Towards Text Formality Annotation with LLMs and Traditional Approaches

Tatiana Anikina
tatiana.anikina@dfki.de

Abstract

Text formality detection is a non-trivial task that has a lot of practical applications. Being able to correctly detect the formality of the text can help to tailor it to a specific audience, making the communication more efficient while adhering to specific situational and cultural norms. In this study we investigate how text formality can be detected via prompting or fine-tuning Large Language Models (LLMs), and compare the results to the scores that we obtain with more traditional methods (e.g., readability score, lexical diversity or syntactic complexity). We experiment with two datasets annotated with formality scores in English and German, and compute various metrics to estimate how well our annotations align with the original scores. Our findings show that fine-tuning or prompting a sufficiently large LM results in strong performance, while readability metrics and syntactic complexity also show competitive results, but lexical diversity is not a good indicator of text formality when measured at a sentence level. Our code and data are publicly available at `https://github.com/tanikina/text-formality`.

1 Introduction

The formality of the text reflects how professional, structured, lexically, and syntactically complex the text is and whether it adheres to standard linguistic conventions. However, determining the formality level depends on several factors, such as vocabulary, sentence structure, length, and even the medium (e.g., academic papers versus casual emails). There is no universal definition of text formality and also no universal metric to measure it because formality largely depends on the domain, genre, and other aspects.

Heylighen et al. [1999] consider expression as being formal *"when it is context-independent and precise (i.e. non-fuzzy), that is, it represents a clear distinction which is invariant under changes of context."* The range of formality levels may depend on the specific setting. For example, communication in the chat or forum typically assumes a different scale of formality compared to communication via email, and the same text can be perceived as more formal in the chat and less formal in the email. Also, detecting formality may require additional context and it remains a somewhat subjective measure since it depends also on the reader's experiences and expectations. In addition, cultural and situational norms affect perceived formality levels, and they need to be taken into account when translating the text or improving the quality and appropriateness through writing assistance tools. In this study we focus on formality annotation using a normalized scale and benchmark several methods on two datasets in different languages (English and German) in an attempt to answer the following research questions:

1. To what extent can we **leverage LLMs** for text formality annotation? Does LLM **prompting** perform well enough to solve this task? How does it compare to **fine-tuning** a smaller LLM?
2. How do **traditional metrics** such as readability scores, lexical diversity, and syntactic complexity perform compared to more modern LLM-based approaches?
3. Can we observe any **trends and differences between the approaches** when benchmarking them on two datasets in different languages (English and German in our study)?

2 Related work

Text formality has been considered an important natural language processing task for a long time. One of the early works by Heylighen et al. [1999] introduced the F-score (formality score) based on the part of speech (POS) information. The key insight of this work is that formal texts rely more on nouns and prepositional phrases, while more informal texts use more pronouns and verbs. Wang et al. [2010] tested several corpus-based methods for deriving real-valued formality lexicons, and evaluated them using relative formality judgments between the word pairs.

Rao and Tetreault [2018] created a large-scale corpus with formality annotations called Grammarly's Yahoo Answers Formality Corpus (GYAFC), and showed that machine translation techniques perform as strong baselines on the style transfer task with respect to formality. The X-FORMAL dataset Briakou et al. [2021] is based on the GYAFC data but covers multiple languages, including French, Italian, and Brazilian Portuguese. Their work shows that state-of-the-art style transfer approaches for formality still have some room for improvement, since they perform similar to simple baselines in the multilingual setting.

Pavlick and Tetreault [2016] worked with the sentence-level formality annotations based on the data collected by Lahiri [2015] and additional crowdsourced annotations, they analyzed how formality varies across different domains, and genres. Pavlick and Tetreault [2016] found that some topics have a clear bias towards formality (e.g. economics), and also formality positively correlates with politeness and negatively with sarcasm. Their work emphasizes that *"formality is one of the most basic dimensions of stylistic variation in language, and the ability to recognize and respond to differences in formality is a necessary part of full language understanding"*.

Some recent works explored formality style transfer and evaluated formality detection in the multilingual setting. E.g., Chawla and Yang [2020] applied a semi-supervised method using a language model-based discriminator to maximize the likelihood of the output sentence being formal, while maximizing mutual information between the source and target styles. Dementieva et al. [2023] benchmarked various text classification models on the formality detection task. They considered several statistical, neural-based, and Transformer-based machine learning methods, and found that character-based BiLSTM models can outperform the Transformer-based ones, but Transformers show better cross-lingual transfer capabilities. This work is the most similar to ours, however, we experiment with different datasets and models, and also the current work explores few-shot LLM prompting with decoder-only models like Llama AI@Meta [2024] and Qwen Team@Qwen [2024].

3 Data and Methodology

This section describes the data used in our experiments, and how they were pre-processed. We also elaborate on the experimental setup with LLM-based approaches and traditional formality detection methods.

3.1 Data

In order to benchmark different approaches for text formality detection we consider two openly available datasets that we call *Pavlick Formality* Pavlick and Tetreault [2016], Lahiri [2015] and *InFormal Sentences* Eder et al. [2023].

3.1.1 Pavlick Formality Dataset

This dataset contains sentence-level formality annotations based on Pavlick and Tetreault [2016]. All sentences are in English and they represent four different genres: news, blogs, email, and question answering forums. The data for news and blogs were collected by Lahiri [2015]. All sentences were annotated by human annotators hired via Amazon Mechanical Turk with scores in the range between -3 and 3, see the examples with the corresponding scores below:

(1) HE LEFT THE YEAR I LEFT SO ITS ALL GOOD! -2.2

(2) We will send a formal request to your scheduling office. 2.4

The original dataset can be accessed through Huggingface[1]. The dataset already has the pre-defined training, validation, and test splits. However, in our experiments we subsample the data to 1,000 samples (randomly sampled) and split them into training (700), validation (100), and test (200) partitions. This is done to test the setting with a limited amount of training data and also to ensure more fair comparison with another dataset that we consider (*InFormal Sentences*) that contains less samples than *Pavlick Formality.* We also re-scale the scores to make them compatible between the datasets and use the range of $[0, 1]$ instead of $[-3, 3]$ where 0 means the lowest formality level and 1 is the highest.

[1] `https://huggingface.co/datasets/osyvokon/pavlick-formality-scores`

3.1.2 InFormal Sentences Dataset

This dataset is based on the work by Eder et al. [2023], and represents a collection of sentences in German from different genres that were annotated by humans on a continuous scale between -1 (most informal) and 1 (most formal). Below are some examples:

(3) aber jerry war mal wieder erster`:^):^):^):^):^):^)` -0.9

(4) Zum Zeitpunkt der Ratifizierung des Maastricht-Vertrags ratifizierte beispielsweise Frankreich diesen mit einer kleineren Mehrheit als Schweden. 0.825

Similarly to *Pavlick Formality* we also sample 1,000 instances from this dataset and split them into training (700), validation (100), and test (200) sets. All scores are also normalized and converted to the range of $[0, 1]$. The training and development partitions are used for the experiments with LLMs, e.g., for fine-tuning XLM RoBERTa or for selecting the few-shot examples in case of prompting.

3.2 Methods

We compare different approaches to detect text formality levels. On the one hand, we use state-of-the-art LLMs for few-shot prompting and fine-tuning, and, on the other hand, we consider traditional approaches such as lexical diversity, syntactic complexity, and readability scores.

3.2.1 LLM-based Methods

State-of-the-art LLMs excel at various language processing tasks, from sentiment analysis and part of speech tagging to very complex tasks such as semantic parsing or coreference resolution. In this study we consider prompting decoder-only LLMs and fine-tuning a smaller encoder-decoder model on the formality detection task.

Prompting

We chose Qwen2.5-7B[2] and Llama3-8B[3] decoder-only quantized models that officially support both English and German. For each dataset we randomly sampled 10 examples with different formality scores and presented them as part of the prompt. The prompt used in our experiments has the following structure:

You are required to annotate given example and assign the text formality level that should be in the range between 0 and 1 (0 means most informal and 1 most formal). Here are some examples: {*demonstrations*} *Now annotate the following sample. Input:* {*input_sample*} *Formality label:*

We also added some post-processing steps to ensure that only floating point numbers in the specified range are considered as valid output. If the model failed to generate a number in the output we retried prompting multiple times and in case all attempts were unsuccessful, we assigned an average value of 0.5 as the final score.

Fine-tuning

We chose a multilingual XLM RoBERTa[4] model with 279M parameters for the fine-tuning experiments. A linear layer was added on top of it to handle the regression task. As loss function we use Huber Loss for robust regression that combines the advantages of Mean Squared Error and L_1 loss, reducing sensitivity to outliers. We found that for a small amount of data we needed to train the model for more epochs with a relatively small learning rate of $2e$-5 and we employed early stopping to avoid overfitting. The maximum amount of epochs was set to 20, and we used AdamW as an optimizer.

3.2.2 Traditional Methods

Heylighen & Dewaele formality score was introduced in Heylighen et al. [1999] and it measures formality of a text based on the frequency of different parts of speech tags. This metric relies on the assumption that formal texts use more nouns and prepositions, while informal ones use more verbs and pronouns. Higher

[2] `https://huggingface.co/Qwen/Qwen2.5-7B-Instruct-GPTQ-Int4`

[3] `https://huggingface.co/TechxGenus/Meta-Llama-3-8B-Instruct-GPTQ`

[4] `https://huggingface.co/FacebookAI/xlm-roberta-base`

scores correspond to higher levels of formality and each part of speech is represented as a percentage of total words.

$$F = \frac{(noun + adjective + preposition + article) - (pronoun + verb + adverb + interjection) + 100}{2}$$

In this work we use a slightly modified version of this score by removing the constants and mapping the score to the same range as in the other experiments, i.e., $F \in [0, 1]$. We use *SpaCy*[5] POS tagging tools for computing the number of words with different parts of speech.

Gunning Fog Index Gunning [1968] is another popular metric that measures text complexity and estimates the U.S. school grade needed to understand the text. It focuses on sentence length as well as the number of complex words that consist of three or more syllables.

$$FogIndex = 0.4 \times (\frac{words}{sentences} + 100 \times \frac{complex\_words}{words})$$

Flesch Reading Ease Flesch [1948] rates readability based on sentence length and word syllables, here higher scores mean easier text.

$$FRE = 206.835 - 1.015 \times (\frac{words}{sentences}) - 84.6 \times (\frac{syllables}{words})$$

Flesch-Kincaid Grade Level Kincaid et al. [1975] also focuses on sentence length and number of syllables per word, but it converts the score into a U.S. school grade level.

$$FKGL = 0.39 \times (\frac{words}{sentences}) + 11.8 \times (\frac{syllables}{words}) - 15.59$$

Automated Readability Index Smith and Senter [1967] is another popular metric that estimates the U.S. school grade level required to understand a text, based on characters per word and words per sentence. Unlike Flesch-based scores (which use syllables), ARI considers the number of characters, words, and sentences.

$$ARI = 4.71 \times (\frac{characters}{words}) + 0.5 \times (\frac{words}{sentences}) - 21.43$$

We use *textstat*[6] library to compute Flesch Reading Ease, Flesch-Kincaid Grade Level, Gunning Fog Index, and Automated Readability Index. All metrics are mapped to the same range of $[0, 1]$ and then averaged.

Lexical Diversity score is computed as a type-token ratio, i.e. number of unique tokens in each sentence divided by the total number of tokens in that sentence.

Syntactic Complexity is calculated as a ratio between the input sentence length and the average sentence length of all formal sentences in the training data (we consider as formal all sentences that have the score $>= 0.75$).

3.3 Evaluation Metrics

We evaluate how well formality annotations obtained with different methods align with the original scores (re-scaled to the same range) using the following metrics: Pearson Correlation, Spearman Correlation, Mean Absolute Error, and Root Mean Squared Error.

Pearson Correlation Pearson [1895] measures linear relationship between the predicted values and the true values. Here x represents the predicted value, y is the actual value of the score, and $\bar{x}$ and $\bar{y}$ correspond to the mean values. This metric ranges from -1 (no correlation) to 1 (very positive correlation).

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2}\sqrt{\sum (y_i - \bar{y})^2}}$$

Spearman Correlation Binet [1904] is a non-parametric measure of monotonic relationships (i.e., whether variables increase/decrease together, not necessarily linearly). It is based on the rank order of

[5] `https://spacy.io`

[6] `https://github.com/textstat`

the values (see the formula below where n is the total number of samples and d_i is the difference between the ranks of predicted and actual values). Similarly to Pearson correlation, this metric ranges from -1 to 1.

$$p = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

Mean Absolute Error Willmott and Matsuura [2005] measures the average magnitude of errors in predictions (regardless of the direction) with lower MAE scores meaning overall better model performance. Here n is the total number of samples, y_i is the actual value, and $\hat{y_i}$ is the predicted value.

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y_i}|$$

Root Mean Squared Error Chai and Draxler [2014] is a square root of the average of squared errors (see the formula below that uses the same notation as above). This metric penalizes large errors more than MAE, although it is also more sensitive to outliers.

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y_i})^2}$$

4 Experiments and Results

We summarize the scores obtained with different metrics in Table 4 for *InFormal Sentences* and Table 4 for *Pavlick Formality.* It is evident from the results that LLM-based approaches perform much stronger than traditional methods and achieve the best scores.

For instance, in case of *InFormal Sentences* we achieve best performance across all metrics with the fine-tuned XLM RoBERTa, followed by the prompting approaches with Llama3-8B which is slightly outperforming Qwen-7B. Among the traditional approaches we observe that average readability scores are the best indicator for the formality level. Interestingly, syntactic complexity shows a very similar pattern and can be also considered as a reliable measure of formality. Heylighen score based on POS tag information is not very competitive, although it still shows positive correlation between the gold and annotated scores. Lexical diversity results in the worst scores and even negative correlation, likely because this metric is not meant for shorter texts that typically do not contain many repetitions, so the type-token ratio is quite high for both formal and informal sentences in this case.

	Pearson	**Spearman**	**MAE**	**RMSE**
Llama-8b prompting	**0.827**	**0.844**	**0.105**	**0.132**
Qwen-7b prompting	0.808	0.823	0.137	0.172
Fine-tuned XLMR	**0.911**	**0.909**	**0.068**	**0.086**
Heylighen score	0.313	0.370	0.193	0.249
Avg. readability score	0.649	0.659	0.168	0.213
Lexical diversity	-0.306	-0.346	0.454	0.509
Syntactic complexity	0.641	0.620	0.178	0.220

Table 1: Evaluation results for different approaches on the *InFormal Sentences* data.

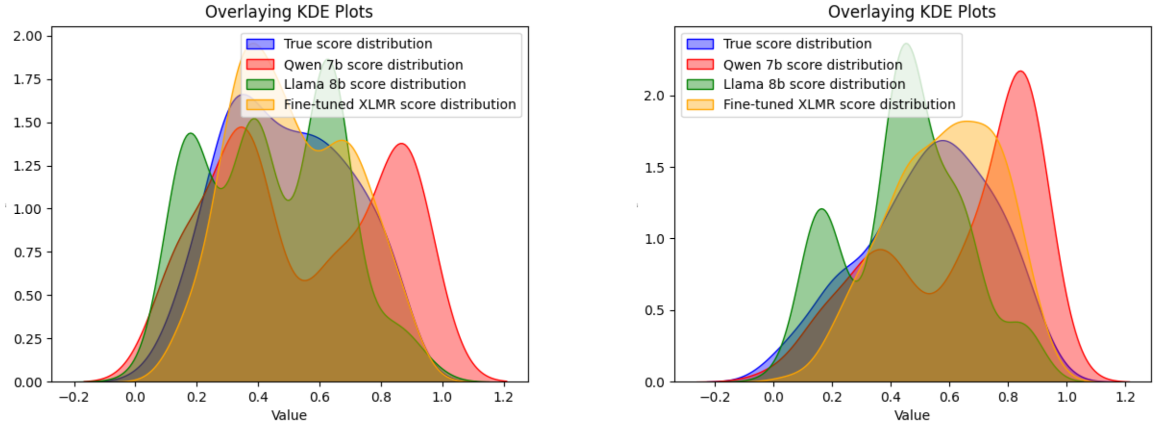
The evaluation results for the *Pavlick Formality* data show a similar trend with LLM-based approaches consistently outperforming traditional methods. However, for English data Qwen exhibits higher Pearson and Spearman correlation scores compared to the Llama model that was the best for *InFormal Sentences.* Overall, few-shot prompting and fine-tuning are the best strategies also for this dataset. Readability scores and syntactic complexity perform on par, followed by Heylighen score, and lexical diversity results in the worst performance and negative correlation.

We also plot the score distribution based on the different tested methods and the gold data using Kernel Density Estimate (KDE) plot, which estimates the probability density function of a continuous variable, and provides a representation of the data distribution. See Figure 1 for the KDE plots based on the LLM-based methods and Figure 2 for the traditional ones. The blue area represents the gold score distribution. We can

	Pearson	Spearman	MAE	RMSE
Llama-8b prompting	0.659	0.657	**0.148**	**0.191**
Qwen-7b prompting	**0.763**	**0.733**	0.158	0.198
Fine-tuned XLMR	**0.749**	**0.740**	**0.124**	**0.157**
Heylighen score	0.363	0.347	0.187	0.243
Avg. readability score	0.552	0.569	0.231	0.279
Lexical diversity	-0.274	-0.330	0.409	0.477
Syntactic complexity	0.537	0.548	0.225	0.274

Table 2: Evaluation results for different approaches on the *Pavlick Formality* data.

observe that LLM-based methods have large overlap with the gold distribution. Interestingly, Llama (green area) has the tendency of assigning lower scores than Qwen (red area) and the distribution has peaks, meaning that the scores concentrate around specific values (e.g., 0.1 and 0.5 if we look at the Llama annotations for the *Pavlick Formality* data).



(a) *InFormal Sentences*

(b) *Pavlick Formality*

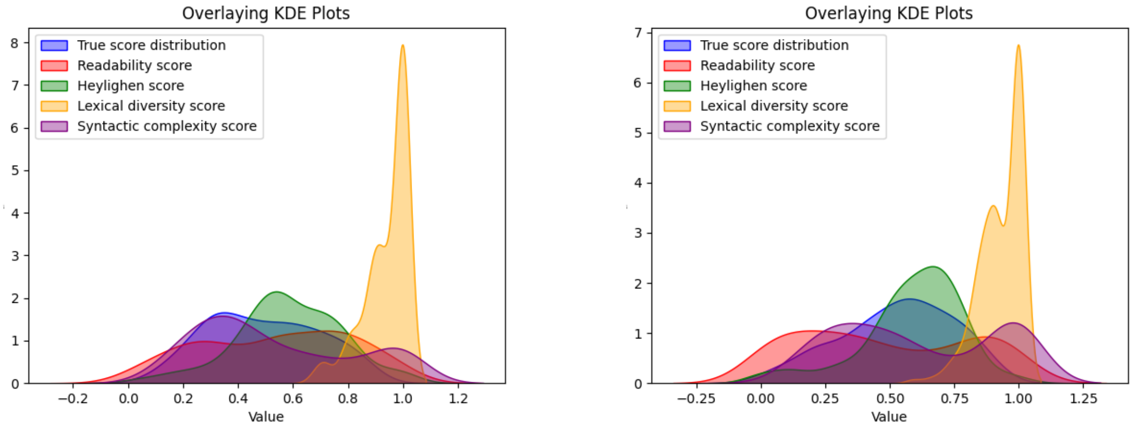
Figure 1: Distribution of scores for the LLM-based methods (fine-tuning and prompting) vs. gold scores.

When we visualize the score distribution for traditional metrics (see Figure 2) it becomes clear why lexical diversity scores do not work well as formality indicators. The distribution is very spiky and all the scores are concentrated in the high range between 0.75 and 1 since most sentences in the dataset do not have any repeated tokens irrespective of their formality levels. Readability and syntactic complexity scores are relatively spread out and the distribution does not have pronounced peaks. For Heylighen measure we can observe that the majority of assigned scores are around 0.6 which makes it a “cautious” metric that typically does not provide very high or very low scores.

Overall, our experiments show that LLM-based approaches can capture formality levels better than the traditional methods. However, they also have some disadvantages. For instance, LLMs require much more computational power, and running Llama or Qwen models with 7 or 8 billion parameters means that we need a GPU, although in this work we experimented with the quantized models that require less memory than the original full precision models.

Also, if we want to fine-tune a model we need to have some training data available, and the availability of data for different domains and languages remains a big challenge. Traditional approaches often use some predefined formulas and statistics that are cheap and fast to compute which can be considered as an advantage. However, some of the formulas also include parameters that are language-specific or corpus-dependent. For instance, if we consider syntactic complexity score and compare the average sentence length of German formal sentences (32.36), it is substantially higher than the corresponding length for English formal sentences (26.02). Also, metrics such as Heylighen score require POS tags, which means that we need to use some external tools that were trained to perform POS tagging in the target language.

On the other hand, traditional metrics are also more interpretable by design because they are often derived



(a) *InFormal Sentences*

(b) *Pavlick Formality*

Figure 2: Distribution of scores for the traditional methods (readability, Heylighen score, lexical diversity and syntactic complexity) vs. gold scores.

based on some formulas and corpus statistics. It would be interesting to see whether using the scores from traditional metrics as additional input to LLMs can change model behavior and potentially speed up the training or even make training more feasible in some low-resource scenarios. Note that one could also use explainability methods such as e.g. token attribution or layer integrated gradient Sundararajan et al. [2017] to improve the interpretability of LLMs because such methods could help us identify which (sub-)tokens LLM consider most important for the formality detection task.

5 Conclusion and Outlook

In this work we explored different approaches towards annotating text formality levels with LLMs and traditional methods. We have found that LLMs are very competitive on this task and can achieve best results either if fine-tuned on a small amount of data (XLM RoBERTa) or prompted in a few-shot fashion to generate the scores without any additional training (Qwen and Llama). Among the traditional scores, average readability scores and syntactic complexity showed consistently good performance, but Heylighen score that is based on POS tag information tends to be less reliable, while lexical diversity exhibits negative correlation with the gold labels. We also found that all metrics behave in a very similar way for both datasets in English (*Pavlick Formality*) and German (*InFormal Sentences*).

We acknowledge the limitations of this project due to the dataset and model selection. It would be interesting to experiment with more models, use zero-shot prompting, or fine-tune LLMs with adapters which would enable efficient fine-tuning even for very large models. Also, in this work we sampled all datasets to $1,000$ instances, and did not differentiate between the domains (e.g., news vs. blogs). However, assessing text complexity and formality levels may vary in difficulty depending on the domain, and this is an important dimension to consider in the future work.

Another direction worth exploring would be to use e.g. the coreset approach Sener and Savarese [2017] to identify the most representative sentences for each coarse-grained formality level, and estimate the cosine similarity of the sentence embedding corresponding to the most representative samples at each formality level and the input sentence (sentence-level embeddings can computed e.g. with SBERT Reimers and Gurevych [2019]). This would work even with relatively small amount of training data available for the coreset selection, and computing cosine similarity would be much cheaper than running inference on LLMs like Llama or Qwen. Besides, if the domain changes we can easily update or extend the coreset examples, which is more efficient than fine-tuning the entire model on the new data.

References

Francis Heylighen, Jean–Marc Dewaele, and Léo Apostel. Formality of language: definition, measurement and behavioral determinants. 1999. URL https://api.semanticscholar.org/CorpusID:16450928.

Tong Wang, Julian Brooke, and Graeme Hirst. Inducing lexicons of formality from corpora. 2010. URL https://api.semanticscholar.org/CorpusID:6742665.

Sudha Rao and Joel Tetreault. Dear sir or madam, may I introduce the GYAFC dataset: Corpus, benchmarks and metrics for formality style transfer. In Marilyn Walker, Heng Ji, and Amanda Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 129–140, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1012. URL https://aclanthology.org/N18-1012/.

Eleftheria Briakou, Di Lu, Ke Zhang, and Joel Tetreault. Olá, bonjour, salve! XFORMAL: A benchmark for multilingual formality style transfer. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou, editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3199–3216, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.256. URL https://aclanthology.org/2021.naacl-main.256/.

Ellie Pavlick and Joel Tetreault. An empirical analysis of formality in online communication. *Transactions of the Association for Computational Linguistics*, 4:61–74, 2016. doi: 10.1162/tacl_a_00083. URL https://aclanthology.org/Q16-1005/.

Shibamouli Lahiri. Squinky! a corpus of sentence-level formality, informativeness, and implicature. *ArXiv*, abs/1506.02306, 2015. URL https://api.semanticscholar.org/CorpusID:7716488.

Kunal Chawla and Diyi Yang. Semi-supervised formality style transfer using language model discriminator and mutual information maximization. In Trevor Cohn, Yulan He, and Yang Liu, editors, *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2340–2354, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.212. URL https://aclanthology.org/2020.findings-emnlp.212/.

Daryna Dementieva, Nikolay Babakov, and Alexander Panchenko. Detecting text formality: A study of text classification approaches. In Ruslan Mitkov and Galia Angelova, editors, *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 274–284, Varna, Bulgaria, September 2023. INCOMA Ltd., Shoumen, Bulgaria. URL https://aclanthology.org/2023.ranlp-1.31/.

AI@Meta. Llama 3 model card. 2024. URL https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md.

Team@Qwen. Qwen2.5: A party of foundation models, September 2024. URL https://qwenlm.github.io/blog/qwen2.5/.

Elisabeth Eder, Ulrike Krieg-Holz, and Michael Wiegand. A question of style: A dataset for analyzing formality on different levels. In Andreas Vlachos and Isabelle Augenstein, editors, *Findings of the Association for Computational Linguistics: EACL 2023*, pages 580–593, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-eacl.42. URL https://aclanthology.org/2023.findings-eacl.42/.

Robbie Gunning. The technique of clear writing. 1968. URL https://api.semanticscholar.org/CorpusID:145838278.

Rudolf Franz Flesch. A new readability yardstick. *The Journal of applied psychology*, 32 3:221–33, 1948. URL https://api.semanticscholar.org/CorpusID:39344661.

Peter Kincaid, Robert P. Fishburne, Richard L. Rogers, and Brad S. Chissom. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. 1975. URL https://api.semanticscholar.org/CorpusID:61131325.

E A Smith and R. Senter. Automated readability index. *AMRL-TR. Aerospace Medical Research Laboratories*, pages 1–14, 1967. URL https://api.semanticscholar.org/CorpusID:38558516.

Karl Pearson. Vii. note on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, 58:240 – 242, 1895. URL https://api.semanticscholar.org/CorpusID:121644161.

Alfred Binet. Spearman the proof and measurement of association between two things; general intelligence objectively determined and measured. *Annee Psychologique*, 11:623–624, 1904. URL https://api.semanticscholar.org/CorpusID:147128166.

Cort J. Willmott and Kenji Matsuura. Advantages of the mean absolute error (mae) over the root mean square error (rmse) in assessing average model performance. *Climate Research*, 30:79–82, 2005. URL https://api.semanticscholar.org/CorpusID:120556606.

Tianfeng Chai and Roland R. Draxler. Root mean square error (rmse) or mean absolute error (mae)? – arguments against avoiding rmse in the literature. *Geoscientific Model Development*, 7:1247–1250, 2014. URL https://api.semanticscholar.org/CorpusID:123118384.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International Conference on Machine Learning*, 2017. URL https://api.semanticscholar.org/CorpusID:16747630.

Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. *arXiv: Machine Learning*, 2017. URL https://api.semanticscholar.org/CorpusID:3383786.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019. URL https://arxiv.org/abs/1908.10084.