.6 million speakers, German is one of the most spoken languages worldwide (Eberhard et al. 2022). The German language community is characterized by an extensive diatopic variation resulting from the formation and expansion of different Germanic tribes, the fine-grained and very dynamic medieval political landscape in German-speaking parts of Europe as well as at least two more prominent waves of standardization attempts (Schmidt 2011). While this variation is a linguistic reality for millions of speakers, it is only marginally covered by publicly available text resources. This can be seen as a medium-dependent phenomenon. There exist a variety of German audio corpora depicting regional variation in the oral domain[1].

When it comes to textual output, however, the number of resources covering this diatopic diversity is limited. Two initiatives for the collection of German dialect texts are, for example, the projects sms4science (Dürscheid & Stark 2011) and ArchiMob (Samardzic et al. 2016), both covering Swiss German texts[2]. The main reason for this relatively small data basis is the hegemonial position of Standard German in written communication. Even though a German native speaker might be proficient in multiple variants of his mother tongue (base dialect, regional dialect, regiolect, standard[3]) and would adapt according to conversational partner and situation, this is in general not true for the written form. For this reason, written dialect, up until recently, was a marginal phenomenon. This might appear obvious as dialects are regionally circumscribed, and the establishment of own writing habits would just limit the communicative range in comparison to an already established and supra-regionally accepted norm. With the emergence of the new media at the turn of the millennium, though, an increasing number of, mainly young, communicators started to write their dialects. This phenomenon is especially prominent in regions with former "medial diglossia" (Ferguson 1959, Kolde 1981) like the German-speaking parts of Switzerland. For example, at the beginning of the nineties Haas

[1] see e.g. Plüss et al. (2022) for Swiss German, Huber et al. (2019) for Austrian dialects and the collection „Oral Corpora at the Institute for the German Language" (Schmidt (2017)) for varieties in Germany

[2] Written dialect data was also collected in the "Synalm" project (Brandner (2015))

[3] For a differentiation of these terms see Schmidt (2011).

(1992) stated that Swiss German is rarely used in the written domain, while Aschwanden (2001), only nine years later, observed that a written form of Swiss German is widely used by new media users. And this to such an extent that she regarded the diglossic division of "standard while writing, dialect while speaking" obsolete.

A similar paradigm shift has taken place in South Tyrolean dialects, a southern Austro-Bavarian variety of German. South Tyrol is the northernmost province of Italy, with a German-speaking population of about 360.000 (official numbers reported by the Autonome Provinz Bozen 2022). Even though political and linguistic conditions differ between the two regions – with South Tyrolean German-speakers being a linguistic minority in Italy – the usage patterns of standard and dialect have developed comparatively similarly within the digital space: South Tyrolean in its written form is the language of choice in the digital communication of younger generations.

While first initiatives to collect a representative sample of South Tyrolean German aimed to describe local peculiarities of Standard German (Abel et al. 2009), in 2016 the DiDi corpus was published by Frey et al. (2016). This corpus for the first time captures the digital communication of South Tyrolean internet users. The authors collected nearly 40.000 Facebook comments, posts and messages from 136 South Tyrolean Facebook users. The users contributed 600.000 tokens mainly written in Standard German, Italian and especially the local dialect as shown in example (1).

(1) STD hel wert schwar bin lei mehr bis muntig do:-([4]
SG Das wird schwer, bin nur noch bis Montag hier:-(
EN That will be hard, I'm only here until Monday:-(
(id: 55211_px0246_c0676)

Moreover, the corpus was enriched with extensive meta-data about the users and each posting. An interesting information on posting level regards the variety the posting is written in. Just taking the German material of the corpus into account this results in the distribution of varieties seen in Table 1.

[4] STD = South Tyrolean Dialect, SG = Standard German, EN = English

German dialect	German standard	German undefined
6863	8145	8180

Table 1: N° of posts by variety

As becomes evident, over a third of all postings are undefined samples that are potentially assignable to one of the two other categories, i.e., standard German or dialect. This relatively high proportion of undefined instances results from the automatic, rule-based labeling procedure the authors applied. Hence, in Glaznieks & Frey (2018: 870) the DiDi authors emphasize that the resource would benefit from a manual revision of undefined samples. Given the uniqueness of the DiDi resource and its great potential for the research of dialects, sociolinguistic patterns, digital writing cultures, the establishment of new writing systems, standardization processes and so on, we wanted to further enhance DiDi. Thus, as suggested by Glaznieks & Frey, the **first aim of the present thesis was to minimize the number of undefined samples**, thereby maximizing the research potential of the corpus. This is not done in a completely manual procedure, but in a semi-automatic re-labeling process: A fourth of the undefined contributions will be labeled manually. With this data, a neural categorizer is trained that will assign the rest of the undefined samples with the output being manually checked at the end. The **second aim is to use the established neural architecture as well as the revised corpus to develop a dialect labeling system on word basis**. In the original corpus whole posts are labeled as dialect or non-dialect. This, however, is partially inapt, especially in mixed cases, which were often just labeled as undefined. With a tagging system on the word-level, it will be possible to label samples in a more fine-grained way. The efficacy of this system will be evaluated on another South Tyrolean dialect corpus, which is part of the MoCoDa project (Marx et al. 2020). In this way, one resource can benefit from knowledge transferred from another.

In order to provide a theoretical framing of dialect writing, describe the resource and implemented systems and give an impression of their potential, the present work is structured into six main parts: Part I will provide an overview of the linguistic situation in South Tyrol covering a short historical wrap-up and linguistic classification. This comprises of a description of the most salient and common features among the local dialects, discussion about the relationship of standard German and the local varieties as well as some considerations about the emergence of written dialect in the new media. At this point it should be noted that in this

work written dialect is only considered in its spontaneous form, written by laymen in a private and spontaneous communication context. This also follows from the resource at hand, as the DiDi corpus does not cover genres like dialect poetry.

The second part will take a closer look at the DiDi corpus. The corpus elements will be analyzed and described statistically to reveal potential biases. Moreover, Part II will describe how the corpus was pre-processed to best fit the needs of the tasks at hand.

Part III will give a short introduction of the principles of neural AI systems. This section will start with an overview of classification in Machine Learning leading up to the benefits of neural nets. Then, the two systems used to solve the tasks at hand will be described in more detail.

Part IV will explain what steps were taken to be able to apply the chosen neural architectures to the tasks. The first section will give a detailed description of how undefined samples were reduced in the corpus as well as an analysis of the results. The second section looks at the follow-up task – the establishment of a word-level labeling mechanism that can differentiate between standard German and dialect words. As there were given no labels on word-level, it will describe the algorithm used to create the training data and which results could be achieved in this way on the MoCoDa data.

Part V starts with potential improvements to the classification pipeline. Then, it will present two potential use-cases for the developed systems as well as the final, extended DiDi resource. Use case 1 will show how DiDi can be used to make inferences about the relationship between the socio-demographic variables, age and gender, and dialect use. In Use case 2 the classification system is applied to gather GPS-referenced dialect data from Twitter, that can be plotted to create a map with dialect data, potentially interesting for dialectometric research.

The work will be finalized in Part VI by ethical considerations regarding the applied systems as well as a final wrap-up.

1.2 A short history of (South) Tyrol

The region in the middle of the Alps historically termed as “Tyrol” (nowadays comprising the federal state Tyrol in Austria and the region Trentino-South Tyrol in Italy) has always been an area of transit and linguistic diversity. Tyrol, representing one of the most important entry points to the Italian peninsula, already during ancient and medieval times was an area of intensive commercial and linguistic exchange. During the era of the Roman Empire the area was

dominated by tribes termed as “Raeti” by Roman historians. These tribes, speaking a language potentially related to Etruscan (Schumacher 1998), were Romanized under the Roman dominion leading to the emergence and dissemination of the Rhaeto-Romanic languages in the region.

After the fall of the Roman Empire, Germanic tribes took possession of the land with Bavarian tribes cementing their hegemonial position eventually. Due to these ethnical shifts in the early Middle Ages, the Rhaeto-Romanic languages were pushed to the periphery and Bavarian varieties became the most influential idioms (Wiesinger & Greule 2019). During the High Middle Ages one Bavarian dynasty in particular managed to expand their influence and territories: the counts of Tyrol. Thus, the whole region became known as Tyrol. After the death of the last duchess of Tyrol, the territory was passed on to the House of Habsburg and, in this way, became part of the Austrian Empire later.

After World War I and the surrender of the Central Powers in 1918 Tyrol was divided: The northern and eastern parts of the region were merged into the federal state of Tyrol belonging to the newly formed First Austrian Republic. The southern parts, i.e., South Tyrol and Trentino (formerly called “Welsch Tyrol”), were integrated into the territory of the Kingdom of Italy. This occupation had severe consequences, especially for the German-speaking population. With the rise of Italian fascism in the 1920s, an era of repression and forced “Italianization” began. The German language was prohibited, German family and place names were translated into Italian, German schools closed. The fascist suppression culminated in an agreement between Hitler and Mussolini in 1939: German speaking South Tyroleans were to decide if they would leave their homes and be resettled in the Third Reich or stay and renounce their German language and culture. 85% of the German population, corresponding to over 200.000 people, decided to leave and become citizens of the Third Reich. However, due to the outbreak of World War II these plans could only be partially realized. By 1943 about 75.000 people had emigrated (Steininger 1997).

After the end of the Second World War the way towards an autonomous province of South Tyrol within the Republic of Italy began. Nearly eighty years later, the Autonomous Province of Bozen/Bolzano–South Tyrol has extensive autonomous rights in sectors like education, administration, and legislation. The German-speaking population of South Tyrol is often termed one of the “best protected linguistic minorities in the world” (see e.g. Baumgartner &

Hechensteiner 2022). The next chapter will take a closer look at the current linguistic situation in the region.

1.3 The linguistic landscape of South Tyrol

Due to the recurrent linguistic shifts and contact between languages described in the previous chapter, South Tyrol nowadays has a very diversified linguistic landscape. In 2001 the Provincial Statistics Institute of South Tyrol reported that 68.27% of the population declared itself as German, 27.42% as Italian and 4.3% as Ladin, a Rhaeto-Romanic language. The language groups are not distributed evenly on the whole territory (ASTAT 2001). Figure 1 provides a good overview of the geographical distribution of language groups:

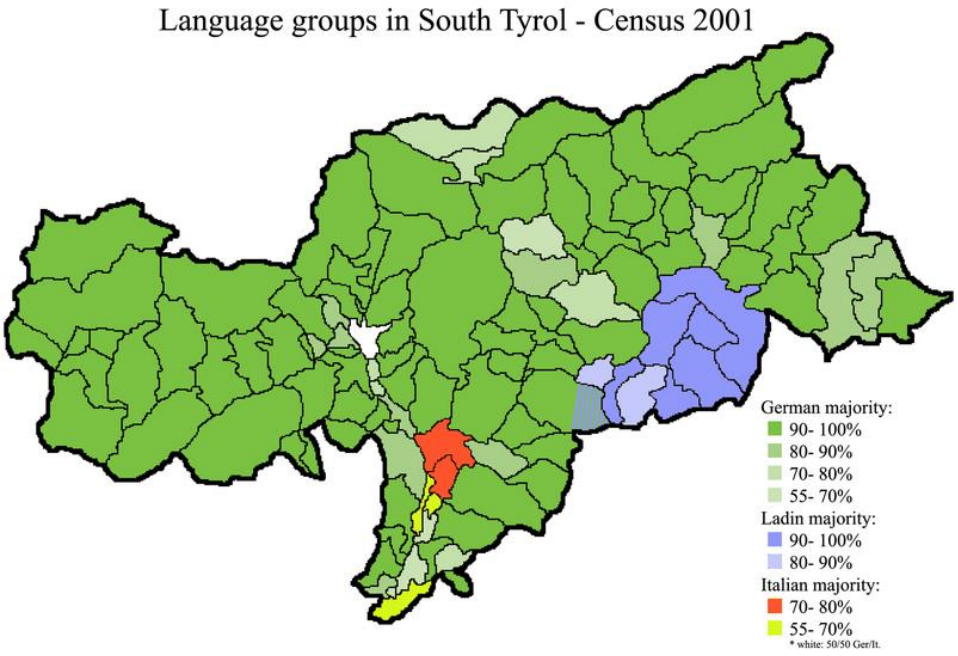

Figure 1: Distribution of Language Groups in South Tyrol[5]

As can be seen, the Ladin language group is concentrated in the lower eastern part of South Tyrol. This is a direct consequence of Rhaeto-Romanic languages being superseded by Germanic settlers and driven to more inaccessible mountainous areas. The Italian language

[5] source: https://commons.wikimedia.org/wiki/File:Language_distribution_in_South_Tyrol,_Italy_2001.png with data from ASTAT (2001)

group, in turn, is particularly dominant in the more urbanized areas, especially the capital of South Tyrol, Bozen/Bolzano. This is mainly due to the settlement of Italians under Mussolini to promote the Italianization of South Tyrol. The fascist regime built a whole new part of town and industrial area for the incoming settlers. In this way the proportion of the Italian speaking population rose from 2.9% in 1910 to 34.3% in 1961 (ASTAT 2011). Despite the fascist settlement program, the majority of South Tyroleans nowadays declare themselves as being part of the German language group. This is reflected by the fact that most of the areas Figure 1 are green. Especially in rural areas German is by far the most spoken language. The following chapters will elaborate on the German varieties spoken in South Tyrol as well as the relationship between German South Tyroleans and the German standard variety.

1.4 The German varieties of South Tyrol

Even though South Tyrol covers a relatively small area geographically, it has a remarkable diversity of German varieties. As was already mentioned in the introduction, the German dialects of South Tyrol belong to the Southern Bavarian dialects. As can be seen in Figure 2, this group of dialects is contiguous to Alemannic dialects (and Rhaeto-Romanic varieties) in the west, and Middle Bavarian dialects in the north and east. In the South, Southern Bavarian borders Italian and Slavic varieties.

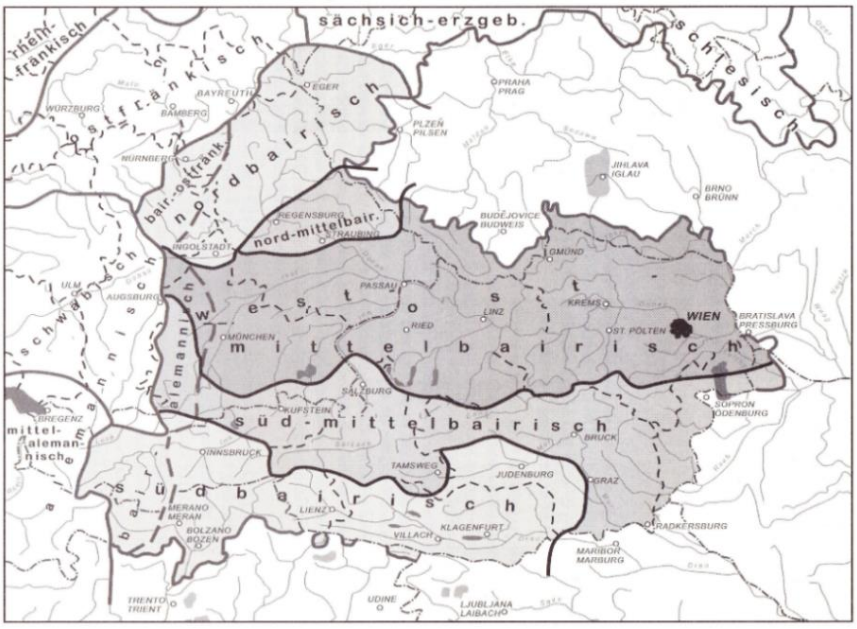

Figure 2: Division of the Bavarian dialect continuum (Pichler-Stainern 2008: 57)

Looking at the Tyrolean dialects themselves, a west-east dividing-line can be drawn. Even though historical Tyrol nowadays is part of two different countries, the most important isoglosses still have a north-south orientation, cutting, in this way, national borders. According to Meraner & Oberhofer (1982) the pre-dominantly German speaking parts of historical Tyrol can be divided into three main areas (see Figure 3): Western, Central and Eastern Tyrol.

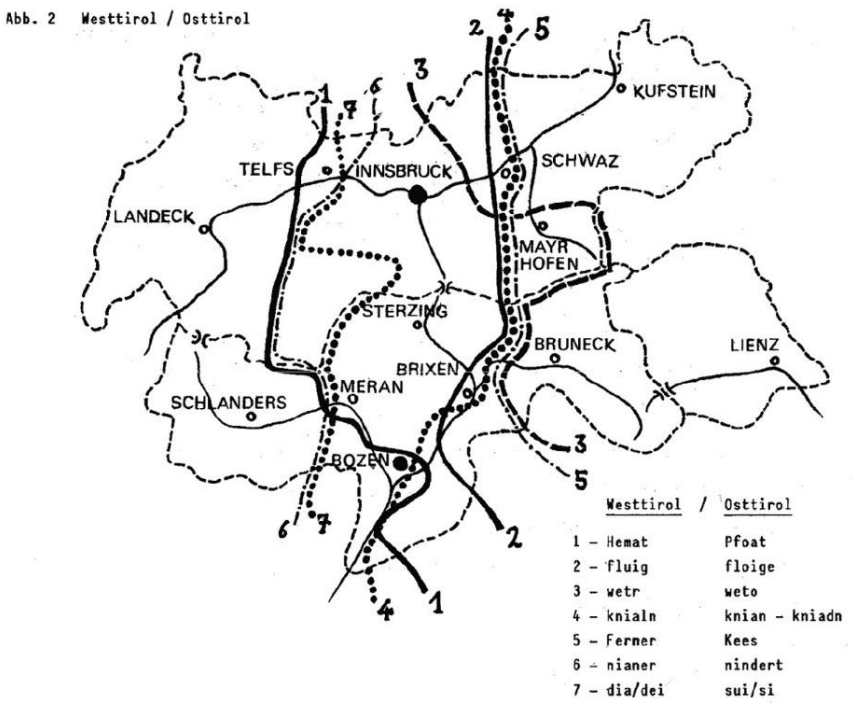

Figure 3: Dialectal division of Tyrol (Meraner & Oberhofer 1982)

Lanthaler (1990) describes the western part as being influenced by the contact situation with Alemannic varieties while the eastern parts show similarities with the adjacent Carinthian dialects. The central part can be seen as transition zone showing features of western and eastern Tyrolean dialects. The examples in (2) show examples taken from DiDi written in an eastern (a), western (b) and central (c) South Tyrolean dialect.

(2) a) STD I schaug zwor irgendwie jedn Tog vorbei, obo olbm la kurz odo mitn Handy, wo i sowieso et gearn longa Texte schreib.

SG Ich schaue zwar irgendwie jeden Tag vorbei, aber immer nur kurz oder mit dem Handy, wo ich sowieso nicht gerne lange Texte schreib.

EN I somehow look by every day, but only for a short moment or with my smartphone, where I don't like to write long texts anyway.
(id: 54631_t0130_m00224; region: Pustertal)

b) STD jo und eis miats mol verstean dass eis radlfohrer u nit fisch seits!
SG ja und ihr müsst mal verstehen dass ihr Radfahrer und keine Fische seid!
EN yes and you have to finally understand that you are cyclists and not fish!
(id: 55216_p06351; region: Vinschgau)

c) STD alzeimer werds es a net hoben... des muaß an ondore Kronkheit sein...
SG alzheimer werdet ihr auch nicht haben, das muss eine andere Krankheit sein…
EN You won't have Alzheimer's… it has to be another disease…
(id: 57086_t0077_m18751; region: Bozen)

It is not possible to give an in-depth linguistic analysis of the instances in (2), but we want to point out a few salient features and differences: All of the instances show the pattern of writing <o> instead of <a> in words like "Tog" ("Tag", "day"), "mol" ("mal", "finally"), "Kronkheit" ("Krankheit", "disease"). This is typical for South Tyrolean dialect writing as we will see in the next chapter. Moreover, all instances exemplify the rich diphthong system, characteristic for South Tyrolean varieties. Examples are words like "gearn" ("gern", "like"), "miats" ("müsst", "have too") and "verstea" ("versteh", "get me"). Another prominent common feature is the personal pronoun "es" ("ihr", "you"), indicating second person plural and appearing in (2b) and (2c). When it comes to differences, typical features of eastern dialects are the replacement of the suffix <er> with <o> as in "obo" ("aber", "but") and "odo" ("oder", "or"). The western and central dialects, on the other side, most often preserve word-final <er> as can be seen in the words "radlfohrer" ("Radfahrer", "cyclists") and "alzheimer"[6]. The eastern dialect is also missing the word-initial nasal sound in the negation particle "et" ("nicht", "not"). In the western and central dialect examples the nasal onset is preserved in the negation particle "nit" (2b) and "net" (2c). The vowel in the negation particle, however, seems to be more similar between the eastern and central dialect, if we compare "nit" (western) with "net" (central) and "et" (eastern).

[6] refer to Glaznieks & Glück (2019) for an analysis of this suffix in the DiDi corpus.

South Tyrol, thus, from a geo-linguistic perspective has a three-part division. However, space is not the only relevant factor of linguistic diversity of German varieties in South Tyrol. Lanthaler (1996) provides a good starting point for an overview of additional factors of linguistic diversification: Referring to the work of Kranzmayer (1956) he first introduces a topologically motivated classification of Tyrolean dialects: "Zentraltirol" (central Tyrol) and "Randtirol" (marginal Tyrol). The term "central Tyrol" captures dialects spoken in the major valleys, characterized by a more dynamic linguistic exchange and leveling-out of specialized, very conservative dialectal features. "Marginal Tyrol", on the other hand, designates more conservative dialects, spoken in remote alpine high valleys.

Lanthaler now argues that this classification, even though valid in the 1950s, is barely transferable to the modern dialectal situation. He describes how even the most conservative Tyrolean base dialects were influenced by modern developments like publicly available broadcasting, tourism and higher mobility due to a better transport infrastructure (Lanthaler 2001). "Marginal Tyrol", thus, is constantly shrinking as more conservative dialectal features are leveled out (see also Lanthaler 1974 for concrete examples). Nevertheless, more conservative dialects are still a linguistic reality for older generations that grew up in a mainly rural, agricultural world. Another diversifying factor, thus, is time: The higher mobility and more intensive linguistic exchange led to the development of regionally more wide-spread variants, especially among younger generations.

Besides the aforementioned factors of diatopic and diachronic factors, diastratic variation also plays an important role in the formation of the linguistic landscape in South Tyrol. The higher mobility, tourism and international communication possibilities (e.g., Austrian and German telestations) also led to a more intensive contact situation with different forms of Standard German. When it comes to the use of these Standard German variants, diastratic and diaphasic effects can be observed in South Tyrol.

On the one hand, in the main city of Bozen/Bolzano, a dialect influenced by standard German phonetic features has evolved. Lanthaler (1996) describes this variety as being used predominantly by the upper classes in the capital. (3) provides an example taken from DiDi.

(3) STD HA HA i bin ganz alleine in <GeoNE> wohn seit Freitag da hann mit an Koleg a ganzes altes Bauernhaus gemiatet

SG HA HA ich bin ganz alleine in <GeoNE> wohn seit Freitag hier hab mit einem Freund ein ganzes altes Bauernhaus gemietet

EN HA HA I am all alone in <GeoNE> living here since Friday and have rented a whole old farmhouse with a friend

(id: 56969_t0150_m00302)

When comparing the postings in (2) and (3) with their respective Standard German transliteration it becomes clear that the variety in (3) is much closer to the standard. It completely lacks, for example, any <o>-for-<a>-writings. Instead, <a> is even kept in words which are clearly influenced by the underlying dialect like the transcription of "hann" (dialect form: "honn") for "habe" ("have"). Other dialectal elements are the undefined articles "an" and "a" as well as the diphthong in "gemiatet" ("gemietet", "rented").

On the other hand, every German speaking South Tyrolean can switch between varying degrees of Standard German and dialect according to the communicative situation, audience and topic. The examples in (4) were all written by the same user (55442) and give an impression of the variational spectrum of South Tyrolean dialect writers.

(4) a) STD Bo meinr Konditioun kim i grod amol bis zur Tir oi, nor muaß i a schun a Stindele roschtn^^

SG Bei meiner Kondition komme ich grad mal bis zur Tür runter, dann muss ich mich auch schon ein Stündchen ausruhen^^

EN The shape I'm in I would just come to the door, then I would have to take a break for one hour^^

(id: 55442_p03293)

b) STD? Servus Leute! Hat jemand Bock am Sunnta 24. November <InstNE> Konzert mitzugiahn?

SG Servus Leute! Hat jemand Bock am Sonntag 24. November auf das <InstNE> Konzert mitzukommen?

EN Hi folks! Is there somebody who wants to go to the <InstNE> concert with me on Sunday 24th November?

(id: 55442_p00338)

c) SG? Perfekt! Danke <PersNE>! Werd ich glatt n paar hundert machen gehn!

EN Perfect! Thank you <PersNE>! I am going to make a few hundred!

(id: 55442_p03293_c5663)

d) SG oh ja. aber ne gute Nachricht gibts für mich schon mal. Morgen gehts ab nach HAUSE!

EN oh yes. but there is good news for me already. Tomorrow I will go HOME!

(id: 55442_p03293)

While (4a) is difficult to understand for readers not proficient in a local vernacular, (4b) is accessible for readers with a proficiency in Standard German. Even though it contains words with a dialectal transcription like “Sunnta” (“Sonntag”, “Sunday”) and “mitzugiahn” (“mitzukommen”, “to come with”), the author of the posting intermingled these elements with Standard German words. This code mixing is the reason why we marked this dialect example with a question mark. (4c) can already be termed a Standard German sentence, though with colloquial style markers like the reduction of “ein” (“a”) to “n”. The wording “machen gehn”, however, could be interpreted as an interference phenomenon. As Paul et al. (2022) observe the construction “infinitive+gehen” undergoes a grammaticalization process in modern Standard German, but is not yet consistently accepted by all speakers/hearers. Even if, to our knowledge, there are no corresponding studies of this grammaticalization process in South Tyrolean dialects yet, in our experience it is very productive and widely accepted in these varieties. Finally (4d) is a sentence in colloquial written standard German and does not offer any clues of the author’s origin.

The present chapter has shown that the dialectal usage patterns in South Tyrol depend on diatopic, and in the broadest sense, diachronic factors. The relationship between dialect and Standard German is determined by diastratic and diaphasic factors. As the relationship of Standard German and South Tyrolean dialects is especially relevant for the establishment of written dialects in the new media, chapter 1.4.2 will describe the dialect-standard continuum in South Tyrol in more detail. Beforehand, the following chapter will shed light on the most salient dialect features of South Tyrolean dialects.

1.4.1 Common features of South Tyrolean dialects

At first, it should be noted that there is no such thing as a universal South Tyrolean dialect. The previous chapter already made clear that the historical and topological conditions led to the formation of very distinct varieties. The lack of conceptual clarity in this case stems from the fact that terms from different disciplines are denotationally mixed. While "South Tyrolean" is a geo-political term, "dialect" is a linguistic one. Geo-political and linguistic spaces often overlap but are rarely congruent. The interaction of geo-political and linguistic factors is also reflected by the fact that the DiDi corpus only contains South Tyrolean samples, even though, from a linguistic perspective, an isolated South Tyrolean dialect area does not exist. The denominations "South Tyrolean" or "written dialect" that are used in upcoming chapters, thus, are simplifications. The reason for this simplification is that a complexity reduction is needed for the models described in Part III and Part IV to build meaningful categorical representations. By breaking down the dialect-standard continuum to a few categories, the problem of sparse data can be circumvented as not every distinguishable sub-variety receives its own label. The task of the neural models presented in this work is not to distinguish between fine-grained varieties, but to build a reliable and generalizable "written dialect" vs "written standard" distinction.

Alber (2020: 40–45) singles out five salient phonetic and two morphological characteristics of Tyrolean dialects. We added examples from DiDi to show that they are also expressed in dialect writing.

1. <u>Phonetical features:</u>
   a) [a] to [ɑ]: The Middle High German anterior vowel [a] turned into a more retracted posterior vowel [ɑ] often denoted as <o> in spontaneous dialect writing:

      e.g. <Kachel>, [ˈkaxl] ("tile") → <Kochl>, [ˈkɑχl]

      (5) STD w**o**r zach **o**br h**o**n sie p**o**ckt
      SG w**a**r krass **a**ber h**a**b sie gep**a**ckt
      EN it was challenging, but I did it
      (id: 55442_p00500_c5728)

b) Unrounding: Vowels commonly referred to as umlaut-vowels, written as <ö> and <ü> in German texts, turn into unrounded vowels, i.e. [i] and [e].

e.g. <Köpfe>, [ˈkœp͡fə] (“heads”) → <Kepf>, [ˈkep͡f]

(6) STD wie geat es passwert f**i**rn internet dahoam f**i**r di gäste?
SG wie lautet das passwort f**ü**r das Internet zuhause f**ü**r die Gäste?
EN what’s the password for the internet at home for the guests?
(id: 56936_t0912_m17718)

c) Complex diphthong system: Due to several sound change processes, Tyrolean dialects have more diphthongs compared to Standard German. Some of these diphthongs are [ɔɐ], [ɛɐ], [ie] and [ua].

e.g. <Schnee>, [ʃneː] (“snow”) → <Schnea>, [ˈʃnɛɐ]
<gut>, [ˈguːt] (“good”) → <guat>, [ˈguat]

(7) STD Sch**ia**ns schn**ea** schepfn in olle!
SG Allen schönes Schneeschöpfen!
EN Happy snow shoveling to all!
(id: 56304_px0013_c0024)

d) s-retraction: Not only word-initial, but also word-internal sibilants preceding a consonant are realized as postalveolar [ʃ] sound.

e.g. <Schwester>, [ˈʃvɛstɐ] (“sister”) → <Schweschter>, [ˈʃvəʃtɐ̯]

(8) STD so a Schwe**sch**ter wia meine gibts lei OAN MOOOL !
SG so eine Schwe**s**ter wie meine gibt es nur EIN MAAAL !
EN a sister like mine exists only ONCE !
(id: 56150_p02934)

e) schwa-apocope and -syncope: The schwa-sound is deleted word-initially and in some cases also word-internally in unstressed syllables.

e.g. <fahre>, [ˈfaːʁə] (“I drive”) → <fohr>, [ˈfɑːʁ]
<gerannt>, [gəˈʁant] (“run”, participle) → <grennt>, [g̥ˈʁɛnt]

(9) STD gegn di mauer bin i net grennt…^^ so an hortn schedl hon i decht widor net...

SG gegen die Mauer bin ich nicht gerannt…^^ so einen harten Schädel hab ich doch wieder nicht…

EN I didn't hit the wall…^^ I don't have that a hard of skull…

(id: 55233_p06292)

2. Morphological features:

a. Special pronouns and inflection suffix for 2nd person plural: Tyrolean and other Bavarian dialects use the pronouns "es" ("you", nominative), "enk" ("you", accusative), "enkere" ("your", possessive) instead of "ihr", "euch" and "euer". Moreover, also the inflectional suffix of the present tense in the 2nd person plural is different, i.e., "-ts" instead of "-t".

e.g. <ihr lach**t**> ("you laugh", 2nd person plural) → <es loch**ts**>

(10) STD i glab net dasses mit 1 pro kopf genua hobs…..

SG ich glaub nicht, dass ihr mit 1 pro kopf genug habt…..

EN I don't think you will have enough with 1 per person

(id: 55844_px0325)

b. Loss of preterit tense: Like in most Upper German dialect varieties preterit tense is replaced with perfect tense. Thus, instead of using inflection to mark past tense, a combination of an auxiliary verb and main participle verb is used.

e.g. <ich lach**te**> ("I laughed") → <I **hon** glocht>

(11) STD ….jo wirklich hon i jetzt glocht

SG ….ja wirklich hab ich jetzt gelacht

EN ….well I really had to laugh just now

(id: 57031_t0620_m01789)

This chapter could only provide a very general overview of the properties of (South) Tyrolean dialects. Syntactic and lexical peculiarities were not considered at all. It suffices, though, to draw a first distinction between Standard German and the local dialects, which is central for analyzing the mistakes the neural models made, in chapter 4.1.3. Moreover, this knowledge is

used to optimize and evaluate the model in chapter 4.2.1. The next chapter will describe the complex dialect-standard continuum in South Tyrol.

1.4.2 Diglossia and the dialect-standard continuum in South Tyrol

Looking at the dialect examples given so far it is evident that South Tyrolean writers do not simply insert dialectal elements into a standard German expression, but consistently transliterate their dialect. This predominance of dialect transliteration in South Tyrolean computer-mediated communication (abbreviated as "CMC" in the following) is by no means an obvious phenomenon. As standard German is the written language German speaking South Tyroleans learn at school, this variety would be the obvious choice. From a writer and reader perspective, writing in dialect requires re-adaptation to a new literacy. This is associated with an increased cognitive effort: Writing in dialect requires the establishment of new word forms and a higher tolerance for non-normative writing as spellings can significantly differ between writers even in the same local variety. It can thus be assumed that writing in dialect increases personal cognitive effort while decreasing efficacy. In short: Consistent dialect writing has a price. To get an understanding why German speaking South Tyroleans are willing to pay this price, it is necessary to take a closer look at the complex relationship between standard German and dialect in South Tyrol.

Under the Austrian Empire, the Tyrolean standard was the standard of imperial-royal authorities. From the middle of the 18th century onwards, this standard was the Modern High German established by Luther extended by Austrian peculiarities, i.e. an Austrian predecessor of modern standard German (see Wiesinger 2014). After the fall of the Austrian empire and the annexation of South Tyrol a whole generation barely received a proper German school education due to fascist Italianization efforts. As Italian was the only language allowed in official schools, standard German was exclusively taught in illegal underground schools, called "catacomb schools". As Lanthaler (2007) describes, this caused a major disruption in the standard German tradition. After the end of fascism in Italy, South Tyrolean authorities would not seek orientation in the Austrian but the Federal German standard, accepting the DUDEN as the most important normative entity. Teachers were urged to consistently speak standard German and to control the standard German expression of their pupils. As the trauma of fascist language policies was still very present and authorities were afraid of a language assimilation process similar to the one in Alsace, the guiding principle was to sustain the cultural and linguistic ties to the international German speaking world. Post-war aid programs initiated by

the Federal Republic of Germany as well as federal German tourists and tv programs reinforced this tendency. This orientation, however, led to a standard German prestige language distant to local phonetic and grammatical patterns. This distance might be one reason for what Lanthaler (1990) describes as a feeling of "alienation", when South Tyroleans use standard German in private oral contexts.

In this way, the linguistic and socio-political developments led to a linguistic situation called diglossia (Ferguson 1959). Standard German in South Tyrol is only used in official contexts like school education, formal speech acts, political and legal situations, and the news. On all other occasions the local dialect is the language of choice. Lanthaler even speaks of a "psychological barrier" that South Tyroleans must overcome when they want to speak standard German in private contexts. Especially in more formal oral contexts, like political communication, this leads to many forms of trade-off phenomena. For example: Their formal office requires politicians to use the standard in spoken communication. However, as this standard is perceived as detached and reserved, South Tyrolean politicians tend to mix dialectal and standard elements creating various transitional varieties. Besides diachronic and diatopic variation this also leads to an extensive diastratic and especially situation-dependent, diaphasic variation of German in South Tyrol. Until recently, these observations were only true for the oral domain. In the written domain, standard German was the predominant variety, regardless of context. As has already been mentioned in the introduction, this habit has radically changed with the emergence of the new media. Nowadays, South Tyrolean dialects are also a written variety. The next chapter will provide a theoretical framing of this change.

1.4.3 Written dialect and the communication model of Koch & Oesterreicher

In the previous chapters it has become clear that South Tyroleans transliterate their dialect on a word for word basis, rather than just inserting single dialectal elements in their written communication. The term "written dialect", however, can be somewhat misleading, because it implies that the written dialect is just the textual equivalent of the spoken dialect. But for spontaneous dialect writing this cannot hold true.

To understand why, we have to take a look at the relationship between standard and dialect and how this relationship is different for the written and oral domain: Schmidt (2011: 149) defines dialect as "the least standard and most local (regionally restricted) full variety". As becomes clear in this definition, the terms "standard" and "dialect" are highly interdependent.

As conceptualized here, dialect is a relational entity, defined by its property of deviating from a theoretical norm. If we apply this definition to the dialect in its spoken variant, it becomes obvious that the concept of dialect as the deviating variety is, above all, a theoretical, methodological consideration: A child learns a dialect as native language before even knowing that there is another, normative variety. A child does not perceive their way of communicating as deviating from this norm. The deviation is conceptually introduced later on, when the child starts to figure out the social implications of both varieties.

Another picture emerges when looking at written dialect. As described in the previous chapter, the German standard is the language of education in South Tyrolean schools. Especially in the written domain the standard orthography is the only system. When South Tyroleans start to write their dialect, the starting point is a very different one: They must adapt a system, originally developed for standard German, to represent their dialect. In contrast to the spoken dialect, the written dialect, thus, cannot be conceptualized as a completely standard-independent variety. In this way, writing dialect is, indeed, a deviation from standard German writing.

To formulate it in slightly exaggerated terms: Written dialect could be termed a standard German variety, while this is not the case for the spoken dialect. It follows that the standard German dimension always plays a role when writing in dialect. A dialect writer, consciously or subconsciously, must decide for each word to what degree he wants to deviate from the internalized writing system. As Müller (2011) describes it, the writer is oscillating between two poles: "standardnah" (close to the standard) and "lautnah" (close to the sound/phonetical impression). The writer can either follow canonical writing patterns, for example if she/he decides to write the word-initial consonant cluster of the word "Stoff" ("tissue") with <st> as the standard rule system demands it, or she/he can stick more closely to the sound perception by realizing it as <scht>, because word-initital <st> is realized [ʃt] in Upper and Middle German dialects and standard German. Which strategy dialect writers choose is not arbitrary[7]. It depends on numerous factors like situation, communication partner, social identity etc..

The communication model of Koch & Oesterreicher (1985) provides a more abstract theoretical framing of this decision process. According to Koch & Oesterreicher, every form of

[7] See also Felder (2015) and Huber & Schwarz (2017) for concrete spelling analyses in Swiss German and South Tyrolean dialect writing.

communication (e.g. telephone conversation, job interview, newspaper article, private letter etc.) can be placed on a continuum of conceptual orality and conceptual literacy. The former the authors also call the pole of immediacy, the latter the pole of distance. To understand this model, the term "conceptual" is crucial.

When referring to orality/immediacy and literacy/distance Koch & Oesterreicher are not referring to the actual medium, i.e. phonic or graphic representation of language. The authors rather argue that "written" and "spoken" can also be conceptualized as theoretical constructs. Even though "written" and the graphic code as well as "spoken" and the phonic code share clear affinities, they are not equivalent. A political speech, for example, even though it is an oral communication form, has features of conceptual literacy. The opposite is true for digital chat rooms. To clarify what they subsume under the concepts of immediacy and distance Koch & Oesterreicher (2012: 450) also specify which features are characteristic for both concepts. This is illustrated in Figure 4:

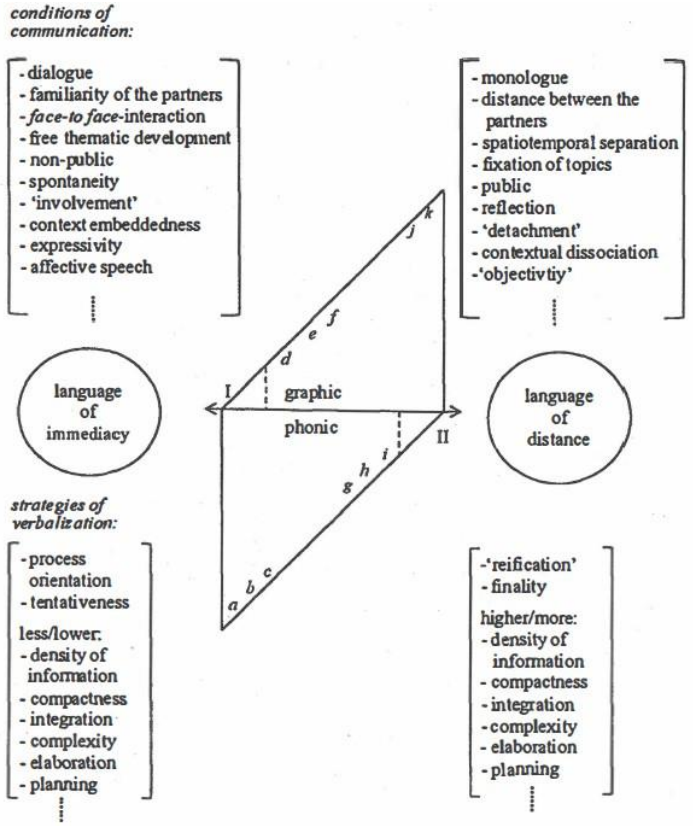

Figure 4: Features of immediacy and distance

As Figure 4 shows, the languages of immediacy and distance are characterized by distinct, mostly complementary features. A communication form very close to the immediacy pole is,

for example, characterized by a higher familiarity of the communication partners interacting face-to-face in a non-public dialogue. By contrast, a distant communication form involves a greater distance between the partners and is realized in a monological, spatio-temporally separated manner. The letters in the figure denote examples for communicative forms on this continuum: *a*, for example, is an intimate conversation, conceptually very close to a language of immediacy. *k* denotes an administrative regulation and is conceptually very close to a language of distance.

In the context of the present work, the model of Koch & Oesterreicher has double relevance because it can answer two central questions: "Why are people starting to write their dialect at all?" and "What influences which pole ("standardnah" or "lautnah) they tend to while writing?"

The answer to the first question lies in the placement of digital chat communication on the immediacy-distance continuum. Looking at the characteristic features in Figure 4 it becomes evident that digital chats fulfill a lot of the immediacy criteria: they are a form of dialogue, the partners are familiar with each other, the conversation is spontaneous etc.. Given this observation and the South Tyrolean diglossia situation described in the previous chapter, the decision to write in dialect seems like a natural choice: As the standard is predominantly used in formal situations and communication forms, it can be termed a language of distance in South Tyrol. The South Tyrolean dialect, on the other hand, is a proto-typical example for a language used in conceptually immediate, close contexts. Therefore, as a digital chat conversation can be seen as a context of closeness, and because distant and close contexts have their own variety in South Tyrol, South Tyroleans intuitively choose the written dialect for text-based-chatting.

These dynamics also play a role in the concrete realization of individual written dialect words. As Müller (2011: 165) observes for Swiss German, young adolescents often tend to use more sound-related spellings. She argues that in this way they distance themselves from the standard they associate with school and the world of adults, a distant world. The Swiss German dialect, on the other hand, is associated with familiarity and closeness and suitable to indicate group membership. Thus, by moving to the "lautnah", phonetical pole in their written dialect realization adolescents writing in Swiss German can emphasize their ingroup-outgroup concept. As the Swiss German and South Tyrolean situations are very comparable when it comes to the diglossic distribution of varieties, similar mechanisms can be assumed for South Tyrolean dialect writing in general.

This chapter concludes Part I, which focused on theoretical considerations. The following, second part enters a more practical discussion given the tasks at hand by describing the DiDi corpus in greater detail.

PART II – The DiDi corpus

2.1 General overview

DiDi is short for "Digital Natives, Digital Immigrants". The research project, conducted by the EURAC Research Institute in South Tyrol, started in 2013 and was completed in 2019. The aim of the project was to implement a corpus-based empirical survey of CMC of German South Tyrolean internet users. Special attention was paid to the influence of age and digital socialization on the communication practices. The resulting resource was first presented in Glaznieks & Stemle (2014).

Digital communication patterns were captured by recruiting South Tyrolean German-speaking Facebook users belonging to different age groups. Through an app, specifically designed for this purpose, participants could choose which of their Facebook content they wanted to share with the researchers. They had the choice to share three different Facebook text types: posts, comments and messages (participants could share texts of either all of the three categories or just one or two). The data collected only included Facebook content written in 2013. Using the app, users were able (and actively asked) to include socio-demographic information like age ("PA_Alter[8]"), gender ("PA_Geschlecht"), rough native dialect localization ("PA_Dialektsprecher_STIR"), education ("PA_Ausbildungsabschluss"), occupation ("PA_Beruf"), place of residence ("PA_Lebensmittelpunkt_STIR"), etc. In a subsequent step, the collected data was enriched with fine-grained linguistic information. On the posting[9] level this includes a "dialect_tag" variable specifying if the post is written in dialect, Standard German or undefined, as well as an automatically generated "language_tag" variable and the corresponding language identification confidence score ("language_langid_confidence_score").

On the word level, among other information, token, lemma, and pos (part-of-speech) tags are given. For words written in South Tyrolean dialect or deviating in any way from the standard German orthography, a standard normalization layer – "norm" – was added. Genuinely South

[8] Actual variable names are given in another font.

[9] The word "posting" (German equivalent is "Beiträge") is used in this thesis as a general term to refer to all Facebook text types (posts, comments, messages) and should be distinguished from the term "post", which only refers to this specific text type.

Tyrolean words, only occurring in the regional dialects, are indicated by an additional label, “stir”. Finally, personal information such as names of users and places were masked to protect privacy, resulting in variables such as “anonym”, “anonym_category”, and “anonym_gender”.

Due to the detailed meta data included in the corpus (the mere socio-demographic information includes 47 variables) only the most important variables could be listed and described in the present and the following chapter. A detailed overview and description of the collected variables on each level (user, postings, tokens) can be found in Appendix A. Before coming to the description of the corpus specifications in the next and subsequent sections it must be said that the corpus we worked with in this thesis is only a sub-part of the whole resource. More specifically, and in line with Glaznieks & Frey (2018), we only considered data from users writing in German. This constitutes 58% of the whole dataset collected (Frey et al. 2016: 3). The reason was that we wanted to focus on the German varieties of South Tyrol in this thesis.

2.2 Pre-Processing the corpus

`Code files: corpusbuilder.py`

We downloaded the DiDi corpus in json-format for a first inspection in December 2021[10]. This version of the corpus was used for all tasks described in this thesis. The downloaded DiDi corpus was organized by Facebook text type. Every text type, i.e., comments, messages, and posts, has its own file. Thread and user information are provided in two additional json-files.

Within the files, the dataset is organized on the posting level. Each posting, be it a Facebook post, comment or message, has its own entry with corresponding meta data. Figure 5 provides a simple example with the text “danke, <PersNE>! :-)” (“thank you, <PersNE>! :-)”):

[10] Repository address: https://clarin.eurac.edu/repository/xmlui/handle/20.500.12124/7

```
{
  "dialect": "de undef",
  "_id": "56950_p05560_c0119",
  "language_corrected_langs": "de",
  "post": "56950_p05560",
  "language_langid": "lt",
  "language_tag": "de",
  "tokens": [
    {
      "comment": "",
      "pos": [
        "PTKANT"
      ],
      "lemma": [
        "danke"
      ],
      "token": "danke",
      "anonym": "",
      "norm": []
    },
    {
      "comment": "",
      "pos": [
        "$,"
      ],
      "lemma": [
        ","
      ],
      "token": ",",
      "anonym": "",
      "norm": []
    },
    {
      "comment": "",
      "anonym_gender": "male",
      "pos": [
        "ADJD"
      ],
      "anonym_category": "firstname",
      "lemma": [
        "<unknown>"
      ],
      "token": "<PersNE>",
      "anonym": "<PersNE>",
      "norm": []
    },
    {
      "comment": "",
      "pos": [
        "$."
      ],
      "lemma": [
        "!"
      ],
      "token": "!",
      "anonym": "",
      "norm": []
    },
    {
      "comment": "",
      "emoticon": "True",
      "pos": [
        "NN"
      ],
      "lemma": [
        "<unknown>"
      ],
      "token": ":-)",
      "anonym": "",
      "norm": []
    }
  ],
  "like_count": "3",
  "created_time": "2013-08-31T11:56:26+0000",
  "user_id": "56950",
  "message": "danke, <PersNE>! :-)",
  "language_langid_confidence": "0.698343562796",
  "fb_text_type": "comments",
  "newlines": ""
}
```


Figure 5: Example of DiDi corpus entry

As can be seen, the posting's meta data also contain the user id. However, the other user meta data is stored separately. For the purposes of this thesis, this data allocation is not ideal. In this way, the characteristics of a given text can only be combined indirectly with the socio-demographic variables of its author. In Machine Learning contexts, though, the relationship between author and output is often crucial. If, for example, we want to sample a training, test and development set, the authors' gender and age might be important dimensions to guarantee balanced and representative samples.

A user-based structure also has the advantage that user-based analyses are easier to perform. For example, if we want to know, if, and in which way, age, gender, or location of a user influence his/her way of writing dialect (see chapter 5.2). For this reason, it was decided to re-design the corpus structure and to subsume all data under their corresponding users. The new structure is depicted in Figure 6.

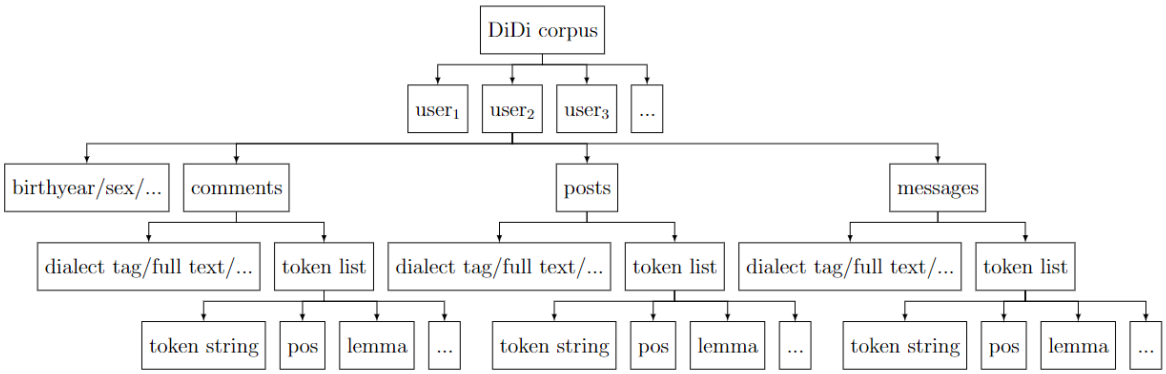

Figure 6: New DiDi corpus structure

The structure was realized in an object-oriented programming paradigm using Python: corpus, users, comments, posts, and messages were conceptualized as separate objects with own methods and organized hierarchically.

The root node is the corpus itself. The corpus object contains a list of user objects. Each user object has its meta data as well as three lists as child nodes, which contain comment, post, and message objects. Each text type object is comprised of meta data about the posting stored in it (e.g., “dialect_tag”, “full text” etc.) as well as a token list. In this token list, all tokens of the posting are separately stored as objects. Each token object heads the token string as well as other information like “pos”, “lemma”, “norm”, etc. To read in the data and store it in the new format, special object methods were implemented.

The different text types were conceptualized as distinct objects because the meta variables, provided for each of them by the original resource, partially differed. Moreover, in the original resource the same variables were not consistently specified for each object. If, for example, a token was not an emoticon, the variable “emoticon” was missing in the token meta data. This saves storage but reduces processability (e.g., it is not feasible to access a variable in an automated loop if it does not exist for every instance). Thus, each object was assigned every possible variable for this object, even if the variable bore no information. In this case, the variable was specified as “n.a.”.

This chapter provided a short overview of how the DiDi resource was pre-processed and restructured to fit the tasks at hand. For a deeper insight into the whole process and architecture please also refer to the comments in the code file. The next chapter presents descriptive key statistics of the DiDi corpus.

2.3 Statistical evaluation

`Code files: 01_1_corpus_analysis_users.ipynb, 01_2_corpus_analysis_texts.ipynb`

The statistical analysis was performed using Python jupyter notebooks. The first of the notebooks listed above mainly focused on user statistics. In total, the corpus contains material of 124 users. Table 2 provides a first overview of important socio-demographic specifications.

Age			**Sex**		**Place of residence**	
Range	*Average*	*Median*	*Male*	*Female*	*South Tyrol*	*Abroad*
14–76	38.975	37	57	67	109	15

Table 2: Socio-demographic user information DiDi corpus

As can be seen in Table 2, the corpus covers a very broad age range with an average age of approximately 39 years and a median of 37. The average and median being close together is an indication of a balanced age distribution. As the demographic report from 2013 (the year of data collection) states (ASTAT 2014: 41), the average age of the South Tyrolean population was 41.8 at that time. Thus, the DiDi corpus users are slightly younger than average, which is to be expected for an internet-based resource. With 57 men and 67 women, women are slightly overrepresented. Their proportion in the corpus is 54%, while in the same age range the proportion of women stood at slightly over 50% in the total population (between 15 and 79 there were 98.13 men to 100 women; ASTAT 2014). 109 users had their place of residence in, 15 outside South Tyrol.

Table 3 shows the native languages of DiDi users.

German	**Italian**	**Ladin**	**Total**
yes	no	no	111
yes	yes	no	11
yes	no	yes	2

Table 3: Native languages of DiDi users

All users in the corpus are native German speakers. 111 of them are German monolinguals. 11 users are natively bilingual with Italian, 2 with Ladin. There are no users who learned all

three languages as a mother tongue. As all users have German as their native language, it is likely they speak a South Tyrolean dialect variety. To have a rough approximation of which dialect variety the DiDi users speak and maybe also write, it is interesting to look at where the users located themselves. Figure 7 shows the localization answers provided by variable "PA_Dialektsprecher_STIR" (we named this variable "german_dialect_region" during corpus re-structuring).

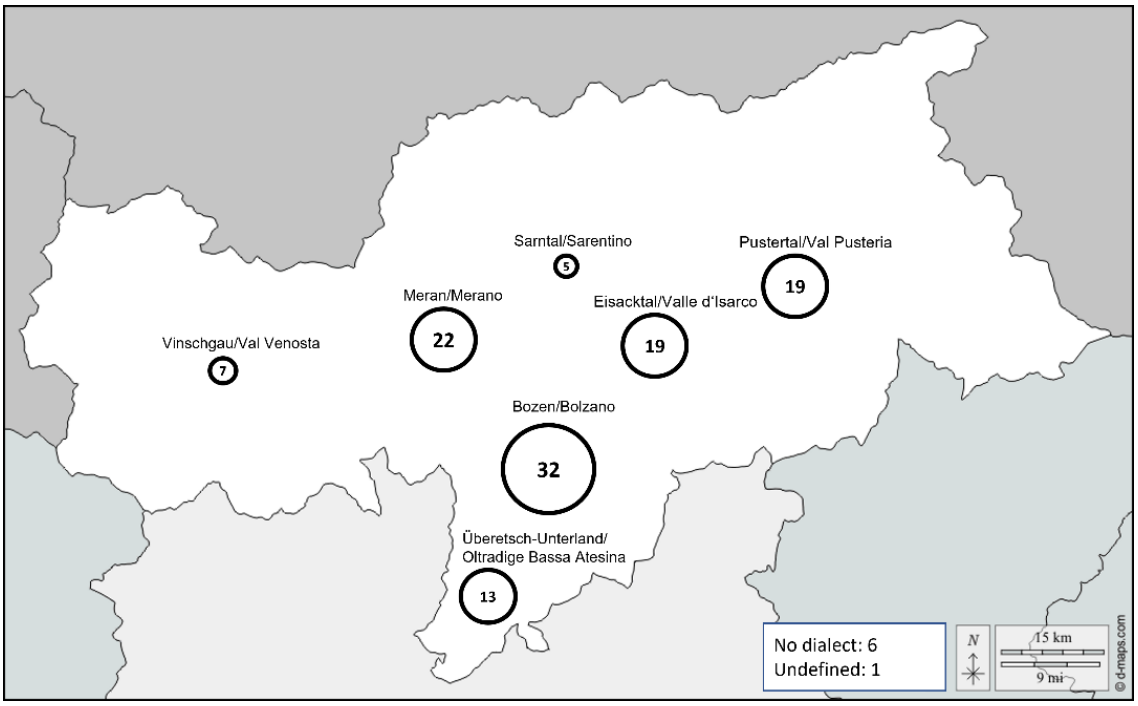

Figure 7: DiDi users per dialect region[11]

As can be seen, the center of South Tyrol is well covered with a focus on the major cities Meran/Merano and Bozen/Bolzano. Also, the East and South are represented sufficiently. Fewer users participated in the South Tyrolean western region Vinschgau/Val Venosta, with only 7 participants representing an area that makes up a quarter of the whole area. However, with about 35.430 inhabitants in 2013, it is also the least populated area of South Tyrol after Sarntal/Sarentino (taking into account the areas listed in the map above; ASTAT 2014: 66). Furthermore, six participants reported not to speak a dialect. 1 case was undefined.

[11] Own representation using https://d-maps.com/carte.php?num_car=262066&lang=de

Another piece of relevant linguistic information in the context of this thesis is the language use of the users on social network sites (SNS). This is covered by the user variable "ZA_Sprachen_SNS". The results of the analysis of this variable are reported in Table 4.

	Absolute Use	**Relative Use**
German dialect	80	0.28
German standard	83	0.30
Italian standard	51	0.18
English	50	0.18
Italian dialect	9	0.03
Ladin	1	0.00
other	7	0.02

Table 4: Languages used by DiDi users on Social Network Sites

As was expected, German is the pre-dominant variety used by DiDi users when communicating on SNS. Standard German and dialect are used almost to the same degree making up two thirds of the languages used online. The high prevalence of the dialect is emphasized by the fact that it is chosen substantially more often on SNS than English or the Italian standard. Italian dialects, Ladin or other languages play only a marginal role.

After this short description of the Didi users' characteristics, we will take a closer look at the postings themselves. Table 5 shows the number of postings per Facebook text type.

N° comments	**N° messages**	**N° posts**	**Total**
3997	14093	5099	23189

Table 5: Number of postings per FB text type

The DiDi corpus contains 23189 postings in total. 60% of these postings are messages, followed by posts and comments. A further analysis revealed that 107 users contributed comments, 104 posts and 54 messages. Only 4 users shared all text types. It is noteworthy that messages represent the most frequent text type, even though shared by the least users.

When we looked at the native languages and online language use of DiDi users in general, it became clear that the pre-dominant language was German. Both standard and dialect German were equally often mentioned as being relevant for online communication on SNS. By

evaluating the variable “dialect_tag” on posting level it is possible to get a clearer picture of the German variety actually used by DiDi users in their postings. Table 6 reports on the results.

	N° de dialect	N° de non-dialect	N° de undef	N° non-de
comments	1399	1255	1343	0
messages	4594	4433	5066	0
posts	870	2457	1771	1
Total	6863	8145	8180	1

Table 6: Number of standard, dialect and non-German postings per FB text type

From Table 6 two things can be derived: Firstly, the corpus nearly exclusively contains German postings. The reason for that is that the corpus authors excluded postings in languages other than German, as their research question only aimed at the CMC communication of German speaking South Tyroleans. Secondly, it also shows that nearly a third of postings are undefined, i.e., it is unclear if they are written in German dialect, standard or another language. This concerns all three text types. This problem will be resolved in the fourth part of this thesis in which these undefined postings were semi-automatically re-assigned to the standard or dialect category.

Up to this point, we only considered meta data available on the user and text-type level. What is missing is a closer look at the token level of the DiDi corpus. We start this evaluation with an analysis of token and type counts in Table 8.

	Comments	Messages	Posts	Total
Tokens	57173	224404	85995	367572
Types	11919	24613	18152	41448

Table 7: Word token and type count per FB text type

In total, the corpus contains 367572 tokens assignable to 41448 distinct word types. An element was considered a token if it had its own token entry in the corpus, without taking into consideration, if it was an actual word, a punctuation marker or any other typographical symbol. It is important to note that the total sum of word types is not just the sum of types per text type. As word types were counted individually for each text type, some word types might appear two or three times (max. once per text type). The sum, though, only considers distinct word types over all text types.

In the context of Neural Machine Learning for NLP it is also relevant how many tokens each data point/posting has. The number of tokens belonging to a data point determines the dimensionality of the resulting vector encoding. The maximal vector length, besides the amount of data to be processed, is one crucial factor of computation time in neural nets. This is why we also evaluated the longest and shortest postings per text type and average posting length.

	Longest	**Shortest**	**Average**
comments	449	1	14.30
messages	1105	1	15.92
posts	2881	1	16.87
Total	2881	1	15.85

Table 8: Minimum, maximum and average number of tokens per FB text type

On average a posting had a mean length of 15.70 tokens. The shortest postings in every text type is comprised of one token, mostly one-word-utterances expressing approval, disapproval, excitement, or greetings. The shortest postings were comprised of standard German as well as dialect postings. The three longest postings, with a maximum length of 2881 tokens, on the other hand, were all written in standard German. They were comprised of a political comment, a newspaper message, and a health-related post.

Another relevant piece of information in the context of this work regards the variables “norm” and “stir” already mentioned above. The variable “norm“ is relevant as it provides a first approximation of how many dialect tokens the corpus might contain. Besides abbreviations and typos in standard German text, all dialect tokens received a normalization layer. Normalization, thus, can be expected to be much more frequent in dialect postings.

	Absolute frequency	**Relative frequency**
comments	15613	27.31
messages	56019	24.96
posts	10883	12.66

Table 9: Absolute and relative frequency of normalized tokens per FB text type

Table 9 shows that normalizations are much more frequent in comments and messages. For comments and messages every fourth word has a normalization, while for posts a normalization is just over half as likely. Looking at Table 6 this emphasizes the assumed correlation between

normalization and dialect writing: While in case of comments and messages the number of dialect and standard postings is very similar, standard postings prevail in case of posts. Given this distribution a maximum number of 82515 tokens – about 20% of the corpus – are potential candidates for dialect words. This holds true with one exception: Tokens with a "stir" variable. "stir" marks words which are genuine to South Tyrolean, not appearing in the standard. Consequently, these tokens do not get a normalization and are disregarded in the counts in Table 9. Thus, to get a more comprehensive picture of potential dialect words, Table 10 looks at the frequency of tokens bearing a "stir" label.

comments	**messages**	**posts**	**Total**
630	2212	286	3128

Table 10: Absolute frequency of tokens with "stir" label per FB text type

It can be seen in Table 10 that besides the 20% normalized tokens, which are potentially dialect, there are another 3128 tokens bearing the "stir" label, which means that they are certainly dialectal. Appendix B contains a word list containing all elements with a "stir" label. That these words indeed are comprised of dialect-specific words can be seen in words like "dorricht", "dofrogg", "darpocks", "dotun", "drreissts", etc.". These words represent verbs with a verbal prefix typical for Bavarian dialects, i.e., "der-" (written as "dor-", "do-", "dar-", "dr-" etc.). This prefix is not so easily transferable into standard German as it developed an own modal meaning component. "Dotun", for example, denotes "managing to be on time".[12]

The present chapter briefly characterized the DiDi corpus by basic statistical metrics. It became clear that the corpus has a relatively balanced and representative distribution from the perspective of age, gender and place of residence of users. All postings contained in the corpus except one are labeled as German. Approximately a third of these postings are written in dialect German, a third in Standard German and the last third is undefined. The next part of this thesis describes the systems that were used to re-classify these undefined samples.

[12] For further information on this topic please refer to Lanthaler (2022), Sonnenhauser (2009), Tessadri (2017)

PART III – Principles of neural AI

3.1 Deep Learning for dialect classification

Two classification tasks are in the focus of this thesis. Both tasks happen on a different textual level. Task 2 involves classification on the word-level, similar to setups like part-of-speech tagging and named entity recognition. Every word is assigned a label. Task 1, on the other hand, involves classification of longer text sequences, i.e., postings, even spanning multiple sentences. In this case, the whole posting will be classified as dialectal, standard German, or undefined. Both tasks are cases of text classification, but while Task 2 deals with the smallest textual unit, a single word, Task 1 deals with text of variable length. As the same system should be applied to both tasks, the first challenge in the context of this thesis was it to find an AI model that can deal with textual input of various lengths and can label text on a word- and paragraph/text-basis. Even though there are numerous AI systems available for text classification (Naïve Bayes, Conditional Random Field, Random Forest, Support Vector Machine, etc.) most of the state-of-the-art results nowadays are achieved with varying neural net models. Moreover, some neural architectures are also adept at solving both word- and paragraph-level classification tasks. This is why a deep learning approach was chosen to solve the tasks at hand.

This decision entailed another choice: which neural architecture to apply. This is by no means a trivial decision as neural nets can significantly differ in their architecture, which depends on the use case of the net. For image classification, for example, convolutional neural nets are the architecture of choice (see e.g. Krizhevsky et al. 2017). In case of text classification, Recurrent Neural Net models like the Long-Short-Term-Memory setup (Hochreiter & Schmidhuber 1997) were dominant for a long time. Recently, so called “Transformer” architectures have conquered the NLP market and deliver most state-of-the-art results. The Transformer architecture was first proposed by Vaswani et al. (2017) and led to a boost in data encoding, parallelization and computation speed, enabling models with hundreds of billions of trainable parameters like the now-famous GPT models (Brown et al. 2020). In the context of this thesis two transformer-based architectures were tested and compared: XLM-RoBERTa and DeBERTa. To get a better understanding of how these models work, it is necessary to take a look behind the scenes and dive deeper into the mechanisms of neural machine learning in general.

Chapters 3.2–3.4, thus, will provide an overview of the idea behind classification in Machine Learning, how Transformer-based neural nets solve such problems and why RoBERTa and DeBERTa were reasonable choices for the tasks at hand.

3.2 Classification in (Neural) Machine Learning

A classification problem in Machine Learning can be broken down to a simple formulation. Given a set of elements $X = x_1, x_2, x_3, \dots, x_k$ and a set of Labels $L = \{l_1, l_2, \dots, l_j\}$ we need to find a function $\gamma \, \varepsilon \, \Gamma \mid \gamma: X \rightarrow L$, mapping each instance x_i to a corresponding label l_j. One example of such a function would just be a random categorizer, that assigns each element a random label. As becomes evident from this, γ cannot just assign any label, but rather, we want it to assign meaningful labels. Meaningful, in this context means that the machine assigns the same label to all $x_i \, \varepsilon \, X$ as a human annotator would. This is called supervised machine learning: we want the machine to learn what humans assume to be the correct categorization.

By assigning each of our data points x_i a corresponding l_j, the mathematical space wherein our $x_i \, \varepsilon \, X$ are allocated can be conceptualized as being divided into categorial areas. The notion of "mathematical space" in these regards must be further specified: Space, as it is understood here, is a multi-dimensional vector space. The dimensions the vector space spans are determined by the features $m_l \, \varepsilon \, M$, i.e., attributes of our x_i. Each feature m_l represents one dimension in our vector space. These dimensions are determined by the values a feature m_l can take on. The concrete value of a feature, $x_{i,l}$, determines the position of the corresponding data point x_i along that dimension. To give a concrete and simple example, Figure 8 depicts a simplified classification problem.

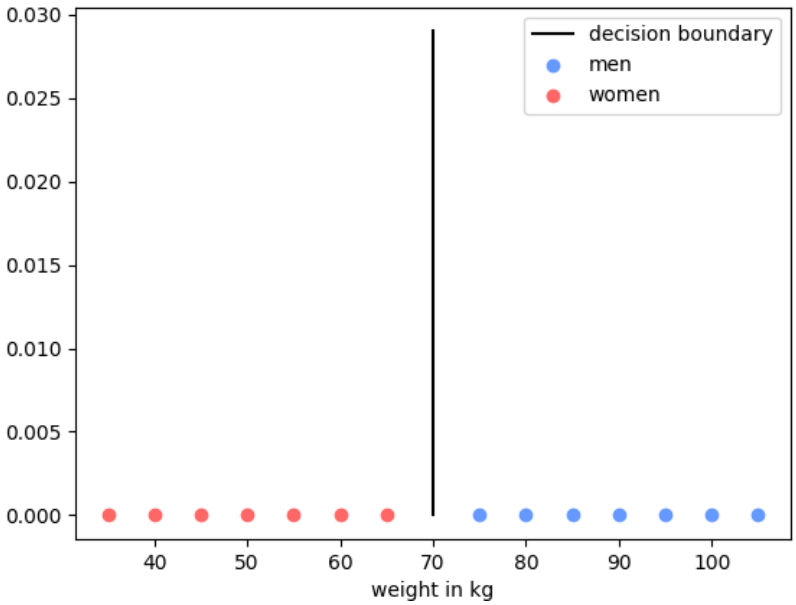

Figure 8: Classifying men and women by weight (based on Staab 2020: 60)

Our example dataset X contains 14 data points x_{1-14}, seven labeled as men ($blue = l_1$) and seven labeled as women ($red = l_2$). These data points are characterized by a single feature $x_{i,1}$, their weight in kilograms. Thus, the size of the feature set M is 1, in mathematical notation $|M| = 1$. The feature weight can take on continuous values from 0 to over 100kg. As we can see, in our simplified example women always weigh less than 70kg and men more than 70kg. The classification function gamma just has to model a linear decision boundary around 70kg to unambiguously classify our data points. Such cases where data points of different classes are not intermixed and are separable by one distinct line or plane are called linearly separable. However, the world is most often more complex than in this example. One outlier suffices to make it such that data points are not unambiguously linearly separable anymore. Such a case is shown in Figure 9, where one man weighs less than and one woman weighs more than 70kg. Consequently, these data points would be wrongly classified.

One way to deal with these cases is to use a probabilistic classification approach: The classification function gamma models a probability distribution over classes, depicted here by the red and blue probability density functions. Using these density functions γ would predict the probability for each class for a data point and choose the class with the highest probability in the end. In this way, our outliers would still be wrongly classified, but the system would be more flexible with regard to non-linearly separable classification scenarios. The example described here is very simple as it involves just one dimension needed for classification. In

realistic Machine Learning scenarios, the number of attributes that characterize data points is much higher. This complexity of features can also lead to very complex probability distributions over classes. Therefore, before the emergence of neural nets, in order to ensure good predictive power, it was crucial to choose meaningful features that enabled a good separability of data points belonging to different classes.

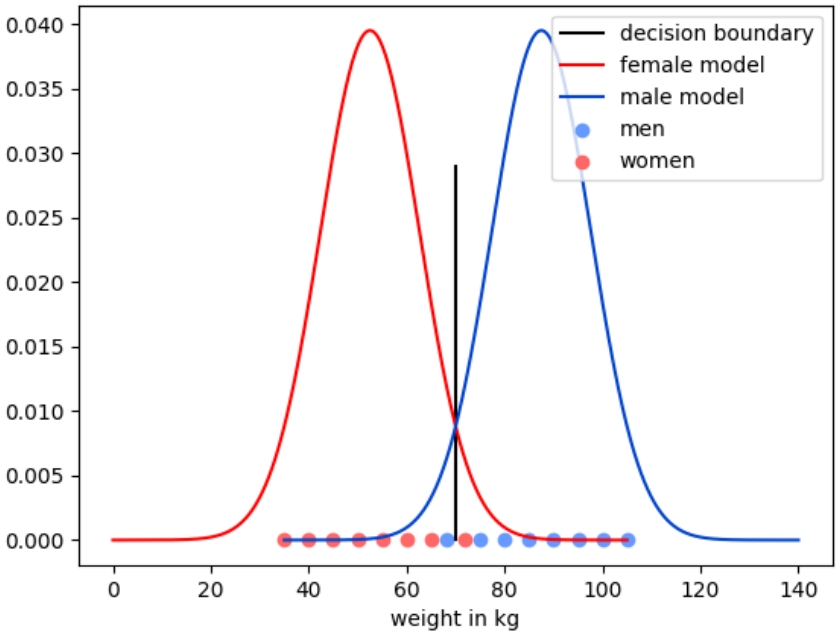

Figure 9: Non-linear classification scenario (based on Staab 2020: 60)

What makes neural nets so powerful is that they make the feature-engineering step superfluous. Neural nets just need the right label for each data point. On the basis of these labels, they automatically derive features from the given data that guarantee the best classification performance. To put it another way: Neural nets can automatically adapt and model very complex probability distributions without knowing any features beforehand. The price we pay for the high performance and efficacy is transparency. The found features are encoded in numerical vectors hardly interpretable by humans. This is a downside shared by all neural net architectures, which is why they are often termed “black box” models. In the context of this thesis this means: Even if the developed systems can distinguish standard from dialect German words/texts, we do not know why, i.e., which features were most relevant for distinction. Further downsides associated with the black box nature of neural nets will be covered in chapter 6.1.

While all neural models share the property of outputting non-transparent numerical vectors, they differ in how these vectors are generated. How the features are engineered, strongly depends on the neural model architecture. The next chapter describes how Transformer-based models approach feature engineering and what makes them superior to earlier models.

3.3 The Transformer

Before Transformer models were established, Recurrent Neural Nets (RNN) were successfully applied to many NLP problems. Figure 10 is a simplified draft of how the encoding of data works in an RNN for a sequence-to-sequence problem (Vu 2019: 44).

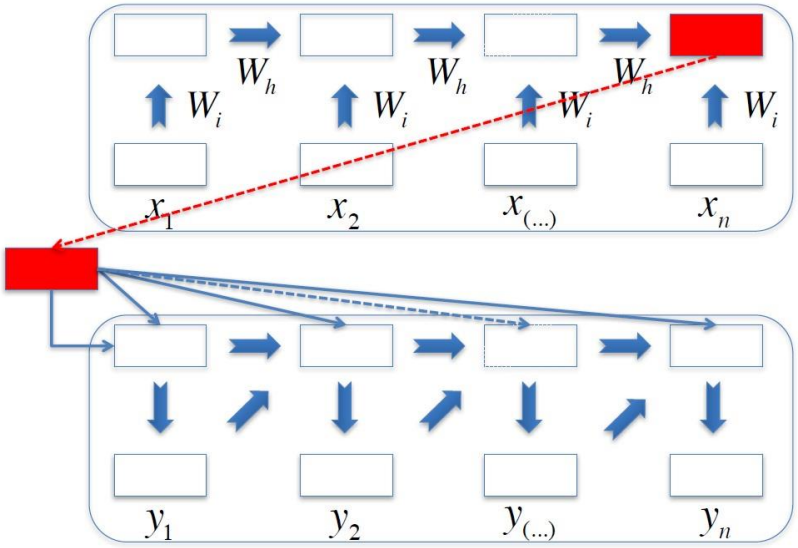

Figure 10: Sequence-to-sequence RNN processing flow

A prominent example for a sequence-to-sequence task is machine translation where a sequence of words in one language has to be transferred to a sequence of words in another language. Recurring to Figure 10, the words to be translated are denoted as x_n, while the translated output words are y_n. An RNN has two parts: an Encoder (upper cell) and a Decoder (lower cell). The Encoder takes in the words to be translated and processes them in sequential order. The representation of every word except the first word is influenced by the representation of its preceding word, which is indicated by the blue arrows. W_i and W_h stand for the matrices with which the word representations are transformed, i.e., multiplied with. As can be seen, the matrices do not differ between encoding steps. This is the reason why these nets are called Recurrent Neural Nets: The same matrix operation is applied recurrently to each word vector

representation of the incoming words[13]. As every previous word is influenced by the representation of the words that precede it, the representation of the last word x_n (red cell) is simultaneously a representation of the whole sentence we want to translate. In the next step, the Decoder starts to decode by outputting the words y_n in the target language on the basis of the previously computed sentence representation in the source language. Every target word, again, is influenced by its preceding target word as well as the general source sentence representation. This short overview of the RNN flow reveals its bottlenecks: The input must be processed sequentially, which lowers computational speed, and the whole source sentence is represented by just one vector. Thus, all of the information of the source sentence has to be compressed into one vector before being decoded. The Transformer architecture provides a solution to these downsides.

The Transformer builds on a mechanism called "Attention" introduced by Bahdanau et al. (2014) and extended by Luong et al. (2015). As shown in Figure 10, during RNN encoding each incoming word is equally influenced by preceding words. The Attention mechanism enables the neural net to assign different weightings to preceding words. Words that are more relevant to the currently encoded or decoded word receive a larger weight. In this way the neural net pays more attention to somehow related words. The Transformer architecture makes optimal use of this mechanism. Figure 11 shows the Transformer setup.

[13] The input to neural nets are not words but their vector representations. There exist different strategies to encode words into vector form such that they can be used as an input for neural nets. One of the most prominent strategies are word embeddings. A good starting point to enter this topic is Mikolov et al. (2013), who describe the logic behind the word2vec word embedding method.

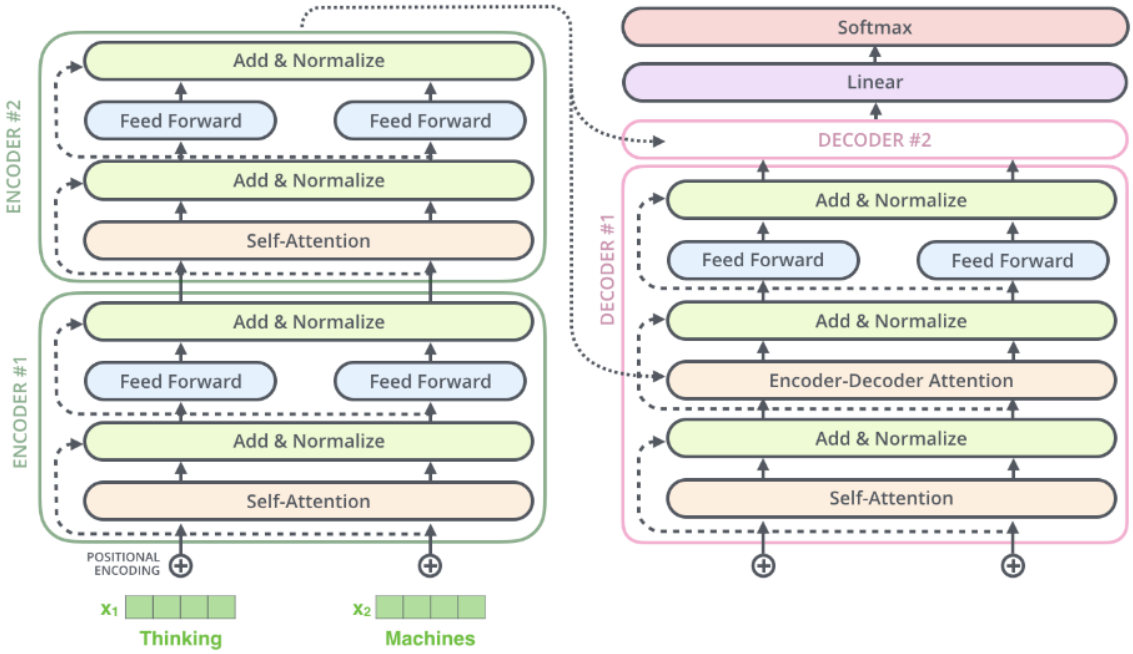

Figure 11: Transformer architecture (taken from Alammar 2018)

Similar to an RNN, the Transformer also has an Encoder (left block) and a Decoder (right block). However, instead of feeding one word after the other into the Encoder, the Transformer is able to process all words at once.

Figure 11 shows the phrase “Thinking Machines” being processed. As already mentioned, the actual words are not being processed, but rather their embeddings, i.e., word representations. Each word embedding has its own separate pathway through the Encoder block. At the beginning a positional encoding is added to each word vector representation. In this way, relevant positional information can be integrated. Then, these representations are fed into a Self-Attention layer. Self-Attention in this regard just means that the words in a phrase or sentence are related to each other, each word “paying attention” to the other words. In a second step, the vectors that are outputted by the Attention layer are normalized. This is done by adding the input representations of the Attention layer to the output representations of the Attention layer. Then, a second normalization step involving a classic feed forward neural net and another normalization step is performed. As becomes evident from Figure 11 as well, all these encoding steps are not performed once but repeated multiple times in subsequent Encoder blocks. The example shown here has two Encoder blocks. In the original paper by Vaswani et al. (2017) six Encoder blocks were used.

The final output of the encoder, then, is fed into the Decoder blocks. A Decoder block is very similar to an Encoder block with three major differences:

1. The input to the first Self-attention layer of the Decoder is just the positional encoding. Thinking of a Machine Translation scenario this makes sense as the remaining input to the Decoder can only be the encoded source sentence, if we want the translated target sentence as our output.
2. The Decoder is not allowed to pay attention to positions that are decoded at a later point in time. If, for example, we want to translate "Thinking Machines" to the German phrase "Denkende Maschinen", the Decoder is not allowed to know that the next decoded word is "Maschinen" when translating "Thinking" to "Denkende". This preview problem appears during training of the neural net, because during training the neural net has to know what our target words are for the computation of the error, i.e. loss function. Thus, during training we have to mask future translations.
3. The other Attention layers of the Decoder implement Encoder-Decoder Attention. This means that during decoding we do not pay attention to the output of the Decoder but to the output of the Encoder. This is signalled by the dotted lines in Figure 11.

Similar to the Encoder setup, the Decoder setup also consisted of 6 blocks in the original paper of Vaswani et al. (2017). The Decoder also encompasses a Linear layer and a SoftMax layer at the end. These last layers have the task of converting the resulting vector representations back into the human-readable words of the target translation. With the knowledge of the last two chapters, it is possible to look at the differences and similarities between RoBERTa and DeBERTa, described in the following chapter.

3.4 RoBERTa and DeBERTa – a comparative overview

As their names already insinuate, the two neural systems tested for the tasks, RoBERTa and DeBERTa, have a common ancestor, namely, BERT. BERT – short for **B**idirectional **E**ncoder **R**epresentations from **T**ransformers – is a Transformer-based neural network established by Devlin et al. (2018). The original task BERT was trained on, is a very common one in NLP – language modelling. Language modelling aims to create a probabilistic representation of the sequences of words in a language: Given a word or a sequence of words a language model predicts which words are likely to follow in the given language.

In comparison to the original Transformer concept, BERT's authors just used the Encoder part of the Transformer architecture to train their language model. The reason for that is that they wanted to train a language model that considered the right and left context of a word simultaneously. Previous transformer language models like ELMO (Peters et al. 2018) were unidirectional, which means that the prediction of a word was only based on the words before the predicted word OR the words after the predicted word, but not the left and right context at the same time. This is similar to the nature of the Decoder in the previous chapter: The Decoder should not attend to words to its right, i.e., words that are still to be translated. Therefore, the authors of BERT introduced an architecture that is solely based on the Encoder. This decision, however, leads to a problem for language modelling.

As described, in language modelling, we predict a word given its context. Looking at the example in Figure 11 this means: First, we predict "Thinking". To predict it, the Attention mechanism would take a look at its context words and find "Machines" there. Thus, the representation of "Thinking" is influenced by "Machines". As words are fed in parallel to the Transformer, the representation of "Machines" will be influenced by "Thinking" at the same time. If we would only have one block this would be no problem. But as we saw, in the Transformer case we stack multiple Encoder blocks on top of each other. If we now pass on the output of the first encoder, i.e., the representations of "Thinking" and "Machines", to the second Encoder, we will have the problem that the representations did already influence each other. If we now take "Thinking" to predict "Machines", the representation of "Thinking" already bears the information that it is followed by "Machines".

The authors of BERT solved this problem by introducing a technique they called "Masked Language Modelling": During the training of the language model, a certain percentage of words is masked, i.e., replaced with a untransparent [MASK] token. The model must predict these masked tokens, which are unknown to the model. In this way the left and right context of a word can be considered without the target word influencing the prediction. This training procedure leads to a high-performing language model.

But BERT does not only work as a language model. The big advantage of BERT, and one reason for its popularity, is that after training it is easily applicable to various NLP tasks. To apply it, for example, to a sentence classification task, we just add a classification layer on top of the pre-trained BERT language model. In this way, the BERT part of the model will create meaningful word or sentence representations, because that is what it was pre-trained on as a

language model. During training, the model just has to learn how to fine-tune the sentence representations in such a way that they are assigned the correct labels. This is an instance of so called "Transfer Learning" with language models and is a widely used framework in Neural Machine Learning to date[14]. DeBERTa and RoBERTa are both based on BERT. They inherit the Transformer architecture as well as the masked training procedure from BERT. However, they apply small modulations to the original model.

RoBERTa is short for "**R**obustly **O**ptimized **BERT** Pretraining **A**pproach" and was introduced by Liu et al. (2019). RoBERTa exploits the full potential of the BERT architecture by hyper-parameter modulation of the original model. This means that RoBERTa improves BERT's training procedure and performance by modulating the training hyperparameters. The RoBERTa model, which was implemented in this thesis, is an extended version of RoBERTa, namely XLM-RoBERTa (Conneau et al. 2019). "XLM" here stands for "Cross-lingual Language Modeling." While RoBERTa was trained and optimized exclusively on English language data, XLM-RoBERTa took a cross-lingual approach and trained on language data from a hundred languages, including German. It was assumed that the tasks at hand could profit from such a multi-lingual background.

DeBERTa on the other hand, stands for "**D**ecoding-**e**nhanced **BERT** with disentangled **A**ttention", and was published by He et al. (2020). Compared to RoBERTa, this model architecture not only optimizes hyperparameters, but changes the actual architecture and training procedure of BERT, achieving, in this way, scores that are superior to the original model in various tasks.

Regarding architecture, DeBERTa mainly improves the way in which positional information is integrated. He et al. (2020) introduce disentangled, i.e., separate matrices for word and relative position encodings. This was implemented by two distinct encoding procedures and attention mechanisms for words and relative positions. Also, in case of BERT, only absolute position encodings were used, and these encodings were added to the word encodings right at the beginning (see Figure 11). DeBERTa, in addition to introducing relative positions, also improves the use of the absolute word position information by integrating this information later in the model: Instead of providing an absolute position in the input layer, DeBERTa integrates

[14] see, among many others, Peng et al. (2019), Mozafari et al. 2020, Qiao et al. (2022) for Transfer Learning examples with BERT

absolute position information just before the prediction, i.e., the SoftMax layer. The authors justify this decision with the assumption "that the early incorporation of absolute positions used by BERT might undesirably hamper the model from learning sufficient information of relative positions" (He et al. 2020: 5).

Secondly, DeBERTa improves BERT performance by a more sophisticated training procedure using a new adversarial training algorithm. In adversarial training, deviant, noisy training samples are added to the training set. In this way overfitting of the model should be prevented leading to a better generalization capacity.

After this comparison of model architectures, the next section will shed light on how both models were implemented for this thesis.

3.4.1 RoBERTa and DeBERTa – implementation details

Both models were not programmed from scratch but implemented using the Huggingface[15] Python library. Huggingface provides an easy-to-use pipeline for the creation and training of neural nets, especially Transformers. We downloaded the pre-trained models "deberta-v3-large[16]" and "xlm-roberta-large[17]" using the Transformer Library. As high computation capacity was needed to train the models, we used GPU infrastructure on Google Colab. Table 11 sums up model specifications.

[15] https://huggingface.co/

[16] https://huggingface.co/microsoft/deberta-v3-large

[17] https://huggingface.co/xlm-roberta-large

	deberta-v3-large	xlm-roberta-large
Authors	He et al.	Conneau et al.
Publication date	2020	2019
Issuing institution	Microsoft	Facebook AI
Model architecture	Transformer	Transformer
N° Encoder layers	24	24
Improvements to BERT	Sophisticated position encoding + adversarial training	More languages + hyperparameter tuning
N° trainable hyperparameters	434 012 160	559 893 507
N° of trained languages	1 (English)	100 (German inc.)
N° of known words	128 001	250 002
Hidden size vector length	1024	1024

Table 11: Model specifications – DeBERTa vs. RoBERTa

As becomes evident from Table 11, both models are similar in size and architecture. The number of Encoder blocks as well as the vector size of the Transformer's SoftMax output are identical. The number of trainable parameters is similar as well, with XLM-RoBERTa being slightly larger than the DeBERTa model.

The main differences, however, primarily lie in the already described design choices. XLM-RoBERTa was trained on 100 different languages while DeBERTa is a mono-lingual English model. This has an effect on the vocabulary size of the models. A Transformer's vocabulary specifies how many different word types a model has been trained on during pre-training. Due to the multi-lingual setup, XLM-RoBERTa has a vocabulary that is nearly double the size of the DeBERTa vocabulary. DeBERTa, on the other hand, has the advantage of a more sophisticated positional encoding mechanism. This setup, i.e., a mono-lingual model with a more complex positional architecture compared to a multi-lingual model, was chosen, because it adds another interesting dimension to this thesis: Does a multi-lingual model perform better due to its cross-lingual pre-training or is it outperformed by a model representing the next generation of Attention architectures?

The next part of this thesis will provide an answer to this question by describing the training process and model performance regarding Task 1. Then, one model will be chosen to solve Task 2, i.e., classification on word basis.

Part IV – Tasks

4.1 Task 1: Undefined reduction

The first task, described in this chapter, involves the training and evaluation of a system that relabels undefined samples in the original DiDi resource. As described in Part II this corresponds to approximately a third of the data, i.e., 8180 postings. The first step in the re-definition process was to create a representative subsample of these undefined postings and to relabel them such that they could be used as a training and development set for supervised training.

4.1.1 Training, validation and test set

`Code files: 02_gold_standard_selection.ipynb`

The statistical evaluation already made clear that the Facebook text types are not equally represented in the corpus: 62% are messages, 21.6% posts and 16.4% comments. Regarding other relevant variables like age, gender and dialect region, the DiDi data has a relatively balanced distribution or at least a distribution reflecting the underlying population (e.g., participants per dialect region). For this reason, the most important dimension to obtain a representative subsample was FB text type. Starting from this variable, a binned sampling procedure was implemented. First, comments, messages and posts were divided into three different sampling groups. Then, each group was filtered for undefined samples. The filtered undefined samples per text type were again grouped by the variables “age”, “sex” and “german dialect region”.

In a last step, 20% of the grouped messages, comments and posts were randomly drawn. In this way, 1636 of 8180 undefined postings were selected, corresponding to 1014 messages, 354 posts and 268 comments.

After the sampling process was completed, the selected subsample had to be re-annotated: Each posting had to be checked to see if it could be reassigned to dialect or standard German or if it was indeed undefinable. To do so, an annotation guideline was written. Then, 100 samples were re-labeled, and the guideline revised[18]. Using this guideline all 1636 undefined

[18] The revised annotation guideline can be found in APPENDIX C (German version only)

samples were re-annotated. Due to the limited time, the only annotator was the author of this thesis, proficient in a local vernacular. Therefore, an inter-annotator agreement or similar measures could not be computed.

After the re-annotation 632 postings, (38%) were labeled as “de dialect”, 533 as “de non-dialect” (33%) and 471 (29%) were still undefined. It can be seen from this distribution that only about a third of the postings originally labeled as “de undefined”, kept this label after re-annotation. This was already a promising indicator that the re-annotation of all undefined samples would lead to a significant reduction of such cases.

In a next step, before the models could be trained, the re-labeled dataset was augmented with another 310 dialect and 409 standard postings from the DiDi corpus. There were multiple reasons for this:

1. 1636 data samples are a relatively small dataset, especially when half a billion parameters have to be fine-tuned in the training process. The addition of another 709 examples led to a bigger dataset of 2355 postings.
2. By adding this amount of new dialect and standard samples, the dataset was more balanced: 942 postings (40%) were dialect postings, 942 (40%) standard postings and 471 (20%) were undefined.
3. Even though two thirds of the samples were re-labeled as dialect or standard German during re-annotation, the representativeness of dialect and standard German postings might not be ideal in a dataset that consists exclusively of formerly undefined samples. They might, for example, consist of particularly short or postings in languages other than German. To guarantee that the models could learn a stable and generalizable representation of dialect and standard German postings, other examples for these categories were added.

Next, the dataset was split into a training and a validation set. The training set is used during training to estimate parameters. It contains 80% of the 2355 postings i.e., 1884 samples. The validation set is used during training to guarantee that the model is not overfitting to the training data and that it generalizes well to unseen data. It contains 20% of the postings, i.e., 471 samples.

At this point, from the 8180 undefined postings in the DiDi corpus 1636 samples had been manually labeled and divided onto test and validation set. The rest of the undefined data, i.e.,

6544 samples, formed the test set. The test set is used for final evaluation of the models. After this short overview of how the data basis for model training was prepared, the next chapter will shed light on the concrete training procedure.

4.1.2 Training DeBERTa and RoBERTa

```
Code files: 03_relabeling_net_deBERTa.ipynb, 04_relabeling_net_roBERTa.ipynb
```

As already described, the pre-trained models were loaded using the Huggingface Transformer library with a keras/tensorflow backend. Both models were trained and validated using the identical training and validation set.

XLM-RoBERTa was trained for 180 epochs using Root Mean Square Propagation as the optimization algorithm as well as Categorical Cross Entropy as the loss function. We used a batch size of 32 and an initial learning rate of 0.00005. Moreover, to prevent the model from overfitting we additionally added a Dropout layer (Srivastava et al. 2014) with a dropout rate of 85%.

The training procedure was very similar in case of DeBERTa: The model was trained for 180 epochs using Root Mean Square Propagation and Categorical Cross Entropy. We used a batch size of 16 and an initial learning rate of 0.00005. Also, in case of DeBERTa, we added a Dropout layer before the classification layer. The dropout rate was 90%. All hyperparameter values were empirically evaluated by re-training the models with different hyperparameter setups.

Eventually, both for RoBERTa as well as DeBERTa, the model version leading to the highest Accuracy on the validation set was chosen. The next chapter will describe how the models performed on the validation and test set.

4.1.3 Performance evaluation and error analysis

The models were evaluated on both the validation as well as the test set. The subsequent sections will give a detailed performance and error analysis for both datasets.

4.1.3.1 Validation set

Table 12 compares the performance of the two models on the validation set using different evaluation metrics[19].

	Precision		**Recall**		**F_1**		**N° samples**
	DeBa[20]	*RoBa*[21]	*DeBa*	*RoBa*	*DeBa*	*RoBa*	
dialect	0.89	0.88	0.87	0.91	0.88	0.89	193
non-dialect	0.89	0.93	0.93	0.93	0.91	0.93	195
undef	0.63	0.67	0.60	0.61	0.62	0.64	83
macro average	0.80	0.83	0.80	0.82	0.80	0.82	471
weighted average	0.85	0.86	0.85	0.87	0.85	0.86	471
	DeBa	*RoBa*					
accuracy	0.85	0.87					

Table 12: Task 1 – model performance on validation set

Values where one model prevailed over the other, were color-coded in green. The color coding reveals a clear pattern: In case of the validation set XLM-RoBERTa consistently outperformed DeBERTa. Only in case of dialect Precision DeBERTa is slightly better.

However, there is no large performance difference between the two models. On average, RoBERTa outperforms DeBERTa by two points. Both models show the highest performance regarding the classes “dialect” and “non-dialect” and by far the worst when classifying cases labeled as undefined. This results in a higher weighted average than macro average as with 83 proponents undefined samples are less than half as frequent than the other two classes, which leads to a lower weighting in the average.

The poor classification performance regarding the undefined cases seems reasonable considering the inherent properties of the three classes. While dialect or standard postings are characterized by two discrete varieties, the nature of undefined samples is inherently

[19] An introduction to the used evaluation metrics can be found in APPENDIX D

[20] DeBERTa

[21] RoBERTa

ambiguous. To put it in more simplified terms: The label "undef" translates to "hard to classify". This is because undefined samples often cover cases, which are:

1. Mixtures of the two German varieties:
   e.g. "jo nochr selbst schuld" ("then, it's his fault")
   → In the local dialect the referential pronoun is not "selbst" but "selber.
2. One-word postings assignable to both varieties:
   e.g. "passt" ("okay")
3. Potential dialect postings only differing from standard by one character:
   e.g. "Jo chillig und du?" ("Yeah, easy and you?")
   → used retracted vowel "o" instead of "a" in first word "Jo".
4. Other languages or German varieties:
   a. "Grüezi! Urlauber; -)" ("Hi! Vacationists; -)")
      → "Grüezi" is a Swiss German greeting.
   b. "hoi guten Morgen, ok, vedrem cosa succede: D buona giornata!" ("Hi, good morning, ok, let's see what happens: D have a good day")
      → Italian-German mixture

The examples listed above were all taken from the validation set. This shows that the mentioned properties of the undefined class also apply to the validation set and explains why undefined samples were the most frequently misclassified. The confusion matrix in Table *13* provides detailed insight into how undefined samples were misclassified.

RoBERTa	*dialect pred*	*non-dialect pred*	*undef pred*	*sum true*
dialect true	175	2	16	193
non-dialect true	4	182	9	195
undef true	20	12	51	83
sum pred	199	196	76	
DeBERTa	*dialect pred*	*non-dialect pred*	*undef pred*	*sum true*
dialect true	168	6	19	193
non-dialect true	4	181	10	195
undef true	16	17	50	83
sum pred	188	204	79	

Table 13: Task 1 – Error confusion matrix of validation set

As is evident, RoBERTa often misclassified undefined samples as dialect. Interestingly, RoBERTa also committed errors in the other direction, i.e., misclassifying dialect examples as undefined. This is rarely the case with the distinction undefined–non-dialect. The fewest errors were made in the distinction between dialect and non-dialect which emphasizes the above considerations about inherent class properties.

The classification performance of DeBERTa shows similar tendencies, also missing the distinction between undefined and dialect samples most often. However, this does not mean that the systems misclassified the identical samples. Regarding samples labeled as undefined in the gold standard, the following could be observed: RoBERTa and DeBERTa both misclassified 22 undefined postings as dialect or non-dialect, i.e., in 22 cases both made the same mistake. In 21 cases, however, only one system was mistaken. These cases are interesting for closer inspection. Some of them are displayed in Table 14.

	full text	**gold labels**	**predicted labels DeBa**	**predicted labels RoBa**
1	jo nochr selbst schuld	de undef	de undef	de dialect
2	Tschetschenen	de undef	de undef	de dialect
3	grüße <PersNE>!	de undef	de non-dialect	de undef
4	Schlimm!!!	de undef	de non-dialect	de undef

Table 14: Task 1 – Misclassified undefined validation samples

It becomes clear that while DeBERTa has a tendency of over-assigning the non-dialect label, RoBERTa over-generalizes the dialect label on undefined postings. In other words: What RoBERTa still rates as undefined, DeBERTa misclassifies as being standard. And what RoBERTa misclassifies as dialect, DeBERTa considers as being undefined.

Thus, in cases where RoBERTa and DeBERTa disagree, their misclassification tendency regard different poles of the dialect-standard continuum. This is confirmed when looking at examples labeled as dialect in the gold standard (Table 15).

	full text	gold labels	predicted labels DeBa	predicted labels RoBa
1	pfiatiii;)	de dialect	de undef	de dialect
2	Des isch Nächstenliebe …	de dialect	de undef	de dialect
3	geht woll oder?	de dialect	de undef	de dialect
4	schualschwänzen!!!!!!!	de dialect	de undef	de dialect
5	Lässige gschicht	de dialect	de undef	de dialect

Table 15: Task 1 – Misclassified dialect validation samples

In case of dialect postings RoBERTa and DeBERTa disagreed in 22 of 193 cases about the correct category[22]. Four of these cases are displayed in Table 15. As can be seen, DeBERTa, again, case seems to be more conservative about the dialect label compared to RoBERTa. Especially in case 1 this is a clear mistake as the posting contains a common greeting phrase in South Tyrolean dialects. Even though "pfiati" is used also in other Bavarian dialects it is clearly marked as dialectal and should be categorized as such. Based on the dialect description in chapter 1.4.1 we can hypothesize why DeBERTa misclassified the other examples in Table 15. All examples are very short and contain only little dialect evidence. However, each posting, has at least one dialect element justifying the dialect label:

- Example 2 has the typical s-retraction of "ist" to "isch" and the replacement of the demonstrative "das" with "des".
- Example 3 contains the affirmative particle "woll", translatable with standard German "doch".
- The one-word expression in 4 is a good example for typical diphthongs and schwa-apocope: "schuleschwänzen" is realized as "schual schwänzen>.
- 5 is also a case of schwa-apocope: "geschichte" is realized as "g schicht ".

So, why did DeBERTa not label these instances as dialect? In case of example 3 this might have to do with the fact that only "woll" is a dialect element, while the other two words have a standard German realization. This is a case of code mixing. As described above, undefined samples are often characterized by such combinations. Looking at examples 2, 4, 5 it is noteworthy that they all contain the umlauted vowel "ä". As rounded umlaut vowels like <ö>

[22] In sum, 162 dialect samples were identically and correctly labeled by both systems. 31 samples were incorrectly labeled: In 9 cases both RoBERTa and DeBERTa assigned the same wrong category. In 22 cases only one of the two systems was mistaken.

and <ü> are unrounded to <e> and <i> in most South Tyrolean varieties, umlaut letters are more frequent in standard German writing. Even though <ä> is not a rounded vowel and is used in dialect writing, DeBERTa might have learned that umlauts in general are a strong indicator for standard writing. Also, the words containing <ä> all have a Standard German realization, which is why it is not implausible to assign these examples to the undefined category.

To sum up, with exception of the first example in Table 15, the labels are open to debate. Even though they were assigned to the dialect category by the annotator, it is not completely incorrect to subsume them under the undefined category. This shows one of the drawbacks of the used evaluation metrics: Precision, Recall, F_1 just tell us how close the system came to human perception, i.e., the labels assigned by the annotator. However, the human categorization might not be the best choice, especially when a label is vague[23]. Thus, even if one system outperforms the other, this does not guarantee a better generalization performance of this system on another set of data. This can be seen when we look at test set performance.

4.1.3.2 Test set

After performance analysis of both systems on the validation set, the next step was to apply them to the test set containing the remaining 6544 undefined samples of the DiDi corpus. In contrast to the validation set, however, these samples were not yet manually re-assigned at this point. Therefore, a semi-automatic re-labeling process was applied.

First, RoBERTa and DeBERTa were used to automatically label the test set. Then, all instances were manually checked for correctness. Special attention was paid to cases where the two systems disagreed. In this way, all the remaining undefined postings in the test set were re-labeled and the standard evaluation metrics could be applied. Table 16 shows DeBERTa and RoBERTa's performance on the test set.

[23] Please also refer to chapters 6.1 for further discussion of this topic.

	Precision		Recall		F_1		N° samples
	DeBa	RoBa	DeBa	RoBa	DeBa	RoBa	
dialect	0.95	0.94	0.82	0.85	0.88	0.90	2884
non-dialect	0.89	0.84	0.90	0.90	0.89	0.87	2208
undef	0.63	0.65	0.78	0.70	0.70	0.67	1452
macro average	0.82	0.81	0.83	0.82	0.82	0.81	6544
weighted average	0.86	0.84	0.84	0.83	0.84	0.84	6544
	DeBa	RoBa					
accuracy	0.84	0.83					

Table 16: Task 1 – model performance on test set

It is evident that we get the inverse picture on the test set: DeBERTa shows a better performance on the test set than RoBERTa. As can be seen when comparing the macro average F_1 and weighted average F_1 of both models, this switch in performance rates can be mainly attributed to the undefined class. DeBERTa performed significantly better when classifying undefined postings. This leads to a higher macro average than weighted average F_1, as the macro average weights all classes equally disregarding their frequency, while the weighted average takes class frequency into account.

The fact that DeBERTa seems to generalize better to the new data can also be seen looking at Accuracy: While DeBERTa's Accuracy on the test set was nearly identical as on validation data, RoBERTa's Accuracy dropped by 4 points compared to the validation set. However, this mere metric-based impression of DeBERTa's higher performance can be misleading as the following considerations will show.

The general mislabeling tendencies, described for the validation set, can also be observed for the test set. Table 17 shows a detailed analysis for each label.

Dialect gold	*RoBA dialect*	*RoBa non-dialect*	*RoBa undef*	*DeBa Total*
DeBa dialect	2208	18	142	2368
DeBa non-dialect	9	13	10	32
DeBa undef	243	40	201	484
RoBa Total	2460	71	353	
Non-dialect gold	*RoBa dialect*	*RoBa non-dialect*	*RoBa undef*	*DeBa Total*
DeBa dialect	7	11	5	23
DeBa non-dialect	7	1863	117	1987
DeBa undef	9	113	76	198
RoBa Total	23	1987	198	
Undefined gold	*RoBa dialect*	*RoBa non-dialect*	*RoBa undef*	*DeBa Total*
DeBa dialect	49	8	43	100
DeBa non-dialect	10	124	80	214
DeBa undef	66	183	889	1138
RoBa Total	125	315	1012	

Table 17: Task 1 – Label assignment counts across systems and classes

The respective gold label is marked in boldface in Table 17. The values in the cells show how often different label combinations appeared across the two systems. The green cells mark the case where both systems agreed on the correct label. As can be seen by comparing the values in the green cells with neighboring values, both systems most often agreed on the correct label.

When it comes to weaknesses, DeBERTa's tendency to over-assign the undefined label is confirmed: In case of postings labeled as "de undef" and "de dialect" in the gold standard, DeBERTa assigned the undefined label over a hundred times more often than RoBERTa (dialect: 484 vs. 353; undefined: 1138 vs. 1012). The over-assignment is also mirrored by the fact that DeBERTa has a high Recall rate for undefined samples of 0.78, but a much lower Precision rate of 0.63. This indicates that "de undef" was frequently assigned, but often incorrectly. This divergence is not so prominent for RoBERTa. In case of RoBERTa the aforementioned dialect label tendency could be confirmed. However, in contrast to the observations on the validation set, this seems not to be a case of "over"-generalization. The fact that RoBERTa's dialect Precision is very high and that it still has a good Recall rate on the

dialect class indicates that the model developed a good representation of what counts as a dialect instance. The data in Table 18 underlines this hypothesis.

	full text	gold labels	predicted labels DeBa	predicted labels RoBa
1	Ja	de undef	de dialect	de non-dialect
2	ja	de undef	de dialect	de non-dialect
3	schön	de undef	de dialect	de non-dialect
4	gut	de undef	de dialect	de non-dialect
5	richtig	de undef	de dialect	de non-dialect
6	Es wor unbeschreiblich...	de undef	de non-dialect	de dialect
7	jo, weiß nur nicht was haha	de undef	de non-dialect	de dialect
8	hoffentlich net unheilbar...:-)	de undef	de non-dialect	de dialect
9	bittebitte...wo denn?	de undef	de non-dialect	de dialect
10	Guten Morgn!	de undef	de non-dialect	de dialect

Table 18: Task 1 – Examples of complete class disagreement

Table 18 shows examples of complete class disagreement, i.e., instances where both systems disagreed with the gold label and with each other. The gold label we compare the predicted labels with is always “de undef”. This setup has multiple advantages: By picking undefined postings we most likely capture difficult classification cases. This effect is further intensified by choosing only cases where both systems predicted another wrong label. In this way we inspect both scenarios, i.e., cases where DeBERTa assumes a non-dialect posting and RoBERTa a dialect posting and vice versa. This adds another interesting dimension: Dialect and non-dialect are normally more easily distinguishable. If we look at undefined, uncertain cases where both systems were mistaken and came to opposing conclusions, this can provide insights into the implicit dialect and standard class representations of the two models.

The first examples, 1–5, cover cases where RoBERTa assigns a dialect label and DeBERTa a non-dialect label. Except example 6 all postings are one-word postings containing clearly standard German words. These postings were labeled as “de undef” in the gold standard just because they differ in only one (1–4) (e.g., “ja” would be realized as “jo” in dialect writing) or no character (5) from a dialect realization, which is in line with the annotation guidelines in APPENDIX B. Nonetheless, RoBERTa’s decision to label these instances as standard German is definitely reasonable as their orthographical form is standard-oriented. DeBERTa’s decision, however, to label these instances as dialect, is clearly wrong. On closer inspection, it became

clear that DeBERTa labeled almost all one-word postings as “de dialect”. This indicates that DeBERTa identified posting length as a relevant classification parameter, which is undesirable when classifying varieties.

Examples 6–10 cover the reverse case: DeBERTa identifies a standard posting and RoBERTa a dialect posting. Among these, 6–8 are again cases of mixed code. The words “jo” and “wor” (standard German “ja” and “war”) show the typical graphical representation of the posterior open vowel [ɑ] described in chapter 1.4.1. As described under (2) the word “net” is a South Tyrolean negation particle, equivalent to Standard German “nicht”. The rest of these postings have a standard-conform realization. This code mixing is the reason why the samples remained undefined in the gold standard. DeBERTa now shows the tendency, already observed on the validation set, to label such ambiguous cases as standard German.

This would be correct if we would base the classification decision merely on the counts of standard-oriented and dialect-oriented word realizations. Standard German and dialect word realizations, however, cannot be seen as equivalent in their importance for classification. As was shown in the introductory part, dialect writing can be seen as a derivation of and deviation from a norm. Such a deviation always represents a communicative decision. By deviating from the norm, the writer wants to signal something. Therefore, if a dialect word is inserted into a standard German posting it has a greater impact on class affiliation than the reverse case. This is why classifying mixed cases as standard is a worse decision, than classifying them as dialect. Considering these examples, RoBERTa’s dialect label tendency, analyzed as over-generalization in the previous chapter, might not represent a dialect bias but just a broader representational concept of dialect.

The described model preferences also apply to 9–10. Posting 9 could be realized in the same way by a standard and dialect writer, which means that they do not provide any hints on the intended variety. As DeBERTa preferentially assigns the non-dialect label and RoBERTa the dialect label this is a perfect example for the model tendencies. The same applies to 10: While “Guten” is a Standard German realization (dialect: “Guaten”), the lacking <e> in the final syllable of “Morgn” (Standard German: “Morgen”) could as well be interpreted as dialect typical schwa-apocope.

To sum up, the general observations made on the validation set also held true on the test set. What became clearer, though, is the fact that while DeBERTa clearly over-generalizes the non-

dialect and especially the undefined label, RoBERTa's representation of what is dialect seems accurate.

4.1.4 Posting classification – conclusions

```
Code files: 05_update_posting_labels.ipynb
```

In summary, Task 1 could be completed successfully. From originally 8180 undefined samples, only about a fourth, i.e., 1920 samples, remained undefined. The rest could be assigned to the dialect or non-dialect class. Even though the process was realized semi-automatically, a manual revision of the undefined labels was done as well. The reason for this was that both models showed a stable though not flawless performance on the validation set. Moreover, the validation set was relatively small and might not have been representative. Therefore, we wanted to evaluate model performance on a bigger hand-labeled dataset, i.e., the rest of the undefined samples.

As reported, this resulted in Accuracy scores around 80%. Such an overlap with human judgement of 80% is significant, but also implies that a fully automatized labeling mechanism would introduce an error in every 5 cases. We also analyzed these mistakes finding that both models had individual and characteristic error patterns. These patterns were evaluated by considering undefined and as such difficult classification cases. The error analysis did not take all data into account, which would not have been feasible in the context of this thesis. This is also the reason why we spoke of labeling "tendencies" in the previous sections. A further evaluation of mistakes will have to clarify if the found patterns can be transferred to the whole dataset.

What also became clear through error analysis is that evaluation metrics must be taken with a grain of salt. Even though RoBERTa seemed to perform worse on the test set, the error analysis revealed that its class representations mirrored human judgement more accurately. The reason is that evaluation metrics just consider error quantity, not quality. But if categories are vague and an intermediate category like "de undef" exists, one mistake often is not equivalent to another. The qualitative analysis revealed that RoBERTa committed more mistakes than DeBERTa (1085 vs. 1051 wrong assignments), but that the errors were often less serious and sometimes even debatable.

In terms of model comparison and choice of architecture this means that for the task at hand, the multi-lingual RoBERTa model outperforms English-only DeBERTa. Thus, pre-training on

different languages has a bigger positive impact on a cross-lingual task than sophisticated position encoding. As the qualitative error assessment could show, this is mainly a consequence of more accurate class representations in case of RoBERTa. A possible explanation for this could indeed be RoBERTa's multi-lingual knowledge: Among many languages, RoBERTa was also trained on standard German. Thus, the model just had to learn the distinction between an already existing and two new classes, while DeBERTa had to build all representations from scratch. This is a challenging task given that the undefined class contains very diverse and noisy samples also bearing features of dialect and Standard German contributions.

Based on the preceding considerations, XLM-RoBERTa was chosen to be applied to the second task. The following chapter, 4.2, will describe how RoBERTa was trained to distinguish between dialect and standard German on word level.

4.2 Task 2: Word-level classification

In the previous part the implementation of a dialect identification system on the posting level was described. The second part of this thesis involves dialect-standard classification on the word level. Similar use cases are, for example, named entity recognition as well as part-of-speech tagging. As chapter 3.3 described Transformer models as RoBERTa and DeBERTa are perfectly suited to solve these kinds of NLP problems as the Attention mechanism enables them to compute meaningful dependencies between words. This chapter describes how a slightly modified version of XLM-RoBERTa is trained and tested. The following chapter sheds light on the first step in this process: the generation of a training, validation and test set.

4.2.1 Training and validation set

`Code files: 06_assign_confidence_scores.ipynb, 07_labels_on_word_level.ipynb`

Task 1, described in chapter 4.1, had the advantage that the labels on posting-level were already provided. This was not the case for word-level classification. DiDi contains all sorts of information on the word-level like a normalization tag, lemma, POS tags etc. However, it is not specified if a word's transcription is standard German, dialect or has a form that could apply to both varieties. Thus, the first step was to establish corresponding labels on word level. As DiDi contains hundreds of thousands of words, it was not feasible to hand-label enough training data for successful optimization. The process to determine whether a word is intended to represent

standard German or dialect, thus, had to be fully automated. This is not a trivial task as word classification in this scenario is strongly context dependent. What we mean by that can be seen when looking at a word like “genau” (“exactly”). “Genau” is variety independent because it is uttered phonetically very similarly in both varieties. A dialect writer would most probably just use the canonical standard German form even when writing his dialect. Therefore, “genau” could be a dialect or a standard German word depending on which variety is intended. If the remaining sentence follows standard German transcription, it is very unlikely that “genau” would be interpreted as dialectal.

On the other hand, if the context words bear clear dialectal features, it is very unlikely that “genau” would be read as a Standard German word. This also applies to orthographical deviations: If in a word like “spieln” (“play”) the second to last letter is missing (canonical form: “spielen”) this could be an indicator for dialect-typical schwa deletion. However, it could also be an omission due to fast typing, a phenomenon frequently found in computer-mediated informal communication[24]. How the deviation must be interpreted, depends on the intended variety of the posting. Luckily, this information is already given in the DiDi corpus and was even refined in the previous task by specifying formerly undefined samples as dialect or standard German. To further profit from this refinement, the original DiDi corpus was updated such that the old labels were replaced with the labels predicted and revised in Task 1.

The simplest approach to transfer this higher-level information to each word within a posting, would have been to just assume that a dialect posting contains only dialect words. However, this would be an oversimplification as the posting in (12) shows:

(12) STD Eheh, i hoff, du frogsch mi nett, ob i die weißn Hoor a in Dutzenden ungebn konn, wie die Johr! ;) Alles wie gehabt.

SG Eheh, ich hoffe, du fragst mich nicht, ob ich die weißen Haare auch in Dutzenden angeben kann, wie die Jahre! ;) Alles wie gehabt.

EN Eheh, I hope you're not asking me to put my white hair in dozens too, like my age! ;) No big changes.

(id: 54616_t0119_m00201)

[24] see, for example, Verheijen (2017)

Comparing the dialect posting in (12) to its standard German transliteration, it becomes clear that the posting is a mixture of varieties. The first sentence is mostly dialectal, the second is standard German. Nevertheless, the whole posting is labeled as dialect in the corpus because the major part is in dialect. But to assume a perfectly dialectal posting and to label all words within the posting accordingly would be misleading.

The solution to this problem was to introduce a new dimension: the degree of dialectality of a posting. This decision was based on the following reasoning: The more dialectal the whole sentence is, the more likely it is that an embedded word can be interpreted as dialect. For this reason, we first split up longer postings into individual sentences. Then, we assigned each sentence a dialectality value. This was done by exploiting another useful property of neural nets. The SoftMax function in the last layer maps the incoming vectors onto logit values for each class. These logits can be interpreted as the degree of confidence of the net regarding a certain class. In this way the values of the SoftMax output can provide an approximation of dialectality or standard orientation.

To receive these values the pre-trained RoBERTa model was used. Each of the split-up sentences was fed into the model and its class predicted. Instead of aiming for the class labels, though, the logit value for each class was retained. In this way, a confidence score for each sentence and class was available. Figure 12 shows some examples of postings with their gold label, predicted label and logit scores.

	text	gold_labels	predicted_labels	predictions
2	<PersNE> und <PersNE>	de undef	de dialect	[0.5853317, 0.33937538, 0.075292945]
15	Dann schauma mal was der SIDO so derzaehlt	de undef	de dialect	[0.9288972, 0.051719043, 0.019383693]
47	Du besser mittwoch gehn!	de undef	de non-dialect	[0.14724858, 0.8448279, 0.007923423]
53	Also obenauf schworze Hoor und unten hell??	de dialect	de non-dialect	[0.3462737, 0.5603544, 0.09337192]
56	Ja is es...fast scho zu viel...hat du eigtl zj...	de undef	de dialect	[0.95875406, 0.0043283207, 0.036917627]
...	...	...	...	...
14842	(y) genau!	de undef	de non-dialect	[0.44175416, 0.5293872, 0.028858691]
14873	Danke!: -)	de undef	de dialect	[0.5285776, 0.4325822, 0.038840123]
14912	7 . Juli - Heiliger Willibald	de undef	de non-dialect	[0.14759333, 0.8453815, 0.0070251874]
14930	ital test mit bruno mars ...	de undef	de dialect	[0.94087607, 0.053794432, 0.005329597]
14941	GUUEETE ORBAAAIIT Pronto per 3 giorni di corso...	de dialect	de non-dialect	[0.0028115832, 0.9961837, 0.0010047205]

Figure 12: Predictions with logit values/confidence scores

The last column represented in Figure 12 shows the logit values outputted by RoBERTa. The first value in this list is to be interpreted as dialect, the second as non-dialect confidence score.

The third value is the logit for the “de undef” label. As can be seen, the examples given here are all instances of wrong classification. The reason why we show these examples here is because they provide a good illustration of potentially low and variable confidence scores. This is indeed the case for most of the above instances. Even the aforementioned word “genau” as well as “danke,” both words with a transcription used in both varieties, appear. Correspondingly, the confidence scores for assigning the dialect or non-dialect category are nearly identical. Such cases should be classified as “de undef”. The logit for this label, however, is even lower, which shows that the model did not fully grasp the intended concept of the undefined category. Nevertheless, Figure 12 shows that using the scores for the dialect and non-dialect label enables us to single out good representatives for both of these classes.

After splitting up the postings in sentences and assigning confidence scores this resulted in 24909 dialect, non-dialect and undefined sentences with label scores as shown in Figure 12. At this point we could have filtered out good representatives for each class and just assign each word in a sentence the sentence label. However, this would still have introduced too much noise to the word labels, especially regarding the dialect class. As we argued in the introductory part, dialect writing can be seen as a variation of a standard German transcription system and writers do not arbitrarily deviate from this norm. In case of dialect writing the deviation has the clear goal to signal the vernacular to create a more personal, conceptually oral communicative space. This also implies that not every written form has to be adapted to reach this communicative goal. Recurring to the example in (12) the wording “in Dutzenden” would actually not be used like that in South Tyrolean dialects as the third case plural is not marked with the flexion suffix “-n”. This is a property of standard German and shows that standard German phrases also appear in dialectal contexts. Consequently, such words would influence the model’s dialect representation unfavorably. For this reason, we further refined the implemented word labeling algorithm by using additional word information contained in the DiDi corpus. This involved the variables displayed in Table 19.

Variable	Variable description	Variable values
lemma	Lemma of the token	Lemma as a string
norm	Normalization in Standard German	Normalized form as a string
stir	Specifies if the token is genuinely South Tyrolean	{'True', True, 'n.a.'}

Table 19: DiDi variables to aid word level classification

The variety information on the sentence level was combined with the variables in Table 19 in the following way to create a refined word-labeling algorithm:

- For all sentence tags: In the case of words from languages other than German, the lemma entry in the corpus is specified as <unknown>. In this case the tag "unknown" is assigned to a word.
- If the sentence tag is "de dialect": If the word entry contains a normalization, i.e., the original word token had to be normalized, the word will be labeled as dialect unless the normalization is not only a matter of uppercase vs. lowercase. A token without a normalization entry is labeled as neutral. Moreover, if the entry of "stir" is "True", i.e., the word is labeled as a genuinely South Tyrolean word, it will be labeled as dialect as well.
- If tag is "de non-dialect": All words are labeled as "non-dialect" unless they are "unknown". The normalization entry is not considered, as typos are also considered as standard German even though the grapheme realization is norm-deviant.
- If tag is "de undef": Words with normalizations are labeled as dialect, those without normalizations as standard German.

It follows from this description that four labels were assigned by the algorithm: "unknown" for foreign or not recognizable words, "dialect" for dialectal words, "non-dialect" for standard German words and "neutral" for words which are transcribed in the same way in both varieties. As can be seen, the tags "unknown" and "neutral" replaced the "de undef" tag from the previous task. The reason for this decision was the fact that "de undef" was a very vague category covering both mixed and foreign postings. The goal of assigning these cases to two distinct labels was to improve the model's representation of what counts as dialect or standard German, and what might represent an intermediate case.

The above description also makes it clear that the decision of which label is assigned is strongly directed by the sentence tag, i.e., assumed variety. In this way, it accounts for the context-dependent nature of the word labeling process. However, even this more sophisticated procedure did not lead to completely satisfying results. Considering the posting in (12) the first sentence would be labeled as shown in Table 20.

unk.	**neutr.**	**dial.**	**dial.**	**neutr.**	**neutr.**	**dial.**	**dial.**
Eheh	,	i	hoff	,	du	frogsch	mi
dial.	**neutr.**	**neutr.**	**dial**	**neutr.**	**dial.**	**dial.**	**dial.**
nett	,	ob	i	die	weißn	Hoor	a
neutr.	**neutr.**	**dial.**	**dial.**	**neutr.**	**neutr.**	**neutr.**	**dial.**
in	Dutzenden	ungebn	konn	,	wie	die	Johr

Table 20: Word labeling algorithm – example sentence

Including punctuation symbols, the sentence in Table 20 contains 24 tokens. The algorithm was able to capture all dialect words correctly. Also, most of the tokens labeled as neutral are indeed neutral, for example punctuation marks and the personal pronoun “du”. Thus, the algorithm’s general performance has already significantly improved. However, the algorithm fails to capture two important elements: “Eheh” is a clear onomatopoetic marker of laughter/smiling and should be labeled as “neutral” in this context. Also, as already mentioned, “Dutzenden” is clearly a standard German word form and should be labeled as such. The latter reveals one of the biggest downsides of the algorithm. The procedure only allows certain label combinations. Even though standard German words do appear in dialect sentences, the algorithm does not allow for this scenario. The same applies to the fact that neutral words can appear in mixed, undefined sentences. In the latter case, however, the algorithm just decides between dialect and standard label. It follows that this setup prevents the model from learning that all label combinations are potentially possible. This is an unwanted bias in the training data.

For this reason, a third refinement step was taken. To capture neutral elements like onomatopoetic expressions, smileys, emojis, and digits simple regex functions were used that captured the relevant strings and labeled them as neutral.

The second issue required a more sophisticated solution: In order to decide whether a neutral word in a dialect sentence is indeed neutral or might be standard German, a basic probabilistic classifier on n-grams was implemented. To do so, the following steps were carried out:

1. Using the confidence scores for dialect and non-dialect sentences, good representatives for both classes were chosen. Also, care was taken that each class contained a similar number of characters. In this way, the dialect sentences contained 88671 tokens consisting of 478420 characters, while the non-dialect sentences contained 76746 tokens consisting of 478251 characters.
2. Each token was divided into its unigrams, bigrams, trigrams and tetragrams.
3. For each n-gram it was counted how often it appeared in both classes. This count was then divided by the overall count of this n-gram. This led to one probability value for each of the two classes, specifying how likely it is that this n-gram occurs in a dialect or standard German word.
4. The classification algorithm, then, takes these probabilities and computes the probability of a word pertaining to one or the other class. To do so, the word is split up into its n-grams, and the logarithmised probability values of the n-grams are summed up for each class.
5. Then, a dialect/non-dialect ratio is computed by dividing the dialect probability sum resulting from the previous step by its non-dialect probability sum.
6. This results in a ratio where a smaller value indicates a higher probability for a dialect word and a higher value a higher probability for a non-dialect word. This has to do with the fact that probabilities were logarithmised. When probabilities are logarithmised, higher probabilities get a smaller negative number. The probability 0.8, for example, gets the value -0.097 while 0.2 gets the value -0.7. If we divide a high dialect probability through a low non-dialect probability this will result in a low score and vice versa.
7. Using these scores, a dialect and a non-dialect threshold was empirically determined. If a word's score was below a certain (empirically tested) threshold, it was assigned the dialect label. If it was above this threshold, it was assigned the non-dialect label. And if it was between the thresholds the neutral tag was assigned.

The n-grams that were most representative for dialect or standard German made sense from a linguistic perspective. This becomes clear if we look at the ten most important n-grams for predicting a dialect or standard word, displayed in Table 21.

	Dialect			Standard German	
	N-Grams	*Score*		*N-Grams*	*Score*
1	('w', 'i', 'a')	0.000224	**11**	('m', 'i', 'c', 'h')	284.8116
2	('w', 'o', 's')	0.000518	**12**	('d', 'i', 'e', 's')	240.5592
3	('m', 'u', 'a')	0.000537	**13**	('m', 'i', 'c')	173.661
4	('h', 'o', 's', 'c')	0.000581	**14**	('h', 'a', 'b', 'e')	154.5808
5	('g', 'e', 'a', 't')	0.000585	**15**	('d', 'i', 'c', 'h')	97.895
6	('o', 'a', 's')	0.000627	**16**	('h', 'a', 'b')	90.63082
7	('h', 'a', 'u', 'g')	0.00063	**17**	('e', 'h', 'e', 'n')	61.53979
8	('m', 'u', 'a', 's')	0.000657	**18**	('i', 'e', 's', 'e')	54.45886
9	('o', 'a', 'n')	0.000768	**19**	('e', 'h', 't')	42.96021
10	('m', 'o', 'l')	0.000991	**20**	('n', 'i', 'c', 'h')	39.66768

Table 21: Most predictive dialect and standard n-grams

If we compare the n-grams in Table 21 with the most salient dialect features described in chapter 1.4.1 we can see that many of these distinctive dialect features are contained. Examples 1, 3, 5, 6, 8 and 9 comprise representations of diphthongs typical for South Tyrolean dialects. Consequently, these diphthong patterns cannot be found when inspecting the standard German n-gram column. The typical posterior open vowel [ɑ], written as <o> by dialect writers, is also discernible in 2, 4 and 10. By contrast, <a> is more characteristic for standard German as can be seen in 14 and 16 which both contain inflected forms of the verb “have” (“haben”). The example in 4 is also interesting from another perspective: The n-gram ('h', 'o', 's', 'c') clearly points to the sequence <sch>, which is commonly used to indicate retracted s in dialect writing.

The examples in 14 and 17 show another salient difference between South Tyrolean dialect and standard German transcription: Dialect writers often omit the character <e> in unstressed suffixes and prefixes – a clear signal for schwa-apocope and -syncope. Accordingly, it is a distinctive feature of standard German if this character is preserved in writing. In these regards, 17 can be interpreted as the typical ending of a standard German infinitive form as in “stehen” (to stand), “gehen” (to walk) and “sehen” (to see), all containing the schwa representation in the final syllable.

Thus, when looking at the ten most distinctive n-grams for both varieties it was already possible to detect 4 out of 5 of the most salient phonetic dialect features. It is even possible to determine the most important features for dialect identification overall. To do so, we took the 100 most predictive dialect n-grams that had at least 250 occurrences in the data. These n-grams were analyzed for salient dialect features following the same procedure as in the above analysis. In this way and by taking into account the overall probability and frequency of each feature, the following ranking could be established:

1. Diphthong
2. Back vowel a
3. Retracted s
4. Schwa deletion

According to the probabilistic classification system the complex diphthong system is most characteristic for dialect writing, followed by using <o> for <a>, the transcription of s-retraction and schwa-deletion. Obviously, this can only be a preliminary analysis that has to be confirmed on other data and by using statistical confidence metrics. Especially, because a feature like unrounding is harder to spot in mere n-grams. Nevertheless, the above ranking coincides with the observations made about the data in Table 21 and provides interesting insights into dialect writing habits. The analysis of the most predictive n-grams also revealed that the probabilistic classification system was able to capture characteristic dialect and standard features.

Therefore, at this point the probabilistic classifier could be integrated into the general word labeling algorithm described above. Within this algorithm it is applied to sentences labeled as dialect or undefined. In case of dialect sentences the probabilistic classifier is applied to words without normalization to decide if they are indeed neutral or rather standard German. In case of undefined sentences, the probabilistic classifier is applied to all words. The reason for this is that undefined sentences are often a mixture of both varieties, and it was assumed that they would profit from a maximally flexible treatment. Sentences labeled as non-dialect in the corpus were still treated as if each of their words is standard German, even if they deviated from the orthographic norm. The reason for that is that we wanted the system to learn that a word is not automatically dialect if it deviates from the canonic standard German representation. This specifically applies to common phenomena in CMC with occurrences like omission of characters, contractions, etc. that can also appear in standard German online texts. By labeling

such deviations as standard German we wanted the system to learn the difference between medium-dependent writing forms and actual dialect patterns.

After the described refinement steps were completed, the word labeling algorithm could be applied to the DiDi data to generate the training data. To guarantee that the system would learn meaningful class representations, only sentences with high dialect or standard confidence scores were chosen. Undefined sentences were filtered to have a length of at least three tokens. In this way, 11464 standard, 8134 dialect and 827 undefined sentences were selected as training data. In summary, these sentences comprised 323572 words labeled by the algorithm. 60% of this dataset was used to train RoBERTa, while the other 40% was used to validate its performance during training time and to choose the best model version accordingly. Chapter 4.2.3 will describe this training process and model specifications in more detail. The next chapter will take a closer look on the test set, the MoCoDa dataset.

4.2.2 Test set: MoCoDa

`Code files: 08_MoCoDa_import.ipynb`

The MoCoDa project (Beißwenger et al. 2020) – short for Mobile Communication Database – aims to build a database of everyday communication by means of electronic short messages and to make it available for university research. Part of the project is also a dataset of WhatsApp messages collected from South Tyrolean users (Alber et al. publication in preparation). This dataset consists of 83 chat conversations with an overall number of 973 messages and 8567 tokens. Each chat conversation comes with meta-information about the number of involved participants and exchanged messages as well as a description of the general topic and the time of creation of each message. Token level pos tags as well as other token specific information – such as if a token is an emoji or a punctuation mark – are provided as well. Most of the chat messages are written in South Tyrolean dialect. However, information about the variety of a message or individual tokens is not given.

We decided to download and use this dataset as test set for our word labeling system, because of three main reasons:

1. The dataset is relatively small. Therefore, we could label all tokens in the dataset by hand, assigning them to the dialect, standard, neutral or unknown class. The labels assigned by hand provide a high-quality gold standard for final system evaluation.
2. By using a dataset from a different source and from a platform with potentially slightly different communication habits and contents (Facebook vs. WhatsApp) it can be

guaranteed that the system did not learn domain specific representations. In this way, we make sure that the system did not over-fit to the training data and is applicable to different datasets.

3. We add another annotation layer, namely, dialect tags, to an existing dataset that can be useful for future research in dialectology.

After pre-processing and labeling the described MoCoDa data, all datasets for training and evaluation were ready. The next chapter will briefly describe the model architecture and training procedure.

4.2.3 Training XLM-RoBERTa for word-level classification

```
Code files: 09_word_labeling_net_roBERTa.ipynb
```

Task 1 showed that XLM-RoBERTa was able to learn more meaningful class representations. Therefore, it was decided to only apply RoBERTa to the second task. This application of RoBERTa to word instead of posting classification only required a small change in architecture. For posting classification, RoBERTa computes a vector for the whole posting using the model-specific [CLS] token. This vector is then fed into the final classification layer that outputs class probabilities.

In case of word classification, instead of using sentence vectors, each individual word vector is retrieved and fed into a classification layer. Thus, the only real modification to the model is to the input of the classification layer. The general model architecture remains unchanged. This is also reflected by the fact that the number of model parameters is nearly identical Figure 13 shows.

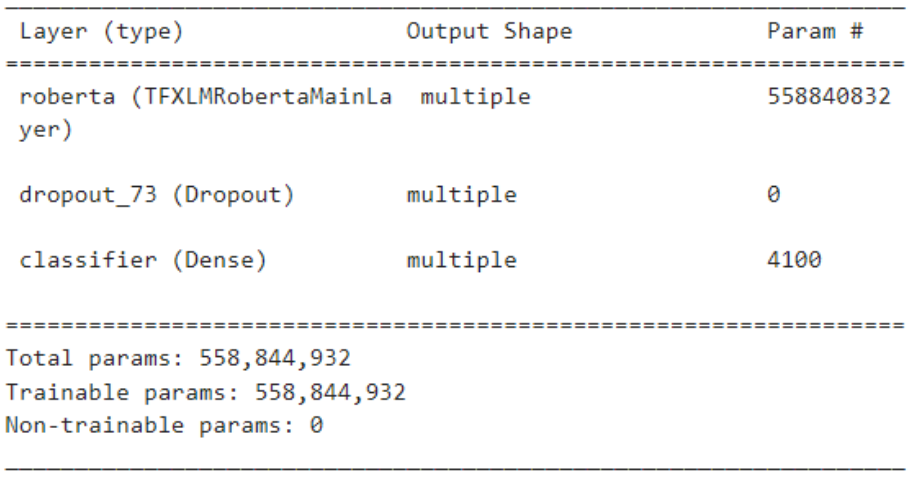

Figure 13: XLM-RoBERTa model summary

As can be seen in Figure 13 the model has over half a billion parameters. Its general setup when loaded with the Huggingface Transformer library is very concise: The first layer of the model is the actual Transformer layer outputting all word vectors of an input sentence at once. This is why the output shape is specified as "multiple". The Transformer layer is followed by a Dropout layer to prevent overfitting and the final layer is a classification head. The word vectors are fed into these layers in parallel, which guarantees efficient computation.

We decided to re-load all model weights from Huggingface from scratch instead of using the model weights learned in the previous task. In this way we wanted to prevent any form of interference effects between tasks which might result from such a transfer learning approach.

The model was trained for 100 epochs. The parameters that performed best on the validation set were stored. The optimizer function remained unchanged starting with a learning rate of $2e^{-5}$ and a weight decay rate of 0.1 for more stable convergence.

The dropout rate was 0.4. This hyperparameter setup was empirically evaluated by re-training the model with different hyperparameter values.

The next chapter will describe model performance on the DiDi validation and the MoCoDa test set and point out error patterns.

4.2.4 Performance evaluation and error analysis

4.2.4.1 DiDi validation set

Before we take a closer look at the model's performance on the validation data, an important detail about RoBERTa's tokenization process should be mentioned: All neural net models convert input words into numbers to be able to perform mathematical operations on them. How words are encoded differs between model architectures. A simple approach would be to assign each word its own index in a pre-compiled dictionary. However, this has the disadvantage that only known words can be assigned a number during prediction time. One solution to circumvent this problem of out-of-vocabulary words is to break down words into their individual characters and to assign each character a number. BERT-derived models like RoBERTa and DeBERTa choose an intermediate approach.

During tokenization these models break down words into frequently occurring sub-word units. For example, the word "Wohnung" (appartement), which occurrs in the training data, is split into two subwords – "Wohn" and "ung". The tokenizer will assign both components the same label that is given to the whole word. During prediction time, the model will predict a label for all subword embeddings of a word, for example "neutral" for "Wohn" and "neutral" for "ung". In most cases these labels will be identical. However, the labels for a word's subwords can also differ.

In these cases, a choice of which of the subword labels will be assigned to the entire word must be made. We decided to take the most frequent subword label as the final word label. If two labels were equally frequent, the label of the first subword was assigned. Accordingly, the following evaluation scores are based on these final labels for entire words, not subword labels. Table 22 shows model performance on the validation set.

	Precision	Recall	F_1	N° samples
dialect	0.985	0.973	0.979	22878
neutral	0.978	0.988	0.983	42248
non-dialect	0.987	0.992	0.989	59954
unknown	0.913	0.810	0.858	3444
macro average	0.966	0.941	0.952	
weighted average	0.982	0.982	0.982	
accuracy	0.982			

Table 22: Task 2 – Model performance on the DiDi validation set

The DiDi validation set comprised 128524 labeled tokens. As is evident from Table 22 most of the tokens were non-dialect tokens, followed by neutral and dialect tokens. The class with the lowest number of representatives was "unknown". When splitting apart the training and validation sets, care was taken to keep the ratio of classes in the two data sets comparable. Thus, "unknown" is also the most infrequent class within the training set. Together with the fact that unknown words cover a variety of cases (unassignable words, words from different languages other than German), which makes classification harder, this might have led to their lower classification score. Nevertheless, an F_1 score of 0.858 is still significant. As the performance scores of the remaining, more frequent classes are all close to 0.99, this leads to an overall model Accuracy of 0.982. This seems like a very high performance rate on the first glimpse, especially when compared with the performance rates of Task 1. However, the expressiveness of these high scores on the validation set is relativized by two factors:

1. The dialect identification system as implemented here is a mixture between sequence labeling and language classification/detection. The main difference to most language classification approaches is that the system operates on individual words instead of whole sentences or texts. A look into state-of-the-art language detection systems shows that these systems all achieve very high Accuracy scores up to 99%. Caswell et al. (2020: 1) even report that language identification is "largely treated as solved in the

literature". While we just distinguished between four classes in our experiments, some of these systems can classify hundreds of languages still achieving good results.

2. The gold standard of the validation set has been automatically created. Even though the previously described word labeling algorithm that generated the training and validation set underwent several refinement steps, it still commits errors. This can be seen when looking at Table 23. The table contains two types of mistakes: 1–5 are cases where the word labeling algorithm labeled dialect words as non-dialect, but the model still assigned the correct label. This is positive from a generalization perspective as the model seems to have built correct class representations, even though the training dataset is noisy. However, such cases limit the meaningfulness of the evaluation metrics, as correctly predicted labels are counted as mistakes. Rows 6–10 contain words all pertaining to the same sentence. Here, the word labeling algorithm and the model both assign the neutral category. However, this is not correct, as the sequence is written in standard German transcription.

	words	**gold labels**	**predicted labels**
1	findesch	non-dialect	dialect
2	okrotzen	non-dialect	dialect
3	iatzzzzz	non-dialect	dialect
4	sigsch	non-dialect	dialect
5	olls	non-dialect	dialect
6	das	neutral	neutral
7	schwedische	neutral	neutral
8	Königshaus	neutral	neutral
9	hat	neutral	neutral
10	offiziell	neutral	neutral

Table 23: Examples of words wrongly labeled by the word labeling algorithm

This shows that the validation set does not provide a solid basis for system evaluation. Thus, the analysis of the test set results will give a better insight into actual model performance.

4.2.4.2 MoCoDa test set

Table 24 shows model performance on the MoCoDa test set.

	Precision	**Recall**	**F_1**	**N° samples**
dialect	0.960	0.929	0.944	4884
neutral	0.896	0.926	0.911	3468
non-dialect	0.399	0.426	0.412	162
unknown	0.250	0.377	0.301	53
macro average	0.626	0.665	0.642	
weighted average	0.919	0.915	0.917	
accuracy	0.915			

Table 24: Task 2 – Performance on the MoCoDa test set

As already mentioned, the MoCoDa test set comprises fewer tokens, which is why all labels of the 8567 tokens could be assigned by hand. This allows a good approximation of the output's overlap with human judgement. As can be seen when looking at Table 24, the system's performance significantly drops when applied to the MoCoDa data. Even though the overall Accuracy of 0.915 is still comparatively high, the macro averaged F_1 drops to 0.642. The reason for this drop, among others, is the very different class distribution of the dataset. While "non-dialect" was the most frequent class in the training and validation dataset, there are only very few non-dialectal words in the test set. This is due to the fact that the MoCoDa dataset was collected by asking undergraduate pedagogy students from the university of Bozen-Bolzano/Brixen-Bressanone to contribute some of their WhatsApp conversations to the project. Since the program is aimed at teachers of German-speaking elementary school, the collected messages are almost exclusively written in German. Among this young target group the dialect is the pre-dominant variety in instant messages. Unknown words also occur very rarely in the collected data. Instead, dialect words and neutral words dominate. It follows that the probability distribution of classes is significantly different when comparing training/validation and test set.

While the likelihood of meeting a non-dialect word in the DiDi dataset was nearly 50%, it is only 2 percent on average in the MoCoDa dataset. In light of this, an F_1 score rate of 42,6% can still be seen as a proof that the model developed a working non-dialect class representation, even if this class representation is not perfect. The same applies to the "unknown" class. This finding is confirmed by a look into the classified data. When looking at the error confusion matrix in Table 25, it can be seen that 20 unknown words as well as 69 non-dialect words were correctly classified by the model (cells in green).

	dialect pred.	**neutral pred.**	**non-dialect pred.**	**unknown pred.**	**Total gold**
dialect gold	4538	276	46	24	4884
neutral gold	168	3212	53	35	3468
non-dialect gold	7	85	69	1	162
unknown gold	15	13	5	20	53
Total pred.	4728	3586	173	80	8567

Table 25: Task 2 – MoCoDa error confusion matrix

All correctly classified unknown words as well as twenty of the correctly classified non-dialect words are listed in the following:

1. Unknowns: elaborato, sowi, rewi, C.C, A.A, bk, easy, peasy, Yessss, _---Keller_, Vediamo, Abla, bk, fisso, them, nice, too, Great, Great, dopo
2. Non-dialect: darf, einer, alarmstufe, landefeuerwehren, klarooo, späteer, über, neue, Lektürebericht, zusammen, gestartet, nägel, geht, keine, ahnung, Ja, Soll, ich, später, Brot, kaufen

As becomes evident, all words in 1 correspond to the intended class concept: Most of them are foreign words, mainly from English and Italian. The remaining words are intransparent abbreviations or words with special characters attached. The words in the non-dialect group also indicate that RoBERTa modeled the non-dialect class correctly. But why is performance on the non-dialect and unknown classes so low then? One part of the answer has already been given and has to do with class distribution. As non-dialect is frequent in the training, but not in the test set, the model over-assigns the non-dialect label to dialect words (blue cell).

The other potential explanation lies within the cells marked in yellow. Remarkably, these cells all lie on the neutral label axes. This means that the highest degree of disagreement between gold labels and predicted labels regarded the neutral class. The explanation for this is two-fold:

On the one hand, the "neutral" label suffers from the same problem as the "de undef" label in the previous task. Even though the concept of "undefined" from Task 1 has been split up onto the labels "neutral" and "unknown" for the present task, "neutral" is still hard to grasp as a concept. In theory, it should capture words which are neither specifically dialect nor standard German, i.e., possible in both varieties. But as dialect writing is not a norm-regulated process what counts as dialect or standard German is a matter of debate.

Firstly, this depends on the writer's base dialect. Some writers might, for example, accept the word "für" ("for") as neutral, because they would also transcribe it like that when writing their dialect. Others would argue that only "fir", showing the typical unrounding of umlauts in South Tyrolean dialects, is the legit dialect form, and "für" is already marked as standard form. In the logic of the word labeling algorithm, "für" was labeled as "neutral" in dialect contexts and "non-dialect" in Standard German contexts. We hoped that RoBERTa in this way would learn to base the assignment decision on the context. However, RoBERTa consistently labels "für" as being "non-dialect", even though in MoCoDa the preposition mostly appears within dialect contexts. This alone led to eight disagreements between predictions and gold labels. Besides non-dialect mistakes, such ambiguous cases also accounted for the majority of the 444 mistakes on the dialect/neutral axis. Most of these errors regarded words that can be labeled both as neutral or dialect depending on the context. The token "wer", for example, in dialect writing is used to denote the pronoun "who" like in standard German. However, it is also often used to denote the 1st person auxiliary verb used in the future tense ("I wer" – "Ich werde"; "I will"). Again, this type of word can be labeled as neutral, standard or dialect depending on its sentence context.

Secondly, there were tokens deviating from the standard spelling that we labeled as "neutral", because it was unclear, if they are a representation of dialect writing or a CMC phenomenon. We already mentioned that schwa-apocope is a common phonetical process in South Tyrolean dialects. Accordingly, dialect writers often omit the character <e> in the final syllable of infinitive verbs like "duschn" ("to shower"), "chilln" ("to relax"), "wissn" ("to know"), etc.[25].

[25] All these examples were taken out of the MoCoDa dataset.

However, so do writers communicating in informal standard German. In the latter, this can be seen as a medium-dependent phenomenon. The same applies to cliticized word forms like “dus” (“you it”) or “wies” (“how it”). Such forms could be equally interpreted as dialect, neutral or standard German.

The second reason, why most mistakes regarded the neutral class, has to do with the automatic word labeling process. Contrary to a manual labeling process, the automatic approach introduces systematic errors. This can lead to whole sequences being assigned to the wrong class. This is especially true for neutral words in dialect sentences as their assignment required an additional probabilistic classifier deciding if they are neutral or have a standard German form instead. This classifier did not always come to the right decision which led to various standard German words being classified as neutral with the consequence that the decision boundary between what is neutral and what is proper standard German becomes even more imprecise.

4.2.5 Neutrality in dialect postings

The preceding chapter mostly addressed the drawbacks of the neutral class in a classification context. However, an F_1 score of over 0.9 is still significant and makes the neutral category a valuable source of information. Especially as the equivocal nature of neutral tokens can provide important insights into dialect writing habits in relation to Standard German orthography. In chapter 1.4.3 we already argued that dialect writers do not transliterate their vernacular arbitrarily but rather seek orientation in the orthographical norm.

A good example can be seen in a phenomenon that is described in Felder (2015): Felder observes that in Swiss German, word-initial and -internal consonant clusters like [ʃt] and [ʃp] are most often not written in a way that fits Swiss German dialect phonetics. Instead of the phonetically more appropriate grapheme patterns <scht> and <schp> dialect writers more frequently used the canonical <st> and <sp> forms. Taking a look at the South Tyrolean data with special attention to the neutral class it becomes evident that this writing habit can also be observed for South Tyrolean dialect writers. The words “Stunden” (“hours”) and “Spaß” (“fun”), for example, have the following variants in postings labeled as “de dialect”, i.e., in dialectal contexts:

Variant	stunden (neutral)	stund (dialect)	stundn (dialect)
Count	13	8	4

Table 26: Spelling variants of "Stunden[26]" ("hours") and "Stress" in dialectal contexts

Variant	spaß (neutral)	spass (neutral)	schpaß (dialect)	soaß (dialect)	spas (dialect)
Count	21	10	1	1	4

Table 27: Spelling variants of "Spaß" ("fun") in dialectal contexts

As can be seen, the neutral, orthographically correct variants are most frequent, even if the intended variety of the posting is dialect. In case of "stunden" it's also interesting to see that even the schwa vowel in the unstressed final syllable is more often preserved than deleted, even though deletion would be more time-efficient from a CMC perspective. The only variant following the dialectal pronunciation is "schpaß". Interestingly, in this case the writer keeps the word-final <ß> grapheme instead of using <s>, which, again, follows standard German spelling conventions. A similar picture emerges when analyzing the mentioned consonant clusters word-internally, exemplified by Table 28 and Table 29.

Variant	gestern (neutral)	gestrn (dialect)	geschtern (dialect)	gesto (dialect)	geschter (dialect)
Count	48	10	7	5	5
Variant	geston (dialect)	gestr (dialect)	gestorn (dialect)	geschtorn (dialect)	
Count	4	2	2	1	

Table 28: Spelling variants "gestern" ("yesterday") in dialectal contexts

[26] All variants were lower-cased.

Variant	versprechen (neutral)	verschprechen (dialect)
Count	3	1

Table 29: Spelling variants of "versprechen" ("to promise") in dialectal contexts

Table 28 shows all variants of the word "gestern" ("yesterday") in dialectal postings. Despite the intended variety being dialect in the word-internal case the orthographically correct, neutral spelling is chosen most often. In summary, for 71 tokens the canonical cluster spelling is preserved, while it is changed to the phonetically closer variant in only 13 cases. This is even more remarkable as the first group also contains tokens where other parts were changed to signal a dialectal reading. "gestrn" and "gestr", for example, show the typical schwa-deletion, while in other cases the schwa is replaced with a full vowel. We could not find a comparably striking example for word-internal [ʃp], but the example of "versprechen" vs. "verschprechen" in Table 29 suggests a similar tendeny for this cluster.

These observations are an indication of a potential standard orientation of dialect writers. We use the term "potential" here as nowadays automatic word correction algorithms are commonly used, especially on mobile devices. The use of such systems limits the validity of any analysis about neutrally written words, as the spelling is not a deliberate choice of the writer anymore. However, the fact that writing variants do occur in the corpus suggests that the availability of text correction does not automatically prevent dialect writers from choosing deviant spelling variants. By this logic, also the choice of a corrected variant is a deliberate decision signaling a certain preference.

Before concluding this chapter, we also want to briefly elaborate on a second interesting feature of the neutral tag. Disregarding punctuation symbols in standard German postings, the neutral category is exclusively assigned to words having a canonical spelling in a dialect posting. This makes the neutral category a good starting point to compare users with diverging spelling habits. To do so, we compared the number of neutral tokens a user contributed to the corpus normalized by the dialect, non-dialect and the overall number of tokens he/she donated. To do so we applied the following formula:

$$\left(\sum_{i=1}^{k} neutral\_tokens_u - \sum_{j=1}^{l} dialect\_tokens_u\right) \times \left(\frac{\sum_{i=1}^{k} neutral\_tokens_u}{\sum_{h=1}^{m} non\_dialect\_tokens_u}\right)$$

For the first term we subtract the sum of dialect tokens from the sum of neutral tokens of a user. This is an indicator of how frequent neutral tokens are in dialectal contexts. As also punctuation, emojis and numbers in standard postings count as neutral, a high frequency of standard postings would artificially increase the first term. For this reason, the second term was added computing the ratio of neutral and non-dialect tokens. This ratio is lower for users where the number of neutral tokens is high because of a high output of standard postings and, therefore, weights the impact of the first term.

In this way, we could approximate the use of neutral tokens in dialect contexts, just using the overall label counts. As such neutral tokens are standard-compliant by definition, they are a good indicator if a writer tends to the orthographical or phonetical pole when transcribing his/her dialect. Using this metric, we established a ranking of users who contributed at least 100 tokens to the corpus and had a remarkably high ratio of neutral tokens in their dialectal postings. Table 30 shows the first and last 10 users[27] of this ranking and their provenance.

[27] Starting with the last rank.

Top 10		Last 10	
user_id	*german_dialect_region*	*user_id*	*german_dialect_region*
54740	Meran	55052	Bozen
56950	Meran	57100	Sarntal
55354	Eisacktal	56409	Pustertal
56150	Bozen	56411	Meran
57040	Bozen	54957	Eisacktal
54625	Bozen	56979	Pustertal
56747	Bozen	56965	Bozen
56503	Meran	56304	Pustertal
57031	Bozen	55206	Meran
55233	Meran	56978	Sarntal

Table 30: Top 10 and last 10 ranks of number of neutral vs. number of dialect tokens

The first thing to notice in Table 30 is that the Top 10 positions are dominated by users stemming from urban centers. Bozen and Meran are the two biggest cities in South Tyrol. In the Eisacktal ("Eisack Valley") also two major cities are located – Brixen and Klausen. The last ten positions, on the contrary, are filled equally by urban and rural areas. These rural areas, i.e., Sarntal and Pustertal are valleys known for their characteristic, more conservative dialects.

The ranking now suggests that writers in an urban environment tend to stick closer to the orthographical norm than writers in rural areas when transliterating their dialect. This first impression is confirmed when comparing the dialectal contributions of the users 56150 and 56409. These users were chosen, because their areas of origin – Bozen and Pustertal – were the most frequent in the Top 10 and Last 10 ranking, respectively, which speaks for a common tendency. Table 31 shows two particularly illustrative examples for the writing patterns of both users when transliterating their dialects.

posting id	dialect region	full text	translation
56150_px1167_c5716	Bozen	i han leider s letschte Mal kon Daweil ghob zum Wandern ... bin in der ärgsten Hitz (30-38°) im Zimmer gsessn und han zun Arbetn ghab.	Unfortunately, I didn't have time to hike last time … sat in my room (30-38°) in the worst heat and had to work.
56409_t0765_m03689	Pustertal	gor ka stimme hon heint kop. zi scham :/ gesto schon nimm viel kop u heint gor kana ma... du vokiahlt?	Had no voice at all today. Shame on me :/ had not much yesterday already and today it was completely gone... you cold?
dialect = red	non-dialect = blue	neutral = green	

Table 31: Two example postings of the users 56150 and 56409

The first thing to notice is that the first posting by user 56150 contains 13 dialect and 8 neutral tokens (disregarding punctuation), while the second posting only contains 15 dialect tokens and 4 neutral tokens. Thus, even though the first posting is longer (24 tokens) than the second (20 tokens) the percentage of dialect tokens is still significantly higher in the latter (54% vs. 75%). On the other side the first posting has twice as many neutral tokens as well as more standard tokens. These quantities show that the first posting tends to the standard pole of dialect writing, the second to the dialect pole. But the difference is also a qualitative one. User 63150 transliterates the auxiliary verb "I have" with "han" while user 56409 uses "hon" for the same form. While "hon" shows the typical transcription for the retracted vowel [ɑ], "han" sticks with the front vowel [a] used in standard German.

Interestingly, user 56150 seems to have the retracted vowel in his phonetic inventory as well, as can be seen looking at the past participle form "ghob" ("gehabt", "had"). But he/she seems to generally prefer the [a] transcription, as "ghob" again changes to "ghab" in the last line of the posting. This is interesting from a linguistic perspective as it implies an active decision process between two concurring forms – a dialectal and a more standard-oriented. By contrast, user 56409 consistently uses <o> to transliterate [ɑ] as can be seen, for example, in "kop". "kop" stands for the same past participle form as "ghob". In standard German this form is

written as “gehabt”. If we compare the two dialectal forms with the standard form it becomes evident, that the transcription used by the Bozen user is closer to the standard while user 56409 chooses a phonetic transcription.

A similar difference can be found when comparing the “-er” suffix user 63150 used in “Zimmer” (“room”), which follows the standard German transcription. The same suffix is transcribed as <o> by user 56409, for example in “vokiahlt” and “gesto” (standard: “verkühlt”, “gestern”; “have a cold”, “yesterday”). This is very typical for written dialect realization in Pustertal and can be found in many other postings from this region as well (see example posting (2)).

Lastly, both postings contain nominalized infinitive verb constructions. User 63150 writes “zum Wandern” (“to hike”) and “zun Arbetn” (“to work”), user 56409 “zi scham” (“to be ashamed”). “Zum Wandern” completely overlaps with the standard German form, especially, as even the bilabial nasal <m> is transcribed to mark the third case. This is remarkable as in this position the bilabial nasal <m> is changed to the alveolar variant <n> in Southern Bavarian dialects (see e.g. Pichler-Stainern 2008: 285). This is visible in the second infinitive construction of user 63150 – “zun Arbetn”. Thus, also in this case user 63150 fluctuates between a standard-oriented and a dialectal transcription. User 56409, on the other hand, again chooses a very standard-deviant, phonetical form – “zi scham”. “zi” already shows a conservative vowel pattern, preserving the Old High German <i> in the infinitive marker (Haspelmath 1989). Moreover, “zi” as well as “scham” show no case marking at all. In case of “scham” this is most probably because of a progressive nasal assimilation effect, where “schamen” is articulated as “scham”.

The only element where user 63150 seems to be more dialect-oriented is “letschte”, where he/she overtly expresses the s-retraction. User 56409 never does so as can be seen in case of “stimme” and “gesto”.

This last example shows that both users fluctuate between the two poles standard and dialect. However, they cover different areas of the spectrum: The transcribing habit of the urban user is much closer to the standard, while the rural user tries to stay closer to the phonetic peculiarities of his vernacular. However, it must be said that this is most probably not only an effect of writing habits. As mentioned in the introduction, in South Tyrol’s main city Bozen exists a dialect form heavily oriented towards standard German, which is seen as a more prestigious variety. This implies that speakers of this dialect also adapt their oral expression accordingly.

In this case the standard-orientation in the written domain would be superseded by the standard-orientation in the oral domain, meaning that the transcription is indeed a phonetical one, just that the phonetic realizations are already close to the standard. This interfering effect is hard to differentiate, but it can be assumed that the written domain definitely amplifies standard-orientation.

4.2.6 Word classification – conclusions

```
Code files: 10_update_words_labels.ipynb
```

RoBERTa had good performance rates on the validation data, while overall Accuracy dropped by 0.07 on the MoCoDa test set. The model managed to classify dialect and neutral words appropriately in most cases. This was not the case for non-dialect words. It was argued that one reason for this performance loss is the fact that the probability distribution over classes significantly differed across datasets. However, our analysis also revealed that another reason lies in the equivocal nature of the neutral category. This is due to several factors: There are tokens which are potentially assignable to dialect, standard German or the neutral category depending on context. In these cases, RoBERTa often failed to assign the context-appropriate label, which is what caused most of the mistakes between the dialect and neutral classes. Mistakes regarding the non-dialect and neutral classes, however, were often due to a procedure-dependent, incorrect ambiguity. This ambiguity was introduced by the automatic word labeling algorithm by assigning standard German words to the neutral class. This class overlap between standard German and neutral words can be seen as the main factor for the significant performance drop on the test set regarding the non-dialect class. Nevertheless, we could also show that it is indeed this ambiguity that makes the neutral class so interesting for linguistic analyses and the evaluation of writing habits between the two concurring poles standard and dialect.

The final step of Task 2 involved integrating the labels to the corpus. Even though the analysis showed that the model output is not flawless, the Accuracy score of over 0.915 on the test set still justifies the decision to add the dialect labels on word basis to the corpus. Thus, in a last step, we let the model predict the dialect tags for all tokens contained in DiDi and added this information to the corpus. In this way 367572 dialect tags were added. Most of these tokens, i.e., 169720, were assigned to the non-dialect class, followed by 126091 words in the neutral class, 62779 dialect and 8982 unknown tokens. In chapter 2.3 we estimated that, given that

about a fifth of all tokens in the corpus has a normalization assigned, also the percentage of dialect tokens would be in this range. The sum of 62779 dialect tokens corresponds to 17 percent of all tokens. This shows that the normalization information is a good approximation of the overall sum of dialect tokens, but also includes other tokens. Thus, the dialect tags we added to the corpus now enable a more fine-grained assessment of written dialect introducing another useful dimension to the resource, which is potentially interesting for answering various research questions. The chapters 5.2 and 5.3 in the next part of this thesis will exemplify two of these potential use-cases for the developed detection systems and the enhanced DiDi corpus. The first chapter of the following Part V will discuss possible improvements.

PART V – Future Perspectives

5.1 Improvements and extensions

5.1.1 Task 1 – Posting classification system

In Task 1 XLM-RoBERTa and DeBERTa were primarily trained to disambiguate undefined postings and to assign these postings to one of two other categories. This classification task can be termed “dialect detection” only to a limited extent. The dataset consisted of mainly German postings with only few tokens originating from other languages. Also, the variation within German was mainly limited to the opposition of South Tyrolean dialects versus standard German. For this reason, the division into dialect, non-dialect and undefined postings sufficed. If we would use the developed system “in the wild”, though, for example to filter out more dialect sentences from social media sites, the system would return very noisy results (see chapter 5.3).

The reason is that language data on the web is much more diverse. The system would meet languages and other dialects it was not trained for. In theory, these instances should be subsumed under the undefined class. However, there are multiple problems with this: Firstly, we do not know how the system generalizes to varieties never seen during training time. It is unlikely that it would label all instances written in unknown languages as undefined. This would result in a high number of False Positives labeled as dialect or non-dialect. Even more so as the undefined category also contains examples of mixed dialect and non-dialect words. This was intended in the task at hand, but is counter-productive in a less controlled data situation. Thus, to be able to apply the first system as a real dialect detection system, a first step would be to introduce more fine-grained labels that replace the undefined category.

One approach could be to use a “mixed” label for cases where dialect is mixed with other varieties and “unknown” for cases which are neither dialect, nor non-dialect or mixed. However, it would be even more interesting to build a full-scale detection system that would exploit existing language detection models and just add South Tyrolean as an own variety. In this way False Positive responses could be minimized when collecting new dialect data. This would open up new challenges, for example, to discriminate Tyrolean dialects from closely related written dialects like Carinthian. Another problem when adding new varieties to a pre-

trained detection system is catastrophic forgetting (Kemker et al. 2018): Neural nets tend to "forget" tasks they were trained on when trained on new tasks. A possible solution is to include various languages in the training data when re-training the network for dialect detection. This would also provide the chance to train DeBERTa on different languages. As described in chapter 4.1.4, DeBERTa most likely suffered from its mono-lingual pre-training, which led to less meaningful class representations. Nevertheless, DeBERTa's performance rates were even higher than RoBERTa's. Thus, a combination of DeBERTa's architecture with a multi-lingual training procedure could lead to an improved detection rate and a system that is applicable to real life data gathering.

From a data- or corpus-perspective it could be beneficial to apply both systems to all postings in the corpus and to check for cases where one or both systems disagree with the provided gold label. The reason is that we still found wrongly labeled instances in the corpus. For example the posting in (13).

(13) STD na du brauchsch 3 jahre Berufserfahrung nach dem Diplom
SG nein, du brauchst 3 Jahre Berufserfahrung nach dem Diplom
EN no you need 3 years of working experience after your diploma
(id: 56969_t0934_m03583)

The posting in (13) is labeled as "de non-dialect", even though it clearly contains the dialect words "na" ("nein", "no") and "brauchsch" ("brauchst", "need"). The label "de undef" or maybe even "de dialect" would be more appropriate for such a mixed case.

5.1.2 Word classification system

The word classification system could as well benefit from a multi-lingual training procedure. This would enable RoBERTa to build a better representation of the unknown class, which was introduced to capture foreign (not German) or intransparent words. Moreover, instead of using one label for all foreign languages, it could be beneficial to allow for multiple language tags – a tag for each language included in the training dataset. A further improvement from data perspective involves an improvement of the labeling algorithm. As we argued, the neutral class suffered from too much overlap between standard and neutral words resulting from the automatic labeling procedure. One improvement option here would be to use another

probabilistic classification framework, for example a Naïve Bayes classifier, or another non-linear classification algorithm (Random Forest, Support Vector Machine) on n-gram basis.

Moreover, the analysis of the neutral label has shown that it has great potential for linguistic analyses, but that its meaningfulness is limited by the fact that also emojis, numbers and punctuation symbol were labeled as neutral. The reason was that we conceptualized neutral as being used universally in both standard and dialect. However, tokens like emojis and numbers in dialect postings are universal, because of extra-linguistic reasons (simplified, universal facial expressions; universal numerical system). This is not true if a word appears in standard spelling in a dialect posting as the present work has shown. An additional label to distinguish between these token classes, thus, would be a further improvement.

Finally, another potential improvement is to exploit transfer learning between Task 1 and Task 2. As we described in chapter 4.2.3. we did not use the weights of the RoBERTa model fine-tuned on posting classification to instantiate word-classification RoBERTa. We argued that this could lead to undesirable biases. However, it is possible that classification on the word-level could benefit from pre-training on a higher level. This would involve a re-training of the Task 1 RoBERTa model with a different classification layer.

Chapter 5.1 has shown that all models presented in this topic as well as the revised DiDi corpus could be further improved. Nonetheless, they already are a useful resource, as the following chapters 5.2 and 5.3 will show.

5.2 Use case 1: DiDi for socio-linguistic dialectology

`Code files: 15_convert_corpus_to_csv_for_user_analysis.ipynb, 16_r_code_user_analysis.R`

This and the next chapter's main purpose is to show possible use cases for the implemented model as well as the resulting enhanced DiDi corpus. The present chapter will elaborate on a socio-linguistic analysis based on the dialect word labels added to DiDi.

As became clear in chapter 2.3, the DiDi project collected various socio-demographic variables of the users contained in the corpus (sex, age, education, occupation, dialect region etc.). Our goal was to investigate the relationship between dialectality and some of these socio-demographic variables. To do so, we first had to find an appropriate operationalization for the dependent variable. We decided on a simple approach, i.e., to assess dialectality by the ratio of dialect tokens a user contributed to DiDi and his/her overall sum of tokens:

$$dialectality = \frac{\sum_{j=1}^{l} dialect\_tokens_u}{\sum_{i=1}^{n} all\_tokens_u}$$

In this way, dialectality was defined as the proportion of dialect words in the overall output of a user. The independent variables tested for their effect on the dependent variable were "age", "sex/gender", "graduation" and "dialect region". With this setup we wanted to test the following four hypotheses:

1. Age: In chapter 1.4.3 we described the model of immediacy and distance by Koch & Oesterreicher as well as that young Swiss adolescents tend to use a more phonetically oriented spelling to signal in-group proximity. Frey & Glaznieks (2018) as well as Glaznieks & Frey (2018) come to the same conclusion by analyzing DiDi data. Thus, we expect to find a similar age effect when using the previously defined dialectality metric.

   Hypothesis I: The higher the users' age the lower their measured dialectality should be.
2. Sex/Gender: The relationship between biological sex and culturally determined gender is complex. In the DiDi corpus the variable is specified as "Geschlecht", a term which spans both concepts in German. When naming the corpus variables, we decided to use the variable name "sex". However, in a socio-linguistic context it is more appropriate to assume gender as an underlying concept. This concept is also what Hansen (2012: 56) refers to when she speaks of "soziales Geschlecht". With regard to this variable she observes that especially young women[28] tend to a more standard-oriented expression in their oral communication. We wanted to test this for the written domain which leads to the second hypothesis.

   Hypothesis II: Women have a lower degree of dialectality, i.e., are more normative in their dialect realization.
3. Graduation: Hansen (2012) also assumes that occupation might be a relevant factor for the degree of dialectality. She hypothesizes that people in communication-oriented jobs often have a higher level of education and frequently use the standard variety in a

[28] The areas under investigation were in the Alemannic border region in the South-West of Germany.

working context. However, her model does not confirm any relationship between occupation and the degree of dialectality. We would like to test a similar hypothesis in this thesis by directly using level of education as independent variable, which in Hansen's model was only indirectly considered by looking at occupation. In line with Hansen's findings in the oral domain, we formulate Hypothesis 3.

Hypothesis III: The level of education does not influence the degree of dialectality.

4. Dialect region: In chapter 4.2.5 we showed that the analysis of the neutral class reveals interesting patterns in dialect writing habits. It was shown that especially users in the main city Bozen seem to have a tendency to use a more standard-oriented dialect writing, while users in the valleys Sarntal and Pustertal show a very phonetical written dialect realization. We wanted to test this observation empirically, which led to Hypothesis 4.

   Hypothesis IV: The dialect area of a user has an effect on his dialectality. Users in Bozen have a lower degree of dialectality, users in Pustertal and Sarntal a higher degree.

To test the hypotheses, we performed a multivariate linear regression analysis. The regression results are displayed in Table 32: Linear regression results.

Characteristic	Beta[1]	SE[2]	p-value
dataf$sex_w	-5.0	2.75	0.070
dataf$age	-0.59***	0.087	<0.001
dataf$`binned_graduation_Low graduation`	7.9*	3.68	0.035
dataf$`binned_graduation_High graduation`	-6.0	3.15	0.059
dataf$german_dialect_region_Bozen	-1.6	3.20	0.6
dataf$german_dialect_region_Pustertal	11**	3.99	0.007
dataf$german_dialect_region_Sarntal	21**	6.83	0.003
R^2	0.406		
Adjusted R^2	0.368		

[1] *p<0.05; **p<0.01; ***p<0.001; ****p<0

[2] SE = Standard Error

Table 32: Linear regression results

Starting with general model specifications our model reached an R^2 value of 0.4 and an adjusted R^2 value of 0.368. This means that the independent variables can explain

approximately 40% of the dependent variable's variance, which is a comparatively high explanatory power in the social sciences domain, especially as the number of predictors is relatively small.

Hypothesis I can be confirmed. Age has a significant effect on dialectality. Following the model, with increasing age, the proportion of dialect tokens drops by a factor of 0.59% per year. This is not meant in the sense of a longitudinal analysis of one person, but only regards the writing habits of the observed user groups. It is not a statement about personal development, but rather a snapshot of the tendencies in the observed community. Consequently, the following holds true: Younger South Tyrolean users are more dialectal in their written expression on Facebook than older users.

Hypothesis II, which regarded the influence of sex/gender on dialectal writing, can also be confirmed: The negative beta factor indicates that being a woman has a negative effect on the degree of expressed dialectality. Being a woman lowers the ratio of dialect tokens versus other tokens by 5%. However, the respective p-value of 0.07 lies slightly above the pre-defined 0.05 margin and indicates only weak evidence for the hypothesis.

To be able to verify Hypothesis III we first classified the users' level of education (LoE) into three categories: low, middle and high graduation LoE[29] and added low and high LoE to the model. Users with middle LoE in this way are indirectly considered by neither having a low nor a high level of education. The model results suggest that Hypothesis 3 cannot be confirmed: While Hansen (2012) did not find an effect caused by occupation – and indirectly education – on dialectality, graduation clearly influences dialectality in our analysis. Considering the p-values this is effect significant for lower levels of education leading to a higher dialectality in written expression. The reverse effect can be observed for a higher educational level, even though the higher p-value of 0.059 suggests that this finding needs more support for final verification.

Hypothesis IV can be partially confirmed. The influence of living in a rural area like Pustertal and Sarntal on an increased dialectality is highly significant. Especially users from Pustertal show a high dialectality in their writing which coincides with the user analysis in chapter 4.2.5 On the other side, users from the capital Bozen show a reduced dialectality. However, the

[29] please refer to APPENDIX E to see how the types of graduation were assigned to categories

p-value of 0.6 indicates that this result could be mere chance and that this trend cannot be conclusively verified.

To summarize, four of seven variables were found to be (highly) significant in the explanation of dialectality. Age was most significant, followed by living in Sarntal or Pustertal, followed by having a low level of education. Having a high LoE as well as female gender were less significant factors. Living in Bozen provided insufficient evidence. The latter factors, thus, need more testing for verification.

To guarantee the validity of the presented results, we performed various statistical tests to check if the presumptions of linear regression hold true (see APPENDIX E).

5.3 Use case 2: RoBERTa for dialectometry

`Code files: 12_Twitter_filter.ipynb, 13_filtered_Twitter_user_anonymization.ipynb`

The previous chapter shed light on how the revised DiDi corpus can be used to verify various sociolinguistic hypotheses for the written dialect in South Tyrol. In the present chapter, we will show how RoBERTa can be applied to gather new data relevant for dialectological research.

As a first step we searched for a good online resource where dialectal contributions can be found. We found that Facebook is indeed a very good resource for South Tyrolean dialect postings. However, Facebook's Terms and Conditions used to prohibit scraping of their pages[30], which is also why the DiDi project used a specially designed web app. By contrast, collecting Twitter data is made possible by the company through an in-house API, which is extensively used by researchers in various fields (for an overview of the legal situation see Kamocki et al. 2022). Therefore, we decided to test RoBERTa on Twitter data. To do so, we first restricted the search space locally by using a 100x100 km square window around the geographical center of South Tyrol (latitude: 46.6594674, longitude: 11.4353947[31]). We then scraped all Tweets in this area posted on Twitter in the time period from September 2010 to January 2019. The result was a dataset with nearly two million collected tweets in various languages. Using RoBERTa for word classification we were able to reduce this number to about 100.000 tweets by taking only tweets with at least 4 predicted dialect tokens. Then, we applied the second system, i.e.,

[30] https://www.facebook.com/legal/terms/previous

[31] https://www.klausen.it/en/enjoyment-region/villanders/sights/the-centre-of-south-tyrol.html

RoBERTa for posting classification. By filtering out tweets which did not receive the “de dialect” label by this system, we got a final result of 23229 tweets which we had to check manually. This dataset was still very noisy, and even contained tweets in languages with another writing system like Russian and Arabic. After the manual revision, the dataset contained 809 tweets which were actually written in (South) Tyrolean dialects. The sequential application of both systems thus resulted in an Accuracy of 3.5%. This shows that the systems still need significant improvement to work in uncontrolled settings.

The 809 tweets were written by 111 different users and come with various meta information. Among others, all of them are GPS geo-tagged. By entering this GPS data into an open-source software, we created the map shown in Figure 14.

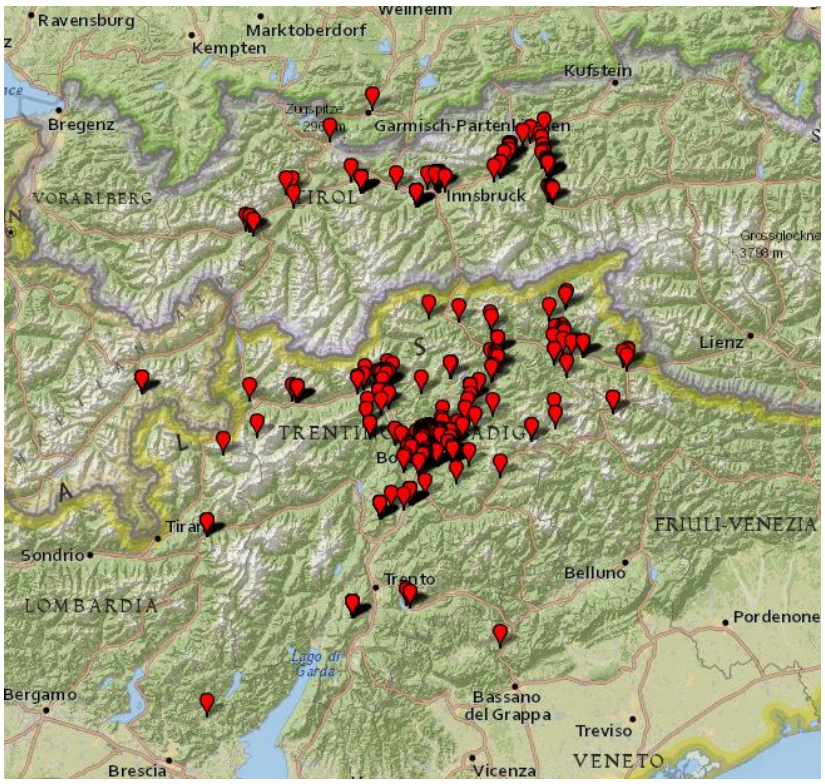

Figure 14: Locations of found dialect tweets

Figure 14 shows that the majority of the found tweets is indeed located in South Tyrol, especially in the region of the capital of Bozen approximately in the center of the map. The rest of the points is mainly found in the federal province of Tyrol in the North. This can have two potential explanations: As there is increased mobility between the two regions there might be South Tyrolean users tweeting in (North) Tyrol. Additionally, as described in the introduction, the two parts are historically and linguistically closely linked, which is why also dialect writing

realizations and practices could strongly overlap. This brings us to another point: Geo-located spontaneously written dialect data bears great potential for dialectology and, more specifically, dialectometry.

The field of dialectometry has evolved in the last decades and tries to identify supra-regional linguistic patterns and isoglosses by extracting and analyzing feature sets of linguistic atlases (see Goebl 2010, Wieling & Nerbonne 2015 for an overview of the field). The compilation of the underlying atlases, however, requires expensive field work where specially trained interviewers have to find and interview reliable subjects, fluent in a sought vernacular. Moreover, only a pre-fixed number of features can be inquired with the used questionnaires. Thus, fieldwork has the advantage of controlled and consistently comparable, relatively clean data but comes with the expense of great time and cost efforts. Geo-referenced spontaneously written dialect data, on the other hand, is less controlled for unwanted biases, but is easily accessible.

As Purschke & Hovy (2019) have shown, it is possible to largely reproduce the traditional dialectological division of German-speaking Europe with this kind of data. The authors used geo-located data from the social media app "Jodel" and encoded this data using neural representation learning. The resulting vectors were grouped using agglomerative clustering with cosine similarity as distance metric. In this way, linguistically related regions such as Bavaria and Austria, Northern German areas and regions in Switzerland were automatically clustered together, even though the authors only performed minimal data pre-processing and cleaning and did not exclude medium-specific terms or languages other than German. Compared with their data the data we collected is already relatively homogenous. Thus, a similar approach as Purschke & Hovy have chosen, might lead to promising and scientifically interesting results. However, more data might be necessary to apply their techniques as the mobility of users otherwise leads to unwanted influences on the regional representation. Plotting just one user, for example, results in the map in Figure 15.

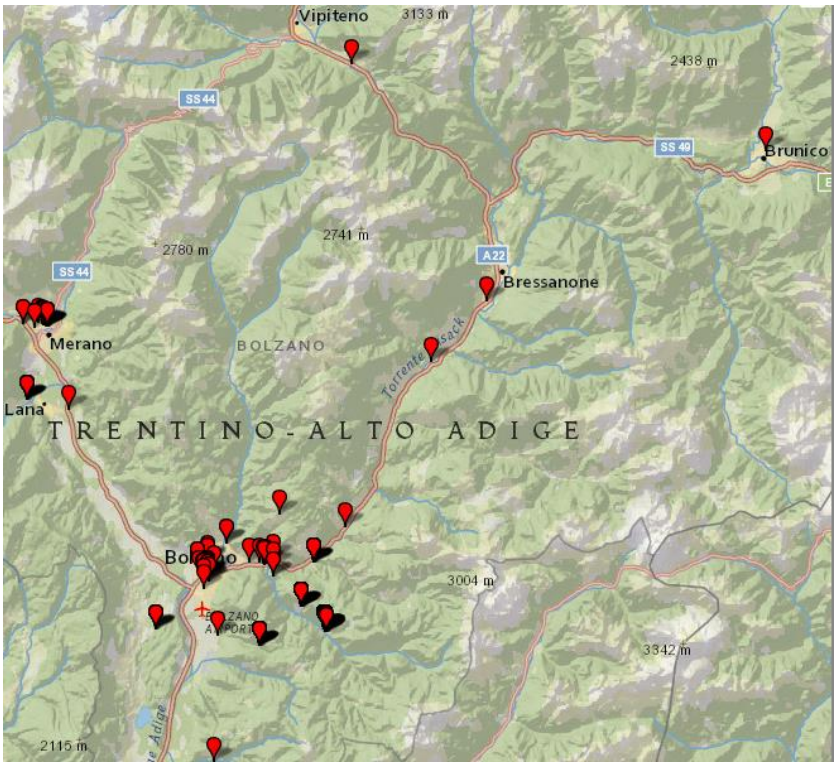

Figure 15: Locations of one collected user

As can be seen, the user moves around South Tyrol and tweeted at least once in the neighborhood of the region's bigger cities Meran/Merano, Bozen/Bolzano, Brixen/Bressanone, Sterzing/Vipiteno and Bruneck/Brunico. Even though a center of gravity is discernible around Bozen/Bolzano and Meran/Merano, Figure 15 is a good example why more data is necessary to level out mobility effects. Nevertheless, we are confident that the approach suggested in Purschke & Hovy (2019) is also transferable to written dialect data and can provide insights into the modern dialectal structure of the Tyrolean area as well.

PART VI – Final considerations

6.1 Ethical implications

Before concluding this thesis, we would like to address some ethical implications arising from the use of neural networks in a classification task. One issue results from the general nature of classification: In a classification task we try to assign observations to discrete categories. But as shown in chapter 1.4, discrete classes actually do not exist for the task at hand. Dialect and standard are just imaginary poles of a continuum that vary along multiple dimensions with diaphasic, diastratic, diatopic and diachronic factors. Classification, thus, is an artificial simplification made for analytical purposes. How closely this simplification mirrors reality is a matter of debate and individual perception. In the analysis of neutral samples in chapter 4.2.5 we already emphasized the inherent ambiguity of the neutral class. But the problem of vagueness to a certain degree affects all classes. Depending on an annotator's background and label definitions she/he could disagree with many of the choices made by the author during the labeling process. For example, an annotator from Pustertal could deny that a word like "ghab" as analyzed in chapter 4.2.5 can count as dialectal form, if it clearly lacks the typical <o> instead of <a> transcription. Or it could be debated if a word like "gehn" in a standard German sentence can really be labeled as non-dialect, because it clearly contains an orthographical variation (canonical form: "gehen") also met in dialect writing.

What makes this discussion even more complex, is the fact that for the labeling of standard postings and words we used the tag "non-dialect", as given in the DiDi corpus. However, it is surely debatable how this term relates to the definition of standard. In this thesis "non-dialect" was conceptualized as "intended standard" – a form of writtenness that aims for an orthographically correct, non-dialect-oriented representation, but also contains spellings that diverge from the norm, for example because of CMC-related phenomena. We tried to establish a consistent annotation process and label definitions. However, we are aware that our definitions might not completely overlap with other conceptualizations of the used tags.

Nevertheless, we have shown in this thesis, that our simplification of reality was appropriate to solve the tasks at hand. The derived labels were also sufficient to confirm some hypotheses in the socio-linguistic literature about specific socio-demographic user variables and their relationship to dialectality. Therefore, they can be used to determine what is more or less

dialectal in the context of the DiDi corpus. However, it would be invalid to use the trained networks as well as the resulting labels, to generally decide what counts as South Tyrolean dialect realization and what as standard German. The models were trained on 136 dialect writers, which is only a minimal proportion of the relevant population. Consequently, the models have only seen a very small subset of the relevant data and can contain systematic biases. Some of them were addressed and analyzed in the chapters 4.1.3 and 4.2.4. But what we also tried to make clear is that it is hard to unveil all biases of a neural network model.

This brings us to the last point: The property that we do not know what a neural net has learnt model-internally, is a matter of intensive debate in the research community and has led to the establishment of a whole new field of research, i.e., "Explainable AI" (see, e.g., Buhrmester et al. 2021; Vollmer et al. 2020). The lack of transparency underlines the fact that the application of deep learning does not and cannot replace human judgement, if we do not want to introduce systematic and potentially harmful biases to our data. To sum up this short discussion, neural networks are useful tools for automating processes, like the labeling of thousands of words for this thesis. But their output still requires human supervision and evaluation.

6.2 Summary and conclusion

The present thesis aimed to enhance the DiDi corpus – a German dialect/standard corpus gathered on Facebook – on two levels:

1. In a first step, samples with undefined variety were re-labeled semi-automatically. We trained two Transformer models, DeBERTa and XLM-RoBERTa, on a hand-labeled subset of undefined samples. These models were used to re-classify the remaining undefined instances as dialect, standard German or, indeed, undefined. In this way, we could reduce the number of such ambiguous samples in the corpus by over 75%. Subsequently, to prove the effectiveness of the implemented systems we re-labeled all undefined samples in the corpus by hand and compared this complete set of gold labels to the predicted labels. This resulted in an Accuracy score of over 80%. Even though an error rate of 1 out of 5 still shows the need for improvement, the error analysis could demonstrate that many disagreements between human annotation and implemented systems were due to the inherent ambiguity of the undefined class and that many mistakes were debatable. Thus, a fully automatized re-labeling procedure would have been feasible. A qualitative analysis of model results showed that especially XLM-

RoBERTa, profiting from its multi-lingual pre-training, developed a functioning representation of classes.

2. The second aim of this thesis involved the generation of word-specific dialect/standard labels. In the original DiDi corpus variety is only specified on the level of whole postings. We wanted to transfer this knowledge to the individual token level. We achieved this goal by first using the revised DiDi corpus to derive a task-specific training set: The dialect tags on the posting level were used to determine if a word's context is dialectal or standard-oriented. This preliminary assumption about the potential variety of a word was further refined by looking at the normalization information, the lemma and a special etymological tag, which were already given in the corpus. Additionally, a probabilistic classifier was integrated into the word labeling algorithm.

   We used the resulting data set to train a XLM-RoBERTa-based model to assign the tags "dialect", "non-dialect", "neutral" or "unknown" to each token in the DiDi corpus. The system reached an Accuracy of 98% on the validation set and 92% on the test set. These scores encouraged us to add the 367572 generated variety tags on the word level to DiDi. As a test set we used South Tyrolean dialect data collected by the MoCoDa database project. This data was hand-labeled to provide a meaningful gold standard.

By performing the above steps, the present thesis contributed to a qualitative improvement of two German written dialect corpora. By applying neural nets it was possible to introduce a more fine-grained distinction between postings and words of different varieties and to transfer this knowledge across resources. The automatized approach most certainly introduced incorrect dialect tags to DiDi as well. However, an analysis of socio-linguistic hypotheses based on the new tags could replicate known relationships between socio-demographic factors and dialectality. We interpret this as a proof for the meaningfulness of the created labels.

This is also underlined by the fact that, by using the implemented systems as a filter for Twitter content, we were able to detect over 800 Tyrolean dialect samples in a dataset of two million regional tweets. As this dataset is still rather small, future research could attempt to improve the systems' detection rate and to gather more spontaneous dialect data online. Purschke & Hovy (2019) have shown the great potential of such automatically gathered resources for the detection of isoglosses and the creation of linguistic maps. It is our vision to analyze Tyrolean varieties with the same approach and in this way to contribute to the

understanding of the regional partition and shifts within Tyrolean and German dialects in general.

References

Abel, Andrea, Stefanie Anstein & Stefanos Petrakis. 2009. Die Initiative Korpus Südtirol. *Linguistik Online* 38, https://bop.unibe.ch/linguistik-online/article/view/502.

Alammar, Jay. 2018. The Illustrated Transformer. (3 April, 2023.)

Alber, Birgit. 2020. *Linguistik des Deutschen, kompakt und kontrastiv.* Verona: Edizione QuiEdit.

Alber, Birgit, Jennifer-Carmen Frey, Aivars Glaznieks, Alexander Glück & Joachim Kokkelmans. publication in preparation. Zum Verhältnis von geschriebenem und gesprochenem Dialekt in WhatsApp-Nachrichten aus Südtirol.

Aschwanden, Brigitte. 2001. Wär wot chätä?' Zum Sprachverhalten deutschschweizerischer Chatter. *Networx Nr. 24* 24, https://www.repo.uni-hannover.de/handle/123456789/2942.

ASTAT. 2001. *South Tyrol in Figures*. Provincial Statistics Institute of South Tyrol.

ASTAT. 2011. *Südtirol in Zahlen*. Provincial Statistics Institute of South Tyrol. https://astat.provinz.bz.it/downloads/jb_2011.pdf. (8 March, 2023.)

ASTAT. 2014. *Demografisches Handbuch Südtirol*. Autonome Provinz Bozen-Südtirol. https://astat.provinz.bz.it/de/aktuelles-publikationen-info.asp?news_action=4&news_article_id=482175. (23 March, 2023.)

Autonome Provinz Bozen. 2022. Autonomie für drei Sprachgruppen. https://autonomie.provinz.bz.it/de/autonomie-fur-drei-sprachgruppen. (6 January, 2023.)

Bahdanau, Dzmitry, Kyunghyun Cho & Yoshua Bengio. 2014. *Neural Machine Translation by Jointly Learning to Align and Translate*. https://arxiv.org/pdf/1409.0473.

Baumgartner, Barbara & Sigrid Hechensteiner. 2022. Wir haben eine Bringschuld: Gespräche zwischen Disziplinen - Der Jurist Marc Röggla und der Asienexperte Günther Cologna im Interview. https://www.eurac.edu/de/magazine/wir-haben-eine-bringschuld. (7 March, 2023.)

Beißwenger, Michael, Marcel Fladrich, Wolfgang Imo & Evelyn Ziegler. 2020. Die Mobile Communication Database 2 (MoCoDa 2). In Konstanze Marx, Henning Lobin & Axel Schmidt (eds.), *Deutsch in Sozialen Medien. Interaktiv–multimodal–vielfältig*, 349–352. de Gruyter.

Brandner, Ellen. 2015. Syntax des Alemannischen (SynAlm). Tiefenbohrungen in einer Dialektlandschaft. *Regionale Variation des Deutschen. Projekte und Perspektiven*. 289–322.

Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever & Dario Amodei. 2020. *Language Models are Few-Shot Learners*. https://arxiv.org/pdf/2005.14165.

Buhrmester, Vanessa, David Münch & Michael Arens. 2021. Analysis of explainers of black box deep neural networks for computer vision: A survey. *Machine Learning and Knowledge Extraction* 3(4). 966–989.

Caswell, Isaac, Theresa Breiner, Daan van Esch & Ankur Bapna. 2020. Language ID in the wild: Unexpected challenges on the path to a thousand-language web text corpus. *arXiv preprint arXiv:2010.14571*.

Conneau, Alexis, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer & Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee & Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dürscheid, Christa & Elisabeth Stark. 2011. sms4science: An International Corpus-Based Texting Project and the Specific Challenges for Multilingual Switzerland. In Crispin Thurlow & Kristine R. Mroczek (eds.), *Digital discourse: Language in the new media* (Oxford studies in sociolinguistics). Oxford, New York: Oxford University Press.

Eberhard, David M., Gary F. Simons & Charles D. Fennig. 2022. *Ethnologue: Languages of the World*, 25th edn. Dallas, Texas: SIL International.

Felder, Samuel. 2015. Korpusgestützte Analyse der Verschriftung des Schweizerdeutschen in SMS. *Networx Nr. 70* 70.

Ferguson, Charles A. 1959. Diglossia. *WORD* 15(2). 325–340.

Frey, Jennifer-Carmen & Aivars Glaznieks. 2018. The Myth of the Digital Native: Analysing language use of different generations on Facebook. *University of Antwerp*. 41–44.

Frey, Jennifer-Carmen, Aivars Glaznieks & Egon W. Stemle (eds.). 2016. *The DiDi Corpus of South Tyrolean CMC Data: A multilingual corpus of Facebook texts*.

Glaznieks, Aivars & Jennifer-Carmen Frey. 2018. Dialekt Als Norm? In Arne Ziegler (ed.), *Jugendsprachen/Youth Languages*, 859–890. de Gruyter.

Glaznieks, Aivars & Alexander Glück. 2019. From the valleys to the World Wide Web: Non-standard spellings on social network sites. In Ciara Wigham & Egon W. Stemle (eds.), *Building computer-mediated communication corpora for socio-linguistic analysis* (Cahiers du Laboratoire de recherche sur le langage 8), 21–45. Presses universitaires Blaise-Pascal.

Glaznieks, Aivars & Egon W. Stemle. 2014. Challenges of building a CMC corpus for analyzing writer's style by age: The DiDi project. *Journal for language technology and computational linguistics* 29(2). 31–57.

Goebl, Hans. 2010. Dialectometry and quantitative mapping. In *Language and Space*, 433–464. Mouton de Gruyter.

Haas, Walter. 1992. Mundart und Standardsprache in der deutschen Schweiz. In *Dialect and Standard Language in the English, Dutch, German and Norwegian Areas*, 312–336. Amsterdam: Nederlandse Akademie van Wetenschappen, Verhandelingen, Afd. Letterkunde, Nieuwe Reeks.

Hansen, Sandra. 2012. Dialektalität, Dialektwissen und Hyperdialektalität aus soziolinguistischer Perspektive. In Sandra Hansen, Christian Schwarz, Philipp Stoeckle & Tobias Streck (eds.), *Dialectological and folk dialectological concepts of space: Current methods and perspectives in sociolinguistic research on dialect change*, 48–74. de Gruyter.

Haspelmath, Martin. 1989. From Purposive To Infinitive—A Universal Path Of Grammaticization. *Folia linguistica historica* 23(Historica-vol-10-1-2). 287–310.

He, Pengcheng, Xiaodong Liu, Jianfeng Gao & Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Hochreiter, Sepp & Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9(8). 1735–1780.

Huber, Christian, Benjamin Fischer & Bernhard Graf. 2019. Corpus of Austrian dialect recordings from the 20th century: A cooperation project. In *Proceedings of the Third International Workshop on the History of Speech Communication Research Vienna*.

Huber, Judith & Christian Schwarz. 2017. SMS-Kommunikation im mehrsprachigen Raum. Schriftsprachliche Variation deutschsprachiger SMS-Nutzer/-innen in Südtirol. *Networx Nr. 70*.

Kamocki, Paweł, Vanessa Hannesschläger, Esther Hoorn, Aleksei Kelli, Marc Kupietz, Krister Lindén & Andrius Puksas (eds.). 2022. *Legal issues related to the use of Twitter data in language research.* Linköping University Electronic Press.

Kemker, Ronald, Marc McClure, Angelina Abitino, Tyler Hayes & Christopher Kanan. 2018. Measuring Catastrophic Forgetting in Neural Networks. *Proceedings of the AAAI Conference on Artificial Intelligence* 32, https://ojs.aaai.org/index.php/AAAI/article/view/11651.

Koch, Peter & Wulf Oesterreicher. 1985. Sprache der Nähe - Sprache der Distanz: Mündlichkeit und Schriftlichkeit im Spannungsfeld von Sprachtheorie und Sprachgeschichte. In Olaf Deutschmann, Hans Flasche, Bernhard König, Margot Kruse, Walter Pabst & Wolf-Dieter Stempel (eds.), *Romanistisches Jahrbuch* (36). Berlin, New York: Walter de Gruyter.

Koch, Peter & Wulf Oesterreicher. 2012. Language of immediacy-language of distance theory and linguistic history: Orality and literacy from the perspective of language.

Kolde, Gottfried. 1981. *Sprachkontakte in gemischtsprachigen Städten: Vergleichende Untersuchungen über Voraussetzungen und Formen sprachlicher Interaktion verschiedensprachiger Jugendlicher in den Schweizer Städten Biel/Bienne und Fribourg/Freiburg i. Ue.* (37). Wiesbaden: Steiner.

Kranzmayer, Eberhard. 1956. *Historische Lautgeographie des gesamtbairischen Dialektraumes.* Wien: Böhlau.

Krizhevsky, Alex, Ilya Sutskever & Geoffrey E. Hinton. 2017. ImageNet classification with deep convolutional neural networks. *Communications of the ACM* 60(6). 84–90.

Lanthaler, Franz. 1974. Systemverändernde Tendenzen in der Mundart des Passeiertales am Beispiel einiger Verbformen. In *Der Schlern* (48), 469–474. Bozen: Athesia.

Lanthaler, Franz. 1990. Dialekt und Zweisprachigkeit in Südtirol. In Franz Lanthaler (ed.), *Mehr als eine Sprache – Più di una lingua: Zu einer Sprachstrategie für Südtirol*. Alpha&Beta.

Lanthaler, Franz. 1996. Varietäten des Deutschen in Südtirol. In Gerhard Stickel (ed.), *Varietäten des Deutschen: Regional- und Umgangssprachen*. IdS-Jahrbuch 1996, 364–383. Berlin – New York: de Gruyter.

Lanthaler, Franz. 2001. Zwischenregister der deutschen Sprache in Südtirol. In Franz Lanthaler & Kurt Egger (eds.), *Die deutsche Sprache in Südtirol: Einheitssprache und regionale Vielfalt*, 137–152. Bozen: Folio Verlag.

Lanthaler, Franz. 2007. The German Language in South Tyrol: Some Sociolinguistic Aspects. In Andrea Abel (ed.), *Aspects of Multilingualism in European Border Regions: Insights and Views from Alsace, Eastern Macedonia and Thrace, the Lublin Voivodeship and South Tyrol*, 220–235. Bozen/Bolzano.

Lanthaler, Franz. 2022. Das Präfix der- im Dialekt des Passeiertales. https://lanthaler.net/artikel/praefix-der-1/. (28 March, 2023.)

Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer & Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Luong, Minh-Thang, Hieu Pham & Christopher D. Manning. 2015. *Effective Approaches to Attention-based Neural Machine Translation*. https://arxiv.org/pdf/1508.04025.

Marx, Konstanze, Henning Lobin & Axel Schmidt (eds.). 2020. *Deutsch in Sozialen Medien.* de Gruyter.

Meraner, Rudolf & Monika Oberhofer. 1982. Zur Mundart in Tirol. In Kurt Egger (ed.), *Dialekt und Hochsprache in Südtirol: Beiträge zum Deutschunterricht in Südtirol*. Bozen.

Mikolov, Tomas, Kai Chen, Greg Corrado & Jeffrey Dean. 2013. *Efficient Estimation of Word Representations in Vector Space*. https://arxiv.org/pdf/1301.3781.

Müller, Christina M. 2011. Dialektverschriftung im Spannungsfeld zwischen standardnah und lautnah: Ein korpuslinguistische Untersuchung der Rubrik "Dein SMS" in der Aargauer Zeitung. In Helen Christen, Franz Patocka & Evelyn Ziegler (eds.), *Struktur, Gebrauch und Wahrnehmung von Dialekt: Beiträge zum 3. Kongress der Internationalen Gesellschaft für Dialektologie des Deutschen*, 155–178. Zürich: praesens.

Paul, Katharina, Maik Thalmann, Markus Steinbach & Marco Coniglio. 2022. Gehen as a new auxiliary in German. *Language Change at the Interfaces: Intrasentential and intersentential Phenomena*. 165–188.

Peng, Yifan, Shankai Yan & Zhiyong Lu. 2019. *Transfer Learning in Biomedical Natural Language Processing: An Evaluation of BERT and ELMo on Ten Benchmarking Datasets*. https://arxiv.org/pdf/1906.05474.

Peters, Matthew E., Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee & Luke Zettlemoyer. 2018. *Deep contextualized word representations*. https://arxiv.org/pdf/1802.05365.

Pichler-Stainern, Arnulf. 2008. *Südbairisch in Laut und Schrift.* Klagenfurt: Verlag Johannes Heyn.

Plüss, Michel, Manuela Hürlimann, Marc Cuny, Alla Stöckli, Nikolaos Kapotis, Julia Hartmann, Malgorzata A. Ulasik, Christian Scheller, Yanick Schraner, Amit Jain, Jan Deriu, Mark Cieliebak & Manfred Vogel. 2022. SDS-200: A Swiss German Speech to Standard German Text Corpus.

Purschke, Christoph & Dirk Hovy. 2019. Lörres, Möppes, and the Swiss - (Re)Discovering regional patterns in anonymous social media data. *Journal of Linguistic Geography* 7(2). 113–134.

Qiao, Yanhua, Xiaolei Zhu & Haipeng Gong. 2022. BERT-Kcr: prediction of lysine crotonylation sites by a transfer learning method with pre-trained BERT models. *Bioinformatics* 38(3). 648–654.

Samardzic, Tanja, Yves Scherrer & Elvira Glaser. 2016. ArchiMob: A Corpus of Spoken Swiss German. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. European Language Resources Association (ELRA).

Schmidt, Jürgen E. 2011. Formation of and change in regiolects and (regional) dialects in German. *Taal & Tongval* 63(1). 143–174.

Schmidt, Thomas. 2017. DGD: Die Datenbank für Gesprochenes Deutsch. *Zeitschrift für germanistische Linguistik* 45(3). 451–463.

Schumacher, Stefan. 1998. Sprachliche Gemeinsamkeiten zwischen Rätisch und Etruskisch. In *Der Schlern* (72), vol. 72, 90–114. Bozen: Athesia.

Sonnenhauser, Barbara. 2009. Zur der-Präfigierung im Bairischen. In Lenka Scholze & Björn Wiemer (eds.), *Von Zuständen, Dynamik und Veränderung bei Pygmäen und Giganten: Festschrift für Walter Breu zu seinem 60. Geburtstag* (Diversitas Linguarum), vol. 25. Bochum: Dr. N. Brockmeyer.

Srivastava, Nitish, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever & Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research* 15(1). 1929–1958.

Staab, Steffen. 2020. *Bayes' Classification* (Machine Learning).

Steininger, Rolf. 1997. *Südtirol im 20. Jahrhundert: Vom Leben und Überleben einer Minderheit*, 4th edn. Innsbruck-Wien: Studienverlag.

Tessadri, Wolfgang. 2017. *Die Entstehung des bairischen Verbalpräfixes der- und dessen synchrone Verwendung in den südbairischen Dialekten Südtirols*. Konstanz: Universität Konstanz Bachelor-Arbeit.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser & Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30.

Verheijen, Lieke. 2017. WhatsApp with social media slang: Youth language use in Dutch written computer-mediated communication.

Vollmer, Sebastian, Bilal A. Mateen, Gergo Bohner, Franz J. Király, Rayid Ghani, Pall Jonsson, Sarah Cumbers, Adrian Jonas, Katherine S. L. McAllister & Puja Myles. 2020. Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics, and effectiveness. *British Medical Journal* 368.

Vu, Thang. 2019. *Deep learning for speech & language processing: RNN* (Deep Learning). Stuttgart.

Wieling, Martijn & John Nerbonne. 2015. Advances in dialectometry. *Annu. Rev. Linguist.* 1(1). 243–264.

Wiesinger, Peter. 2014. *Das österreichische Deutsch in Gegenwart und Geschichte.* Wien: Lit-Verlag.

Wiesinger, Peter & Albrecht Greule. 2019. *Baiern und Romanen: Zum Verhältnis der frühmittelalterlichen Ethnien aus der Sicht der Sprachwissenschaft und Namenforschung.* Tübingen: Narr Francke Attempto Verlag.

List of Figures

List of Tables

APPENDICES

APPENDIX A

User Meta

Variable	Description	Values appearing in the corpus
Completed Date	Not specified in DiDi documentation	dates in the range from 14.03.2014 to 10.05.2014
User ID	Id of user	unique ID in the range from 54616 to 57311
Inboxpermission	Permission to access FB messages	{0, 1}
LanguageCode	Not specified in DiDi documentation	{'de', 'it_IT'}
Locale	Not specified in DiDi documentation	{'de_DE', 'en_GB', 'en_US', 'fr_FR', 'it_IT'}
PA_Alter	Age of user in the year 2013	ages in the range from 14 to 76
PA_Ausbildungsabschluss	Graduation	{'Berufsausbildung (Lehre)', 'Fachhochschulabschluss, abgeschlossene Ausbildung an einer Meisterschule bzw. Berufs- oder Fachakademie', 'Fachoberschule mit Matura', 'Hochschulabschluss', 'Oberschule 3.Klasse', 'Oberschule mit Matura', 'Pflichtschulabschluss', 'einen anderen Abschluss', 'no data', 'ohne Pflichtschulabschluss', 'technische Fachausbildung nach Matura'}
PA_Beruf	Occupation	{'ArbeitnehmerIn', 'FreiberuflerIn', 'RentnerIn, PensionistIn', 'SchuelerIn', 'StudentIn', 'nicht erwerbstätig / arbeitslos', 'no data'}
PA_Dialektsprecher_STIR	Specification of the South Tyrolean dialect region if the participant's L1 is German	{'Bozen', 'Eisacktal', 'Meran', 'Pustertal', 'Sarntal', 'Ueberetsch-Unterland', 'Vinschgau', 'kein Dialekt', 'no data'}
PA_Dialektsprecher_ITA	Specification of the Italian dialect region if the participant's L1 is Italian	{'kein Dialekt', 'no data', 'trentino'}
PA_Geburtsjahr	Year of Birth	Birthyears in the range from 1937 to 1999
PA_Geschlecht	Sex	{'m', 'w'}
PA_Lebensmittelpunkt_STIR	Information whether center of life is in South Tyrol	{'0', '1'}

PA_L1_andere	Specification of L1 if participant has an L1 other than German, Italian or Ladin	{'0', 'en', 'no'}
PA_L1_Deutsch	Specifies if L1 is German	{'1'}
PA_L1 Italienisch	Information whether L1 is (also) Italian	{'0', '1'}
PA_L1_Ladinisch	Information whether L1 is (also) Ladin	{'0', '1'}
PA_Schule	Specification of school type if participant is still pupil	{'Fachoberschule oder berufsbildende Oberschule: z.B. Handelsoberschule, Oberschule für Geometer', 'Oberschule: z.B. Humanistisches Gymnasium, Realgymnasium, Pädagogisches Gymnasium', 'no data'}
Postpermission	Permission to access FB posts	{0, 1}
RespondentID	Id of user	unique ID in the range from 54616 to 57311
StartDate	Not specified in DiDi documentation	dates in the range from 14.03.2014 to 10.05.2014
ZA_Frequenz_Blog	Specification of frequency of writing blogs	{'min. 1x pro Monat', 'min. 1x pro Tag', 'min. 1x pro Woche', 'nie', 'no data'},
ZA_Frequenz_Email	Specification of frequency of writing emails	{'min. 1x pro Monat', 'min. 1x pro Tag', 'min. 1x pro Woche', 'nie', 'no data'},
ZA_Frequenz_Foren	Specification of frequency of writing in forums	{'min. 1x pro Monat', 'min. 1x pro Tag', 'min. 1x pro Woche', 'nie', 'no data'},
ZA_Frequenz_IM	Specification of frequency of writing instant messages	{'min. 1x pro Monat', 'min. 1x pro Tag', 'min. 1x pro Woche', 'nie', 'no data'},
ZA_Frequenz_Mikroblog	Specifiction of frequency of writing microblogs	{'min. 1x pro Monat', 'min. 1x pro Tag', 'min. 1x pro Woche', 'nie', 'no data'},
ZA_Frequenz_SNS	Specifiction of frequency of writing on social network sites	{'min. 1x pro Monat', 'min. 1x pro Tag', 'min. 1x pro Woche', 'nie', 'no data'},
ZA_Geraete_PC	Does user use PC or notebook for surfing the internet	{'0', '1', 'no data'}
ZA_Geraete_Smartphone	Does user use Smartphone for surfing the internet	{'0', '1', 'no data'}
ZA_Geraete_Spielkonsole	Does user use gaming console for surfing the internet	{'0', '1', 'no data'}

ZA_Gerate_Tablet	Does user use table for surfing the internet	{'0', '1', 'no data'}
ZA_Internetnutzung_Ausbildung	Specification of occasion when the internet is used: education	{'0', '1', 'no data'}
ZA_Internetnutzung_Arbeit	Specification of occasion when the internet is used: work	{'0', '1', 'no data'}
ZA_Internetnutzung_Dauer	Specification of the period of active internet usage	Hours in the range from 0 to 19 + no data
ZA_Internetnutzung_Freizeit	Specification of occasion when the internet is used: free time	{'0', '1', 'no data'}
ZA_Internetnutzung_Information	Specification of occasion when the internet is used: information	{'0', '1', 'no data'}
ZA_Internetnutzung_Kalenderjahr	Specification since when user is using the internet	Years in the range from "before 1995" to 2013 + no data
ZA_Internetnutzung_Kommunikation	Specification of occasion when the internet is used: communication	{'0', '1', 'no data'}
ZA_Internetnutzung_Spiel	Specification of occasion when the internet is used: gaming	{'0', '1', 'no data'}
ZA_Internetnutzung_Unterhaltung	Specification of occasion when the internet is used: entertainment	{'0', '1', 'no data'}
ZA_Internetnutzung_Haushalt	Specification of the year in which the user’s house was connected to the internetfirst time	Years in the range from 'before 1995' to 2013 + no data
ZA_Sprachen_Blog	Specification of languages used on blogs	{'Deutsch_Dialekt', 'Deutsch_standardsprache', 'Englisch', 'Italienisch_standardsprache'}
ZA_Sprachen_Email	Specification of languages used for E-Mails	{'Deutsch_Dialekt', 'Deutsch_standardsprache', 'Englisch', 'Italienisch_Dialekt', 'Italienisch_standardsprache', 'Ladinisch', 'Spanisch', 'andere'}

ZA_Sprachen_Foren	Specification of languages used on forums	{'Deutsch_Dialekt', 'Deutsch_standardsprache', 'Englisch', 'Italienisch_standardsprache', 'Ladinisch'}
ZA_Sprachen_IM	Specification of languages used for instant messages	{'Deutsch_Dialekt', 'Deutsch_standardsprache', 'Englisch', 'Italienisch_Dialekt', 'Italienisch_standardsprache', 'Ladinisch', 'andere'}
ZA_Sprachen_Mikroblog	Specification of languages used on microblogs	{'Deutsch_Dialekt', 'Deutsch_standardsprache', 'Englisch', 'Italienisch_standardsprache', 'andere'}
ZA_Sprachen_SNS	Specification of languages used on social network sites	{'Deutsch_Dialekt', 'Deutsch_standardsprache', 'Englisch', 'Italienisch_Dialekt', 'Italienisch_standardsprache', 'Ladinisch', 'andere'}

Comment Meta

Variable	**Description**	**Values appearing in the corpus**
dialect	Semi-automatically assigned variety of the written text	{'de dialect', 'de non-dialect', 'de undef'}
comment_id	Id of the posting	unique comment id in the format '57149_px0863_c5550'
language_corrected_langs	Corrected language label of the main language	{'de', 'en', 'it', 'es', 'fr', 'la', 'other'}
post	Id of post comment pertains to	unique post id in the format '57085_p03697'
language_langid	Language as assigned by the identification system	{'af', 'am', 'az', 'br', 'ca', 'cs', 'cy', 'da', 'de', 'en', 'eo', 'es', 'et', 'eu', 'fi', 'fo', 'fr', 'ga', 'hu', 'id', 'it', 'la', 'lb', 'lt', 'mg', 'mt', 'nb', 'nl', 'nn', 'no', 'oc', 'pl', 'pt', 'ro', 'se', 'sl', 'sv', 'tr', 'ug', 'vo', 'zu'}
language_tag	Final language label of the main language	{'de'}
like_count	Number of likes the post has	like counts in the range of 0 to 14

created_time	Time the text was created	time specifications in the format '2013-01-11T11:55:44+0000'
user_id	Id of the posting's author	User ids in the range from 54616 to 57311
message	Full text of the post	full text as a string
language_langid_confidence	Confidence score as given by the identification system	confidence scores in the range of 0 to 1
fb_text_type	Specifies if it is a comment, post or message	{'comments'}
newlines	Count of newlines	count of newlines in the range 1 to 348

Message Meta

Variable	Description	Values appearing in the corpus
dialect	Semi-automatically assigned variety of the written text	{'de dialect', 'de non-dialect', 'de undef'}
message_id	Id of the posting	unique comment id in the format '56411_t0088_m02456'
language_corrected_langs	Corrected language label of the main language	{'de', 'en', 'it', 'es', 'fr', 'other'}
thread	Id of thread messasge pertains to	unique thread id in the format 't0080'
language_langid	Language as assigned by the identification system	{'xh', 'ms', 'sv', 'ga', 'vo', 'af', 'an', 'mt', 'rw', 'et', 'se', 'pl', 'id', 'en', 'lb', 'de', 'pt', 'ht', 'si', 'sl', 'sw', 'la', 'sk', 'ro', 'cs', 'ug', 'br', 'no', 'nb', 'oc', 'eu', 'lt', 'tr', 'hu', 'it', 'eo', 'nn', 'ar', 'zu', 'gl', 'da', 'ca', 'sq', 'nl', 'ps', 'he', 'hr', 'es', 'fi', 'fo', 'cy', 'tl', 'fr', 'mg'}
language_tag	Final language label of the main language	{'de'}
created_time,	Time the text was created	time specifications in the format '2013-12-19T20:42:01+0000'
user_id	Id of the posting's author	User ids in the range from 54616 to 57285
message	Full text of the post	full text as a string

language_langid_confidence	Confidence score as given by the identification system	confidence scores in the range of 0 to 1
fb_text_type	Specifies if it is a comment, post or message	{'messages'}
newlines	Count of newlines	count of newlines in the range 0 to 1105

Post Meta

Variable	Description	Values appearing in the corpus
dialect	Semi-automatically assigned variety of the written text	{'de dialect', 'de non-dialect', 'de undef', 'non-de'}
language_corrected_langs	Corrected language label of the main language	{'de', 'en', 'es', 'it', 'fr', 'pt', 'la', 'other'}
attachment_link	True if post contains link	{'False', 'True'}
attachment_picture	True if post contains picture	{'False', 'True'}
language_manually_annotated	Language tag manually revised	{'n.a.'}
application_namespace	Application with which the post was created (mobile, website)	{'solitaireblitz', 'mfg_at_app', 'change-org', 'handelszeitung', 'frasiitaliano', 'bildqgtv', 'fbipad_', 'ilmeteodellumore', 'guestions', 'yt-fb-app', 'selectivetwitter', 'europartytown', 'fb_blackberryten', 'sueddeutsche', 'twitter', 'hipstamatic', 'fbandroid', 'barfussmagazin', 'test_mobile', 'video', 'avaaz-org', 'diewelthd', 'dolomitenraetsel', 'eventfotos', 'blingee', 'linksalpha', 'smilesforyou', 'get-spotify', 'business_xing', 'likes', 'suchtpravention-sauf', 'runtastic', 'photosi_general', 'fanpageit', 'og_vimeo', 'bbplatform', 'n.a.', 'appme_chat', 'fbiframes_one', 'indiegogo', 'ifthisthenthat', 'gsocialize', 'dropboxdropbox', 'fbiphone', 'friendshugs', 'regalaunarosa', 'soundcloud'}
likes	Like count	like counts in the range of 0 to 25
created_time	Time the text was created	time specifications in the format '2013-06-22T19:04:31+0000'
message	Full text of the post	full text as a string

language_langid_confidence	Confidence score as given by the identification system	confidence scores in the range of 0 to 1
fb_text_type	Specifies if it is a comment, post or message	{'posts'}
updated_time	Last revision point	time specifications in the format '2013-06-22T19:04:31+0000'
newlines	Count of newlines	count of newlines in the range 0 to 2844
post_id	Id of the posting	unique post id in the format '57090_p06583'
user_id	Id of the posting's author	User ids in the range from 54616 to 57311
language_langid	Language as assigned by the identification system	{'xh', 'ms', 'sv', 'jv', 'vo', 'af', 'mt', 'et', 'se', 'pl', 'id', 'en', 'lb', 'de', 'pt', 'la', 'sk', 'ro', 'am', 'br', 'no', 'nb', 'eu', 'lt', 'tr', 'hu', 'it', 'ko', 'ku', 'eo', 'nn', 'da', 'ca', 'sq', 'nl', 'lv', 'hr', 'es', 'fi', 'cy', 'tl', 'fr', 'sl'}
language_tag	Final language label of the main language	{'de'}
comments	Number of comments under post	count of comments under post in the range from 0 to 26
shares	Number of times post was shared	count of shares in the range from 0 to 132
application_name	Unambiguous marker for the application	{'Facebook for Windows Phone', 'Mobile', 'Frasi', 'Sauftirol / Alcol Adige?', 'XING', 'Bild QGTV', 'Fanpage', 'Hipstamatic', 'Pages', 'Fanpage App 1', 'Og_likes', 'Notes', 'PhotoSì', 'Solitaire Blitz', 'Il Meteo dell Umore', 'Links', 'mitfahrgelegenheit.at', 'Hotgags.net', 'Status', 'Facebook for Every Phone', 'Rätsel der Dolomiten', '@Hugs', 'Events', 'SoundCloud', 'mobileblog', 'Windows Phone', 'Photos', 'BlackBerry', 'Runtastic.com', 'Party Town', 'SZENE1 - Da ist die Party!', 'Blingee Cards, Birthdays & More!', 'Likes', 'Facebook Exporter for iPhoto', 'Facebook for iPad', 'AppMe', 'Video', 'Commentarist', '@Smiles', 'Vimeo', 'Change.org', 'Widget Share Log App', 'Regala una rosa', 'iPhoto ', 'BlackBerry Smartphones App', 'minddrive', 'Handelszeitung.ch', 'Gigya Socialize', 'Selective Tweets', 'Twitter', 'Dropbox', 'Spotify', 'Questions', 'OS X', 'barfuss.it', 'Share_bookmarklet',

		'Indiegogo', 'Facebook for iPhone', 'n.a.', '"Welt HD"-iPad-App der "Welt"', 'Facebook for BlackBerry 10', 'iOS', 'YouTube', 'LinksAlpha.com', 'Avaaz.org', 'Süddeutsche.de', 'IFTTT', 'Facebook for Android'}
icon	Not specified in documentation	{'False', 'True'}
attachment_type	Type of linked element (picture, video etc.)	{'checkin', 'link', 'photo', 'status', 'swf', 'video'}
status_type	Type of post	{'added_photos', 'added_video', 'app_created_story', 'mobile_status_update', 'n.a.', 'shared_story'}

Token Meta

Variable	Description	Values appearing in the corpus
comment	Space for comments if needed	Token specific comments by the corpus authors in string format
pos	Part-of-speech tag of the token	{'VAIMP', '$(', 'ADJD', 'FM', 'TRUNC', '$,', 'KOUI', 'PWAT', 'APZR', 'KON', 'PWAV', 'PTKVZ', 'NN', 'PDS', 'PTKANT', 'PPOSS', 'PRELS', 'PRELAT', 'VAINF', '$.', 'NE', 'VMFIN', 'VMINF', 'PTKZU', 'VAPP', 'XY', 'VVINF', 'KOKOM', 'APPO', 'VVFIN', 'PRF', 'ADV', 'VVPP', 'CARD', 'VAFIN', 'PPOSAT', 'VVIMP', 'PWS', 'VMPP',

		'ADJ', 'ADJA', 'PAV', 'PTKNEG', 'PIAT', 'ITJ', 'PDAT', 'PPER', 'ART', 'PIS', 'APPR', 'PTKA', 'VVIZU', 'APPRART', 'KOUS'}
lemma	Lemma of the token	Lemma as a string
token	String of the token	Token as a string
anonym	Specifies if the token is anonymized	{", '<InstNE>-Weihnachtsfeier', '<InstNE>-schülerInnen', '<PersNE>und', '<InstNE>-Besuch', '<mail>', '<InstNE>-Fans', '<GeoADJA>reise', '<GruppeNN>-Delegation', '<InstNE>-Zeltlager', '<GruppeNN>fete', '<PersNE>-bär', '<InstNE>-Gala', '<PersNE>-Ordnung', '<GeoNE>tour', '@<InstNE>', '\ufeff<InstNE>', "<PersNE>'", '<InstNE>server', '<GeoNE>-brugg', "s'<GeoNE>", '<GeoADJA>Fuaßbollplotz', '<InstNE>Zeit', '<link>', '<InstNE>party', '<PersNE>kuschlobnt', '<PersNE>-Vorfahren', '<GeoNE>oder', '<GeoNE>', '<InstNE>fest', '<tel>', '<GeoNE>aufentholt', '<InstNE>-Kirche', '<PersNE>', '<GeoNE>-city', '<GeoADJA>gegend', 'hihi-<PersNE>', '<GeoNE>-Aufenthalt', '<GeoNE>-<GeoNE>', '<InstNE>gala', '<InstNE>-Erasmusbüro', '<GeoADJA>', '<PersNE>-', '<InstNE>-Ausbildnerin', '<InstNE>parkplatz', '<GeoNE>/<InstNE>', '<InstNE>ausbildung', '<PersNE>-Humor', '<InstNE>-Beschriftung', '<GeoNE>wanderung', '<GruppeNN>ausbildung', '<GeoADJA>bergen', '<GeoADJA>-Schotz', '@<GeoNE>', '<GruppeNN>', "<GeoNE>'s", '<PersNE>-Woche', '<PersNE>s', '<InstNE>-Mailserver', '<InstNE>-Webseite', '-<GeoNE>', '<InstNE>-Webserver', '<GeoNE>-City', '<InstNE>-adresse', '<GeoNE>s', '@<PersNE>', '<InstNE>-Frühstück', '<GruppeNN>organisation', '<InstNE>-Server', '<GeoNE>tal', '<GeoNE>-Brugg', '<InstNE>', '<PersNE>für', '<GeoNE>-<GeoNE>-<GeoNE>', '<InstNE>-Ausbildung', '<PersNE>-Witz', '<XXX>', '<InstNE>-Damen', '@<GruppeNN>', "<PersNE>'s", '<GeoNE>-Klettersteig', '<PersNE>mäher', '<GruppeNN>kollegin', '<InstNE>-Chef', '@<GeoNe>', '/<GeoNE>/'}
norm	Normalization in Standard German	Normalized form as a string

stir	Specifies if the token is genuinely South Tyrolean	{'True', True, 'n.a.'}
at	Specifies if token is @-mention	{'at_group', 'at_organisation', 'at_person', 'at_place', 'at_topic', 'n.a.'}
iter_graph	Does the token contain repetitions of graphemes	{'True', 'n.a.'}
at_topic	Specifies if token is @-mention of topic	{True, 'n.a.'}
emoticon	Specifies if token is an emoticon	{'True', True, 'n.a.'}
at_organisation	Specifies if token is @-mention of organisation	{True, 'n.a.'}
at_place	Specifies if token is @-mention of place	{True, 'n.a.'}
anonym_category	Specifies category of anonymization	{'firstname', 'initial', 'lastname', 'n.a.', 'nick'}
iter_emoji	Does the token contain repetitions of emojis	{'True', 'n.a.'}
anonym_gender	Specifies of anonymization regards gender	{'', 'f', 'female', 'm', 'male', 'n.a.'}
hyperlink	Specifies if token is a hyperlink	{True, 'n.a.'}
at_group	Specifies if token is @-mention of a group	{True, 'n.a.'}
language	Specifies the language of the token	{'de', 'en', 'es', 'it', 'n.a.', 'pt'}
acronym	Specifies if token is an acronym or cmc-related abbreviation	{'', True, 'n.a.'}
hashtag	Specifies if token is a hashtag	{False, True, 'n.a.'}
iter_emoticon	Does the token contain repetitions of emoticons	{'True', 'n.a.'}
at_person	Specifies if token is @-mention of a person	{True, 'n.a.'}
emoji	Specifies if token is an emoji	{'True', True, 'n.a.'}
iter_punct	Does the token contain repetitions of punctuation characters	{'True', 'n.a.'}

APPENDIX B

Genuinely South Tyrolean words in the DiDi corpus (tag: “stir”)

{'dorrichtets', 'heipinggl', 'knittl', 'Nutz', 'dorricht', 'feli', 'alm', 'Dertuas', 'kamott', 'dofrogg', 'en', 'darpocks', 'lättn', 'sebm', 'ament', 'dotun', 'Hoi', 'MITTOGSRASCHTERL', 'Hem', 'OLM', 'spack', 'em', 'tschopnn', 'hm', 'disel', 'drrichtn', 'Törggelen', 'Hel', 'Grottn', 'hoi', 'drreissts', 'haxn', 'wischen', 'Flotto', 'Wollei', 'tschari', 'keks', 'säääääften', 'rear', 'Gitsch', 'Ooooolm', 'assi', 'hell', 'Pfiete', 'walsch', 'röötn', 'hintrisch', 'schleinen', 'Sem', 'Poppile', 'Dr', 'dortuaschs', 'LAUSGITSCH', 'ingaling', 'LEI', 'gitsch', 'Pollo', 'strialen', 'drfetzn', 'lai', 'dr', 'dohupft', 'schiachste', 'ala', 'PFIAT', 'detun', 'leibelen', 'Schiarcher', 'di', 'walesch', 'teifln', 'hwm', 'Grante', 'la', 'gikinst', 'infahl', 'letzt', 'letzen', 'hoihoi', 'Letzer', 'lamma', 'lausgpitsch', 'HEL', 'Lai', 'sirig', 'Marende', 'tatschen', 'Lepskörbl', 'säften', 'letzer', 'spackt', 'gmerglt', 'kastl', 'rern', 'tochtl', 'tratzen', 'doricht', 'klumpert', 'maren', 'oschti', 'dortun', 'marenessn', 'ollm', 'gneatig', 'hetz', 'pfiatiii', 'fock', 'leimer', 'nutz', 'HOI', 'krautwalsche', 'olbn', 'unkeksn', 'Multifunktionshirnkastl', 'hal', 'se', 'Hetz', 'dertun', 'drricht', 'le', 'derrichtn', 'zemm', 'Perloggn', 'holmitog', 'laimer', 'gissilar', 'Glasler', 'oooolm', 'zem', 'postkastl', 'pfieti', 'laimear', 'derkuglen', 'lei', 'sel', 'der', 'nagglt', 'Watten', 'Gitschn', 'sell', 'Touta', 'toerggelen', 'stollerhausn', 'gemaungelet', 'frustsäften', 'Di', 'dertuasch', 'kinzen', 'Dor', 'sierig', 'schiachn', 'Ollm', 'giahn', 'dorgeben', 'letzes', 'ente', 'dortua', 'A', 'sacheler', 'fra', 'welchtig', 'Foller', 'hem', 'Sell', 'Sel', 'stuff', 'sl', 'La', 'pfiati', 'derpocki', 'Pfieti', 'ketz', 'olbm', 'drtuas', 'walscher', 'pfinsta', 'dafragt', 'ingekentet', 'Klunz', 'drkeil', 'seggo', 'letza', 'TOTA', 'ingekenntet', 'letzr', 'domochsch', 'holmittog', 'derrichten', 'schliafn', 'Oschpele', 'kamotte', 'leimor', 'Stoller', 'Grausbiren', 'olm', 'letz', 'Hoiiii', 'puff', 'polla', 'oschpele', 'dorfrog', 'dorrichtn', 'Haita', 'kommenlei', 'Geheimfachl', 'felli', 'Zulln', 'marende', 'SEM', 'Puff', 'lolli', 'datuas', 'Pingl', 'schnöllor', 'enten', 'getakkelt', 'Hell', 'ummrgmerglt', 'suser', 'die', 'dopockat', 'tschorgg', 'doladn', 'gitschn', 'siri', 'Ospele', 'pinggl', 'derrichts', 'dofrog', 'wem', 'feih', 'Hoihoi', 'törgelen', 'Olm', 'Knitl', 'stoller', 'Haxn', 'hel', 'poppi', 'Poppa', 'amerst', 'schlein', 'Pfiatenk', 'ement', 'Gitschen', 'rearen', 'briafkastl', 'mittogsraschterl', 'schleimen', 'Zem', 'hoila', 'dorrichtet', 'törggelen', 'entn', 'in', 'Seggo', 'dopockn', 'oschpela', 'Lei', 'zadertian', 'ospele', 'leibele', 'tschoppn', 'Ratscher', 'Hax', 'nutza', 'SELL', 'schluzig', 'lafa', 'pfiatiiii', 'fuganzn', 'doschifft', 'Groggln', 'Ho', 'tschopp', 'SCHLEIN', 'drschifft', 'semm', 'schintr', 'mittogs', 'dem', 'fuder', 'Säwarguat', 'In', 'ANTRISCH', 'giriern', 'Letzn', 'allm', 'zulln', 'pfiatiiiii', 'schlatzn', 'sota', 'Plent', 'hetzig', 'Scherm', 'Tata', 'Popile', 'Se', 'Hoila', 'stirgen', 'glasl', 'Gschoff', 'panporgo',

'engelebengele', 'pfiat', 'drrichtet', 'a', 'OLLM', 'hl', 'Poppo', 'DELLOSCHTIA', 'getutscht', 'kuppelet', 'dorgaling', 'dopockt', 'bock', 'schiech', 'dertian', 'Heipinggl', 'Butzies', 'im', 'Hutsche', 'Nel', 'dopock', 'gian', 'ausnbexn', 'derrichtesch', 'fockelotti', 'zm', 'Stickln', 'SIRI', 'opm', 'lri', 'sem', 'drtian', 'Tota', 'Marend', 'dersel', 'frei', 'tata', 'tatti', 'Poppele', 'raschterle', 'dertua', 'Haxe', 'gutzelts', 'hemm', 'aipalandern', 'zomderstellt', 'ingalign', 'säftn', 'Die', 'letzn', 'watter', 'dor', 'gschamig', 'lettig', 'rasterle', 'aposchto', 'derricht', 'gach', 'letze', 'Hirnkastl', 'Glasl', 'alben', 'Pfiati', 'olben', 'seeem', 'Letz', 'seletwegen'}

APPENDIX C

Richtlinien für die (Re-)Annotation von Beiträgen des DiDi-Korpus

Zielsetzung: Letztendlicher Anwendungsfall des Korpus ist das Training neuronaler Netze. Die nachfolgenden Richtlinien ergeben sich aus dieser Zielsetzung.

1. Standard und Dialekt:

- Beiträge, die eine Mischform standard- und Dialektwörtern beinhalten, werden insofern als „de dialect“ eingestuft als die Länge der zu Dialektwörtern gehörenden Zeichen die der zu Fremdwörtern gehörenden übersteigt bzw. die vermeintlichen standard-Wörter auch im dialektalen Schreibmodus so ausgedrückt werden würden. Bei 50/50 muss von Fall zu Fall entschieden werden.
- Ein-Wort-Beiträge werden nur dann als „de dialect“ oder „de non-dialect“ gekennzeichnet, wenn sie unzweifelhaft als dialektal oder eben nicht analysiert werden können.
- Beiträge, die grundsätzlich im standard verfasst sind, die jedoch einzelne Elemente enthalten, die Dialekt sein könnten, die aber auch auf Schreibfehler zurückzuführen sein könnten (z. B. „bessr“), werden als „de non-dialect“ gekennzeichnet.
- Beiträge, die keinen Rückschluss darauf zulassen, ob sie nun Dialekt oder standard sind, da die Zeichenketten beidem zugeordnet werden könnten, werden als „de undef“ gekennzeichnet.
- Wörter, die verhashtaggt sind, werden als eigenständige Wörter betrachtet und je nach Fall als Dialekt oder standard bzw. als „dialect“ oder „non-dialect“ gekennzeichnet.

Wichtiges Kriterium bei der Zuordnung von standard- und Dialekt-Labeln ist der Vokalismus. Umlaute, wie sich im deutschen standard gebräuchlich sind, werden im Dialekt (in den meisten Varietäten – Ausnahme Sarntal) nativ nicht verwendet.

Ist der Beitrag zwar tendenziell im standard ausgedrückt, enthält aber zahlreiche Fehler, so wird er als „de undef“ gekennzeichnet (z. B. „ja nauterlich bringe ihn wieder im Proberaum... danke!“)

2. Dialekt, Fremdwörter und Internetphänomene:

- Beiträge, die eine Mischform aus Fremd- und Dialektwörtern beinhalten, werden insofern als „de dialect“ eingestuft als die Länge der zu Dialektwörtern gehörenden Zeichen die der zu Fremdwörtern gehörenden signifikant übersteigt.
- Beiträge, die anderen deutschen Dialekten zugeordnet werden können bzw. nur dialektal eingefärbt sind, werden als „de undef“ eingeordnet.
- Ein-Wort-Beiträge werden nur dann als „de dialect“ gekennzeichnet, wenn sie unzweifelhaft als dialektal analysiert werden können und nicht nur Wortmaterial anderer Varietäten dominiert.
- Wörter, die verhashtaggt sind, werden als eigenständige Wörter betrachtet und je nach dem als Dialekt oder nicht bzw. als „de dialect“ oder „de undef“ gekennzeichnet.
- Beiträge, die Links enthalten, werden unter Nicht-Beachtung des Links gekennzeichnet.

3. Anonymisierungen:

- Um einen Beitrag mit Anonymisierungen einer Kategorie zuordnen zu können, muss ohne Zweifel klar sein, welcher Varietät er angehört.

Die beschriebenen Prinzipien gelten ebenso für das Labeln von Wörtern für Task 2.

APPENDIX D

Introduction to evaluation metrics

The evaluation metrics used to assess model performance in this thesis are: Precision, Recall, F_1 and Accuracy. In the following we will give a short introduction into these metrics. Precision, Recall and F1 score are interdependent measures defined as follows:

$$Precision = \frac{TP}{TP+FP} \qquad Recall = \frac{TP}{TP+FN} \qquad F_1 = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

TP stands for "True Positives", which is the number of overlapping labels between given gold standard and the automatically predicted labels. For the tasks at hand, this would mean: A certain posting is labeled as "dialect" and the system predicts "dialect". FN stands for "False Negatives". In case of False Negatives we expect a certain label to be predicted, but it is not predicted. This would apply if our gold standard would specify a posting as "dialect", but it is labeled as undefined by our system. Lastly, FP stands for "False Positives" and is the reverse case: Based on our gold standard we do not expect a label to be predicted, but is predicted nevertheless. We could, for example, have a posting that is labeled as undefined in our gold standard but is predicted to be "dialect" nevertheless. As becomes clear from these considerations what counts as True Positive, False Negative and False Positive is a matter of perspective and depends on the label we are currently looking at. This shall be illustrated by the following example: If a posting is labeled as "dialect" in the gold standard, but predicted as "non-dialect" by our system, it is a False Negative considering all true dialect labels and a False Positive considering all true non-dialect labels. Thus, TPs, FPs and FNs have to be counted for each label individually. The following descriptions will provide an intuition of the three metrics we can calculate using TPs, FPs and FNs: Precision, Recall and F_1 score.

Given a certain label, Precision is the number of True Positives divided by the sum of True Positives and False Positives. Thus, if we have, for example, 500 samples predicted as "dialect", Precision tells us how often this prediction was correct. Recall, on the other hand, is the number of True Positives divided by the sum of True Positives and False Negatives. If, for example, we have 500 samples in our gold standard labeled as "dialect", Recall would tell us how many of them our system actually recognized as being "dialect". The F_1 metric is just the harmonic mean of Precision and Recall.

An intuitive example of Precision and Recall would be the following: Imagine, we have a pond with 10 salmons and 20 trouts in it. A fisherman, who does not like to eat trouts, wants to fish only salmons. When he starts to fish, Precision will tell us, how often he fished out a salmon when he takes out a fish. If, for example, he catches 10 fish, 8 being trouts and 2 being salmons, this would correspond to a salmon Precision of 25%. Recall, on the other hand, will tell us how many of the salmons in the pond he caught already. Sticking to the example above, with 2 of 10 salmons caught, this corresponds to a salmon Recall of 20%. The salmon F_1 score would, then, be $(2 \times 0.25 \times 0.2) \div (0.25 + 0.2) = 0.22$.

As became clear, TPs, FPs and FNs are label-specific. Thus, also Precision, Recall and F_1 score can be computed per label. Imagine a second fisherman sitting at the pond and having the same fishing result, who prefers trout over salmon. His trout Precision would be 80%, Trout Recall 40% and F_1 53,33%.

In a second step, to get a metric of model performance over all categories, we can calculate different sorts of averages of Precision, Recall and F_1. In this thesis two sort of averages were used: macro average and weighted average. Macro average is computed by simply averaging over all Precision, Recall or F_1 values. The macro average Precision for the fish example would be $(0.25 + 0.8) \div 2 = 0.525$. This means, that together the two fishermen would have a Precision of 52.5% regarding their respective favorite fish. The weighted average, on the other side, also considers the distribution of classes. This makes sense if classes are not distributed equally. In the above fish example trouts are twice as frequent as salmons. Thus, the chance to catch a salmon is not equivalent to catching a trout. The weighted average Precision takes this into consideration by multiplying the Precision values with the number of class members. The weighted average Precision for the fish example is: $\big((0.25 \times 10) + (0.8 \times 20)\big) \div 30 = 0.62$. As can be seen, in case of unbalanced classes, the macro average allows less frequent classes the same influence on the average as more frequent classes, while in case of the weighted average the importance of less frequent classes is reduced.

In the literature also a third sort of average can be found, i.e., micro average. Micro average Precision is computed by summing up all True Positive counts over classes in a first step. Then this sum is divided by all False Positives over classes. Regarding the fish example: $\big((2 + 8) \div (10 + 10)\big) = 0.5$. If we would compute the micro average Recall and F_1 score the result would also be 0.5. This is because the micro average always computes the proportion of

correctly classified observations, i.e., correctly fished fish. This corresponds to the definition of another metric, too – Accuracy. Accuracy is defined as:

$$\frac{TP + TN}{TP + TN + FP + FN}$$

Therefore, when it comes to micro average the following holds true:

$$MicroPrecision = MicroRecall = MicroF_1 = Accuracy$$

APPENDIX E

Assignment of graduation to level of education

- graduation_low (4): ohne Pflichtschulabschluss; Berufsausbildung (Lehre); Oberschule 3.Klasse; Pflichtschulabschluss
- graduation_middle (3): Fachoberschule mit Matura; Oberschule mit Matura; technische Fachausbildung nach Matura
- graduation_high (2): Fachhochschulabschluss, abgeschlossene Ausbildung an einer Meisterschule bzw. Berufs- oder Fachakademie; Hochschulabschluss
- graduation_else (2): einen anderen Abschluss; no data)

Linear regression – assumption tests

The contribution of the predictors gender, age, educational level and place of residence to the prediction of dialect use was investigated by calculating a multiple regression. Educational level was divided into three categories, high, middle and low, while in place of residence the three locations Bozen (urban area), Sarntal and Pustertal (rural areas) were considered. To calculate the criterion variable the percentage of dialectal tokens in relation to the total tokens per person was computed.

An examination of the residuals using a QQ plot (see Figure 16) showed that they followed a normal distribution (normality).

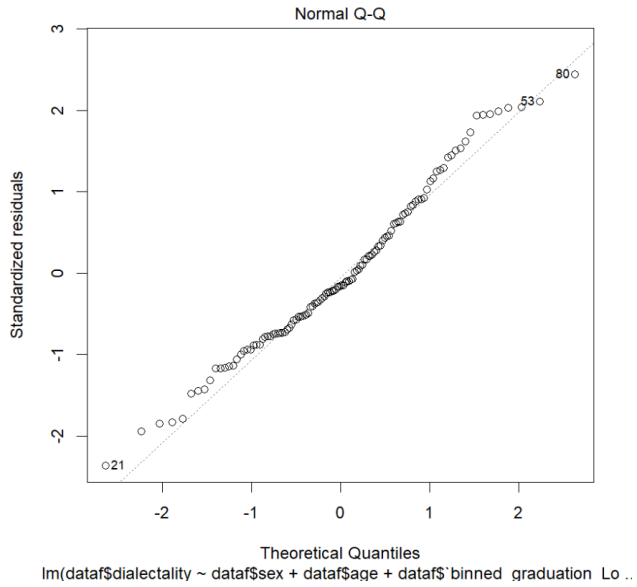

Figure 16: QQ plot of model residuals

However, it was found that the homoskedasticity condition was not met, as shown in Figure 17.

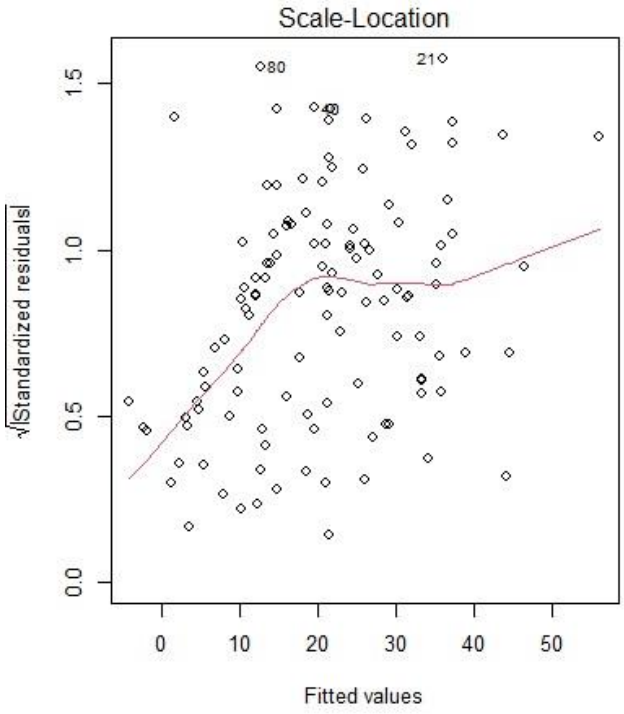

Figure 17: Scale-Location plot to check for homoscedasticity

To address this issue, robust standard errors were calculated using HC 3.

Furthermore, an examination of the variance inflation factors (VIFs) was conducted to test for possible multicollinearity among the predictors. It was found that the highest VIF value was 1.40 (low education level), indicating that there is no strong multicollinearity and that the predictors, thus, are not highly correlated.