Ice and Fire: Dataset on Sentiment, Emotions, Toxicity, Sarcasm, Hate speech, Sympathy and More in Icelandic Blog Comments

Steinunn Rut Friðriksdóttir[1], **Annika Simonsen**[1], **Atli Snær Ásmundsson**[1], **Guðrún Lilja Friðjónsdóttir**[1], **Anton Karl Ingason**[1], **Vésteinn Snæbjarnarson**[2,3], **Hafsteinn Einarsson**[1]
[1]University of Iceland, [2]University of Copenhagen, [3]Miðeind ehf
{srf2, ans72, asa71, glf2, antoni, hafsteinne}@hi.is, vesn@di.ku.dk

Abstract

This study introduces "Ice and Fire," a Multi-Task Learning (MTL) dataset tailored for sentiment analysis in the Icelandic language. It encompasses a wide range of linguistic tasks, including sentiment and emotion detection, as well as the identification of toxicity, hate speech, encouragement, sympathy, sarcasm/irony, and trolling. With 261 fully annotated blog comments and 1,045 comments annotated in at least one task, this contribution marks a significant step forward in the field of Icelandic natural language processing. The dataset provides a comprehensive resource for understanding the nuances of online communication in Icelandic and an interface to expand the annotation effort. Despite the challenges inherent in subjective interpretation of text, our findings highlight the positive potential of this dataset to improve text analysis techniques and encourage more inclusive online discourse in Icelandic communities. With promising baseline performances, "Ice and Fire" sets the stage for future research to enhance automated text analysis and develop sophisticated language technologies, contributing to healthier online environments and advancing Icelandic language resources.

Keywords: Sentiment Analysis, Icelandic Language Resources, Multi-Task Learning

1. Introduction

With the rise of social media and other online platforms where people can express their thoughts and opinions, a challenge has arisen where inappropriate behavior is on the rise (Saha et al., 2023). Comment sections can contain prejudice and harmful content targeted at specific individuals or groups, even to the extent of qualifying as hate speech. Victims of online toxic attacks are more likely to engage in conversations and reply in a toxic manner (Aleksandric et al., 2022). This is further amplified by the observation that content generated by hateful users tends to spread faster and farther and reach a wider audience (Mathew et al., 2019). With the surge of data produced online daily, automatic methods are needed to detect and monitor toxic and hateful behaviors as manual inspection is time-consuming and costly. Various approaches exist for text analysis in this regard, among which are sentiment analysis and hate speech detection.

Our work introduces the first sentiment analysis dataset for Icelandic intended for Multi-Task Learning (MTL). Text extracts in the dataset have been labeled for 8 broad tasks relating to sentiment analysis. The initiative is motivated by the speculation that to truly understand the complexity of human communication in text, a multifaceted approach is required that includes not only sentiment analysis but also emotion detection and other nuanced aspects of language. Previous research is increasingly leaning towards a Multi-Task Learning (MTL) framework, which offers a more integrated and efficient way to handle interconnectedness in text analysis tasks. Studies such as Huang et al. (2013), Plaza-del Arco et al. (2022) and Tan et al. (2023) demonstrate the efficacy of MTL in enhancing the accuracy of sentiment analysis, emotion detection, and even sarcasm understanding in high-resource languages. These studies illustrate the benefits of addressing multiple related tasks simultaneously, leveraging shared insights to improve overall model performance. However, the application of MTL beyond English remains limited, with only a handful of studies, like those by Sane et al. (2019); Srivastava et al. (2020); Plaza-del Arco et al. (2021) and Ghosh et al. (2023), exploring its potential in languages such as Spanish and Hindi-English code-mixed texts. These efforts reveal the significant improvements MTL can bring to sentiment analysis and emotion detection tasks, even in complex, code-mixed scenarios. However, the scarcity of annotated, high-quality datasets for languages besides English remains a major obstacle.

The contributions of our paper are as follows:

Annotation framework We present our framework for annotating a broad family of sentiment analysis tasks for a given passage of text. In doing so, we move away from the one-sided view of classical single-label classification towards a more holistic viewpoint. We have implemented the annotation framework as a web application.

TRAC-2024 Workshop, pages 73–84

Ice and Fire, the Icelandic sentiment corpus We showcase the utility of our framework by annotating and releasing a much-needed multi-task sentiment analysis dataset for the low-resource language Icelandic. The dataset[1], which we have named "Ice and Fire", includes blog comments that have been annotated for 8 main tasks: sentiment analysis, toxicity detection, hate speech detection, emotion detection, encouragement and sympathy detection, constructive feedback detection, sarcasm/irony detection, and troll detection. Each main task contains several components, adding up to 20 subtasks overall. To the authors' knowledge, this is the first sentiment analysis dataset released for Icelandic that can be used for MTL purposes. Our dataset has the potential to be used to train language models that understand the subtleties of human communication as well as to train multi-dimensional reward models applicable to reinforcement learning with human feedback.

Model Evaluation To establish baselines, we train and evaluate Icelandic BERT models in representative tasks to evaluate performance. We further evaluate performance using GPT-4 and see a modest improvement in some categories and a lower performance in others.

2. Background

Sentiment analysis is the process of analyzing text to discern the sentiment underlying the words, aiming to understand the attitudes, opinions, and emotions expressed, a techique also referred to as opinion mining (Pang et al., 2008). This task usually involves labeling the polarity of a text with labels such as 'positive', 'neutral' and 'negative'. Closely related to this is *emotion detection*, which identifies the specific emotions being expressed in the text. This task commonly makes use of the six main types of emotions as proposed by Ekman (1992) as labels, namely 'fear', 'happiness', 'sadness', 'surprise', 'disgust', and 'anger' with 'contempt' sometimes included as well. Sentiment and emotion are closely related in that it is possible to sort most emotional states into either positive or negative. For example, 'happiness' can be considered a positive emotion, while 'fear' can be considered negative. Other related text classification tasks include toxicity, sarcasm and hate speech detection. For example, sarcastic sentences are often misclassified in text classification as positive when they should be classified as negative (Ghosh et al., 2023; Tan et al., 2023). Therefore, an ideal text classifier would need to have a grasp of all of these interconnected nuances in order to get the best result.

While the value of sentiment analysis is well-recognized for English, the journey for Icelandic and similar low-resource languages is only just beginning. At the time of writing, few studies have been published on sentiment analysis in Icelandic, although it was highlighted as an important topic in the first Icelandic Language Technology Programme (Nikulásdóttir et al., 2020). To the authors' knowledge, there have been only two previous contributions to single-task sentiment analysis for Icelandic, namely a paper by Ilyinskaya et al. (2023) and a bachelor thesis by Arndal et al. (2023). Ilyinskaya et al. (2023) used sentiment analysis on Icelandic Twitter posts to investigate the impact of geohazards on the mental health of the Icelandic population. They manually annotated 636 Icelandic tweets that contained earthquake- and eruption-related keywords with the labels 'negative sentiment', 'positive sentiment', or 'neutral statement'. Additionally, they automatically labeled a larger portion of tweets using a language model (Snæbjarnarson and Einarsson, 2022) that was fine-tuned for classification using the manually labeled data. Initial results showed good accuracy, with accuracy ranging from 69% to 71% and F1 scores from 69 to 71.

In their bachelor's thesis, Arndal et al. (2023) translated 50,000 English IMDb reviews, labeled as either positive or negative based on reviewer scores (where 1-4 stars was deemed negative and 5-10 stars positive), into Icelandic using Google Translate and Vélþýðing from Miðeind (Símonarson et al., 2021). They used the resulting data to train the first openly available Icelandic sentiment analysis models. They evaluated their models on movie reviews originally written in Icelandic that they found on Twitter and a movie-reviewing blog that they labeled in the same fashion as the English IMDb dataset. Their models obtain 89-93% accuracy in the binary sentiment analysis task on the Icelandic movie reviews, which is close to the performance of English models on the original IMDB dataset.

Similar to previous work in Icelandic, most studies tackled annotation tasks individually in the past. Recognizing the limitations of single-task approaches, which often led to isolated models that could not leverage the interconnectedness of text, the recent trend has shifted towards employing an MTL framework. In machine learning, the MTL framework is a strategy that enhances learning and generalization by simultaneously tackling related tasks, leveraging the shared knowledge and domain insights from each task's training data to improve the performance of all tasks involved (Caruana, 1997). As mentioned in the introduction, an

[1] https://huggingface.co/datasets/hafsteinn/ice_and_fire

early study by Huang et al. (2013) demonstrated the benefits of combining sentiment and topic analysis of English tweets using a Multi-Task Multi-Label (MTML) classification approach. Their findings showed that MTML produces a higher accuracy of both sentiment and topic analysis, but the approach is especially beneficial for topic analysis. Further advancing the MTL framework, Plaza-del Arco et al. (2022) explored the potential of enhancing hate speech and offensive language detection in English tweets by integrating sentiment analysis, emotion analysis, and target identification and employing a BERT-based MTL model. Their research concluded that MTL with emotion, sentiment, and target identification can be an effective approach for offensive speech detection systems for social media platforms. The correlation between sentiment analysis and sarcasm detection was explored by Tan et al. (2023), who found that understanding sarcasm could significantly enhance sentiment analysis in English tweets.

As evidenced by the aforementioned work, the literature has largely focused on English. However, there have been recent efforts to bring other languages into the domain. Several studies have been done on MTL for text classification in Hindi-English code-mixed language. Ghosh et al. (2023) applied cross-lingual contextual embeddings and a transfer learning strategy to sentiment and emotion detection in Hindi–English tweets. In their study, they manually annotated 20,000 instances of Hindi–English tweets from the SentiMix dataset that already have sentiment labels with emotion labels. Their method outperforms both single-task models and previous multitask methods, achieving notable improvements in F1 scores for sentiment and emotion detection tasks. Srivastava et al. (2020) presented a Hindi-English code-mixed dataset of 1001 tweets that express opinions annotated across multiple dimensions, such as aggression, hate speech, emotion arousal and figurative language usage. For English, Bengali and Hindi, Safi Samghabadi et al. (2020) integrated multi-task learning to a BERT-based model, which classifies texts into different aggression classes. Their analysis showed the two tasks, aggression and misogyny identification, were related, as shown by co-occurrences across labels.

These studies highlight the importance of MTL in sentiment analysis, underscoring the need for high-quality annotated data and models that can accurately interpret a wide range of linguistic contexts.

3. Methods

3.1. Data source

The dataset is composed of comments and blog posts from the website blog.is. As a selection criteria, the top 400 blogs were used, and posts with at least 1 comment were scraped along with the comments. For annotation, 5% of the comments were randomly selected, resulting in ≈50 thousand comments that were ordered randomly for annotation. Each comment on posts from the top 400 blog sites was thus equally likely to be selected for annotation.

As one of the country's longest-standing and still operational blog services, the source website serves as a valuable resource. Managed by a company that operates both a web media outlet and a newspaper, the platform predominantly features blogs that express opinions about current affairs. This synergy fosters a wealth of opinionated commentary, enriching the site with diverse viewpoints and discussions. This data is in the public domain and the released dataset does not contain author signatures.

3.2. The Annotation Interface

Figure 1 presents the annotation interface, designed as a crowdsourcing web application, in operation. At the upper portion, the annotator has the option to choose among various annotation tasks. For any selected comment, the interface allows the annotator to access preceding comments and the related blog post, providing the necessary context for accurate annotation. After submitting an annotation, the system automatically navigates to the next comment that has not been annotated in the chosen task but with a small probability of navigating to a comment that has been annotated once by another annotator. Additionally, at the interface's lower section, buttons are available for the annotator to review the guidelines and track their progress, indicating the number of completed annotations for each task. During the annotation process, annotators focused on performing annotations for single tasks. This means that comments that are fully annotated are likely annotated by different annotators.

We release the annotation framework as open-source software with this publication, accessible on Github[2].

3.3. Annotation Tasks

Three annotators, two women and one man each holding a bachelor's degree in Icelandic, annotated

[2] https://github.com/Haffi112/multi_task_annotation

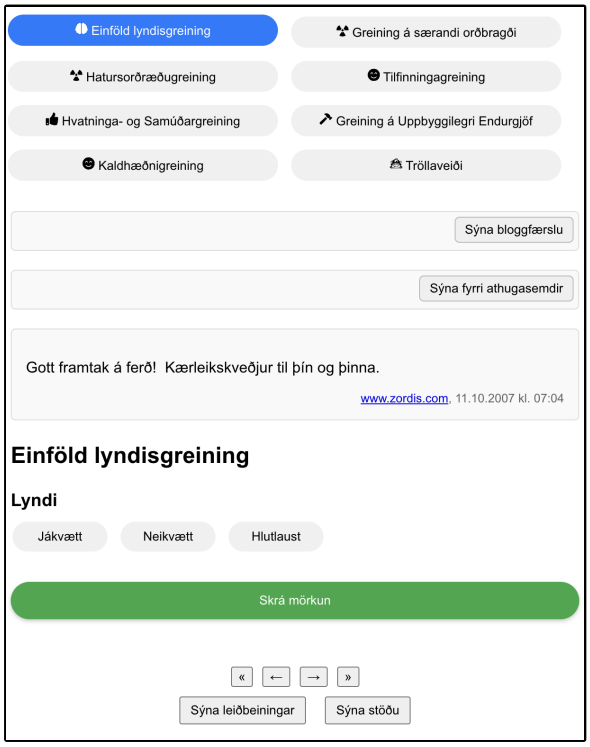

Figure 1: The annotation interface in use. In this case, for sentiment analysis.

the comments across eight distinct tasks. The annotators did not annotate all the task at the same time, instead, each task was annotated separately and each task was acompanied by the corresponding annotation guidelines. Furthermore, the annotator could view previous comments and the blog post in case further context was required to perform the annotation. This information was logged upon submission, i.e., for each annotation, we have information on whether prior comments or the blog post were open.

The tasks used for annotation were the following:

Sentiment analysis In this task, the annotator had to label whether a comment was positive, negative, or neutral. This was a multiclass task, i.e., the annotator could select a single label.

Toxicity detection This task was based on the toxicity detection task as described in Zampieri et al. (2019). For a given comment, the annotator labeled whether it was toxic or not. Toxic comments might for instance involve curse words, rudeness towards the interlocutor or general offensive behavior. If the comment was toxic, the annotator labeled whether it was intentional or unintentional. For intentional toxic remarks, the annotator needed to specify if it was directed towards a group or an individual.

Hate speech detection This task was based on the annotation scheme introduced by Basile et al. (2019). First, the annotator labeled whether the comment included hate speech or not. We refer to hate speech as it is defined by Article 233 (a) of the Icelandic penal code, further discussed in Section 5, i.e. threats, defamation or denigration on the basis of nationality, color, race, religion, sexual orientation, disabilities or gender identity. If the hate speech label was assigned, then the annotator needed to say towards whom it was directed (immigrants, religion, disabled, women or queer), whether it was directed towards a group or an individual, and finally, whether it was aggressive or not.

Emotion detection This task was inspired by the work of Demszky et al. (2020), but for the sake of simplicity, it was decided to start with the expanded basic emotions of Ekman (1992) (fear, happiness, sadness, surprise, disgust, and anger) along with contempt (Ekman and Heider, 1988), indignation, and neutrality.

Encouragement and sympathy detection This task was based on the work of Sosea and Caragea (2022). In this task, the annotator had to label whether a comment was encouraging or not and whether it was sympathetic or not.

Constructive feedback detection was based on the task introduced by Kolhatkar et al. (2020). In this task, the annotator labeled whether they agreed or not with what the comment said. They then labeled constructive and non-constructive properties of the comment in a multilabel manner. Finally, the annotator needed to say whether the comment was constructive or not overall.

Sarcasm/irony detection was based on the work of Ptáček et al. (2014). The aim was to label whether a comment included sarcasm or not. An "unclear" label was also included.

Troll detection was a task where the annotator needed to label whether a troll wrote a comment or not. A troll was defined as a person deliberately trying to provoke an emotional reaction from others, usually under an apparent pseudonym.

3.4. Inter-Annotator Agreement

We computed inter-annotator agreement using Krippendorff's Alpha (Krippendorff, 2018). We used the implementation by Castro (2017) with a nominal metric. For computing agreement in multilabel

tasks, we viewed them as separate binary annotation tasks and computed agreement for each label separately.

4. Results

Our dataset consists of 261 comments that have been fully annotated for all tasks and 1,045 comments that have been annotated in at least one task. We show the number of comments that have been annotated for a given number of tasks in Figure 2 and the contribution of each annotator in Figure 3.

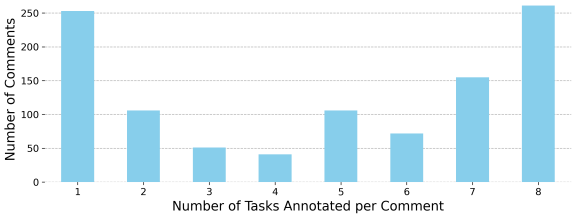

Figure 2: Distribution showing how many comments were labeled for how many tasks.

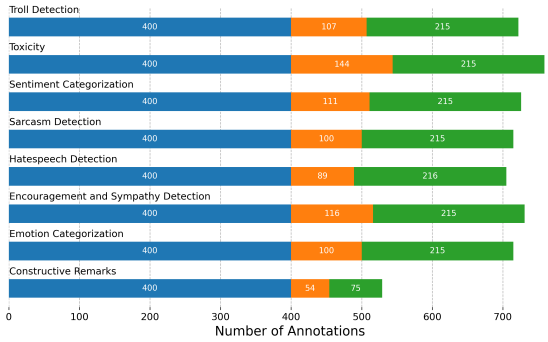

Figure 3: Number of annotations in each task per annotator. Blue corresponds to annotator 1, orange to annotator 2 and green to annotator 3.

Table 1 provides an overview of the reliability and agreement levels across multiclass tasks within our dataset that had sufficiently many double annotations. We observed varying levels of agreement among annotators across different tasks. Notably, the task of Sentiment Categorization yielded the highest Krippendorff's alpha coefficient (0.58), indicating a relatively high level of agreement. Conversely, Sympathy Detection demonstrated the lowest agreement with an alpha of 0.08, suggesting substantial discrepancies in annotator perceptions. The other tasks, including Hate speech Detection (0.49) and Toxicity - Offensive Language Detection (0.54), showed moderate agreement levels. These agreement scores reflect the complexity and subjectivity inherent in annotating blog comments, particularly when discerning nuanced concepts such as sarcasm, encouragement, and sympathy. To model the annotator, we release the annotator ID along with the dataset.

Task	#	$\neq$	α
Sentiment Categorization	125	34	0.58
Toxicity Detection	73	9	0.54
Hate speech Detection	77	2	0.49
Sarcasm Detection	75	10	0.44
Encouragement Detection	117	18	0.38
Troll Detection	58	8	0.22
Constructive Remarks	30	11	0.21
Sympathy Detection	117	13	0.08

Table 1: Agreement in multiclass annotation tasks. The table shows the number of double annotated examples (#), disagreements ($\neq$), and Krippendorff's alpha values (α).

Table 2 shows the annotator agreement of multilabel tasks through a binary representation of the labels. Krippendorff's alpha revealed significant variability in agreement across labels. Some of the labels occurred infrequently in double annotated examples, so agreement values should not be taken to generalize. For the emotion categorization, some of the labels occurred frequently enough to warrant discussion. The value for the happiness label is 0.75, indicating moderate reliability. The alpha values for other labels with occurrence in at least 30 comments were 0.48 for neutral and 0.24 for indignation. We note that indignation was added after the annotation had started.

The distribution of labels for the sentiment categorization task is shown in Figure 4. We observe a somewhat balanced distribution of sentiment with negative and neutral labels, each being around 50% more common than positive labels.

The distribution of labels in emotion detection is shown in Figure 5. The most common label chosen is neutral, but we see a great number of examples representing happiness, anger and indignation. Indignation was a label we added specifically in this task due to the nature of the discussion in the dataset.

The distributions of labels for the constructive feedback detection task are shown in Figure 6. The comments are quite balanced with respect to whether they are considered constructive overall, but in most cases, they do not include any constructive or non-constructive properties.

The distribution of labels for the hate speech detection task are shown in Figure 8. We observe a relatively infrequent occurrence of hate speech in the comments annotated. This rarity may be due to general civility or due to bloggers or moderators removing such comments as they oppose the content policy on the blogging platform.

Label	#	≠	α
Constructive Remarks - Unconstructive Properties (30 double annotations)			
Not relevant	2	2	-0.02
Is provocative	12	5	0.62
Is unsubstantial	11	11	-0.20
No non-constructive characteristics	18	10	0.33
Does not respect the views and beliefs of others	10	8	0.18
Is sarcastic	4	3	0.36
Constructive Remarks - Constructive Properties (27 double annotations)			
Targets specific points	7	4	0.52
Provides evidence	1	0	1.00
Contributes something substantial to the conversation and encourages dialogue	6	5	0.20
No constructive characteristics	19	6	0.55
Provides a solution	1	1	0.00
Provides a personal story or experience	3	2	0.47
Emotion Categorization (128 double annotations)			
Disgust	4	4	-0.01
Sadness	4	2	0.66
Anger	29	18	0.47
Neutral	73	33	0.48
Enjoyment/Happiness	30	10	0.75
Indignation	36	28	0.24
Contempt	15	12	0.29
Fear	4	3	0.39
Surprise	12	10	0.25

Table 2: Agreement for multilabel annotation tasks. The table shows the number of comments in double annotated examples containing the label (#), disagreements (≠), and Krippendorff's alpha values (α).

The label distribution of the sarcasm detection task is shown in Figure 9. Sarcasm is relatively rare in the dataset, and it is often unclear whether the comment is intended to be sarcastic or not.

The label distribution of the troll detection task is shown in Figure 10. Trolls are relatively rare in the dataset, which might be due to content policies. It is also often not clear whether a commenter is trolling or not, especially since they are not necessarily anonymous.

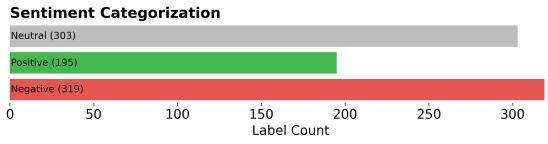

Figure 4: Label distribution for the sentiment categorization task.

In the annotation interface, the annotators can view previous comments and the blog post. Whether they were open was logged upon submission to indicate whether the annotator had required more context to perform the task. Figure 11 shows the fraction of the time this was done for each task, revealing that annotators generally did not require

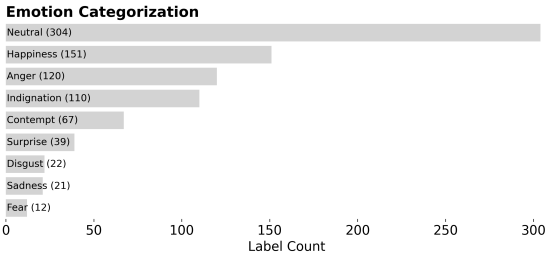

Figure 5: Label distribution for the emotion detection task.

additional context to perform the annotation. *Hate Speech - Other* was a bit of an outlier, and it refers to the extra annotation tasks performed when hate speech was detected. The annotators reported that hate speech often required more context as it referenced the previous comments or blog post, but with the actual hate being in the comment itself.

4.1. Baseline Single-Task Results

To accompany the dataset and encourage its use, we release some non-hyper parameter tuned baselines for a selection of the task. We fine-tune an Icelandic BERT model (Snæbjarnarson et al., 2022)

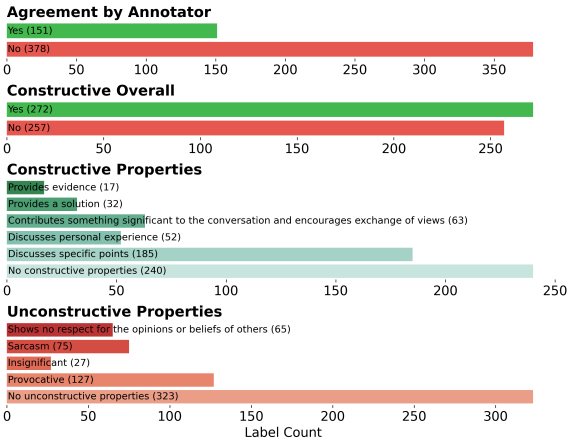

Figure 6: Label distributions for the constructive feedback detection task.

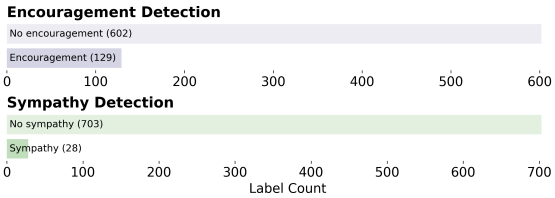

Figure 7: Label distributions for the encouragement and sympathy detection tasks.

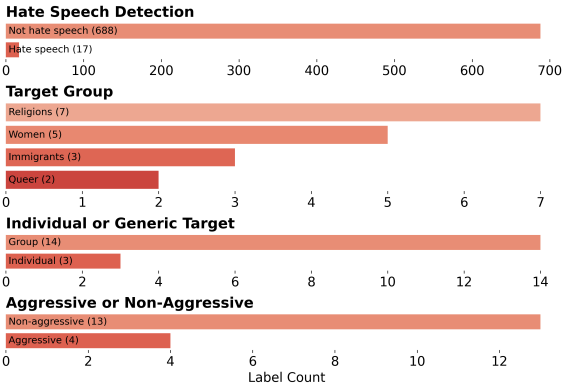

Figure 8: Label distributions for the hate speech detection tasks.

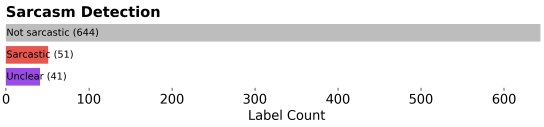

Figure 9: Label distribution for the sarcasm detection task.

on the tasks emotion, sarcasm, sentiment, non-constructive properties, toxicity, agreement by annotator, encouragement. Since data points are limited for some of the categories, we aggregate some of them together. For *agreement by annotator* we use the labels *'yes'* (121) and *'no'* (309). For *emotion*, we use the label *'neutral'* (209) and aggregate the others as *'not neutral'*. For *toxic*, we use the labels *'toxic'* (142) and *'not toxic'* (618). For *non-constructive* (256) we use the label *'not non-constructive'* and aggregate the others as *'non-constructive'*. Finally, for *sentiment* we use the labels *'positive'* (159), *'negative'* (258) and *neutral* (256). We fine-tune all models for 5 epochs on a single task at a time using a learning rate of 2e-5, a batch size of 16, and a weight decay of 0.01 with the AdamW optimizer. We report the macro-F1 and accuracy results in Table 3. All figures are calculated using tenfold cross-validation. The intervals given are the standard error.

For an LLM evaluation, see Section A in the Appendix.

5. Discussion

The Ice and Fire Dataset: A Nuanced Approach to Sentiment Analysis In this work, we introduced the Ice and Fire dataset, the first Multi-Task Learning (MTL) resource for sentiment analysis in Icelandic, encompassing a comprehensive suite of annotation tasks, including basic sentiment analysis, emotion detection, sarcasm, encouragement, and troll detection. This initiative is motivated by the complexity and multifaceted nature of human communication, advocating for a nuanced approach that extends beyond traditional sentiment analysis to incorporate a broader spectrum of communicative cues. Our findings reveal a diverse range of sentiments and emotions present in online discourse, with a notable prevalence of neutral and negative sentiments. This reflects the critical and often contentious nature of online discussions. The baseline results for single-task models provide a benchmark for future research, highlighting the challenges in accurately capturing the subtleties of human communication, particularly for nuanced tasks like emotion detection and non-constructive comment identification.

Insights and Recommendations for Future Annotation Efforts The variation in agreement levels across tasks underscores the subjective nature of interpreting text, especially for nuanced tasks such as sarcasm and sympathy detection. The imbalanced label distribution and the forced-choice scenario without a "skip" option likely contributed to reduced annotator consistency. These insights suggest that future annotation efforts could benefit from improved guidelines, the inclusion of a skip option, and consensus-building phases to enhance annotation reliability, particularly for subjectively interpreted tasks. To ease the annotator's task, we

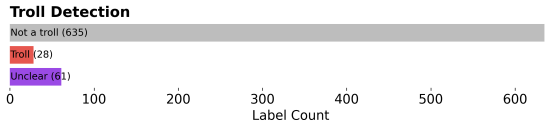

Figure 10: Label distribution for the troll detection task.

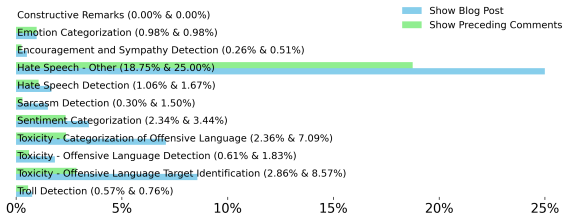

Figure 11: Distribution showing how often the annotators viewed the blog post or prior comments for each annotation task. 'Hate speech - Other' refers to the three tasks following hate speech detection.

recommend experimenting with dividing multilabel tasks into individual binary classification tasks. For emotion classification, this approach would be cognitively less taxing and would enable the annotator to concentrate on a single emotion at a time during the annotation process.

Challenges in Accommodating Annotator Perspectives Building on the need for streamlined annotation tasks, we also face the challenge of accommodating the annotator's perspective amidst the multifaceted nature of online communication.

We acknowledge that while our annotators identified problematic comments to the best of their abilities, the nature of these annotations is inherently subjective. As discussed by Curry et al. (2024), "we must take care not to treat conflicting responses equally. If a minority with the necessary lived experience (e.g. to recognise misogyny) disagree with the majority who don't, that matters". They further argue that the difference between hate and offence must be taken into account when examining hate speech and we agree on this point. The relatively small number of identified hate speech in our dataset should be considered from this perspective, as identifying toxicity is more in line with that of identifying offence while labeling something as hate speech requires a thorough reasoning and undeniable hate is not often present in our data.

Detecting sarcasm in written text presents inherent challenges, as intentions can be obscured by the author's stylistic choices, such as excessive punctuation, which may alter the perceived meaning. The delineation of hate speech within the scope of this study is confined to expressions targeting nationality, color, race, religion, sexual orientation, disabilities, and gender identity, leaving statements against political ideologies, for example, outside its purview. The relevance of annotator agreement on sentiment often becomes moot in instances where the sentiment is neutral or non-controversial, such as generic greetings, leading to a default classification of disagreement in ambiguous cases. Moreover, the interpretation of encouragement encompasses a spectrum from genuine support to sarcastic or hostile remarks, highlighting the complexity of sentiment analysis. The distinction between online trolls and overtly toxic individuals, particularly when using their real names, raises questions about the nature of online identities and their impact on communication. Additionally, the adequacy of basic emotional categories to encapsulate complex sentiments, such as schadenfreude or passive aggression, is limited, suggesting a need for nuanced labeling practices. Ambiguity in sentiment analysis is further compounded in longer texts, where shifts in tone may necessitate a more nuanced approach to determining the overall sentiment. This complexity underscores the intricacies of annotating sentiment in online discourse, where clarity and context are paramount.

Task	Accuracy	F1
Toxicity	0.836 ± 0.010	0.646 ± 0.034
Sarcasm	0.950 ± 0.003	0.487 ± 0.001
Encouragem.	0.827 ± 0.013	0.644 ± 0.028
Sentiment	0.719 ± 0.010	0.723 ± 0.010
Emotion	0.655 ± 0.011	0.524 ± 0.032
Agreement	0.721 ± 0.010	0.419 ± 0.003
Non-constr.	0.635 ± 0.017	0.435 ± 0.030

Table 3: Baseline results for selected tasks in the Ice and Fire dataset.

Potential Benefits for Icelandic Society Models trained on our dataset hold potential benefits for Icelandic society, particularly in addressing hate speech and other harmful online behaviors. In Iceland, hate speech is implicitly covered under Article 233 (a) of the penal code (Government of Iceland, 1940):

Anyone who publicly mocks, defames, denigrates or threatens a person or group of persons by comments or expressions of another nature, for example by means of pictures or symbols, for their nationality, colour, race, religion, sexual orientation or gender identity, or disseminates such materials, shall be fined or imprisoned for up to 2 years.

This article serves as the foundation for the blog platform's rules, potentially accounting for the minimal hate speech identified in our annotation effort. However, while hate speech seems to be criminalized in Iceland, it is rarely enforced, and preventa-

tive measures are lacking. In 2023, the Council of Europe's anti-racism (ECRI) body called for a more strategic and coordinated approach to tackle hate speech in Iceland (Council of Europe, 2023). This was a response to the work completed by a Governmental Working Group against Hate Speech that was appointed by the Prime Minister in 2022. Based on their work, the Prime Minister presented a proposal for a parliamentary resolution on the Government's action plan against hate speech in 2023. ECRI, therefore, recommended that the authorities reinforce their responses against hate speech by implementing the action plan against hate speech, with particular emphasis being placed on effective ways to tackle online racist and LGBTI+-phobic hate speech. Currently, there are no automated methods available that can effectively identify Icelandic hate speech. This lack of resources becomes apparent when considering the amount of negative and toxic comments on some Icelandic discourse platforms, as manual moderation can only catch a limited amount of such content. It is, therefore, our hope that our contribution can help to foster a more inclusive and respectful online discourse, especially for Icelandic, where the resources so far have been limited.

Applications Beyond Hate Speech Detection Models trained on this dataset have applications beyond hate speech detection. They can be employed to analyze individual online behavior in relation to the tasks presented in this work. This approach has the potential to provide valuable insights into the study of history at a large scale, as demonstrated by previous research (Michel et al., 2011). Moreover, text-based approaches have been used to infer various user characteristics, such as age and gender (Nguyen et al., 2014), well-being (Jaidka et al., 2020), or even the presence of depression (De Choudhury et al., 2013). Models trained on the tasks in this work can be used to investigate how online discourse evolves over time or in response to specific topics. By leveraging the capabilities of models trained in the tasks, researchers can explore the dynamics and trends within online communities at a scale that complements traditional manual analysis methods. While the effectiveness of automated methods has been established for English (Schwartz and Ungar, 2015), our dataset enables the application of such techniques to Icelandic, a less-resourced language. This opens up new possibilities for studying large volumes of Icelandic text data, offering insights into the unique characteristics and evolution of online discourse within the Icelandic-speaking community.

Future Directions: Active Learning and Multi--Dimensional Reward Models Looking ahead, integrating models trained on our dataset into active learning workflows could significantly improve the efficiency of annotation efforts to grow the dataset, especially for rare label classes. This approach would prioritize human annotation efforts on the most informative or ambiguous examples, thereby enhancing model performance with minimal additional annotation work. We posit that organizing this as a crowdsourcing effort could prove advantageous, particularly in mitigating annotator bias in tasks reliant on subjective assessment. Additionally, the potential for training multi-dimensional reward models for Reinforcement Learning with Human Feedback (RLHF) is promising. Such models could lead to the development of Icelandic language models that are not only sensitive to the nuances of language but also capable of adapting their responses based on human feedback. Applications could range from more effective automated monitoring tools for social media to emotionally intelligent and culturally aware Icelandic chatbots.

6. Conclusion

In sum, the "Ice and Fire" dataset represents a significant step forward in the study of sentiment analysis and MTL, especially for a low-resource language like Icelandic. Despite challenges in annotator agreement for more subjective tasks, the varied performance across different communicative categories reflects the depth and complexity of the dataset. The baseline results from fine-tuning an Icelandic BERT model on the dataset underscore the dataset's utility and the potential of NLP technologies in Icelandic. For an LLM evaluation, we saw a further improvement in all categories, except sarcasm detection and agreement detection. The dataset opens new avenues for research into the complex interplay of sentiment, emotion, and other communicative aspects in online discourse, with the potential to contribute meaningfully to Icelandic society and beyond.

7. Acknowledgements

Steinunn Rut Friðriksdóttir was supported by The Ludvig Storr Trust no. LSTORR2023-93030 and The Icelandic Language Technology Programme. Annika Simonsen was supported by The European Commission under grant agreement no. 101135671. Vésteinn Snæbjarnarson acknowledges support from the Pioneer Centre for AI, DNRF grant number P1.

8. Bibliographical References

Ana Aleksandric, Sayak Saha Roy, and Shirin Nilizadeh. 2022. Twitter users' behavioral response to toxic replies. *arXiv preprint arXiv:2210.13420*.

Birkir Finnbogi H. Arndal, Eysteinn Örn Jónsson, and Ólafur Aron Jóhannsson. 2023. Evaluating icelandic sentiment analysis models trained on translated data. Bachelor's thesis, Reykjavík University, Reykjavík, Iceland. Department of Computer Science.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Rich Caruana. 1997. Multitask learning. *Machine learning*, 28:41–75.

Santiago Castro. 2017. Fast Krippendorff: Fast computation of Krippendorff's alpha agreement measure. https://github.com/pln-fing-udelar/fast-krippendorff.

Council of Europe. 2023. European Commission against Racism and Intolerance Report on Iceland (sixth monitoring cycle). Technical report, Council of Europe. Accessed: February 2024.

Amanda Cercas Curry, Gavin Abercrombie, and Zeerak Talat. 2024. Subjective *Isms*? On the Danger of Conflating Hate and Offence in Abusive Language Detection. *arXiv preprint arXiv:2403.02268*.

Munmun De Choudhury, Scott Counts, and Eric Horvitz. 2013. Social media as a measurement tool of depression in populations. In *Proceedings of the 5th annual ACM web science conference*, pages 47–56.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. Goemotions: A dataset of fine-grained emotions. *arXiv preprint arXiv:2005.00547*.

Paul Ekman. 1992. An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200.

Paul Ekman and Karl G Heider. 1988. The universality of a contempt expression: A replication. *Motivation and emotion*, 12(3):303–308.

Soumitra Ghosh, Amit Priyankar, Asif Ekbal, and Pushpak Bhattacharyya. 2023. Multitasking of sentiment detection and emotion recognition in code-mixed hinglish data. *Knowledge-Based Systems*, 260:110182.

Government of Iceland. 1940. General penal code of iceland, nr. 19/1940. Government of Iceland. Accessed: February 2024.

Shu Huang, Wei Peng, Jingxuan Li, and Dongwon Lee. 2013. Sentiment and topic analysis on social media: a multi-task multi-label classification approach. In *Proceedings of the 5th Annual ACM Web Science Conference*, WebSci '13, page 172–181, New York, NY, USA. Association for Computing Machinery.

E. Ilyinskaya, V. Snæbjarnarson, H. K. Carlsen, and B. Oddsson. 2023. Brief communication: Small-scale geohazards cause significant and highly variable impacts on emotions. *Natural Hazards and Earth System Sciences Discussions*, 2023:1–12.

Kokil Jaidka, Salvatore Giorgi, H Andrew Schwartz, Margaret L Kern, Lyle H Ungar, and Johannes C Eichstaedt. 2020. Estimating geographic subjective well-being from twitter: A comparison of dictionary and data-driven language methods. *Proceedings of the national academy of sciences*, 117(19):10165–10171.

Varada Kolhatkar, Nithum Thain, Jeffrey Sorensen, Lucas Dixon, and Maite Taboada. 2020. Classifying constructive comments. *arXiv preprint arXiv:2004.05476*.

Klaus Krippendorff. 2018. *Content analysis: An introduction to its methodology*. Sage publications.

Binny Mathew, Ritam Dutt, Pawan Goyal, and Animesh Mukherjee. 2019. Spread of hate speech in online social media. In *Proceedings of the 10th ACM conference on web science*, pages 173–182.

Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K Gray, Google Books Team, Joseph P Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, et al. 2011. Quantitative analysis of culture using millions of digitized books. *science*, 331(6014):176–182.

Dong Nguyen, Dolf Trieschnigg, A Seza Doğruöz, Rilana Gravel, Mariët Theune, Theo Meder, and Franciska de Jong. 2014. Why gender and age prediction from tweets is hard: Lessons from a crowdsourcing experiment. In *25th International Conference on Computational Linguistics (COLING 2014)*, pages 1950–1961. Dublin City

University and Association for Computational Linguistics.

Anna Björk Nikulásdóttir, Jón Guðnason, Anton Karl Ingason, Hrafn Loftsson, Eiríkur Rögnvaldsson, Einar Freyr Sigurðsson, and Steinþór Steingrímsson. 2020. Language technology programme for icelandic 2019-2023. *arXiv preprint arXiv:2003.09244*.

Bo Pang, Lillian Lee, et al. 2008. Opinion mining and sentiment analysis. *Foundations and Trends® in information retrieval*, 2(1–2):1–135.

Flor Miriam Plaza-del Arco, Sercan Halat, Sebastian Padó, and Roman Klinger. 2022. Multi-task learning with sentiment, emotion, and target detection to recognize hate speech and offensive language.

Flor Miriam Plaza-del Arco, M. Dolores Molina-González, L. Alfonso Ureña-López, and María Teresa Martín-Valdivia. 2021. A multi-task learning approach to hate speech detection leveraging sentiment analysis. *IEEE Access*, 9:112478–112489.

Tomáš Ptáček, Ivan Habernal, and Jun Hong. 2014. Sarcasm detection on Czech and English Twitter. In *Proceedings of COLING 2014, the 25th international conference on computational linguistics: Technical papers*, pages 213–223.

Niloofar Safi Samghabadi, Parth Patwa, Srinivas PYKL, Prerana Mukherjee, Amitava Das, and Thamar Solorio. 2020. Aggression and misogyny detection using BERT: A multi-task approach. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 126–131, Marseille, France. European Language Resources Association (ELRA).

Punyajoy Saha, Kiran Garimella, Narla Komal Kalyan, Saurabh Kumar Pandey, Pauras Mangesh Meher, Binny Mathew, and Animesh Mukherjee. 2023. On the rise of fear speech in online social media. *Proceedings of the National Academy of Sciences*, 120(11):e2212270120.

Sushmitha Reddy Sane, Suraj Tripathi, Koushik Reddy Sane, and Radhika Mamidi. 2019. Stance detection in code-mixed Hindi-English social media data using multi-task learning. In *Proceedings of the Tenth Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 1–5, Minneapolis, USA. Association for Computational Linguistics.

H Andrew Schwartz and Lyle H Ungar. 2015. Data-driven content analysis of social media: A systematic overview of automated methods. *The ANNALS of the American Academy of Political and Social Science*, 659(1):78–94.

Haukur Barri Símonarson, Vésteinn Snæbjarnarson, Pétur Orri Ragnarsson, Haukur Páll Jónsson, and Vilhjálmur Þorsteinsson. 2021. Miðeind's wmt 2021 submission. *arXiv preprint arXiv:2109.07343*.

Vésteinn Snæbjarnarson and Hafsteinn Einarsson. 2022. Cross-lingual QA as a stepping stone for monolingual open QA in Icelandic. In *Proceedings of the Workshop on Multilingual Information Access (MIA)*, pages 29–36, Seattle, USA. Association for Computational Linguistics.

Vésteinn Snæbjarnarson, Haukur Barri Símonarson, Pétur Orri Ragnarsson, Svanhvít Lilja Ingólfsdóttir, Haukur Jónsson, Vilhjalmur Thorsteinsson, and Hafsteinn Einarsson. 2022. A warm start and a clean crawled corpus - a recipe for good language models. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4356–4366, Marseille, France. European Language Resources Association.

Tiberiu Sosea and Cornelia Caragea. 2022. EnsyNet: A dataset for encouragement and sympathy detection. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5444–5449, Marseille, France. European Language Resources Association.

Arjit Srivastava, Avijit Vajpayee, Syed Sarfaraz Akhtar, Naman Jain, Vinay Singh, and Manish Shrivastava. 2020. A multi-dimensional view of aggression when voicing opinion. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 13–20, Marseille, France. European Language Resources Association (ELRA).

Yik Yang Tan, Chee-Onn Chow, Jeevan Kanesan, Joon Huang Chuah, and YongLiang Lim. 2023. Sentiment analysis and sarcasm detection using deep multi-task learning. *Wireless personal communications*, 129(3):2213–2237.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Predicting the type and target of offensive posts in social media. *arXiv preprint arXiv:1902.09666*.

A. LLM Multi-Task Results

For comparison, we also evaluate an LLM, the GPT-4-turbo model (
textttgpt-4-turbo-2024-04-09), on the dataset. The LLM annotates the data in the same manner as the fine-tuned baseline model and the results are shown in Table 4.

To compute accuracy, we resolve annotator conflicts using the following rules: Agreement task: Majority vote, with "No" on a tie. Emotion task: "Neutral" if all labels are neutral, "Emotion detected" if at least one annotator assigned an emotion. Encouragement task: "Encouragement" if at least one annotator assigned it, "No encouragement" otherwise. Non-constructive feedback detection task: "Non-constructive feedback" if at least one annotator assigned that label, "No non-constructive feedback" otherwise. Sarcasm detection task: "Sarcasm" if at least one annotator assigned that label, "No sarcasm" otherwise. Sentiment task: Conflicts resolved with the "Neutral" label. Toxicity task: "Toxic" if at least one annotator used that label, "Not toxic" otherwise.

Task	Accuracy	Δ
Toxicity	0.860	+0.024
Sarcasm	0.886	-0.064
Encouragem.	0.859	+0.032
Sentiment	0.781	+0.062
Emotion	0.723	+0.068
Agreement	0.608	-0.113
Non-constr.	0.763	+0.128

Table 4: Accuracy for GPT-4-turbo on the Ice and Fire dataset along with an absolute comparison to the performance of the baseline model (Δ).