Publish or Hold? Automatic Comment Moderation in Luxembourgish News Articles

Tharindu Ranasinghe[♡], **Alistair Plum**[♢], **Christoph Purschke**[♢], **Marcos Zampieri**[♠]
[♡]Aston University, Birmingham, UK
[♢]University of Luxembourg, Esch-sur-Alzette, Luxembourg
[♠]George Mason University, Fairfax, VA, USA
t.ranasinghe@aston.ac.uk, alistair.plum@uni.lu
christoph.purschke@uni.lu, mzampier@gmu.edu

Abstract

Recently, the internet has emerged as the primary platform for accessing news. In the majority of these news platforms, the users now have the ability to post comments on news articles and engage in discussions on various social media. While these features promote healthy conversations among users, they also serve as a breeding ground for spreading fake news, toxic content, and hate speech. Moderating or removing such content is paramount to avoid unwanted consequences for the readers. However, apart from a few notable exceptions, most research on the automatic moderation of news article comments has dealt with English and other high-resource languages. This leaves under-represented or low-resource languages at a loss. Addressing this gap, we perform the first large-scale qualitative analysis of more than one million Luxembourgish comments posted over the course of 14 years. We evaluate the performance of state-of-the-art transformer models in Luxembourgish news article comment moderation. Furthermore, we analyse how the language of Luxembourgish news article comments has changed over time. We observe that machine learning models trained on old comments do not perform well on recent data. The findings in this work will be beneficial in building news comment moderation systems for many low-resource languages.

1 Introduction

In recent years, the Internet has revolutionised how individuals access and consume news. With the popularity of smart devices such as phones and tablets, the Internet has emerged as the primary medium for acquiring news and information (Kwak et al., 2010). People often share news articles on social media using these devices and discuss them with their friends. At the same time, news websites also allow users to post comments and discuss stories (Zannettou et al., 2017).

While the inclusion of comment sections provides users with a platform to engage in constructive discussions regarding news stories, these discussions can also devolve into the expression of offensive remarks and hate speech (Erjavec and Kovačič, 2012; Davidson et al., 2017; Chowdhury et al., 2020). Furthermore, malicious users can exploit discussion platforms to intentionally spread misinformation, often in the form of fake news, to mislead and provoke readers (Risch and Krestel, 2018; Yanagi et al., 2020). The wide spread of inappropriate comments motivates the use of content moderation to avoid further undesirable consequences.

Moderating comment sections is a difficult task, mainly due to how widely the content can range, including fake news (Patwa et al., 2021) and various forms of offensive speech (Risch and Krestel, 2018; Napoles et al., 2017; Zampieri et al., 2019a; Weerasooriya et al., 2023). Detecting these varied types of content is difficult for humans alone, and in addition, the sheer number of comments that can be generated by any comment section makes manual moderation an overwhelming and costly task (Djuric et al., 2015). Many approaches in NLP are dedicated to identifying fake news (Yanagi et al., 2020; Nguyen et al., 2020), hate speech (Mollas et al., 2022), and related phenomena. However, as is often the case, these approaches focus on English and other high-resource languages (Schmidt and Wiegand, 2017). With the increasing prevalence of smart devices, a significant number of individuals prefer to express their thoughts and opinions in their native languages. Consequently, there is a pressing demand for systems that can cater to each language. Unfortunately, the lack of language resources poses a significant challenge in developing such systems, particularly for low-resource languages (Zampieri et al., 2022; Gaikwad et al., 2021).

Proceedings of Recent Advances in Natural Language Processing, pages 968–978
Varna, Sep 4–6, 2023
https://doi.org/10.26615/978-954-452-092-2_104

In this paper, we experiment with automatic content moderation for Luxembourgish, a West Germanic language spoken by around 400,000 people, primarily in Luxembourg. We use state-of-the-art multi- and cross-lingual language models, as well as a recently released model for Luxembourgish specifically. Using a dataset provided by the main news broadcaster in Luxembourg, we trained a number of models to predict whether a given comment should be archived or not, according to the internal policy of the dataset provider. As such, this presents the first real evaluation of such an approach in the field of automatic content moderation, as well as its sub-tasks, for Luxembourgish. Additionally, it has been demonstrated that the case of Luxembourgish is unique, offering resources for research but being under-represented in research (Adda-Decker et al., 2008; Purschke, 2020).

This paper answers two research questions:

- **RQ1** - How do the state-of-the-art transformer models perform in automatic content moderation in Luxembourgish?
- **RQ2** - What is the validity of the content moderation models trained on old data?

The remainder of this paper is structured as follows. Section 2 presents an overview of related work in the field. Section 3 describes the dataset used for the experiments, followed by a description of the employed methodology in Section 4. The results of the experiments are presented in Section 5. Finally, Section 6 offers our future plans as well as concluding remarks.

2 Related Work

Automatic content moderation is a challenging and interesting task which has attracted the attention of the NLP community for many years. Content moderation involves a number of sub-tasks in NLP, mainly including racism and hate speech detection, as well as fake news detection and irony and sarcasm detection.

Offensive Content Detecting and classifying offensive content has been studied extensively both for news comments and social media posts. Early approaches have applied traditional machine learning classifiers to the task, while more recent work has applied neural networks (Schmidt and Wiegand, 2017; Ranasinghe et al., 2019). Most of the datasets and approaches have been based on English (Salminen et al., 2018). Nevertheless, research is also conducted on Croatian (Shekhar et al., 2020; Ljubešić et al., 2018), Estonian (Shekhar et al., 2020), German (Assenmacher et al., 2021), Korean (Moon et al., 2020), and Slovene (Ljubešić et al., 2018) on detecting offensive content in news media comments. There is also a rise in shared tasks on the topic, notably SemEval 2019 Task 6 (OffensEval), which treated the identification and categorisation of offensive language on social media for English, attracting over 800 teams with 115 final submissions (Zampieri et al., 2019b). Moreover, there have been shared tasks for various languages, including German (Struß et al., 2019), Bangla (Kumar et al., 2020), Hindi (Modha et al., 2022), as well as multilingual (Zampieri et al., 2020) and code-mixed (Chakravarthi et al., 2020; Satapara et al., 2023) settings.

Misinformation Misinformation detection in news media comments is another sub-task that has caught the attention of the NLP community, as many malicious users exploit discussion platforms to spread misinformation intentionally (Risch and Krestel, 2018). However, not much work has been done on detecting misinformation in news media comments (Sharma et al., 2019). On the other hand, there have been several works on misinformation detection in social media posts, which are also focused on English and other high-resource languages (Uyangodage et al., 2021). However, fake news detection remains a complex task in NLP (Ali et al., 2022). While various current architectures have been trained for this task, it is said that these approaches require more complex ensembles of architectures to accurately predict fake news segments, particularly shorter ones (Ali et al., 2022).

Resources for Luxembourgish In general, Luxembourgish is said to be under-represented in NLP, particularly because it is a relatively small language, especially compared to its linguistic neighbours, French and German. This can be attributed to the relatively recent development of the written domain in Luxembourgish that has largely been fostered by the advent of social media. However, resources are steadily increasing. Gierschek (2022) developed a state-of-the-art pipeline for sentiment analysis based on the same dataset as our study. Purschke (2020) published a pipeline for

the automatic orthographic correction of text data,[1] i.a. based on correction data from spellchecker.lu, an online spellchecking tool for Luxembourgish.[2] Additionally, the Luxembourgish Online Dictionary (LOD) recently launched an open API to its lexical resources.[3] Lothritz et al. (2021) introduced an intent classification dataset for Luxembourgish, which contains 1006 instances divided into 28 different intents related to banking requests such as opening/closing a bank account or ordering/blocking a credit card. The Winograd Natural Language Inference task which is part of the GLUE benchmark collection (Wang et al., 2018) contains more than 750 instances in Luxembourgish. With recent advances in neural networks, there now exists a language model for Luxembourgish, LUXEMBERT (Lothritz et al., 2022), which we also use for the purposes of this paper. With LUXEMBERT, Lothritz et al. (2022) introduced several language resources for Luxembourgish, including part-of-speech tagging, named entity recognition and news classification. At the time of writing, there is no published research on work related to content moderation in Luxembourgish.

3 Data

The dataset used for the purposes of this paper was provided by the RTL media group, the largest news provider in Luxembourg. The dataset provided stems from their own news platform,[4] which has existed since 2008 and is the only news offering that is entirely in Luxembourgish. Given the recent expansion of Luxembourgish into the written domain and the central role of RTL in the country's media system, for many Luxembourgers, the RTL news platform has been one of their main points of contact with written Luxembourgish, apart from private messaging. Against this backdrop, our data represents not only the largest collection of written texts in Luxembourgish currently available, but also a crucial source for studying the development of written Luxembourgish in real time.

For the purposes of this paper, we work exclusively with user comments, comprising over one million comments posted on around 61,000 news articles over the course of a 14 year time-span, starting in 2008. Each comment includes manual content moderation information provided by a number of dedicated content moderators over the years, with labels assigned according to a step in the moderation process. While the label *published* should be clear, three others indicate that the given comment has been moderated or archived (and there may be other moderation steps to be taken). We treat these three labels here as *archived* (meaning not published). It should be made clear at this point, that for the years 2008-2010 all comments are labelled *published*. This is an error in our iteration of the dataset and has resulted in this data being excluded.

Year	Archived	Published
2011	1766	53368
2012	10791	81795
2013	10592	76835
2014	12368	65723
2015	8213	46239
2016	8548	57959
2017	14690	51686
2018	14988	77898
2019	18049	74404
2020	44810	142654
2021	28352	70368
2022	19280	61482
Sum	**192447**	**860411**

Table 1: Number of instances per year in the dataset labelled as *archived* or *published*.

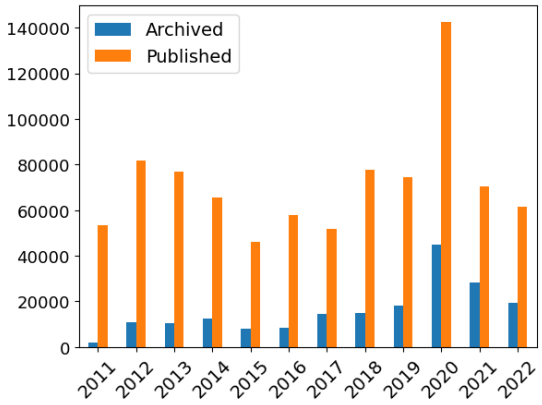


Figure 1: Proportion of labels over the years.

Table 1 shows the proportion of labels overall, and for each year in the dataset. We observe roughly the same proportion each year, which is also highlighted by Figure 1. We see here also that roughly each year the same number of comments are made, with the exception being 2020, the first

[1] https://github.com/questoph/spellux/
[2] https://spellchecker.lu
[3] https://lod.lu/api/doc
[4] https://rtl.lu

year of the COVID-19 pandemic, where there were almost double the number of comments than usual.

In terms of preprocessing, the comments have to be cleaned of special characters, incorrect encodings and markup language. Since the platform has undergone some changes in its technical implementation, various markup standards are represented and need to be removed. In addition, various text encodings need to be converted to Unicode, and special characters and embedded content need to be removed. All preprocessing steps were carried out in a dedicated Python pipeline.

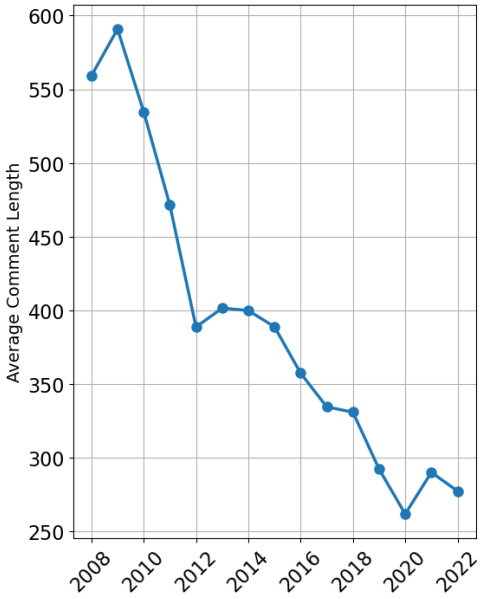


Figure 2: Average comment length over the years.

The mean comment length is 352 characters, with the median lying at 220 characters. The shortest comment is one character in length, with the longest comment being 34,597 characters in length. Figure 2 shows the average length of comments over the years represented in the dataset, which highlights the fact that the comment length has gone down by almost 50% since 2008. Interestingly, the lowest average comment length was recorded in 2020, the same year that has by far the highest number of comments on a yearly basis.

Luxembourg is a multilingual country, with German, French and Luxembourgish recognised as official languages, although with different domain allocations in administration and everyday practice (Horner and Weber, 2008). While French and German are the main administrative languages, Luxembourgish has the status of the national language. French is the language of legislation, and German serves as the language for alphabetisation. It also holds, for historical reasons, an important position in print media, whereas Luxembourgish has only recently developed from a predominately spoken into a written variety that is suitable for all social domains (Gilles, 2019). Furthermore, due to the country's migration and industrial history, Portuguese and Italian are considered important minority languages. Nowadays, cross-border commuting and the international workforce in the finance industry put pressure on the traditional language regime, with French and English gaining more ground. This complex multilingualism is, of course, reflected in the corpus, with instances of code switching on the comment level, but also answers in French or German to Luxembourgish comments are not uncommon in the dataset.

To investigate the language representation further, we processed all comments with the *langdetect* package available for Python.[5] As Luxembourgish is not available for this package, we used a custom profile, which has been trained previously for the recognition of Luxembourgish, based on the RTL news articles (Purschke, 2020). Detection accuracy for Luxembourgish works reliably (100%) using a random sample of 1,000 texts. For non-Luxembourgish texts, accuracy is around 96% for texts longer than 200 characters, but drops to 64% for short texts that do not offer many language-specific patterns.

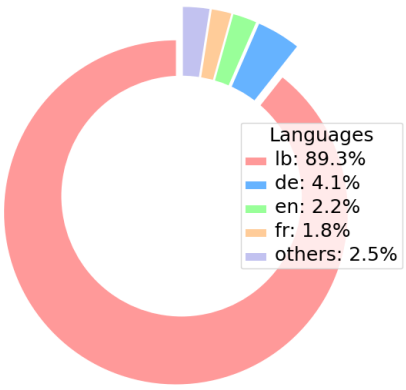


Figure 3: Languages represented in the dataset.

Figure 3 shows the percentage of the top four languages detected automatically in the dataset, with all others grouped together. Although these results are not necessarily representative: Luxembourgish language detection is an area of ongoing research and can often be misclassified as French (due to

[5] https://pypi.org/project/langdetect/

many loan words) and German (due to the two being closely related). In addition, the full list of detected languages comprises about 30 languages, including Languages such as Chinese, which are not very likely, although it should not be dismissed entirely. Further analysis has shown that many labels are assigned based on one word, hinting again at mislabelling.

4 Methodology

To investigate the research questions posed in Section 1, we carried out the following steps. First, the data was processed and cleaned. Next, we trained various language models on the task of classifying the comments into two groups. Following this, we experimented with the composition of the training set, limiting it to certain years and testing the effectiveness on the most recent year.

4.1 Encoder Transformers

We first experimented with encoder transformers, which have provided excellent results in various NLP tasks, including text classification (Li et al., 2022). From an input sentence, they compute a feature vector $\boldsymbol{h} \in \mathbb{R}^d$, upon which we built a classifier for the task.

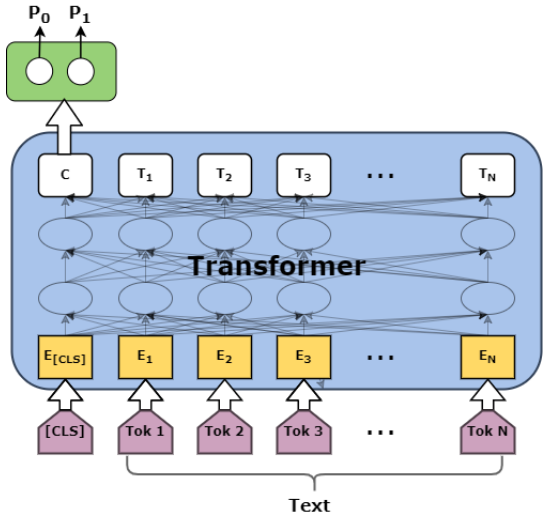


Figure 4: A schematic representation of the transformer models in classification (Ranasinghe and Zampieri, 2020).

For this task, we implemented a softmax layer, i.e., the predicted probabilities are $\boldsymbol{y}^{(B)} = \text{softmax}(W\boldsymbol{h})$, where $W \in \mathbb{R}^{k \times d}$ is the softmax weight matrix, and k is the number of labels which in our case is two. This architecture is depicted in Figure 4. We employed a batch size of 32, Adam optimiser with learning rate $2\text{e}{-5}$, and a linear learning rate warm-up over 10% of the training data. During the training process, the parameters of the transformer model, as well as the parameters of the subsequent layers, were updated. The models were evaluated while training using an evaluation set that had one-fifth of the rows in data. We performed early stopping if the evaluation loss did not improve over three evaluation steps. All the models were trained for three epochs. From this type of transformer, we experimented with BERT-BASE-MULTILINGUAL-CASED (Devlin et al., 2019), XLM-ROBERTA-BASE (Conneau et al., 2020) and XLM-ROBERTA-LARGE (Conneau et al., 2020). All of these models have been used widely in multilingual text classification (Ranasinghe and Zampieri, 2021). In addition to them, we also used LUXEMBERT (Lothritz et al., 2022), which is trained specifically on Luxembourgish. We trained the models using a cluster of ten NVIDIA RTX A6000 48GB GPUs. All the pre-trained transformer models we used for the experiments are available on HuggingFace (Wolf et al., 2020).

4.2 Text-to-text Transformers

We also experimented with several state-of-the-art text-to-text transformers, which treat all tasks as text generation problems. These transformers have provided excellent results in text classification tasks (Bulla et al., 2023; Sabry et al., 2022; Ni et al., 2022). They do not rely on a classification layer (Raffel et al., 2020) and have a flexible input-output format. The input texts to the model were the comments, and output texts were labelled *Archived* if the text is archived and *Published* if they are published, as shown in Figure 5. We used a batch size of 16, Adam optimizer with learning rate $1\text{e}{-4}$, and a linear learning rate warm-up over 10% of the training data and trained the models over ten epochs. From this type of transformer, we experimented with MT5-BASE (Xue et al., 2021), MT5-LARGE (Xue et al., 2021), BYT5-BASE (Xue et al., 2022) and BYT5-LARGE (Xue et al., 2022). MT5 models support Luxembourgish. On the other hand, byt5 models follow a tokenizer-free approach and are more suitable for tasks involving code-switching and code-mixing (Xue et al., 2022). We trained the models using a cluster of ten NVIDIA RTX A6000 48GB GPUs.

	Archived			Published			Weighted Average			
Model	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**F1 Macro**
XLM-R BASE	0.68	0.15	0.24	0.79	0.97	0.87	0.78	0.76	0.73	0.56
XLM-R LARGE	0.67	0.17	0.26	0.79	0.97	0.87	0.78	0.77	0.75	0.57
MBERT	0.58	0.06	0.12	0.77	0.97	0.86	0.72	0.77	0.70	0.49
LUXEMBERT	0.60	0.08	0.15	0.78	0.98	0.87	0.73	0.77	0.70	0.51
MT5 BASE	0.61	0.06	0.11	0.77	0.98	0.87	0.74	0.77	0.69	0.49
MT5 LARGE	0.64	0.10	0.15	0.78	0.98	0.87	0.75	0.76	0.72	0.51
BYT5 BASE	0.65	0.17	0.27	0.79	0.97	0.87	0.76	0.78	0.73	0.57
BYT5 LARGE	0.67	0.20	0.31	0.79	0.98	0.88	0.77	0.78	0.74	**0.59**
ALL ARCHIVED	0.23	1.00	0.37	0.00	0.00	0.00	0.05	0.23	0.08	0.18
ALL PUBLISHED	0.00	0.00	0.00	0.76	1.00	0.86	0.59	0.76	0.66	0.43

Table 2: Results for content moderation with default settings. For each model, Precision (P), Recall (R), and F1 are reported on all classes, and weighted averages. Macro-F1 is also listed.

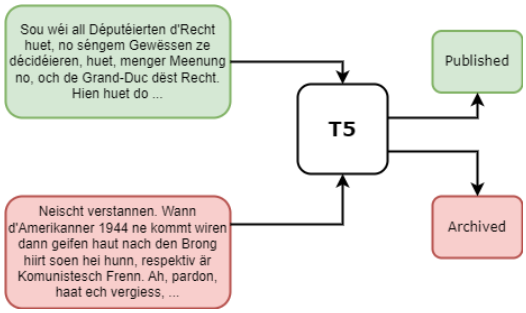


Figure 5: A schematic representation of the text-text transformer models in classification (Raffel et al., 2020).

5 Results

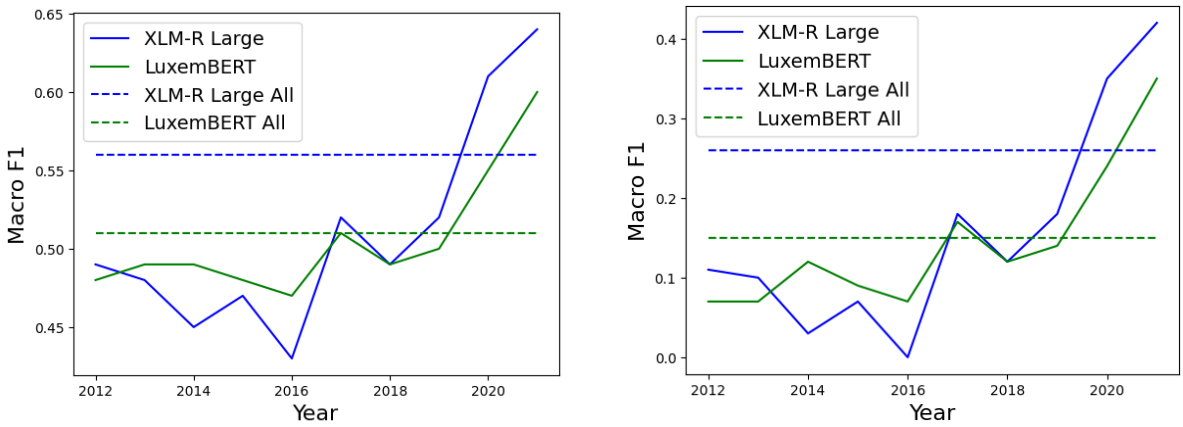
We first concatenated all the comments from 2011-2021 as the training set. The comments from 2022 were considered as the test set. We trained all the models described in Section 4 under this setting. The results of these models are shown in Table 2. As the label distribution is highly imbalanced, we evaluate and compare the performance of the different models using the Macro F1-score. Furthermore, a classifier that can correctly identify both classes would protect freedom of expression while moderating the unwanted texts. We further report per-class Precision (P), Recall (R), F1-score (F1), and weighted averages. Finally, we compare the performance of the models against simple majority and minority class baselines.

As can be seen in Table 2, most of the state-of-the-art transformer models perform reasonably well in automatic content moderation in Luxembourgish. We can see that all models perform significantly better than simple majority and minority class baselines. BYT5 LARGE (Xue et al., 2022) model performed best by giving a 0.59 Macro F1 score, closely followed by XLM-R LARGE (Conneau et al., 2020), BYT5 BASE (Xue et al., 2022), and XLM-R BASE (Conneau et al., 2020).

Interestingly, the LUXEMBERT model (Lothritz et al., 2022), which was built on Luxembourgish text, did not perform well compared to other models in this task. Models such as XLM-R, which do not support Luxembourgish, outperform LUXEMBERT. We assume that this can be due to two reasons; (*i*) the texts used to train the models are heavily code-switched and code-mixed. XLM-R models have an advantage over this. This is further confirmed by the superior performance of BYT5 models. BYT5 models follow a tokenizer-free approach and, therefore, perform well in code-switched and code-mixed texts. (*ii*) XLM-R models provide stronger models compared to LUXEMBERT. Overall, we can see that it is advantageous to use XLM-R rather than language-specific LUXEMBERT.

All the models we experimented with performed poorly in identifying the *Archived* class. The best model, BYT5 LARGE, only had an F1 score of 0.31 for the *Archived* class. Scores of the *Published* class were better and consistent across the models. We assume that identifying *Archived* comments is challenging for machine learning models, as there are many reasons why a comment could have been archived, including but not limited to the sub-tasks of content moderation mentioned in Section 2. It is clear that this requires more research input and some insight into the moderation policy.

The BYT5-LARGE model took approximately 155 hours on an NVIDIA RTX A6000 48GB GPU



(a) Macro F1 score change with model training year

(b) F1 score for archived class change with model training year

Figure 6: F1 score change with model training year. Dotted line shows the result from Table 2 for each model, where the models were trained on all the instances from 2008-2021.

to train. The XLM-R LARGE model took 83 hours, and LUXEMBERT only took 44 hours to train on the same GPU. Therefore, even though BYT5-LARGE provided the best result for our task, it is not the most computationally efficient model.

With these results, we answer **RQ1:** How do the state-of-the-art transformer models perform in automatic content moderation in Luxembourgish? We showed that several transformer models perform fairly well in the task. However, the models do not provide impressive results, and this task requires more attention from the NLP community for low-resource languages such as Luxembourgish.

Validity of the content moderation models trained on old data In order to answer our **RQ2**, we changed our training data. We kept the testing set similar to the above experiment by having all the instances from 2022 as the test set. In the first experiment, we only had instances from 2012 as the training set and trained transformer models using a similar configuration we mentioned in Section 4. We repeated the experiments for 2012, 2014, 2015 and up to 2021. As the instances from 2008-2011 did not have any archived instances, we dropped these years from our experiments. We only conducted these experiments for LUXEMBERT and XLM-R LARGE, as BYT5 models were computationally expensive. Figure 6a shows the variation of the macro F1 score and Figure 6b shows the variation of the F1 score of the *Archived* class with each training year.

As can be seen in the graphs, models trained on recent years' data provided better results in content moderation. Most of the models trained before 2015 provided very poor results when evaluated on 2022 data. However, the models trained on recent data, especially after 2019, provided promising results and performed better than earlier models, which were trained on all data from 2012-2021. As shown in Figure 6b, the F1 score for the *Archived* class followed a similar pattern. However, we noticed that the results for the *Published* class do not change with respect to the year.

With this, we answer our **RQ2**, the models trained on old data do not perform well on recent data for content moderation. Models trained on recent data performed better than models trained on data that includes both old and recent data. While this finding is against the popular belief that more data can lead to better results, we acknowledge the fact that the models trained on more related data can perform well in content moderation.

6 Conclusion

In this paper, we presented the first study on automatic comment moderation in Luxembourgish News Articles. Our study involved a comprehensive qualitative analysis of over one million Luxembourgish comments spanning a period of 14 years. The main objective was to evaluate the performance of various state-of-the-art multilingual, cross-lingual, and language-specific transformer models in the task of content moderation. Among these models, BYT5 LARGE (Xue et al., 2022) emerged as the best model, indicating that its tokenizer-free approach is particularly advantageous for handling the code-mixed and code-switched nature of Luxembourgish news comments.

While the transformer models overall produced satisfactory results, there remains significant room for improvement, especially when it comes to the *Archived* class. Additionally, our findings revealed that machine learning models trained on old data exhibit poor performance when applied to recent data on content moderation.

Our findings in this study will be beneficial for researchers working on automatic content moderation in low-resource languages. In future work, we hope to enhance the interpretability of the recommended machine learning models to better assist human content moderators in their decision-making process. By pursuing these avenues, we aim to contribute towards the advancement of automatic content moderation techniques while ensuring their alignment with human moderation needs.

Acknowledgments

We would like to thank the anonymous RANLP reviewers who have provided us with constructive feedback to improve the quality of this paper.

The computational experiments in this paper were conducted on the Aston EPS Machine Learning Server, funded by the EPSRC Core Equipment Fund, Grant EP/V036106/1.

References

Martine Adda-Decker, Thomas Pellegrini, Eric Bilinski, and Gilles Adda. 2008. Developments of "Lëtzebuergesch" Resources for Automatic Speech Processing and Linguistic Studies. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Abdullah Marish Ali, Fuad A. Ghaleb, Bander Ali Saleh Al-Rimy, Fawaz Jaber Alsolami, and Asif Irshad Khan. 2022. Deep Ensemble Fake News Detection Model Using Sequential Deep Learning Technique. *Sensors*, 22(18).

D Assenmacher, M Niemann, K Müller, M Seiler, D M Riehle, and H Trautmann. 2021. RP-Mod&RP-Crowd: Moderator-and crowd-annotated german news comment datasets. In *NeurIPS Datasets and Benchmarks*.

Luana Bulla, Aldo Gangemi, and Misael Mongiovi'. 2023. Towards Distribution-shift Robust Text Classification of Emotional Content. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8256–8268, Toronto, Canada. Association for Computational Linguistics.

Bharathi Raja Chakravarthi, Anand Kumar M, John P McCrae, B Premjith, KP Soman, and Thomas Mandl. 2020. Overview of the track on HASOC-Offensive Language Identification-DravidianCodeMix. In *FIRE (Working notes)*, pages 112–120.

Shammur Absar Chowdhury, Hamdy Mubarak, Ahmed Abdelali, Soon-gyo Jung, Bernard J Jansen, and Joni Salminen. 2020. A multi-platform Arabic news comment dataset for offensive language detection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6203–6212.

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 the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the international AAAI conference on web and social media*, volume 11, pages 512–515.

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

Nemanja Djuric, Jing Zhou, Robin Morris, Mihajlo Grbovic, Vladan Radosavljevic, and Narayan Bhamidipati. 2015. Hate Speech Detection with Comment Embeddings. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15 Companion, page 29–30, New York, NY, USA. Association for Computing Machinery.

Karmen Erjavec and Melita Poler Kovačič. 2012. "You Don't Understand, This is a New War!" Analysis of Hate Speech in News Web Sites' Comments. *Mass Communication and Society*, 15(6):899–920.

Saurabh Sampatrao Gaikwad, Tharindu Ranasinghe, Marcos Zampieri, and Christopher Homan. 2021. Cross-lingual Offensive Language Identification for Low Resource Languages: The Case of Marathi. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 437–443, Held Online. INCOMA Ltd.

Daniela Gierschek. 2022. *Detection of Sentiment in Luxembourgish User Comments*. Ph.D. thesis, University of Luxembourg.

Peter Gilles. 2019. 39. Komplexe Überdachung II: Luxemburg. Die Genese Einer Neuen Nationalsprache. In Joachim Herrgen and Jürgen Erich Schmidt, editors, *Sprache und Raum - Ein internationales Handbuch der Sprachvariation. Volume 4 Deutsch*, pages 1039–1060. De Gruyter Mouton, Berlin, Boston.

Kristine Horner and Jean Jacques Weber. 2008. The Language Situation in Luxembourg. *Current Issues in Language Planning*, 9(1):69–128.

Ritesh Kumar, Atul Kr. Ojha, Shervin Malmasi, and Marcos Zampieri. 2020. Evaluating Aggression Identification in Social Media. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 1–5, Marseille, France. European Language Resources Association (ELRA).

Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. 2010. What is Twitter, a Social Network or a News Media? In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, page 591–600, New York, NY, USA. Association for Computing Machinery.

Qian Li, Hao Peng, Jianxin Li, Congying Xia, Renyu Yang, Lichao Sun, Philip S. Yu, and Lifang He. 2022. A Survey on Text Classification: From Traditional to Deep Learning. *ACM Trans. Intell. Syst. Technol.*, 13(2).

Nikola Ljubešić, Tomaž Erjavec, and Darja Fišer. 2018. Datasets of Slovene and Croatian Moderated News Comments. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 124–131, Brussels, Belgium. Association for Computational Linguistics.

Cedric Lothritz, Kevin Allix, Bertrand Lebichot, Lisa Veiber, Tegawendé F. Bissyandé, and Jacques Klein. 2021. Comparing MultiLingual and Multiple Mono-Lingual Models for Intent Classification and Slot Filling. In *Natural Language Processing and Information Systems*, pages 367–375, Cham. Springer International Publishing.

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 the Thirteenth Language Resources and Evaluation Conference*, pages 5080–5089, Marseille, France. European Language Resources Association.

Sandip Modha, Thomas Mandl, Gautam Kishore Shahi, Hiren Madhu, Shrey Satapara, Tharindu Ranasinghe, and Marcos Zampieri. 2022. Overview of the HASOC Subtrack at FIRE 2021: Hate Speech and Offensive Content Identification in English and Indo-Aryan Languages and Conversational Hate Speech. In *Proceedings of the 13th Annual Meeting of the Forum for Information Retrieval Evaluation*, FIRE '21, page 1–3, New York, NY, USA. Association for Computing Machinery.

Ioannis Mollas, Zoe Chrysopoulou, Stamatis Karlos, and Grigorios Tsoumakas. 2022. ETHOS: a multi-label hate speech detection dataset. *Complex & Intelligent Systems*, 8(6):4663–4678.

Jihyung Moon, Won Ik Cho, and Junbum Lee. 2020. BEEP! Korean Corpus of Online News Comments for Toxic Speech Detection. In *Proceedings of the Eighth International Workshop on Natural Language Processing for Social Media*, pages 25–31, Online. Association for Computational Linguistics.

Courtney Napoles, Joel Tetreault, Aasish Pappu, Enrica Rosato, and Brian Provenzale. 2017. Finding Good Conversations Online: The Yahoo News Annotated Comments Corpus. In *Proceedings of the 11th Linguistic Annotation Workshop*, pages 13–23, Valencia, Spain. Association for Computational Linguistics.

Van-Hoang Nguyen, Kazunari Sugiyama, Preslav Nakov, and Min-Yen Kan. 2020. Fang: Leveraging social context for fake news detection using graph representation. In *Proceedings of the 29th ACM international conference on information & knowledge management*, pages 1165–1174.

Jianmo Ni, Gustavo Hernandez Abrego, Noah Constant, Ji Ma, Keith Hall, Daniel Cer, and Yinfei Yang. 2022. Sentence-T5: Scalable Sentence Encoders from Pretrained Text-to-Text Models. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1864–1874, Dublin, Ireland. Association for Computational Linguistics.

Parth Patwa, Mohit Bhardwaj, Vineeth Guptha, Gitanjali Kumari, Shivam Sharma, Srinivas PYKL, Amitava Das, Asif Ekbal, Shad Akhtar, and Tanmoy Chakraborty. 2021. Overview of CONSTRAINT 2021 Shared Tasks: Detecting English COVID-19 Fake News and Hindi Hostile Posts . In *Proceedings of the First Workshop on Combating Online Hostile Posts in Regional Languages during Emergency Situation (CONSTRAINT)*. Springer.

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

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *J. Mach. Learn. Res.*, 21(1).

Tharindu Ranasinghe and Marcos Zampieri. 2020. Multilingual Offensive Language Identification with Cross-lingual Embeddings. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5838–5844, Online. Association for Computational Linguistics.

Tharindu Ranasinghe and Marcos Zampieri. 2021. An Evaluation of Multilingual Offensive Language Identification Methods for the Languages of India. *Information*, 12(8).

Tharindu Ranasinghe, Marcos Zampieri, and Hansi Hettiarachchi. 2019. BRUMS at HASOC 2019: Deep learning models for multilingual hate speech and offensive language identification. In *Working Notes of FIRE 2019-Forum for Information Retrieval Evaluation*, pages 199–207.

Julian Risch and Ralf Krestel. 2018. Delete or not Delete? Semi-Automatic Comment Moderation for the Newsroom. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 166–176, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Sana Sabah Sabry, Tosin Adewumi, Nosheen Abid, György Kovács, Foteini Liwicki, and Marcus Liwicki. 2022. HaT5: Hate Language Identification using Text-to-Text Transfer Transformer. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7.

Joni Salminen, Hind Almerekhi, Milica Milenković, Soon-gyo Jung, Jisun An, Haewoon Kwak, and Bernard Jansen. 2018. Anatomy of Online Hate: Developing a Taxonomy and Machine Learning Models for Identifying and Classifying Hate in Online News Media. *Proceedings of the International AAAI Conference on Web and Social Media*, 12(1).

Shrey Satapara, Prasenjit Majumder, Thomas Mandl, Sandip Modha, Hiren Madhu, Tharindu Ranasinghe, Marcos Zampieri, Kai North, and Damith Premasiri. 2023. Overview of the HASOC Subtrack at FIRE 2022: Hate Speech and Offensive Content Identification in English and Indo-Aryan Languages. In *Proceedings of the 14th Annual Meeting of the Forum for Information Retrieval Evaluation*, FIRE '22, page 4–7, New York, NY, USA. Association for Computing Machinery.

Anna Schmidt and Michael Wiegand. 2017. A Survey on Hate Speech Detection using Natural Language Processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*.

Karishma Sharma, Feng Qian, He Jiang, Natali Ruchansky, Ming Zhang, and Yan Liu. 2019. Combating Fake News: A Survey on Identification and Mitigation Techniques. *ACM Trans. Intell. Syst. Technol.*, 10(3).

Ravi Shekhar, Marko Pranjić, Senja Pollak, Andraž Pelicon, and Matthew Purver. 2020. Automating News Comment Moderation with Limited Resources: Benchmarking in Croatian and Estonian. *Journal for Language Technology and Computational Linguistics (JLCL)*, 34(1).

Julia Struß, Melanie Siegel, Josef Ruppenhofer, Michael Wiegand, and Manfred Klenner. 2019. Overview of GermEval Task 2, 2019 Shared Task on the Identification of Offensive Language.

Lasitha Uyangodage, Tharindu Ranasinghe, and Hansi Hettiarachchi. 2021. Can Multilingual Transformers Fight the COVID-19 Infodemic? In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1432–1437, Held Online. INCOMA Ltd.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Tharindu Cyril Weerasooriya, Sujan Dutta, Tharindu Ranasinghe, Marcos Zampieri, Christopher M Homan, and Ashiqur R KhudaBukhsh. 2023. Vicarious Offense and Noise Audit of Offensive Speech Classifiers. *arXiv preprint arXiv:2301.12534*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. 2022. ByT5: Towards a Token-Free Future with Pre-trained Byte-to-Byte Models. *Transactions of the Association for Computational Linguistics*, 10:291–306.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Yuta Yanagi, Ryohei Orihara, Yuichi Sei, Yasuyuki Tahara, and Akihiko Ohsuga. 2020. Fake News Detection with Generated Comments for News Articles. In *2020 IEEE 24th International Conference on Intelligent Engineering Systems (INES)*, pages 85–90.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019a. Predicting the Type and Target of Offensive Posts in Social Media. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short*

Papers), pages 1415–1420, Minneapolis, Minnesota. Association for Computational Linguistics.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019b. SemEval-2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (OffensEval). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 75–86, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Marcos Zampieri, Preslav Nakov, Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Hamdy Mubarak, Leon Derczynski, Zeses Pitenis, and Cagri Coltekin. 2020. SemEval-2020 Task 12: Multilingual Offensive Language Identification in Social Media (OffensEval 2020). In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1425–1447, United States. Association for Computational Linguistics.

Marcos Zampieri, Tharindu Ranasinghe, Mrinal Chaudhari, Saurabh Gaikwad, Prajwal Krishna, Mayuresh Nene, and Shrunali Paygude. 2022. Predicting the type and target of offensive social media posts in Marathi. *Social Network Analysis and Mining*, 12(1):77.

Savvas Zannettou, Tristan Caulfield, Emiliano De Cristofaro, Nicolas Kourtelris, Ilias Leontiadis, Michael Sirivianos, Gianluca Stringhini, and Jeremy Blackburn. 2017. The Web Centipede: Understanding How Web Communities Influence Each Other through the Lens of Mainstream and Alternative News Sources. In *Proceedings of the 2017 Internet Measurement Conference*, IMC '17, page 405–417, New York, NY, USA. Association for Computing Machinery.