Variation is the Norm: Embracing Sociolinguistics in NLP

Anne-Marie Lutgen[1], **Alistair Plum**[1], **Verena Blaschke**[2,3], **Barbara Plank**[2,3], **Christoph Purschke**[1]
[1]University of Luxembourg, Esch-sur-Alzette, Luxembourg
[2]MaiNLP, LMU Munich, Germany
[3]Munich Center for Machine Learning (MCML), Munich, Germany
anne-marie.lutgen@uni.lu

Abstract

In Natural Language Processing (NLP), variation is typically seen as noise and "normalised away" before processing, even though it is an integral part of language. Conversely, studying language variation in social contexts is central to sociolinguistics. We present a framework to combine the sociolinguistic dimension of language with the technical dimension of NLP. We argue that by embracing sociolinguistics, variation can actively be included in a research setup, in turn informing the NLP side. To illustrate this, we provide a case study on Luxembourgish, an evolving language featuring a large amount of orthographic variation, demonstrating how NLP performance is impacted. The results show large discrepancies in the performance of models tested and fine-tuned on data with a large amount of orthographic variation in comparison to data closer to the (orthographic) standard. Furthermore, we provide a possible solution to improve the performance by including variation in the fine-tuning process. This case study highlights the importance of including variation in the research setup, as models are currently not robust to occurring variation. Our framework facilitates the inclusion of variation in the thought-process while also being grounded in the theoretical framework of sociolinguistics.

Keywords: Variation, Sociolinguistics, Luxembourgish

1. Introduction

In structural linguistics, variation is often perceived as a disturbing factor and discarded in grammatical descriptions to emphasise the constant and invariable elements of a language (Berruto, 2004). With the advent of sociolinguistics, the study of language variation in social contexts (Wodak et al., 2011) has become a central focus of linguistic research, mirroring the fundamental role of variation in language and the production of social meaning (Eckert, 2016).

Similarly to structural linguistics, variation in Natural Language Processing (NLP) is also often seen as noise (in the signal) and a practical nuisance (for processing) (Nguyen et al., 2021). We argue that sociolinguistic insight should be a part of the NLP research setup, since language variation and its role in constructing social meaning are not an exception but a characteristic of language and could therefore improve the performance of language models as well as their representation of linguistic diversity.

We develop a framework to combine the sociolinguistic classification of language variation with domains of application in NLP. This framework includes guidelines to understand the status and function of a linguistic entity (language or variety; see Section 3) and the dimensions of variation linked to it. The technical dimension illustrates five steps in language modelling where variation has an impact. By combining these two sides, we achieve a precise understanding of the linguistic entity researched and how it is linked to the technical implementation of modelling. This also leads to practical solutions for how to handle variation in a processing pipeline, as a (socio)linguistic description of the researched entity will identify where problems on the technical side may occur.

We illustrate the use and effectiveness of our framework on a case study of orthographic variation in Luxembourgish and how it impacts the performance of fine-tuned classification models. We conduct an experiment where we fine-tune a Luxembourgish BERT model (Lothritz et al., 2022) and mBERT (Devlin et al., 2019) on 8 different classification tasks in Luxembourgish. To compare performance with and without orthographic variation, we destandardise or normalise training and test data. Additionally, we fine-tune the models on a combined version of the dataset which includes the standard and non-standard version.

The results not only show large discrepancies in the performance of models tested on non-standard data but also of models that are trained on data with large amounts of variation. With an extensive understanding of the sociolinguistic situation, we are able to provide a possible solution to include variation in the process and improve the performance with the combined method.

2. Language, Variety, and Variation

In most contexts, both everyday practice and research, we take the existence of languages as

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 11332–11344
11–16 May 2026.

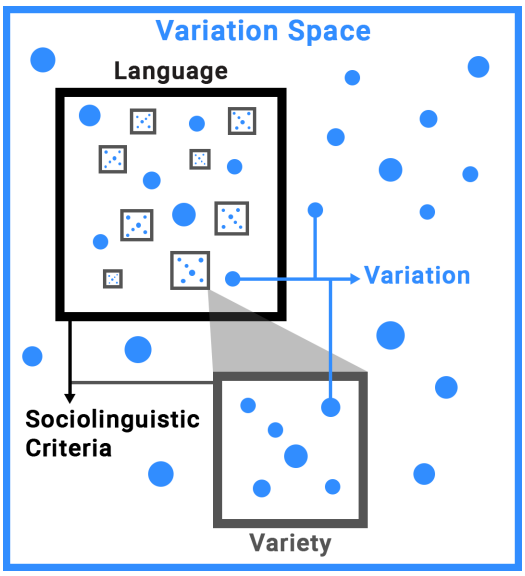

Figure 1: Illustration of the container metaphor for language and variety.

something natural. We differentiate one language from another, and within a language, maybe one dialect from the next. We do so based on our personal experience and the knowledge acquired in practice. From a linguistic point of view, however, what is a language (and what is not) is not evident, but we need to define why we call a particular way of speaking and writing a "language" or a "dialect" (Cutler et al., 2025). Hence, before we elaborate on the social and technical dimensions of our framework, we need to establish notions of the terms variation, variety and language. In doing so, we mainly draw on variationist literature dealing with regional variation.

2.1. Defining Varieties and Languages

To understand how we define linguistic entities, we present a simple model of the "variation space" from a social interactionist point of view (Schmidt, 2010). Figure 1 illustrates the variation space, including linguistic variants (blue points) as well as delimitated containers within this space representing *languages* and *varieties* (Grieve et al., 2025). To establish these containers, we discuss sociolinguistic criteria based on which varieties and languages can be defined (Purschke, 2019).

The ways in which people speak and write varies in multiple dimensions: regionally, socially, historically, between generations, stylistically, etc. (Berruto, 2010). The sum of all different ways of speaking and writing defines what we call the "variation space", i.e., the totality of all existing linguistic variants. Within this space, the continued communication of speakers creates (social) orders of interaction and, hence, orders of linguistic variation. To capture these linguistic orders, Schmidt (2010) defines three levels of *synchronisation* to reflect the fact that speakers structure the variation space by synchronising their ways of speaking/writing with other speakers (based on their ways of living). On the micro level, individual interactions between speakers create shared (= synchronised) linguistic repertoires, e.g., a couple developing a form of private register shared only between them. Repeated interactions between larger groups of people (social, regional, national, etc.) synchronising their ways of speaking/writing define the meso level, i.e., speakers inside the group talk more to each other than they do to outsiders and, hence, develop a shared way of speaking/writing (e.g., a regional dialect). If all speakers within a large community (such as a state) synchronise their communications toward one normative way of speaking/writing, this defines the macro level, best compared to a national or linguistic *norm* (e.g., a standard language).

Using the concept of synchronisation, we can now understand how speakers (and linguists) delimitate entities within the variation space (Lameli, 2013). Starting from the meso level, we assume that continued interactions within groups lead to stabilised ways of speaking/writing distinct from those of other groups, e.g., two neighbouring regional dialects. The distinction between these socially synchronised (internal order) yet (socio)linguistically distinct (external border) ways of speaking/writing invokes the concept of *variety* as used in linguistics (Berruto, 2004), that is, a way of speaking/writing characteristic for a group of speakers and different from other groups' ways of speaking/writing, based on social, regional, stylistic, etc. criteria. As such, dialects are examples of varieties, as are sociolects or languages.

The question of who defines and labels which linguistic entity in what way (a dialect, a language) depends on the criteria used to delimitate that entity from other entities within the variation space (Ramberg and Røyneland, 2025). Purschke (2019) proposes a list of linguistic systemic and individual subjective criteria by which a group of variants used by a group of speakers (= a socially synchronised way of speaking/writing) can be described and defined *as* a linguistic entity. Those criteria include linguistic differences between two entities, group-related norms and social functions of entities for the linguistic side, and perceived differences, language preferences/attitudes and individual norm concepts for the subjective side. For our framework, we use an adapted version of these (see Section 3) as a catalogue of criteria which can be used to describe a linguistic entity based on its (socio)linguistic qualities. In this way, we provide a checklist for NLP purposes to describe and eval-

uate linguistic entities as cultural constructs, not naturally occurring things.

2.2. Why is Variation Important?

By focusing on the standard varieties of languages, in NLP, variation often seems to be a disruptive, 'non-standard' factor (Plank, 2016). However, as established in Section 2.1, variation is the very fundament of language and social interaction, and the acceptability or correctness of variants are the product of linguistic and social processes of standardisation. Moreover, non-standard variants carry social meaning, i.e., they are indicative of regional, social, national, etc. identities (Cercas Curry et al., 2024).

Variation is part of a social semiotic system that conveys the entire range of social concerns of a community (Eckert, 2012). Essentially, people express social meaning through variation. Moreover, variation can be a feature indexing ideology, stance or belonging. Even orthographic variation carries not only linguistic meaning (in the sense of correct/incorrect) but also social meaning (in the sense of indexing identities; Sebba, 2007). So when people vary their language use, they express who they are, where they belong, and how they want to be seen by others.

In NLP, variation is often treated as a problem to solve, and as noise in the data to be normalised (Eisenstein, 2013; Al Sharou et al., 2021). However, normalising language holds the assumption that there exists a default norm which is often not the case, e.g., if we consider non-standardised languages or varieties (such as dialects). Further, normalisation often takes the standard variety and formal register as a default (Doğruöz and Sitaram, 2022), therefore erasing the social meaning from the original text.

3. Sociolinguistic NLP Framework

This section introduces our analytical framework. We first discuss nine sociolinguistic criteria of varieties using the example of Luxembourgish in Section 3.1. Then in Section 3.2, we discuss five essential steps in language modelling and how they relate to the sociolinguistic dimension.

3.1. Sociolinguistic Criteria

As discussed in Section 2.1, we define varieties/languages as containers within the variation space (Figure 1). We now provide a list of sociolinguistic criteria to describe linguistic entities and, hence, delimitate them as varieties/languages based on those criteria.

The sociolinguistic setting describes the socio-pragmatic context where this entity is set. This includes multilingualism (individual, societal) and forms of language contact with other entities.

Luxembourg has three official languages, Luxembourgish (national language), French (legislative language) and German, and is therefore a multilingual society. Luxembourgish has a long history of language contact with German and French as well (Gilles, 2023). Additionally, half of the population are foreign residents mostly of French, Portuguese, Italian, Belgian and German descent (STATEC, 2024).

Institutional support describes the political status of the entity and efforts to its societal anchoring via language policy.

Luxembourgish was established in law in 1984 and is today the national language of the country, following a constitutional reform in 2023. Linguistically, it derives from a Moselle-Franconian dialect. The Centre for the Luxembourgish Language (ZLS), a state run institution, is in charge of the official dictionary and orthography.

Structural independence describes the structural linguistic differences between the entity and neighbouring ones (e.g., adjacent dialect or roofing standard variety). The degree of linguistic independence from other entities and its linguistic relations to those often define the difference between labelling an entity as a language or a variety of a language (Auer, 2013).

Luxembourgish is considered an "Ausbausprache" (Kloss, 1967). In linguistics, it was long seen as a variety of German (Moselle-Franconian), which has developed into a language, covering all functional domains of a standard variety. While there is still regional variation in Luxembourgish, Gilles (1999) finds an advanced state of dialect levelling that resolves the former regional dialects into a national variety with small (lexical, phonological, grammatical) remnants of variation.

The degree of codification describes how advanced the orthographic, grammatical and lexical standardisation of an entity is. Standardised languages often represent varieties with high social prestige, act as the main literary language and often are institutionally regulated (Bird, 2022).

Luxembourgish used to be a mainly spoken language but has gained ground in the written domain in the past 25 years. The language is not fully standardised yet (Gilles, 2023), but an official orthography has long existed, last updated in 2019 (Zenter fir d'Lëtzebuerger Sprooch, 2019). However, since Luxembourgish is not systematically taught in schools, the population has little knowledge of

the official rules, resulting in a broad spectrum of variation in written texts (Gilles, 2023).

Domain specificity describes where the entity is used and what functions are attributed to it.

Luxembourgish multilingualism is characterised by a functional differentiation by social domains. French is the legislative language and important in private business contexts, while public institutions rely on Luxembourgish (parliament, administration). German has its main domains in the traditional print media and in literacy training (Purschke, 2020). Luxembourgish is the preferred language in communications among locals and a language of social integration, but given the highly diverse population, the situational choice of a language depends mostly on individual preferences and linguistic repertoires. Luxembourgish is used mostly in informal domains verbally and in writing.

School education describes the anchoring of an entity in education contexts, that is school curricula or foreign language learning.

In Luxembourg, German, and lately French, are the official languages for literacy training, with Luxembourgish as an additional language of instruction. In secondary education, German and French serve as the main languages of instruction, although Luxembourgish is informally also present. Nevertheless, there is little to no formal Luxembourgish education at school. There is also high demand for Luxembourgish language courses by immigrants and cross-border workers (Sattler, 2021).

Communicative range describes the size of the speaker group and how useful this language is in social practice.

Luxembourgish is a small language with around 400.000 speakers and is mostly spoken in Luxembourg, with small pockets in the neighbouring regions due to socio-economic mobility (cross-border commuters) and historical connection (the Belgian region called “Luxembourg” that borders Luxembourg). Luxembourgish has a central position as language of social integration but is currently losing ground in the language regime – due to the influx of foreign residents – despite having growing numbers of speakers (Fehlen et al., 2023). In contact with speakers of closely related varieties, i.e., Moselle Franconian in Germany, Luxembourgers most often switch to German, highlighting the cultural difference and, hence, limiting the communicative range of Luxembourgish.

Attitudes and ideologies are fundamental to social practice, shaping how people use, perceive, and evaluate language (Purschke, 2020). Especially in multilingual contexts, attitudes towards different languages reveal much about social dynamics and ideological tensions in a speech community.

In Luxembourg, attitudes towards the different languages reflect the complexity of the societal multilingualism. Purschke (2020) shows a clear hierarchy of language preferences in daily life, with Luxembourgish being the preferred language in practice for Luxembourgish native speakers. Additionally, there is a close connection between language, nation and national identity (Purschke, 2025). French is still considered the prestige variety in most contexts, while English is becoming more important. German, while losing ground in the language regime, is preferred over French in younger generations, also considering its typological closeness with Luxembourgish.

3.2. NLP Domains

If we consider variation to be a part of any variety and therefore a part of language modelling, embracing the sociolinguistic dimension adds a new perspective on known problems in NLP. We have isolated *five distinct domains* in modelling language where variation and different varieties can introduce complexities. In each section we list related work on Luxembourgish as our case study, giving an overview on recent advances.

Data – Knowing your data and its sociolinguistic dimensions allows for a better understanding of the varieties and variation present. For example, Kreutzer et al. (2022) show that five commonly-used web-crawl corpora (CCAligned, ParaCrawl, WikiMatrix, OSCAR, mC4) are of questionable quality, by identifying languages with no usable text, languages with a very low amount of usable text, and languages with wrong or ambiguous languages codes. Especially low-resource languages are concerned in this quality issue as language identification is especially challenging for language varieties (van der Goot, 2025). Lau et al. (2025) also highlight how common mislabelling in language data is and how this directly influences the performance of language models.

For Luxembourgish, automatic language identification is difficult since the language is structurally close to German and exhibits a high amount of French borrowing (Gilles, 2023), resulting in low accuracy for language identification. For instance, at first glance, the sentences tagged as Luxembourgish in mC4 (Xue et al., 2020) and OSCAR (Ortiz Suárez et al., 2019) are often German, Dutch or French.

Data – Data selection should be influenced by which variety should be represented during training, and not only by the amount of data available. The specific sociolinguistic situation is also an important factor to account for. For instance, Ramponi (2024) shows the shortcomings for Italian varieties since NLP focuses more on the amount of resources instead of taking the cultural context into account, like culture preservation, language learning, and intergenerational transmission. He argues for a more responsible speaker-centric approach for varieties in order to preserve language varieties of Italy. The amount of data available is one of many factors to account for while selecting the data for language modelling.

For Luxembourgish, different types of training data exist which are either closer to the orthographic standard or contain a high amount of variation (Plum et al., 2025). While selecting the training data, we need to be aware of the content of the training data. A high amount of variation in the training data has an impact on the preprocessing and training, which we show below in Section 4.

Preprocessing – Unicode character normalisation specifically for diacritics should be consistent. Gorman and Pinter (2025) show that Unicode inconsistencies lead to performance deterioration.

Diacritics are also widely used in Luxembourgish. They are part of the orthography, specifically to distinguish between different vowel qualities (Zenter fir d'Lëtzebuerger Sprooch, 2019).

Preprocessing – Normalisation aims to transform non-standard spelling into a standardised form (Han and Baldwin, 2011; van der Goot, 2019). However, by normalising text, we remove rich social signals which are present due to sociolinguistic variation (Nguyen et al., 2021).

For Luxembourgish, two normalisation tools are available: *spellux* (Purschke, 2020), a normalisation pipeline, and a neural normalisation model trained with variation-infused parallel data (Lutgen et al., 2025). Both perform similarly, however, Lutgen et al. (2025) argue that normalisation is not effective to smooth training data. In the context of our case study in Section 4 we show the impact normalisation has on fine-tuning.

Modelling – Tokenisation is sensitive to language variation, and its impact on down-stream tasks depends on the robustness or sensitivity needed to account for variation (Wegmann et al., 2025). Subword segmentation for German varieties for instance is challenging as well and does not correspond to a meaningful representation of the data. This is a stark contrast to the German standard variety, as tokenisation difficulties for German varieties lead to low performance in down-stream tasks (Blaschke et al., 2023).

For Luxembourgish, the impact of data with a high or low amount of variation on tokenisation still needs to be researched. However, we expect differences in tokenisation between the orthographic standard training data and the training data with a high amount of variation.

Modelling – Pre-training strategies for different varieties must be carefully considered depending on the available data. Modern LLMs are trained on diverse data, including multiple languages and high-noise data, i.e. variation (Grieve et al., 2025). However, the controlled study of variation for specific, low-resource varieties remains crucial for robustness, fairness, and understanding model limitations. Hedderich et al. (2021) illustrate different strategies for pre-training while working on low-resource languages. One strategy for working with multiple language varieties is adding language labels during training. This enables more cross-lingual transfer as Conneau and Lample (2019) show for XLM-R and Liu et al. (2020) for mBART. Another strategy is hyper-parameter optimisation based on the amount of training data. The amount of data per language is often regulated (up/downscaled) using a smoothing strategy. Usually, multinomial smoothing is used, where for example XLM-R has a lower upscaling rate (Conneau et al., 2020) in comparison to mBERT (Devlin et al., 2019). Knowing your training data, the importance of accurate language labels, the amount of variation in the training data and smoothing parameters impact pre-training strategies.

For Luxembourgish, Plum et al. (2025) train a T5 model while combining the Luxembourgish training data with French and German data from similar domains. The results show a clear performance boost in comparison to multilingual models, which often use incorrectly tagged non-Luxembourgish training data in the pre-training corpus. Similarly, LuxGPT (Bernardy, 2022) was trained using transfer learning from German. Understanding the prevalence of language contact and the impact on the language with syntactical, lexical and morphological variation make this an informed strategy for upscaling the training data yielding good results. In comparison, Lothritz et al. (2022) use synthetic Luxembourgish for data augmentation to train LuxemBERT also yielding good results.

Modelling – Strategies for fine-tuning for data with a high amount of variation depend on the specific situation and if the data is intended for a classification or generation task. Generally, fine-tuning of pre-trained language models has been shown to be unstable especially for small datasets (Du and

Nguyen, 2023) which is often the case for lower-resource languages. For generative tasks, like neural machine translation, different varieties and variation also pose a challenge. Zampieri et al. (2020) illustrate different strategies used for machine translation for varieties. One strategy, for instance, allows for a shared subword-level vocabulary which enables orthographic and morphological variation learning between related languages.

For Luxembourgish, Plum et al. (2025) present the generative evaluation dataset LuxGen, which encompasses tasks for standard orthographic and non-standard Luxembourgish. This mix enables them to evaluate models on multiple Luxembourgish varieties. For classification tasks, Lothritz et al. (2022) present multiple datasets for classification tasks, however, a fine-grained evaluation on the impact of variation in those tasks is not part of the evaluation. We will show the impact of variation on those classification tasks in Section 4. Ranasinghe et al. (2023) benchmark various language models on the task of comment moderation, which encompass a high degree of variation.

Evaluation – Evaluation of non-standard varieties poses a major challenge as evaluation metrics are not robust to non-standardised varieties, specifically for generation. Aepli et al. (2023) show that translation metrics are not reliable for evaluating Swiss German translations, and propose changes to improve robustness. Sun et al. (2023) illustrate similar findings and conclude that existing metrics for machine translation prioritise dialect similarity over semantics. In both cases, human and automatic evaluations are compared.

In Luxembourgish, Plum et al. (2025) also evaluate the LuxGen tasks not only with BLEU, as this metric is not suitable for language with a high degree of variation, but also qualitatively. Lutgen et al. (2025) use quantitative and qualitative evaluation metrics to evaluate and compare a ByT5-based normaliser and the normalisation pipeline *spellux* (Purschke, 2020) for Luxembourgish. By using behavioural performance tests for different correction types, Lutgen et al. (2025) present a fine-grained evaluation on normalisers and show the strengths and weaknesses for each approach.

Usage – The usage of language technologies is becoming more common, therefore the user preferences, especially for lower-resource varieties, should be a concern (Markl et al., 2024). For instance, Blaschke et al. (2024) surveyed German dialect speakers to research language preferences for AI and found that respondents were more interested in potential NLP tools that work with dialectal in- rather than output.

For Luxembourgish, a similar study has not been performed yet. A study could focus on preferred language use, considering the complex multilingual situation and the varying degrees of adherence to the orthography in the written domain.

Usage – Safety issues for low-resource varieties can be seen in jailbreaking (Upadhayay and Behzadan, 2025), as using low-resource languages and languages with a high amount of code switching are effective for jailbreaking attacks. Lent (2025) also shows a high security risk for mono- and multilingual models with low-resource languages.

Safety issues for Luxembourgish have not been researched yet. One could investigate the impact of a high or low degree of variation in the prompts.

3.3. How to Use this Framework?

The goal of the framework we have presented is to systematically include sociolinguistics in NLP. Given the social and linguistic complexity of language, understanding the sociolinguistic perspective on language and how this interacts with language modelling in NLP is essential, both for model performance and a comprehensive representation of linguistic diversity. We argue that sociolinguistic knowledge should inform NLP and, hence, should become an integral part of the research setup. We further demonstrate the usefulness of our framework empirically in a study on orthographic variation in Luxembourgish.

4. Luxembourgish as a Case Study

In this section we illustrate how we use the sociolinguistic criteria to inform NLP experimentation. As a case study, we analyse how orthographic variation impacts the performance of fine-tuned models for classification tasks for Luxembourgish. The targeted NLP dimensions are data, preprocessing, modelling (fine-tuning) and evaluation. In the process of analysing variation we have first carefully evaluated the sociolinguistic context, as illustrated in Section 3.1, in order to understand how sociolinguistic criteria impact orthographic variation. The study is informed by the framework's sociolinguistic analysis (Section 3) and investigates the impact of variation on specific NLP pipelines.

We consider two varieties of Luxembourgish, the more formal written orthographic standard Luxembourgish called **standard** and the informal written Luxembourgish that allows for a wide range of orthographic variation called **non-standard**. Therefore, since we know our data and which variety it represents, we are able to preprocess it for our intended use. We quantitatively evaluate the performance difference of language models for the

same downstream tasks in standard Luxembourgish and in non-standard Luxembourgish. Further, we experiment with a **combined** fine-tuning dataset that encompasses both the standard and the non-standard data of the same datasets to include variation in the training and evaluate the performance. Our fine-tuning setup allows for a fine-grained evaluation of each fine-tuned model (standard, non-standard and combined) on the standard, non-standard and combined variety of the datasets. By including the sociolinguistic perspective, we not only bring to light the performance differences for different varieties on the same task. We are also able to present a possible solution by combining the standard and non-standard data in order to improve the performance for each evaluated fine-tuned model.

4.1. Datasets

To evaluate the models on various downstream tasks, we use datasets that are already available for Luxembourgish. Every task is defined as a classification task, either token or sequence based. All the following tasks are from Lothritz et al. (2022), namely: intent classification (IC), winograd natural language inference (WNLI), part-of-speech tagging (POS), named entity classification (NER) and sentiment classification (SC). Additionally topic classification (TC) (Adelani et al., 2023) and comment moderation (CM) (Ranasinghe et al., 2023) are part of our setup. We always use the same hyper-parameters as Lothritz et al. (2022) but the data differed for TC as we used (Adelani et al., 2023) dataset. For CM, we used the same hyper-parameter and size of the dataset as for the SC task in Lothritz et al. (2022). More detailed explanation on the tasks including the fine-tuning hyper-parameters can be found in Lothritz et al. (2022); Adelani et al. (2023); Ranasinghe et al. (2023) and are reported in the Appendix.

Variant	Task	WER	CER
Destandardised	IC	78.09	14.12
	NER	63.78	12.94
	POS	65.61	13.00
	WNLI	83.23	16.02
	TC	64.57	9.89
Normalised	CM	28.81	4.97
	SC	28.80	4.72

Table 1: WER (in %) and CER (in %) between pairs of standard and non-standard variety sentences.

4.2. Normalisation & Destandardisation

In order to manipulate the datasets of the downstream tasks and inject different amounts of variation into the data, we use normalisation and destandardisation. Normalisation transforms non-standard forms into standard forms with a normalisation pipeline *spellux* (Purschke, 2020). Destandardisation instead aims to inject variation, and is performed with the destandardisation algorithm based on Lutgen et al. (2025). The algorithm is based on data provided by *Spellchecker.lu*[1], a semi-automatic spellchecking website frequently used in Luxembourg. The destandardisation algorithm includes different variants for each word as well as frequency information for user corrections for each variant. This creates a real-life dictionary of spelling variants per lemma, including their frequency of use. This dictionary is then used to replace words with a variant based on the frequency of use. This approach is considered superior to adding random character replacements to the data (generating synthetic data), as this captures real variation patterns in Luxembourgish.

The following tasks are written in the orthographic standard and are thus destandardised to form the non-standard variant of the same task: IC, NER, POS, WNLI and TC. Both SC and CM are based on online comments on news articles and are therefore the non-standard variant by default. SC and CM are normalised to form the standard variant of the task.

Additionally both the standard and non-standard variant of the same datasets were combined to form the **combined** variant of the same dataset. This data manipulation is performed for the train, test and dev set to fine-tune the models on each variant and evaluate them on each variant.

To quantify how much variation is introduced, we calculated word error rate (WER), character error rate (CER) and normalised character of the standard and non-standard variant of the datasets which can be seen in Table 1. The destandardisation has a mostly uniform pattern and between 60% and 80% of the words are changed as the WER describes. However, we can also see that the normalisation process is not changing the data drastically since the WER is around 28%.

4.3. Models

For this experiment, we chose BERT-based models to compare the standard and non-standard version of the classification tasks. Our two models are *LuxemBERT* (Lothritz et al., 2022) and *mBERT* (Devlin et al., 2019) as we chose one model trained for Luxembourgish and one multilingual model to compare the performance. Further, encoder-based models are the best choice for classification tasks (Weller et al., 2025; Ojo et al., 2025). We repeat the experiments with five random seeds and measure standard deviation and

[1] https://spellchecker.lu

therefore the stability of the fine-tuning process.

Model		std	n-std	comb
LBERT	std	57.97 ± 2.18	46.48 ± 1.75	52.60 ± 0.74
LBERT	n-std	44.88 ± 2.27	47.67 ± 2.18	46.48 ± 1.78
LBERT	comb	68.56 ± 1.05	65.81 ± 1.64	67.22 ± 1.25
mBERT	std	26.71 ± 4.93	21.91 ± 4.49	24.64 ± 4.69
mBERT	n-std	13.35 ± 2.82	19.07 ± 2.21	16.51 ± 2.07
mBERT	comb	45.37 ± 1.33	46.47 ± 2.72	46.02 ± 2.01

Table 2: IC – Weighted F1 (± std)

Model		std	n-std	comb
LBERT	std	87.52 ± 0.04	75.73 ± 0.14	81.36 ± 0.07
LBERT	n-std	86.90 ± 0.11	83.03 ± 0.07	84.85 ± 0.08
LBERT	comb	87.60 ± 0.08	83.03 ± 0.01	85.18 ± 0.01
mBERT	std	85.02 ± 0.52	67.68 ± 0.54	76.34 ± 0.48
mBERT	n-std	84.61 ± 0.41	81.43 ± 0.70	82.98 ± 0.55
mBERT	comb	85.12 ± 0.06	80.84 ± 0.07	82.93 ± 0.06

Table 3: POS – Weighted F1 (± std)

Model		std	n-std	comb
LBERT	std	64.42 ± 0.57	64.08 ± 0.41	64.26 ± 0.44
LBERT	n-std	63.14 ± 0.37	64.04 ± 0.66	63.60 ± 0.41
LBERT	comb	64.28 ± 0.05	64.54 ± 0.06	64.37 ± 0.06
mBERT	std	55.28 ± 7.39	55.41 ± 7.64	55.35 ± 7.52
mBERT	n-std	59.38 ± 1.14	60.50 ± 1.14	59.96 ± 1.06
mBERT	comb	60.36 ± 0.09	60.68 ± 0.05	60.52 ± 0.06

Table 4: CM – Weighted F1 (± std)

4.4. Results

In the following we discuss the results of three tasks more closely, one sequence classification task (IC), one token classification task (POS) and one sequence classification task for which the normalisation process was used (CM). The results of the remaining tasks are in the Appendix. The results are shown in Table 2, 3 and 4. The tables show the weighted F1 score for each model (LuxemBERT and mBERT) trained on the different datasets (standard, destandard and combined) and then tested on the different datasets (standard, destandard and combined).

First, in nearly all tasks the model fine-tuned exclusively on the non-standard datasets is the worst performing category. In the IC task for instance, shown in Table 2, the LuxemBERT model fine-tuned on standard data has an F1 score of 57.97% for the standard test set but the same model performs a lot worse on the non-standard test set with a score of 46.48%. The non-standard model variant for the same task also shows a performance drop on the standard test set (44.88%). Only a slightly better performance for the non-standard test set (47.67%) is visible even though this model is fine-tuned on the same variant-infused data as the test set. On nearly all tasks, a similar trend as this one can be observed.

We can also see this trend on the mBERT results. Additionally, the performance is generally lower for all the tasks compared to the LuxemBERT variant, showing that pre-training with more in-language data is beneficial for the performance of fine-tuned tasks on a specific language.

Secondly, the fine-tuned combined model is one of the best performing models in nearly all of the tasks. If we look at the IC task again, Table 2, the combined variant of LuxemBERT has by far the best performance in all three test sets (standard, non-standard and combined). This shows that including variation in the standard data for the training setup can not only benefit the performance of the model for the non-standard data but also the standard variant. However, Luxembourgish has a high degree of variation in the language even in orthographically standardised text therefore it makes sense that this experimental setup is also beneficial for the standard variant. Yet, the combined fine-tuned model seems to have the most impact on sequence classification tasks. In token classification tasks like POS, shown in Table 3, we can only observe a marginal difference. But generally, when variation is part of the training set the performance is higher for the destandard test set.

Third, the results for both tasks that were normalised for the CM experiment (Table 4) show nearly no difference in performance indicating that the current normalisation process is not that successful to make a significant difference. This also aligns with the WER and CER in Table 1. Although normalisation is typically seen as the solution to work with text including a large amount of variation (van der Goot et al., 2021; van der Goot and Çetinoğlu, 2021; Plank et al., 2020) we can clearly see in our experimental setup that incorporating linguistic variation in addition to the standard form within the training data can yield even more improvement overall in the down-stream tasks. Furthermore, the social meaning that those variants carry are also included in the model training. This results in a more diverse training data and can even be varied depending on the social context of specific tasks.

5. Conclusion

This paper proposes a framework to systematically include sociolinguistics in NLP. We show how these sociolinguistic criteria can be used in the case of Luxembourgish and illustrate performance challenges for different varieties of Luxembourgish. By contrasting the orthographic standard and a variation-infused variant of the same data we show how differently the fine-tuned models perform for each variant. Additionally, we show a possible route for improving the performance and robustness by including variation in the training process by combining standard and non-standard data.

This framework is not only intended for small languages and varieties like Luxembourgish but is universally applicable. The Luxembourgish case study served as a proof-of-concept and the framework is suitable for every language. Given the social and linguistic complexity of language, the sociolinguistic criteria give an informed overview of the variety researched. Our framework could be practically applied in a sort of "sociolinguistic" language card in datasets. This would give an extensive overview of the specific varieties included in the dataset and could identify possible problems in an NLP research setup.

6. Limitations

This research involves experiments on Luxembourgish language data, where we normalised and destandardised the data. The destandardisation process does not cover the entirety of the variation space in Luxembourg as it only covers the variants included in the data used by the algorithm. We acknowledge that the linguistic coverage of our datasets may not fully reflect the linguistic diversity of Luxembourgish.

7. Acknowledgements

This research was supported by the Luxembourg National Research Fund (Project code: C22/SC/117225699) and the ERC Consolidator Grant DIALECT 101043235.

The experiments reported in this paper were conducted on the MeluXina high-performance computing infrastructure, an allocation granted by the University of Luxembourg on the EuroHPC supercomputer hosted by LuxProvide.

We would also like to thank Rob van der Goot, Peter Gilles, Emilia Milano, Felicia Körner, Lou Pepin, Nils Rehlinger and Mélanie Wagner for their invaluable input and Jacques Spedener for the illustration.

8. Bibliographical References

Noëmi Aepli, Chantal Amrhein, Florian Schottmann, and Rico Sennrich. 2023. A Benchmark for Evaluating Machine Translation Metrics on Dialects without Standard Orthography. In *Proceedings of the Eighth Conference on Machine Translation*.

Khetam Al Sharou, Zhenhao Li, and Lucia Specia. 2021. Towards a Better Understanding of Noise in Natural Language Processing. In *Proceedings of RANLP*.

Peter Auer. 2013. *Dialect Divergence at the State Border*, pages 295–309. Peter Lang Verlag, Bruxelles, Belgium.

Laura Bernardy. 2022. A Luxembourgish GPT-2 Approach Based on Transfer Learning. Master's thesis, University of Trier.

Gaetano Berruto. 2004. The problem of variation. *The Linguistic Review*, 21(3-4):293–322.

Gaetano Berruto. 2010. *13. Identifing dimensions of linguistic variation in a language space*, pages 226–241. De Gruyter Mouton, Berlin, New York.

Steven Bird. 2022. Local Languages, Third Spaces, and other High-Resource Scenarios. In *Proceedings of ACL*.

Verena Blaschke, Christoph Purschke, Hinrich Schuetze, and Barbara Plank. 2024. What Do Dialect Speakers Want? A Survey of Attitudes Towards Language Technology for German Dialects. In *Proceedings of ACL*.

Verena Blaschke, Hinrich Schütze, and Barbara Plank. 2023. Does Manipulating Tokenization Aid Cross-Lingual Transfer? A Study on POS Tagging for Non-Standardized Languages. In *Proceedings of VarDial*.

Amanda Cercas Curry, Zeerak Talat, and Dirk Hovy. 2024. Impoverished Language Technology: The Lack of (Social) Class in NLP. In *Proceedings of LREC-COLING*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of ACL*.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Celia Cutler, Unn Røyneland, and Sebastijan Vrzić. 2025. *Language Activism: The Role of Scholars in Linguistic Reform and Social Change*. Cambridge University Press, Cambridge.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.

A. Seza Doğruöz and Sunayana Sitaram. 2022. Language Technologies for Low Resource Languages: Sociolinguistic and Multilingual Insights. In *Proceedings of SIGUL*.

Yupei Du and Dong Nguyen. 2023. Measuring the Instability of Fine-Tuning. In *Proceedings of ACL*.

Penelope Eckert. 2012. Three Waves of Variation Study: The Emergence of Meaning in the Study of Sociolinguistic Variation. *Annual Review of Anthropology*, 41:87–100.

Penelope Eckert. 2016. *Variation, meaning and social change*, page 68–85. Cambridge University Press.

Jacob Eisenstein. 2013. What to do about bad language on the internet. In *Proceedings of NAACL-HLT*.

Fernand Fehlen, Peter Gilles, Louis Chauvel, Isabelle Pigeron-Piroth, Yann Ferro, and Etienne Le Bihan. 2023. RP2021 N°8 – Linguistic Diversity on the Rise. Technical report, STATEC and University of Luxembourg, Luxembourg.

Peter Gilles. 1999. *Dialektausgleich im Lëtzebuergeschen: Zur phonetisch-phonologischen Fokussierung einer Nationalsprache*. Niemeyer, Tübingen, Germany.

Peter Gilles. 2023. Luxembourgish. In Sebastian Kürschner and Antje Dammel, editors, *Oxford Encyclopedia of Germanic Linguistics*. Oxford University Press, Oxford.

Kyle Gorman and Yuval Pinter. 2025. Don't Touch My Diacritics. In *Proceedings of NAACL-HLT*.

Jack Grieve, Sara Bartl, Matteo Fuoli, Jason Grafmiller, Weihang Huang, Alejandro Jawerbaum, Akira Murakami, Marcus Perlman, Dana Roemling, and Bodo Winter. 2025. The sociolinguistic foundations of language modeling. *Frontiers in Artificial Intelligence*, Volume 7 - 2024.

Bo Han and Timothy Baldwin. 2011. Lexical Normalisation of Short Text Messages: Makn Sens a #twitter. In *Proceedings of ACL*.

Michael A. Hedderich, Lukas Lange, Heike Adel, Jannik Strötgen, and Dietrich Klakow. 2021. A Survey on Recent Approaches for Natural Language Processing in Low-Resource Scenarios. In *Proceedings of NAACL-HLT*.

Heinz Kloss. 1967. Abstand Languages and Ausbau Languages. *Anthropological Linguistics*, 9(7):29–41.

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Ortiz Suarez, Iroro Orife, Kelechi Ogueji, Andre Niyongabo Rubungo, Toan Q. Nguyen, Mathias Müller, André Müller, Shamsuddeen Hassan Muhammad, Nanda Muhammad, Ayanda Mnyakeni, Jamshidbek Mirzakhalov, Tapiwanashe Matangira, Colin Leong, Nze Lawson, Sneha Kudugunta, Yacine Jernite, Mathias Jenny, Orhan Firat, Bonaventure F. P. Dossou, Sakhile Dlamini, Nisansa de Silva, Sakine Çabuk Ballı, Stella Biderman, Alessia Battisti, Ahmed Baruwa, Ankur Bapna, Pallavi Baljekar, Israel Abebe Azime, Ayodele Awokoya, Duygu Ataman, Orevaoghene Ahia, Oghenefego Ahia, Sweta Agrawal, and Mofetoluwa Adeyemi. 2022. Quality at a Glance: An Audit of Web-Crawled Multilingual Datasets. *Transactions of ACL*, 10:50–72.

Alfred Lameli. 2013. *Strukturen im Sprachraum: Analysen zur arealtypologischen Komplexität der Dialekte in Deutschland*, volume 54 of *Linguistik – Impulse und Tendenzen*. De Gruyter, Berlin, Boston.

Mingfei Lau, Qian Chen, Yeming Fang, Tingting Xu, Tongzhou Chen, and Pavel Golik. 2025. Data Quality Issues in Multilingual Speech Datasets: The Need for Sociolinguistic Awareness and Proactive Language Planning. In *Proceedings of ACL*.

Heather Lent. 2025. Beyond Weaponization: NLP Security for Medium and Lower-Resourced Languages in Their Own Right.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual Denoising Pre-training for Neural Machine Translation. *Transactions of ACL*, 8:726–742.

Cedric Lothritz, Bertrand Lebichot, Kevin Allix, Lisa Veiber, Tegawende Bissyande, Jacques Klein, Andrey Boytsov, Clément Lefebvre, and Anne Goujon. 2022. LuxemBERT: Simple and Practical Data Augmentation in Language Model Pre-Training for Luxembourgish. In *Proceedings of LREC*.

Anne-Marie Lutgen, Alistair Plum, Christoph Purschke, and Barbara Plank. 2025. Neural Text Normalization for Luxembourgish Using Real-Life Variation Data. In *Proceedings of VarDial*.

Nina Markl, Lauren Hall-Lew, and Catherine Lai. 2024. Language Technologies as If People Mattered: Centering Communities in Language Technology Development. In *Proceedings of LREC-COLING*.

Dong Nguyen, Laura Rosseel, and Jack Grieve. 2021. On learning and representing social meaning in NLP: a sociolinguistic perspective. In *Proceedings of NAACL-HLT*.

Jessica Ojo, Odunayo Ogundepo, Akintunde Oladipo, Kelechi Ogueji, Jimmy Lin, Pontus Stenetorp, and David Ifeoluwa Adelani. 2025. AfroBench: How Good are Large Language Models on African Languages? In *Findings of ACL*.

Pedro Javier Ortiz Suárez, Benoit Sagot, and Laurent Romary. 2019. Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures. Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7) 2019. Cardiff, 22nd July 2019, pages 9 – 16, Mannheim. Leibniz-Institut f"ur Deutsche Sprache.

Barbara Plank. 2016. What to do about non-standard (or non-canonical) language in nlp. In *KONVENS*.

Barbara Plank, Kristian Nørgaard Jensen, and Rob van der Goot. 2020. DaN+: Danish Nested Named Entities and Lexical Normalization. In *Proceedings of COLING*.

Alistair Plum, Tharindu Ranasinghe, and Christoph Purschke. 2025. Text Generation Models for Luxembourgish with Limited Data: A Balanced Multilingual Strategy. In *Proceedings of VarDial*.

Christoph Purschke. 2019. *Vom Sprechen zur Sprache. Versuch über die variationslinguistische Praxis des Begrenzens*, pages 9–30. De Gruyter, Berlin, Boston.

Christoph Purschke. 2020. Attitudes Toward Multilingualism in Luxembourg. A Comparative Analysis of Online News Comments and Crowdsourced Questionnaire Data. *Frontiers in AI*, 3:536086.

Christoph Purschke. 2025. Discourse Figures in the Luxembourg Language Debate (2015–2020). *Zeitschrift für Sprachvariation und Soziolinguistik*, 1(2):37–53.

Bjørn T. Ramberg and Unn Røyneland. 2025. Norm at play. *Sociolinguistica*, 39(2).

Alan Ramponi. 2024. Language Varieties of Italy: Technology Challenges and Opportunities. *Transactions of ACL*, 12:19–38.

Tharindu Ranasinghe, Alistair Plum, Christoph Purschke, and Marcos Zampieri. 2023. Publish or Hold? Automatic Comment Moderation in Luxembourgish News Articles. In *Proceedings of RANLP*.

Anna-Sabrina Sattler. 2021. *Curriculumentwicklung in einer mehrsprachigen Gesellschaft: Das Beispiel Luxemburg*. Ph.D. thesis, Unilu - University of Luxembourg, Esch-sur-Alzette, Luxembourg.

Jürgen Erich Schmidt. 2010. *12. Language and space: The linguistic dynamics approach*, pages 201–225. De Gruyter Mouton, Berlin, New York.

Mark Sebba. 2007. *Introduction: society and orthography*, page 1–9. Cambridge University Press.

STATEC. 2024. *Luxembourg in Figures 2024*. STATEC, Luxembourg City.

Jiao Sun, Thibault Sellam, Elizabeth Clark, Tu Vu, Timothy Dozat, Dan Garrette, Aditya Siddhant, Jacob Eisenstein, and Sebastian Gehrmann. 2023. Dialect-robust Evaluation of Generated Text. In *Proceedings of ACL*.

Bibek Upadhayay and Vahid Behzadan. 2025. Tongue-Tied: Breaking LLMs Safety Through New Language Learning. In *Proceedings of the 7th Workshop on Computational Approaches to Linguistic Code-Switching*.

Rob van der Goot. 2019. MoNoise: A Multi-lingual and Easy-to-use Lexical Normalization Tool. In *Proceedings of ACL (System Demonstrations)*.

Rob van der Goot. 2025. Identifying Open Challenges in Language Identification. In *Proceedings of ACL*.

Rob van der Goot and Özlem Çetinoğlu. 2021. Lexical Normalization for Code-switched Data and its Effect on POS Tagging. In *Proceedings of EACL*.

Rob van der Goot, Alan Ramponi, Arkaitz Zubiaga, Barbara Plank, Benjamin Muller, Iñaki San Vicente Roncal, Nikola Ljubešić, Özlem Çetinoğlu, Rahmad Mahendra, Talha Çolakoğlu, Timothy Baldwin, Tommaso Caselli, and Wladimir Sidorenko. 2021. MultiLexNorm: A Shared Task on Multilingual Lexical Normalization. In *Proceedings of W-NUT 2021*.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355.

Anna Wegmann, Dong Nguyen, and David Jurgens. 2025. Tokenization is Sensitive to Language Variation. In *Findings of ACL*.

Orion Weller, Kathryn Ricci, Marc Marone, Antoine Chaffin, Dawn Lawrie, and Benjamin Van Durme. 2025. Seq vs Seq: An Open Suite of Paired Encoders and Decoders.

Ruth Wodak, Barbara Johnstone, and Paul Kerswill. 2011. *The SAGE Handbook of Sociolinguistics*. SAGE Publications Ltd, London.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mt5: A massively multilingual pre-trained text-to-text transformer. In *North American Chapter of the Association for Computational Linguistics*.

Marcos Zampieri, Preslav Nakov, and Yves Scherrer. 2020. Natural language processing for similar languages, varieties, and dialects: A survey. *Natural Language Engineering*, 26(6):595–612.

Zenter fir d'Lëtzebuerger Sprooch, editor. 2019. *D'Lëtzebuerger Orthografie*. Zenter fir d'Lëtzebuerger Sprooch, Stroossen.

9. Language Resource References

David Ifeoluwa Adelani and Hannah Liu and Xiaoyu Shen and Nikita Vassilyev and Jesujoba O. Alabi and Yanke Mao and Haonan Gao and Annie En-Shiun Lee. 2023. *SIB-200: A Simple, Inclusive, and Big Evaluation Dataset for Topic Classification in 200+ Languages and Dialects*.

Lothritz, Cedric and Lebichot, Bertrand and Allix, Kevin and Veiber, Lisa and Bissyande, Tegawende and Klein, Jacques and Boytsov, Andrey and Lefebvre, Clément and Goujon, Anne. 2022. *LuxemBERT: Simple and Practical Data Augmentation in Language Model Pre-Training for Luxembourgish*. European Language Resources Association.

Christoph Purschke. 2020. Attitudes Toward Multilingualism in Luxembourg. A Comparative Analysis of Online News Comments and Crowdsourced Questionnaire Data. *Frontiers in AI*, 3:536086.

Ranasinghe, Tharindu and Plum, Alistair and Purschke, Christoph and Zampieri, Marcos. 2023. *Publish or Hold? Automatic Comment Moderation in Luxembourgish News Articles*.

10. Appendices

The dataset description for each classification task is presented in Section 10.1, with respective sizes and hyperparameter settings, shown respectively in Tables 5 and 6. Section 10.2 contains the results for topic classification, WNLI, NER and sentiment classification (shown in Tables 7, 8, 9 and 10), which were not presented in the main results section.

10.1. Datasets

Task	Train	Dev	Test
IC	698	149	159
NER	4298	459	770
POS	4278	460	388
WNLI	568	63	136
CM	6000	1000	2000
SC	1299	185	364
TC	701	99	204

Table 5: Number of sentences in the training, development, and test sets for each task.

Task	batch size	LR	epochs
IC, WNLI, CM, SC	16	5^{-5}	5
NER, POS	16	5^{-5}	3
TC	16	2^{-5}	5

Table 6: Hyperparameter settings for each task, inspired by Lothritz et al. (2022)

Intent classification (IC) entails detecting the intent or goal of a text. It is a multi-class classification task that was created by Lothritz et al. (2022) and consists of a Banking Client Support Dataset.

Topic Classification (TC) is part of the SIB-200 dataset and is annotated with topic labels. The topics include science/technology, travel, politics, sports, health, entertainment and geography (Adelani et al., 2023).

Winograd Natural Language Inference (WNLI) is an inference task and part of the GLUE benchmark (Wang et al.). Lothritz et al. (2022) translated the dataset to Luxembourgish and it consists of two sentences and a label (1 or 0) for each pair.

Part-of-Speech Tagging (POS) is a word classification task that assigns a word class to each token. This dataset was annotated automatically with the spaCy pipeline and checked by a human annotator (Lothritz et al., 2022).

Named Entity Recognition (NER) was also created by Lothritz et al. (2022) and is the same dataset as the POS one. It was manually annotated with five labels: Person, Organisation, (natural) Location, Geopolitical Entity, and Miscellaneous.

Sentiment Classification (SC) is a manually annotated subset of online comments on RTL and is annotated by the sentiment of the comment, so either positive, negative or neutral.

Comment Moderation (CM) is a subset of online comments from RTL, who have manually moderated the comments. The task is to classify whether a comment was archived or published (Ranasinghe et al., 2023).

10.2. Results

Model		std	n-std	comb
LBERT	std	74.75 ± 2.10	66.13 ± 1.78	70.62 ± 1.39
LBERT	n-std	61.84 ± 2.11	56.61 ± 1.54	59.59 ± 1.24
LBERT	comb	81.16 ± 1.07	77.47 ± 1.26	79.32 ± 1.14
mBERT	std	72.74 ± 6.07	66.60 ± 8.30	69.77 ± 6.91
mBERT	n-std	75.35 ± 4.52	75.48 ± 4.47	75.43 ± 4.41
mBERT	comb	82.74 ± 0.94	82.09 ± 0.83	82.43 ± 0.57

Table 7: TC – Weighted F1 (± std)

Model		std	n-std	comb
LBERT	std	50.31 ± 0.71	50.74 ± 2.22	50.57 ± 1.25
LBERT	n-std	50.31 ± 0.71	50.74 ± 2.22	50.57 ± 1.25
LBERT	comb	50.31 ± 0.71	50.74 ± 2.22	50.57 ± 1.25
mBERT	std	52.66 ± 2.80	50.49 ± 1.15	52.10 ± 1.47
mBERT	n-std	52.80 ± 2.48	51.38 ± 1.83	52.30 ± 1.52
mBERT	comb	54.86 ± 3.46	52.00 ± 3.70	54.06 ± 2.91

Table 8: WNLI – Weighted F1 (± std)

Model		std	n-std	comb
LBERT	std	65.62 ± 0.00	60.10 ± 0.01	62.84 ± 0.00
LBERT	n-std	65.00 ± 0.00	62.41 ± 0.00	63.67 ± 0.00
LBERT	comb	65.94 ± 0.03	63.52 ± 0.03	64.71 ± 0.02
mBERT	std	68.23 ± 0.00	55.82 ± 0.00	61.92 ± 0.00
mBERT	n-std	68.60 ± 0.01	64.84 ± 0.02	66.71 ± 0.01
mBERT	comb	66.82 ± 0.01	63.81 ± 0.01	65.35 ± 0.01

Table 9: NER – Weighted F1 (± std)

Model		std	n-std	comb
LBERT	std	62.37 ± 0.66	62.40 ± 0.53	62.40 ± 0.52
LBERT	n-std	62.94 ± 0.50	64.04 ± 0.83	63.49 ± 0.48
LBERT	comb	63.37 ± 0.04	64.13 ± 0.13	63.75 ± 0.08
mBERT	std	55.78 ± 1.53	54.51 ± 1.24	55.16 ± 1.36
mBERT	n-std	55.15 ± 1.14	55.55 ± 1.18	55.36 ± 1.07
mBERT	comb	56.83 ± 0.15	57.73 ± 0.29	57.29 ± 0.20

Table 10: SC – Weighted F1 (± std)