The Overlooked Lengthening Words for Sentiment Analysis

Anonymous ACL submission

Abstract

Recent advancements in Large Language Models (LLMs) have catalyzed pivotal discussions about their proficiency in sentiment analysis and informal styles. This work focuses on lengthening words, a distinctive but often overlooked nuanced expression. Although around 5.8% of documents in 4 public datasets contain lengthening words, there is no ready-to-use dataset for this nuanced expression. To evaluate the performance and interpretation of SOTA LLMs for stretched words, we introduce **Lengthening**, the first weakly labeled sentence-level dataset consisting of 850k samples with lengthed words for sentiment analysis. Comprehensive experiment results involving 5 LLMs and 5 domains reveal the sentiment-expressive value of lengthening words from both sentence and word levels. This study highlights the importance of lengthening words for online social media, where short text and nuanced communication prevail. Our results show that pretrained LLMs can reach better performances than zero-shot ChatGPT4 after fine-tuning on the Lengthening dataset but might be insufficient in explainability. With an explainable instruction dataset consisting of only 1000 samples, instruct-tuned LLaMA2 (13B-chat-hf) can reach the same level of performance and explainability as ChatGPT4.

1 Introduction

Informal styles are prevalent on social media platforms, where people use nuanced expressions to share opinions and emotions personally and engagingly. Previous research has explored various informal styles such as Meme (Lin et al., 2024; Sharma et al., 2023), Emoji (Peng et al., 2023; Barbieri et al., 2018; Reelfs et al., 2022), slogan (Iwama and Kano, 2018; Misawa et al., 2020) and abbreviation (Gorman et al., 2021). However, it remains a challenge for NLP models to understand the sentiment in nuanced and subtle linguistic expressions, which require a deep contextual and cultural understanding (Zhang et al., 2023b). This work delves into one specific informal expression - lengthening words, which refers to the linguistic phenomenon where additional characters are added to the standard spelling of a word to enhance or alter its conveyed meaning (Brody and Diakopoulos, 2011; Kalman and Gergle, 2014). Based on previous research, we generalize the concept of word lengthening and divide it into two types: Letter Repetitive (e.g., 'loooove') and Punctuation Repetitive (e.g., 'love!!!!'). Our study finds that an average of 6% documents possess lengthening expression among 4 public datasets (Table 1) where some of them may include more than 13% such documents. Prior research has leveraged lengthening words to augment sentiment lexicons (Brody and Diakopoulos, 2011). (Schnoebelen et al., 2012; Gray et al., 2020) describe this nuanced linguistic as expressive words that are often used to emphasize or exaggerate the underlying meaning of the root word. However, they are more focused on linguistic and statistical analysis. The sentiment value of lengthening words for up-to-date techniques still lacks systematic measurement and confirmation.

Large Language Models (LLMs) such as RoBERTa (Liu et al., 2019), T5 (Raffel et al., 2019), ChatGPT4 (OpenAI, 2023), and LLaMA2 (Touvron et al., 2023), exhibit impressive performance in sentiment classification tasks (Wang et al., 2023). However, we still know little about the boundaries of performance and interpretation of LLMs for sentiment analysis with lengthening words because lengthening words are often overlooked (Go, 2009; Nguyen and Nguyen, 2017; Abdul-Mageed and Ungar, 2017; Ali et al., 2019). There is no ready-to-use dataset designed for this informal style. The case study of document-level sentiment analysis in Figure 2 shows that state-of-the-art (SOTA) pretrained LLMs can reach 90% overall accuracy for documents with lengthening words. However, their

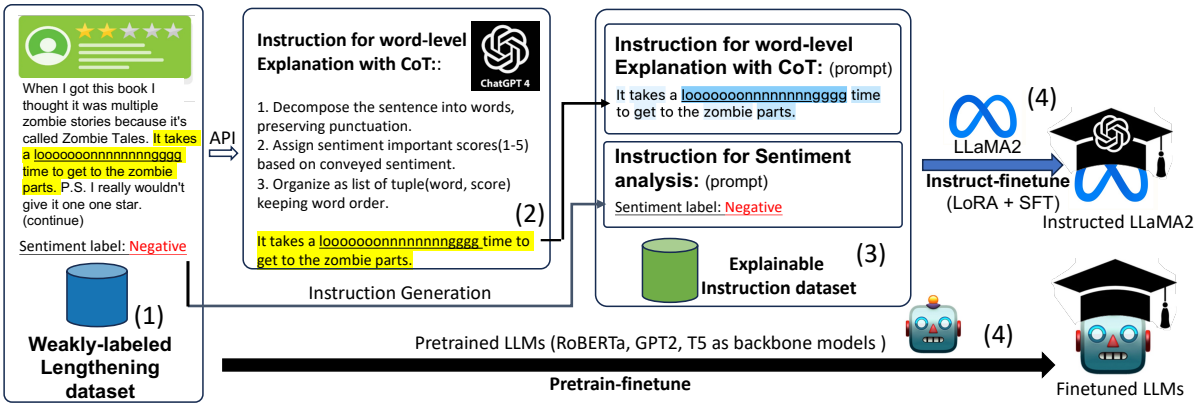


Figure 1: An overview of our framework for lengthening words.

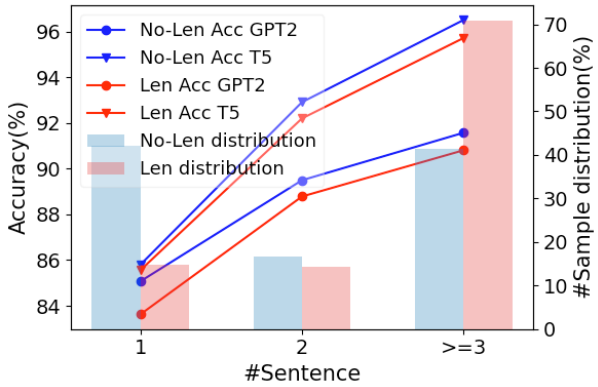


Figure 2: Comparision of document-level zero-shot performance for GPT2(large) and T5 (base) model. We randomly sample 30k documents with/without lengthening words.

performance experiences a sharp drop to lower than 86% for short documents with one sentence. This leads us to question whether LLMs can understand the sentiments contained in nuanced expressions. They may ignore text fragments with lengthening words but still can make accurate sentiment inferences relying on linguistic redundancy. For example, in Figure 3, the model successfully predicts the sentiment label of a sentence with the Letter Repetitive style lengthening word. However, the saliency map reveals a disproportionate emphasis on words irrelevant to sentiment expression. It's worth asking the question if LLMs are "right for the wrong reasons" (McCoy et al., 2019) when processing lengthening words.

To address these issues, we introduce **Lengthening**, a weakly labeled sentence-level dataset with 860k samples featuring word lengthening based on 4 public datasets. Furthermore, we unify the saliency method and the prompting strategy with chain of thought to get saliency maps, which represent the model's ability to interpret the sentiment value of lengthening words. Finally, we compare ChatGPT4 with LLaMA2, and other state-of-the-art (SOTA) pre-trained transformer-based models and focus on their zero-shot and fine-tuned performance and explanation on **Lengthening**. The main findings of this evaluation are as follows:

1. We call attention to a distinctive informal style - lengthening word - for sentiment analysis. We introduce **Lengthening**, a weakly labeled sentences-level dataset featuring lengthening styles based on 4 public datasets. It's the first benchmark with 850k samples that focus on comprehensively evaluating and improving LLMs' performance and explainability for lengthening words.

2. We unify the saliency method and prompt strategy to evaluate LLMs' word-level interpretation for lengthening words. Our results show that pre-trained models (Roberta, GPT2, and T5) can reach better performances than zero-shot ChatGPT4 after fine-tuning on the Lengthening dataset but might be insufficient in explainability. With an explainable instruction dataset consisting of only 1000 samples, we can instruct-tune LLaMA2 (13B-chat-hf) to reach the same level of performance and explainability as ChatGPT4.

3. We conduct comprehensive sentence-level sentiment analysis experiments across 5 domains and 6 LLMs. Our results confirm the sentiment-expressive value of lengthening words from both sentence and word levels. Our findings shed light on the importance

Domain	Source	#Documents	Len Rate(%)	#Samples	Label Type
Books	Amazon Review	51,312k	5.41%	255k	5 stars
Electronics	Amazon Review	20,994k	4.32%	212k	5 stars
Restaurants	Yelp	6,990k	11.28%	315k	5 stars
Social Media	Twitter	1,600k	13.36%	44k	binary
Hotels	TripAdvisor	879k	10.01%	24k	5 stars
ALL	—	80,873k	5.80%	850k	binary

Table 1: Data sources, lengthening rate and samples distribution after downsampling of **Lengthening** benchmark

of lengthening words in online social media, where short text and nuanced communication prevail.

2 Related Work

Lengthening Words The study of (Kalman and Gergle, 2014) shows that lengthening words is a written emulation of nonverbal spoken cues in user-generated content and proposed a method to reduce stretched words to their root words. (Gray et al., 2020) generated normalized forms for different lengthening styles. For example, variations like 'looove' and 'looooove' have same normalized form 'lo+ve', and 'looovvve' normalized to 'lo+v+e'. They have the same root word 'love'. (Brody and Diakopoulos, 2011) underscore the importance of accurately interpreting lengthening words for sentiment analysis. We further generalize the concept of lengthening to two different styles: Punctuation Repetitive and Letter Repetitive. Based on previous research, we introduce **Lengthening** benchmark and detailed in section 3.1.

Model Explanation , particularly the importance of input features, has been extensively explored in recent years. Studies such as (Ribeiro et al., 2016; Li et al., 2017; Zhu et al., 2023; Ray Choudhury et al., 2022; Atanasova et al., 2020) have employed saliency-based methods for pretrained-finetune (Wei et al., 2021) language models. Prompting methods(Zhong et al., 2023; Wang et al., 2023) with Chain of Thought (CoT) (Wei et al., 2023) are proposed for instruction-based LLMs, including LLaMA2 (Touvron et al., 2023) and ChatGPT4. Our work unifies the saliency method and prompt strategy to evaluate LLMs' word-level interpretation for lengthening words (section 3.2).

3 Methodology

In this section, we describe the technical details of our framework. As shown in Figure 1, our work for lengthening words consists of 4 major parts, including (1) generation of **Lengthening**, a weakly labeled sentence-level dataset for sentiment analysis; (2) prompting ChatGPT4 with CoT for word important score; (3) generation of the explainable instruction dataset; (4) instruction tuning LLaMA2 and fine-tuning pre-trained LLMs.

3.1 Generation of Lengthening benchmark

To comprehensively evaluate and pinpoint the usage of lengthening words in online user reviews and social media platforms, we select four public datasets covering five distinct domains: Amazon Review(Ni et al., 2019) (Books & Electronics); a snapshot of Yelp(Yelp, 2021) (Restaurants) data from Feb 16, 2021; Twitter dataset (Go, 2009) (Social media); TripAdvisor(Li, 2020) (Hotels). We present details of datasets in Table 1, including the ratio of documents containing lengthening words and sample size used in this work.

We extract sentences with Lengthening words and label them with the following steps. First, we segment documents into individual sentences using an open-source tool based on regular expressions[1], merging sentences with fewer than five words to preserve meaningful text fragments. Next, we utilize regex patterns to identify and locate lengthened words (Gray et al., 2020). Sentences with lengthening words will be extracted and assigned the sentiment label of their associated documents. We assign each lengthening word its root word (Kalman and Gergle, 2014) based on American English [2]. Moreover, for each sentence with lengthened words, a non-lengthened sentence from the

[1] https://stackoverflow.com/questions/4576077/how-can-i-split-a-text-into-sentences

[2] https://pyenchant.github.io/pyenchant/tutorial.html

Domain	Lengthening Style	POS	Lengthening Word	Root Word	Label	Example Sentence
Books	Punctuation	Noun	book!!!!!	book!	1	Do yourself a favour a read this book!!!!!
Electronics	Letter	Verb	loooove	love	1	I loooove my new phone case.
Restaurants	Punctuation	Adj	amazing!!!!!	amazing!	1	We are from Seattle and this coffee is amazing!!!!!
Twitter	Letter	Adv	SOOOO	SO	0	SOOOO bummed i'm going to miss sam's party tonight.
Hotel	Punctuation	Noun	year.....	year...	0	I am looking to go back next year.....

Table 2: Samples from **Lengthening** Benchmark

Metric	Acc(%)		F1(%)	
Backbone Model	Len	Non-Len	Len	Non-Len
RoBERTa (large)	86.06	84.13	**90.74**	89.37
GPT2 (medium)	79.71	77.45	85.86	84.15
T5(base)	83.42	81.82	88.83	87.77
LLaMA2 (13B)	77.52	74.55	83.86	81.36
ChatGPT4	**86.49**	**86.21**	90.72	**90.67**

Table 3: Overall zero-shot accuracy and F1 score for sentences with lengthening words(Len) and sentences without lenghtening words(Non-len). The **best** results of columns are highlighted in bold, and the second best results are underlined.

same document is randomly chosen for comparison (we only consider documents containing two or more sentences). Finally, we apply downsampling to top domains and lengthening types to balance the data distribution. All samples are transformed into binary classifications with 1-2 stars as negative and 4-5 stars as positive (we don't consider samples with 3 stars). We divided the dataset into training (70%), validation (20%), and test sets (10%).

3.2 Evaluating LLMs' word-level interpretation with saliency score

Given a transformer-based pre-trained model (denoted as f), consider dataset D containing an input sentence x containing a lengthened word w with a sentiment label as y.

Prompting with Chain of Thought for Instruction LLMs In the prompt construction, we explicitly incorporate the CoT method to demonstrate the reasoning process as shown in Figure 1 part (2). In contrast with the occlusion method, which gets word importance scores from the model's loss function, we prompt LLMs with CoT to output importance scores(ranging from 1 to 5) for each word, as shown in the following:

$$S(x) = f(x|\text{content}, \text{prompt}) \quad (1)$$

Saliency-based method for pre-trained LLMs We choose the occlusion-based method to determine the importance scores of individual words within a sentence. This method involves systematically obscuring each word w_i to observe the absolute change in the loss function, which serves as an indicator of the word's significance. The word important score $S(x, y)$ for each words in sentence x with ground truth label y is written as

$$S(x, y) = \{|L_f(x, y) - L_f(x - w_i, y)|, w_i \in x\} \quad (2)$$

where $L_f(x, y)$ represents the loss value by model f for the input pair (x, y).

We first apply min-max normalization to the initial weight score S, followed by an L_1 normalization. This process results in the normalized score S_{norm}. The expressive value for the lengthening word w is then calculated using the equation 3.

$$s_{ev}(x, w, y) = S_{norm}(x, w, y)[j_{(w_j=w)}] \quad (3)$$

Here, $s_{ev}(x, w, y)$ represents expressive value (normalized importance score) of the lengthening word w within sentence x and its corresponding label y. We get S_{ev} by applying Equation 4 to all sample pairs to assess the model's overall capability in explaining the expressive value of lengthening words. Complementing, we also employ accuracy and F1-score to measure overall performance in sentiment analysis tasks. Those metrics provide a comprehensive and systematic evaluation of LLMs' performance and explainability for the Lengthening benchmark.

$$S_{ev} = E_{(x,w,y)\sim\mathcal{D}}\, s_{ev}(x, w, y) \quad (4)$$

3.3 Generation of explainable instruction dataset

We sample 1000 samples from the training set and annotated them with word importance scores obtained by prompting ChatGPT4 with CoT, as detailed in Section 3.2. Utilizing these word importance scores alongside ground truth sentiment labels, we generated an explainable instruction dataset as shown in Figure 1 (3). We randomly selected 2000 samples from the test set to evaluate the performance of LLaMA2-13B and ChatGPT4.

Backbone Model	RoBERTa (large)				GPT2 (medium)				T5 (base)			
Metric	Acc(%)		F1(%)		Acc(%)		F1(%)		Acc(%)		F1(%)	
Domain / Group	Len	No-Len	Len	No-Len	Len	No-Len	Len	No-Len	Len	No-Len	Len	No-Len
Books	**87.63**	86.56	**92.94**	92.27	81.13	81.15	88.79	88.84	86.53	86.70	92.30	92.43
Restaurants	**88.36**	85.84	**91.94**	90.15	82.47	78.96	87.36	84.63	85.32	83.01	89.67	88.08
Electronics	**84.81**	82.10	**89.10**	86.98	76.67	73.15	82.10	78.96	80.66	77.77	85.80	83.55
Twitter	66.74	**67.11**	**69.10**	68.57	67.03	66.41	66.38	65.98	65.47	64.45	66.55	65.22
Hotel	**84.37**	83.66	**89.25**	88.78	77.04	74.96	83.28	81.88	81.17	80.46	86.76	86.50

Table 4: Zero-shot accuracy and F1 score for sentences with/without lenghthening words(Len/Non-Len) under different domains, we only report RoBERTa-large, GPT2-medium and T5-base because test samples limited for ChatGPT and LLaMA-13B-chat-hf. The **best** results of rows are highlighted in bold, and the second best results are underlined.

		RW	RS
(1)	La cabra falls prey to the common mistake of having waaayyy to many heavy IPA's and high ABV's.	5	4
(2)	La cabra falls prey to the common mistake of having waaayyy to many heavy IPA's and high ABV's.	1	1
(3)	La cabra falls prey to the common mistake of having waaayyy to many heavy IPA's and high ABV's.	3	2

Figure 3: Example for Human evaluation, we use the word important score of 'waaayyy' to represent the model's comprehension of the expressive value of the lengthened word in this sentence. RW and RS represent the score of reliability in word and sentence level.

This sample size aligns with the standards in previous studies (Wang et al., 2023; Zhang et al., 2023a; Deng et al., 2022).

3.4 Instruct-tuning and Fine-tuning LLMs

We instruct-tune LLaMA2-13B with the explainable instruction dataset and utilize LoRA (Hu et al., 2021) and Supervised fine-tuning (SFT) (von Werra et al., 2020). Specifically, we set the LoRA attention dimension as 64, employ 4-bit quantization for weight storage, and utilize float16 precision for computational processes for LLaMA2. We fine-tune pre-trained models with backbones as RoBERTa-large, GPT2-medium, and T5-base with 33% of the train and validation set and use cross-validation to identify the optimal checkpoints, subsequently evaluating these models on the whole test set.

4 Experiments

4.1 Experiment Setup

In our study, we select five SOTA pre-trained language models (LMs) encompassing diverse sizes and architectural frameworks for comprehensive assessment, including SiEBERT[3] (Hartmann et al., 2023), a RoBERTa-large based model fine-tuned on diverse text sources; Finetuned LMs with backbone as GPT2-medium[4] , T5-base[5]; LLaMA2-13B-chat-hf[6](For convenience we use LLaMA2-13B in the later content.) and ChatGPT4[7]. Our interactions with ChatGPT4 are conducted through the OpenAI API, specifically using the gpt-4 version. We use Transformer library (Wolf et al., 2020) for accessing others.

4.2 Are lengthening words important for sentence-level sentiment analysis?

Zero-shot performance We assess models' zero-shot performance for sentence-level sentiment analysis with accuracy and F1 scores as metrics. Specifically, we compare two groups: sentences containing lengthening words (Len) and those without (No-len). The result is shown in Table 3. Interestingly, the Len group consistently outperforms the No-Len group in accuracy and F1-score across all models. This evidence indicates sentences with lengthening words correlate with documents' sentiment polarities better than sentences without lengthening words. This finding shows that lengthening words make sentences carry heightened sentiment values,

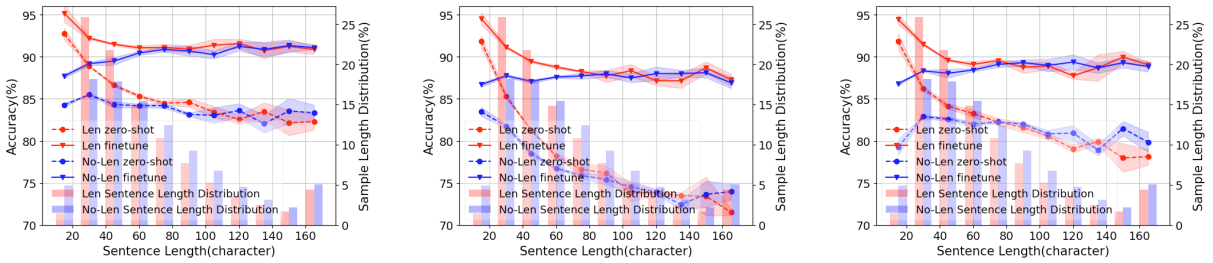
[3]https://huggingface.co/siebert/sentiment-roberta-large-english

[4]https://huggingface.co/michelecafagna26/gpt2-medium-finetuned-sst2-sentiment

[5]https://huggingface.co/mrm8488/t5-base-finetuned-imdb-sentiment

[6]https://huggingface.co/meta-llama/Llama-2-13b-chat-hf

[7]https://platform.openai.com/playground?mode=chat&model=gpt-4



(a) RoBERTa (large) (b) GPT2 (medium) (c) T5 (base)

Figure 4: Comparison of accuracy between the Len (lengthened) and Non-len (non-lengthened) groups using zero-shot and fine-tuned models by sentence length. The lines represent average values, and the shaded regions illustrate the standard deviation across the three runs. Both results across 3 models with and without finetuning show a convergence between the Len and No-Len groups when the sentence character length is around 80.

Backbone Model	S_{ev}	Acc(%)	F1(%)
RoBERTa (large)	0.21	86.06	**90.74**
GPT2 (medium)	**0.40**	79.71	85.86
T5 (base)	0.18	83.42	88.83
LLaMA2 (13B)	0.18	77.52	83.86
ChatGPT4	0.37	**86.49**	90.72

(a) Zero-shot performance

RoBERTa (large)	0.23	**91.54**	**94.58**
GPT2 (medium)	0.29	89.30	93.20
T5 (base)	0.23	89.87	93.54
LLaMA2 (13B)	**0.40**	89.80	93.29

(b) Finetuned performacne

Table 5: Comparision performances between zero-shot and finetuned models for Lengthening benchmark. The **best** results of columns are highlighted in bold, and the second best results are underlined.

reinforcing their importance in sentiment analysis. As shown in Table 4, this superiority of lengthening words in expressing sentiment is consistently observed across diverse domains and models, underscoring the generalization of this finding.

Fine-tuned performance We compare fine-tuned performance with zero-shot performance and show results in Table 5. We observe that ChatGPT4 has the best zero-shot performance, and all fine-tuned models can reach better performances than this best zero-shot. Fine-tuned RoBERTa-large has the best performance and reaches 91.54% accuracy among fine-tuned models.

Furthermore, we study the accuracy of zero-shot and finetuned RoBERTa-large, GPT2-medium, and T5-base in relation to sentence character length. The results are presented in Figure4. After fine-tuning with the Lengthening dataset, we observe obvious improvements for all models under all length ranges. Sentences with lengthening words consistently exhibit higher accuracy both in zero-shot and fine-tuned results. This trend reinforces our conclusion that lengthening words has significant value for sentiment analysis. It is noteworthy that the discrepancy in accuracy between the Len and No-Len groups is most pronounced for short sentences within 80 characters, where more than 70% samples in this range. This highlights the critical need to focus on lengthening words within the realm of social media, which is predominantly characterized by short and informal expressions.

4.3 Using saliency score to evaluate interpretation of LLMs for lengthening words at word-level

LLMs' understanding of the expressive value of lengthened words can be reflected in their ability to allocate appropriate importance scores to these words within, which is like human intuitive sense. We use S_{ev} proposed in 3.2 to measure models' overall interpretation for the test set of Lengthening. The results of zero-shot and fine-tuned experiments are presented in Table 5. Our results show that ChatGPT4 has the best zero-shot interpretation of lengthening words. Although GPT2-medium has the highest zero-shot S_{ev} score, the low accuracy and F1 score suggest an insufficient understanding of lengthened words. This can be further supported by the drop of S_{ev} score after fine-tuning for GPT2-medium in Table 5 (b). We observe that fine-tuned RoBERTa-large, T5-base, and instructed LLaMA2-13B can better interpret lengthening words. The improvement in accuracy and saliency score after fine-

Lengthening Style	Punctuation Repetitive			Letter Repetitive		
Backbone Model / Metric	S_{ev}	Acc(%)	F1(%)	S_{ev}	Acc(%)	F1(%)
RoBERTa (large)	0.24	86.85	**91.32**	0.10	83.47	88.79
GPT2 (medium)	**0.50**	80.76	86.72	0.07	76.28	82.87
T5 (small)	0.20	84.14	89.40	0.11	81.03	86.90
LLaMA2 (13B)	0.17	77.86	84.30	0.23	76.34	82.22
ChatGPT4	0.39	**86.92**	91.08	**0.31**	**84.99**	**89.41**

(a) Zero-shot performance

RoBERTa (large)	0.26	**91.94**	**94.86**	0.16	**90.22**	**93.65**
GPT2 (medium)	0.34	89.94	93.65	0.12	87.20	91.68
T5 (small)	0.26	90.45	93.95	0.13	87.97	92.17
LLaMA2 (13B)	**0.42**	90.08	93.50	**0.33**	88.80	92.54

(b) Finetuned performacne

Table 6: Comparision performances between zero-shot and finetuned models with two lengthening styles. The **best** results of columns are highlighted in bold, and the second best results are underlined.

tuning for RoBERTa-large and T5-base indicates we can use resource-efficient LMs processing informal styles when labeled samples are enough. However, their explainability may still be lower compared with zero-shot ChatGPT4. In practice, this is important because we do not want to use expensive APIs or resource highly demanded LLMs for sentiment analysis. Surprisingly, after instruction-tuned with only 1000 samples from the explainable instruction dataset, the instructed LLaMA-13B can match with zero-shot ChatGPT4 in accuracy, F1 score, and explainability.

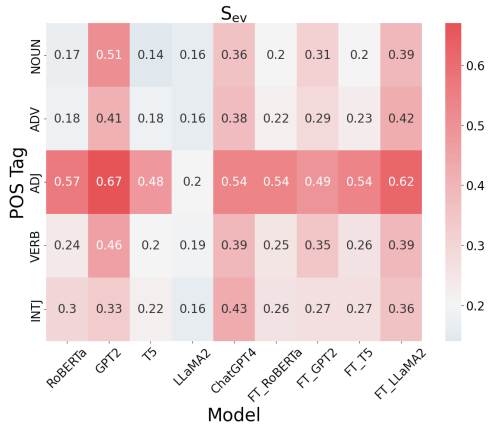
4.4 Analysis of Lengthening Styles

In this section, we compare models' performance for two lengthening styles. The results are presented in Table 6. We observe that samples in the Punctuation Repetitive style have higher scores in 3 evaluation metrics. This gap is consistent among all models in zero-shot and fine-tuned results. This finding indicates that the Punctuation Repetitive style plays a more pronounced role in sentiment expression. Additionally, it remains a low level of S_ev among fine-tuned RoBERTa-large, GPT2-medium, and T5-small models for Letter Repetitive syles while compared with ChatGPT4 and Instructed LLaMA2-13B. This implicitly suggests that understanding this nuanced expression style remains challenging for them.

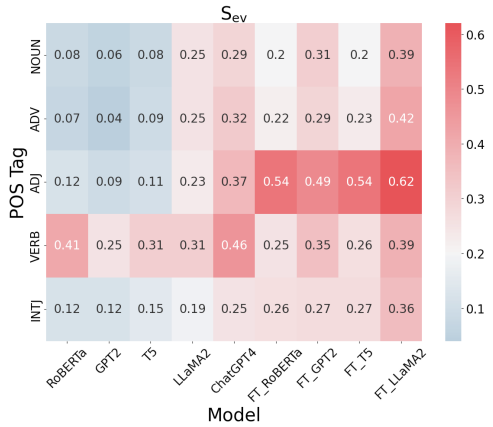
We further analyze models' word-level interpretation by Part-of-Speech tagging (POS Tag). We get POS of lengthened words by replacing them with root words and parse them with a deep learning-based tool TweebankNLP (Jiang et al., 2022). We attribute the POS tag of root words back to its corresponding lengthened words. As shown in Figure 4.1, we add 'FT' before a model's name to represent the fine-tuned version. We observe the Instructed LLaMA2-13B has a similar S_{ev} score distribution as ChatGPT4 among all fine-tuned models. This strengthens our conclusion that it can be a replacement for zero-shot ChatGPT4 to interpret informal expressions with limited computation resources and a few labeled samples. RoBERT-large, GPT2-medium, and T5-base reach the same level of S_{ev} of ChatGPT4 when POS Tag is ADJ and INTJ. There is still space for further improvement for other POS Tags such as NOUN, ADV, and VERB. This finding can guide data sampling and fine-tuning for future work.

4.5 Human evaluation

We also evaluate the model's qualities in interpreting sentences at the word level via a human-validated test, which is done by human judgment. As shown in Figure 3, for each sentence, we visualize the saliency map with corresponding words. We use RW and RS to represent reliability at the word level and reliability at the sentence level. The score range of RW and RS is 1-5, where 5 score means best. We randomly selected 200 samples and presented results in Figure 6. Models are ordered according to the average RW score. Results show that zero-shot ChatGPT4 has the best explain-



(a) Punctuation Repetitive



(b) Letter Repetitive

Figure 5: Comparision of S_{ev} score across Lengthening Styles and POS Tags. We present results for the top 5 POS Tags: Noun, Adverb, Adjective, Verb, and Interjection. The model prefixed with 'FT' represents the fine-tuned version.

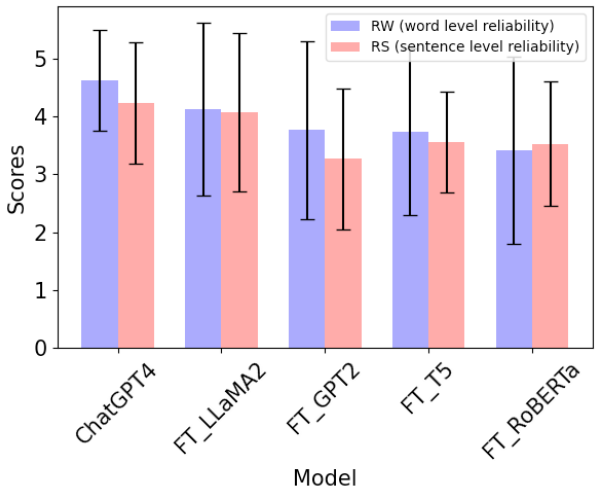


Figure 6: Results of human evaluation, we evaluate the word level (RW) and sentence level (RS) reliability for models' interpretation results. The model prefixed with 'FT' represents the fine-tuned version.

ability compared with fine-tuned and instruct-tuned models. Models' RW score order is the same as S_{ev} in Table 5 (b) order shows the effectiveness of our proposed method in Section 3.2. However, the RS score order is different from the RW score. This indicates there is a need to measure the model's interpretability from the sentence level.

5 Conclusion

In this paper, we propose a framework to evaluate the sentiment-expressive value of informal text and the explainability of LLMs for nuanced linguistic styles. Specifically, our focus centers on sentence-level sentiment analysis for lengthening words. We introduce **Lengthening**, a weakly labeled sentence-level dataset with 850k samples with lengthed words for sentiment analysis. We prompt ChatGPT4 with CoT for the word important score and generate an explainable instruction dataset with 1000 samples. Finally, we instruct-tune LLaMA2 on the explainable instruction dataset and fine-tune various pre-trained LLMs on **Lengthening**. We conduct comprehensive experiments involving 5 LLMs and 5 domains and reveal the sentiment-expressive value of lengthening words from both sentence and word levels. We highlight its importance for online social media, where short text and nuanced communication prevail. Our results show that pre-trained models (Roberta, GPT2, and T5) can reach better performances than zero-shot ChatGPT4 after fine-tuning on the Lengthening dataset but might be insufficient in explainability. With an explainable instruction dataset consisting of only 1000 samples, we can instruct-tune LLaMA2 (13B-chat-hf) to reach the same level of performance and explainability as ChatGPT4.

6 Limitation

We acknowledge the limitations that present opportunities for future research. Firstly, we can improve the quality of the explainable instruction dataset with human correction. This can further improve the performance and interpretation of instructed LLaMA2. Secondly, we can instruct-tune T5 with the explainable instruction dataset and compare it with existing results. We leave this work for future work. Thirdly, although we chose 200 samples for human evaluation, it can further support our conclusion with more samples and annotators for this section.

References

Muhammad Abdul-Mageed and Lyle Ungar. 2017. EmoNet: Fine-grained emotion detection with gated recurrent neural networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 718–728, Vancouver, Canada. Association for Computational Linguistics.

Farman Ali, Daehan Kwak, Pervez Khan, Shaker El-Sappagh, Amjad Ali, Sana Ullah, Kye Hyun Kim, and Kyung-Sup Kwak. 2019. Transportation sentiment analysis using word embedding and ontology-based topic modeling. *Knowledge-Based Systems*, 174:27–42.

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2020. A diagnostic study of explainability techniques for text classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3256–3274, Online. Association for Computational Linguistics.

Francesco Barbieri, Miguel Ballesteros, Francesco Ronzano, and Horacio Saggion. 2018. Multimodal emoji prediction. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 679–686, New Orleans, Louisiana. Association for Computational Linguistics.

Samuel Brody and Nicholas Diakopoulos. 2011. Coooooooooooooollllllllllllll!!!!!!!!!!!!!! using word lengthening to detect sentiment in microblogs. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 562–570, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Xiang Deng, Vasilisa Bashlovkina, Feng Han, Simon Baumgartner, and Michael Bendersky. 2022. What do llms know about financial markets? a case study on reddit market sentiment analysis.

Alec Go. 2009. Twitter sentiment classification using distant supervision.

Kyle Gorman, Christo Kirov, Brian Roark, and Richard Sproat. 2021. Structured abbreviation expansion in context. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 995–1005, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tyler J. Gray, Christopher M. Danforth, and Peter Sheridan Dodds. 2020. Hahahahaha, duuuuude, yeeessss!: A two-parameter characterization of stretchable words and the dynamics of mistypings and misspellings. *PLOS ONE*, 15(5):e0232938.

Jochen Hartmann, Mark Heitmann, Christian Siebert, and Christina Schamp. 2023. More than a feeling: Accuracy and application of sentiment analysis. *International Journal of Research in Marketing*, 40(1):75–87.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models.

Kango Iwama and Yoshinobu Kano. 2018. Japanese advertising slogan generator using case frame and word vector. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 197–198, Tilburg University, The Netherlands. Association for Computational Linguistics.

Hang Jiang, Yining Hua, Doug Beeferman, and Deb Roy. 2022. Annotating the tweebank corpus on named entity recognition and building nlp models for social media analysis. *In Proceedings of the 13th Language Resources and Evaluation Conference (LREC)*.

Yoram M. Kalman and Darren Gergle. 2014. Letter repetitions in computer-mediated communication: A unique link between spoken and online language. *Computers in Human Behavior*, 34:187–193.

Jiwei Li. 2020. Hotel review dataset. Access date: 2023-10-11.

Jiwei Li, Will Monroe, and Dan Jurafsky. 2017. Understanding neural networks through representation erasure.

Hongzhan Lin, Ziyang Luo, Bo Wang, Ruichao Yang, and Jing Ma. 2024. Goat-bench: Safety insights to large multimodal models through meme-based social abuse. *arXiv preprint arXiv:2401.01523*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

Shotaro Misawa, Yasuhide Miura, Tomoki Taniguchi, and Tomoko Ohkuma. 2020. Distinctive slogan generation with reconstruction. In *Proceedings of Workshop on Natural Language Processing in E-Commerce*, pages 87–97, Barcelona, Spain. Association for Computational Linguistics.

Huy Nguyen and Minh-Le Nguyen. 2017. A deep neural architecture for sentence-level sentiment classification in twitter social networking. *CoRR*, abs/1706.08032.

Jianmo Ni, Jiacheng Li, and Julian McAuley. 2019. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 188–197, Hong Kong, China. Association for Computational Linguistics.

OpenAI. 2023. GPT-4 technical report.

Letian Peng, Zilong Wang, Hang Liu, Zihan Wang, and Jingbo Shang. 2023. Emojilm: Modeling the new emoji language. *arXiv preprint arXiv:2311.01751*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *CoRR*, abs/1910.10683.

Sagnik Ray Choudhury, Anna Rogers, and Isabelle Augenstein. 2022. Machine reading, fast and slow: When do models "understand" language? In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 78–93, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Jens Reelfs, Timon Mohaupt, Sandipan Sikdar, Markus Strohmaier, and Oliver Hohlfeld. 2022. Interpreting emoji with emoji. In *Proceedings of the Fifth International Workshop on Emoji Understanding and Applications in Social Media*, pages 1–10, Seattle, Washington, USA. Association for Computational Linguistics.

Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. ""why should I trust you?"": Explaining the predictions of any classifier. *CoRR*, abs/1602.04938.

Tyler Joseph Schnoebelen, P. Eckert, D. Jurafsky, C. Potts, and J. R. Rickford. 2012. *Emotions are relational: Positioning and the use of affective linguistic resources*. Ph.d. thesis, Stanford University. Submitted to the Department of Linguistics.

Shivam Sharma, Siddhant Agarwal, Tharun Suresh, Preslav Nakov, Md. Shad Akhtar, and Tanmoy Chakraborty. 2023. What do you meme? generating explanations for visual semantic role labelling in memes. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'23/IAAI'23/EAAI'23. AAAI Press.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, and Shengyi Huang. 2020. Trl: Transformer reinforcement learning. `https://github.com/huggingface/trl`.

Zengzhi Wang, Qiming Xie, Zixiang Ding, Yi Feng, and Rui Xia. 2023. Is chatgpt a good sentiment analyzer? a preliminary study.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2021. Finetuned language models are zero-shot learners. *CoRR*, abs/2109.01652.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models.

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.

Yelp. 2021. Yelp dataset. Yelp. Retrieved Feb 16, 2021.

Boyu Zhang, Hongyang Yang, and Xiao-Yang Liu. 2023a. Instruct-fingpt: Financial sentiment analysis by instruction tuning of general-purpose large language models.

Wenxuan Zhang, Yue Deng, Bing Liu, Sinno Jialin Pan, and Lidong Bing. 2023b. Sentiment analysis in the era of large language models: A reality check.

Qihuang Zhong, Liang Ding, Juhua Liu, Bo Du, and Dacheng Tao. 2023. Can chatgpt understand too? a comparative study on chatgpt and fine-tuned bert.

Kaijie Zhu, Jindong Wang, Jiaheng Zhou, Zichen Wang, Hao Chen, Yidong Wang, Linyi Yang, Wei Ye, Yue Zhang, Neil Zhenqiang Gong, and Xing Xie. 2023. Promptbench: Towards evaluating the robustness of large language models on adversarial prompts.