Engagement-driven Persona Prompting for Rewriting News Tweets

Reshmi Pillai and Antske Fokkens and Wouter van Atteveldt
Vrije University Amsterdam
De Boelelaan 1105, 1081HV Amsterdam
Correspondence: r.pillai@vu.nl

Abstract

Text style transfer is a challenging research task which modifies the linguistic style of a given text to meet pre-set objectives such as making the text simpler or more accessible. Although large language models have been found to give promising results, text rewriting to improve audience engagement of social media content is vastly unexplored. Our research investigates the performance of various prompting strategies in the task of rewriting Dutch news tweets in specific linguistic styles (formal, casual, and factual). Apart from zero-shot and few-shot prompting variants, wit h and without personas, we also explore prompting with feedback on predicted engagement. We perform an extensive analysis of 18 different combinations of Large Language Models (GPT-3.5, GPT-4o, Mistral-7B) and prompting strategies on three different metrics: ROUGE-L, semantic similarity, and predicted engagement. We find that GPT-4o with feedback and persona prompting performs the best in terms of predicted engagement for all three language styles. Our results motivate further exploration of applying prompting techniques to rewrite news headlines on Twitter to align with specific style guidelines.

1 Introduction

Rewriting text to match specific guidelines or conventions has garnered research interest due to its potential in diverse applications such as content optimization, text simplification, or social media content creation (Lample et al., 2019) (Shu et al., 2024). Specifically, text style transfer is defined as the task of rewriting an original sentence to match a prescribed stylistic norm, while maintaining the semantic intent (Toshevska and Gievska, 2021). Pre-trained language models such as BERT and GPT have shown superior performance in the style transfer task for text data from various domains (Shu et al., 2024; Oh et al., 2024) .

It is a key concern for journalists to garner reader interactions or engagement in the form of likes, quotes, retweets, comments, etc. for tweets shared on social media outlets. People increasingly consume journalistic news via a link posted to social media rather than from the app or web site of the media organization (Newman et al., 2023). This prompts media companies to prioritize social media in their dissemination strategy (Welbers and Opgenhaffen, 2019). Rewriting news headlines to incorporate social media-specific stylistic norms when posted on social media outlets has been shown to increase user engagement in case of international (Park et al., 2021) and Dutch (Lamot et al., 2022) news outlets. Despite the emergence of Large Language Models as a promising technique for various text rewriting tasks, their potential in news tweet rewriting has, to our knowledge, not been explored so far.

In the backdrop of this research gap, we investigate ways in which prompt-driven LLMs can be used to rewrite news tweets in specified language styles. In addition, we pose and address the following research question:

RQ1: How does style transfer impact the predicted engagement of news tweets?

To address these, we collected tweets authored by national and regional news media organizations in the Netherlands. Firstly, we extensively augment the dataset with several hand-crafted textual features to comparatively understand the characteristics of tweets authored by national and regional media. For the purpose of this research, we consider three different language styles: *formal* with sophisticated vocabulary, *factual* which present matter-of-fact updates, and *casual* which are conversational and informal with simpler vocabulary. We use GPT-3.5, GPT-4o, and Mistral-7B as language models to rewrite the given tweets according to the defined language styles. We use combinations of the zero-shot and few-shot prompting strategies with and

Proceedings of the 31st International Conference on Computational Linguistics, pages 8612–8622
January 19–24, 2025.

without personas using these LLMs to generate stylistic retweets. In addition, we also experiment with providing a feedback of predicted engagement on the rewrites to the LLM and prompting it to perform the rewrite the tweets again considering the feedback information. We report and discuss the performance of the combinations of models and techniques with metrics such as ROUGE-L, semantic similarity, and predicted engagement of the rewritten tweets. We find that casual style rewriting results in higher predicted engagement. Feedback prompting with persona has superior performance compared to the other prompting techniques in terms of the predicted engagement metric. We also report the human evaluation of a sample of rewritten tweets, in terms of adherence to the specified style and preservation of the content (meaning).

Our contributions are the following:

1. A dataset of tweets from five major national and regional newspapers in the Netherlands, over a period of three years.
2. Implementation and evaluation of a pipeline to rewrite news tweets with persona prompting using Large Language Models.[1]

The rest of the paper is organized as follows. Section 2 describes the related work in relevant areas that provide the background to our present work. The data set, experimental setup, and evaluation metrics are presented in detail in Section 3. Section 4 reports the results, and Section 5 discusses the analysis of the results, limitations of the current study, and potential future work directions. Section 6 concludes the paper.

2 Related Work

2.1 Feature Analysis of News Tweets

Using Twitter as a news sharing outlet has resulted in an evolution of journalistic norms to incorporate certain social-media-specific norms such as personalization and virality. The stylistic types (e.g. brief updates, opinionated) of the news headlines on Twitter can alter whether the audience perceives the tweet to be news or not (Moon and Hadley, 2014; Araujo and van der Meer, 2020; Kwak et al., 2010). The features of news tweets have been characterized in the context of fake news detection (Nyow and Chua, 2019; Verma et al., 2021), crisis (e.g. COVID-19) news propagation (Son et al., 2019), and event detection (Hossny et al., 2020). The topics and themes inherent in tweets have been widely explored, whereas those specifically originating from news media organizations are not. Previous studies that analyze the topics in tweets by news media often focus on specific events or issues such as COVID-19 (Han et al., 2021), climate change (Dahal et al., 2019) or the Black Lives Matter movement (Giorgi et al., 2022). Analysis of broader-range social media news articles, though relatively rare, has also shown insightful findings (Aldous et al., 2019).

2.2 Engagement of News Tweets

Factors influencing twitter engagement have been explored in various domains such as Canadian public health (Slavik et al., 2021) consumer brands (Han et al., 2019) and scholarly articles (Fang et al., 2022), with varying results.

In addition to structural elements, the semantic features of tweets have previously been explored in research with respect to engagement, such as conversation toxicity (Salehabadi et al., 2022), veracity and offensive labels (Papakyriakopoulos and Goodman, 2022), and readability metrics (Gkikas et al., 2022).

Though digital journalism through online news portals and social media has been steadily growing in the last few decades, research interest in factors determining the online engagement received by these articles is relatively recent.

An early experimental study on online news engagement focused on users' perceptions of aesthetic and structural presentation of portals together with their emotional reactions to the content of the report (O'Brien, 2011). Such studies relied on subjective assessment by the users, collected often through semi-structured interviews. In contrast, recent studies consider the reaction tools (e.g. likes, comments) on social media and online news articles as a quantitative metric to gauge engagement.

Media companies specifically target social networks by posting (short) texts to maximize the 'newsworthiness' or 'shareworthiness' of their content (Trilling et al., 2017). To gauge what is interesting to the audience, a number of 'news values', or predictors of newsworthiness in social media headlines, have been proposed and tested, including emotionality, negativity, proximity, the presence of elites and celebrities (Eilders, 2006; Harcup and O'neill, 2017), formatting style (Janét et al., 2022) and humor forms (Brugman et al., 2022).

[1] data and code available at request to the authors

2.3 Rewriting Tweets

Text rewriting has been explored in recent research both as a standalone task with a distinct goal and as a pipeline step in downstream NLP tasks such as parsing and machine translation. Rewriting could have different goals like simplification (Maddela et al., 2020) or inclusivity. An example of the latter is to rewrite a gendered sentence in English with its gender-neutral variant (Sun et al., 2021). Another purpose of rewriting could be to prevent the leakage of sensitive information about privacy (Xu et al., 2020).

Text style transfer is a challenging variant of text rewriting, in which additional/alternative stylistic elements are incorporated while maintaining factual and semantic correctness. Rewriting text with style transfer has recently garnered increased attention with the promising results from using LLMs. Augmented zero-shot learning has been proposed as an approach to rewrite with non-standard styles (e.g. scary) (Shu et al., 2024).

2.4 Personas as a Prompting Tool for LLMs

Persona prompting has emerged as a strategy to guide the responses from an LLM by specifying characteristics aligned with an identity. Along with other prompt engineering tools such as pattern catalog and chain-of-thought, persona prompting enhances the capabilities of LLMs and has been evaluated in several tasks such as question answering (Olea et al., 2024), dialogue generation (Pu et al., 2023), and translation (He, 2024). Role play prompting has been found to improve LLM zero-shot reasoning capability across 12 different benchmarks (Kim et al., 2024). In addition to generating text, personas have been used in rendering LLMs as judges / assessors for language tasks such as summarization (Dong et al., 2024).

Although our research draws motivation from these preceding studies on social media engagement and text rewriting, we identify and address the research gap in terms of these specific factors: 1) prompting large language models has, to our knowledge, not been evaluated for social media text rewriting tasks so far. 2) we incorporate information of predicted engagement class of the rewritten tweets in the prompt to guide the LLMs.

3 Methodology

3.1 Dataset Collection

We collected a dataset of tweets from the official Twitter accounts of selected Dutch newspapers during a period of three years, from 01.01.2020 to 01.01.2023 (using snscrape[2]). We chose ten newspapers for this study, five regional newspapers and five national ones, based on the number of Twitter followers, as mentioned in their official accounts.

Overall, the data set consists of 5,43,927 tweets (3,07,005 tweets from regional newspapers and 2,36,922 tweets from national newspapers).

The collected data consisted of the following: Tweet ID, date and time of publication of the tweet, indicators of engagement (number of retweets, number of likes, number of quotes, number of replies) and the tweet text. We create a new feature 'totalEngagement' which is the sum of the engagement indicators (likes, retweets, quotes, and replies). We also add a feature 'normalizedEngagement' by dividing the totalEngagement by the corresponding newspapers' followers count in 100k, to account for the possible increase of interactions by the sheer number of people reading it. We further defined a binary variable (with values 'high' or 'low') to indicate whether or not the normalizedEngagement value of the given tweet is in the fourth quartile. This is a heuristic to indicate whether the engagement received by the tweet, after adjustment to the size of the newspaper audience on Twitter, can be considered 'high.'

In order to explore the dataset, we further did extensive feature engineering and augmented it with new textual features. Although these features are not directly used in the current research, we describe them below to provide a complete description of our dataset.

We identified hashtags and URLs in tweets using regular expressions. Duplicate tweets and the tweets that were found to be empty after removing hashtags, URLs, and mentions were excluded from the dataset. Twenty-two additional features were created in four distinct feature groups from the available data.

1) **Engagement** 'totalEngagement' which is the sum of the engagement indicators (likes, retweets, quotes and replies), 'normalizedEngagement' by dividing the totalEngagement by the corresponding newspapers' followers count in 100k, to account

[2]https://github.com/JustAnotherArchivist/snscrape

FeatureGroup	Higher for High Engagement	Higher for Low Engagement
NER	**places,persons (p<0.001)**	**organizations p<0.05**
Emotions	**fear, positive words, negative words fear, sadness, surprise, joy (p<0.001)**	**anger, anticipation (p<0.001)**
Text Complexity	**Question Words, pronouns (p<0.001)**	**characters per word**

Table 1: Statistical analysis of text-based features in high and low engagement corpus collected from Twitter handles of national and regional news media in the Netherlands

for the possible increase of interactions by the sheer number of people reading it, 'boolEngagement' a binary variable (with values 'high' or 'low') to indicate whether or not the normalizedEngagement value of the given tweet is in the fourth quartile.
2) **Named Entities** using the Dutch Spacy Named Entity Recognition module (number of places, persons and organizations in a tweet, each divided by the number of words),
3) **Emotion and Polarity Words** using LiLah emotion lexicon (Ljubešić et al., 2020) (number of positive, negative polarity words and anger, joy, fear, sadness, disgust, anticipation, and surprise emotion words in a tweet, each divided by the number of words),
4) **Text Style and Complexity** using PyPI readability module (number of word types, question words, pronouns, characters per word, Gunning complexity score).

We analyzed, using Mann-Whitney U test, whether there is a statistically significant difference between the text-based features in the tweets belonging to the first and fourth quartile of normalized engagement. The key results are summarized in Table 1. As seen in the Table 1, we found statistically significant differences in feature values from the features: Named Entity (NER), Polarity and Emotion, and Text Style and Complexity between the corpus of tweets which received high and low engagement.

3.2 Engagement Prediction Model

The objective of our experimental setup is to rewrite the given tweet in specific styles. The rewritten tweets are evaluated on the basis of specific metrics: one being predicted engagement as a boolean variable (indicating whether or not the tweet is predicted to have high engagement). Here, we define 'high' as being in the fourth quartile of the normalized engagement, calculated in the dataset described in Section 3.1.

For the purpose of this evaluation, we use a pretrained BERT model, and fine-tune it to the task of engagement prediction of tweets. Specifically, we use TwHIN-BERT (twhin-bert-base) which is a pretrained language model trained on 7 billion tweets from 100+ languages. We choose TwHIN-BERT for the engagement prediction because 1) it is a multilingual model and hence is suitable for our dataset of Dutch tweets and 2) it is shown to have superior performance over generic BERT-based models in the downstream task of engagement prediction of tweets in several languages including Dutch.

To train the twhin-bert-base model, we first construct a subset of 10,000 tweets from the dataset described in Section 3.1, with roughly equal number of high engagement and low engagement tweets, to make sure the training set is balanced with both classes.

This trained model (with F1-score 0.821) is used for the prediction tasks mentioned Sections 3.3 and 3.4.

3.3 Pipeline

The key objective of this methodological pipeline is to rewrite the given news tweets so as to increase engagement on a social media platform. We focus on rewriting tweets in the following styles: 'Formal', 'Factual', or 'Casual'. We limit the scope of our experiments to these styles.

For the context of this research, we define these three language styles of interest as follows:

Formal : formal style maintains a structured tone, features possibly higher text complexity and uses a more advanced vocabulary, avoids contractions, and adheres strictly to grammatical rules.
Example in Dutch: *De bosbranden die het westen van de Verenigde Staten al gedurende drie weken teisteren, hebben woensdag een ernstige escalatie doorgemaakt.*
Translated to English: *The wildfires that persistently ravaged the western United States for a period of three weeks have, as of Wednesday, experienced a marked escalation in intensity.*

Factual : objective and neutral reporting of facts; without opinions, perspectives or descriptive words, focusing solely on conveying the core information in a possibly concise manner.
Example in Dutch: *Het aantal mensen in de Chinese stad Lanzhou dat besmet is met de dierziekte brucellose is opgelopen tot 6620*
Translated to English : *The number of people in the Chinese city of Lanzhou infected with the animal disease brucellosis has risen to 6,620*

Casual : maintains an informal tone, might use colloquial expressions, contractions or conversational language, aiming to relate to the audience
Example in Dutch: *Hoe versla je Nils van der Poel? Die vraag is na de 10 kilometer in Stavanger dit weekend alleen maar prangender geworden.*
Translated to English: *How do you beat Nils van der Poel? That question has only gotten more urgent after the 10 kilometers in Stavanger this weekend.*

For the task of rewriting tweets, we experiment with multiple combinations of LLMs and prompting techniques. We use zero-shot, few-shot prompting with and without personas. Specifically, we generate rewritten versions in formal, factual and casual linguistic styles, with the following prompt versions:
ZeroShotNoPersona The LLM is given the task of rewriting the given tweet in three specified styles, with the common goal of receiving high engagement on Twitter. The definition of each style is provided in the prompt.
FewShotNoPersona The LLM is given the task of rewriting the given tweet in three different styles. The definition of each style is provided in the prompt. In addition, 5 examples of tweets in each style, which received high engagement on Twitter is also provided for few-shot learning.
FeedbackNoPersona This version consists of two sequential stages. In the first step, the LLM is provided few-shot prompting, where five examples for each language style are given as input. The LLM then generates three rewritten tweets, one each in formal, factual and casual style. In the second stage, a BERT-based model described in Section 3.2 is employed to predict the engagement level of the generated tweets, categorizing them into "high" or "low" engagement classes. This engagement feedback is included as additional context in a follow-up prompt, and the LLM is instructed to rewrite the tweets, leveraging the engagement classification to refine its output. This iterative process aims to optimize tweet generation by incorporating task-specific feedback into the prompt engineering cycle.

Further, we also apply each of these prompting strategies to generate the rewritten tweets, but with a persona included in the prompting (**ZeroShotPersona**, **FewShotPersona**, **FeedbackPersona**).
The persona used for this purpose is given below:
You are a journalist at a news media organization in the Netherlands. You have a degree in journalism and several years experience in digital media, with Dutch language content. Your goal is to rewrite the give tweet {Tweet} in the specified style {Style}. The objective of this rewrite is to connect with news audiences and increase engagement metrics in the form of likes, retweets, quotes and replies in Twitter, while maintaining factual integrity. You are adaptable and able to shift in language styles, factual, formal or casual, based on the requirement.

Thus, we apply 6 different prompting strategies using 3 LLMs to rewrite tweets in 3 specified language styles. We evaluate these strategies according to the performance metrics explained in the Section 3.5. We use three LLMs GPT-3.5-turbo-0125 (mentioned as GPT-3.5 elsewhere in the paper), GPT-4o, and Mistral 7B. The GPT-3.5 and GPT-4o were chosen as candidate models because of the superior performance in text rewriting tasks (Norberg et al., 2023; Shu et al., 2024). We also acknowledge the relevance of smaller, open-source models considering the limited resources of newsrooms, especially regional and local newsrooms, which are the potential users of such tweet rewriting systems. We decided to use Mistral-7B as such a model.

A summary of the various techniques used for rewriting is shown in Figure 1.

3.4 Evaluation metrics

We evaluate the generated stylistic rewrites for each language style and evaluate them using the following metrics.

Lexical similarity: To evaluate the generated tweet rewrites in the formal, casual, and factual styles, we utilize the ROUGE (Recall-Oriented Understudy for Gisting Evaluation) score. ROUGE scores have been used extensively in the assessment of text rewriting/summarization tasks (Shu

RewriteStyle	Model	prompting technique	ROUGE-L	Similarity	predEngagement
Formal	GPT-3.5	fewshotPersona	0.63	0.77	0.54
	GPT-4o	feedbackPersona	0.56	0.86	0.57
	Mistral-7B	feedbackPersona	0.87	0.34	0.54
Casual	GPT-3.5	feedbackNoPersona	0.33	0.73	0.63
	GPT-4o	feedbackPersona	0.32	0.76	0.67
	Mistral-7B	feedbackPersona	0.34	0.72	0.64
Factual	GPT-3.5	feedbackPersona	0.52	0.81	0.52
	GPT-4o	feedbackPersona	0.52	0.84	0.56
	Mistral-7B	feedbackPersona	0.51	0.82	0.53

Table 2: Performance comparison of the different LLMs and prompting strategies in the task of rewriting tweets. The rewrites are compared in terms of ROUGE-L, semantic similarity and percentage of increase in the predicted engagement

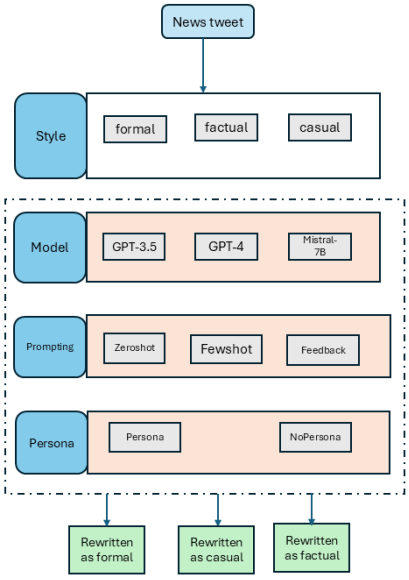

Figure 1: Summary of the methodological setup: combination of different LLMs, prompting strategies are used to rewrite the given tweet in specified text style

et al., 2024; Barbella and Tortora, 2022). ROUGE measures n-gram overlaps, and focuses on recall, measuring how many n-grams in the reference (in our context, the original tweet), are present in the rewritten tweet. We use ROUGE-L (Longest Common Subsequence) as the metric. We consider the ROUGE-L to provide insights into the rewrites generated by the different techniques and style options, and as a metric to what extent the generated rewrites adhere to the content of the original tweet. However, a stylistic rewrite might use different words while preserving the intent; hence we require different metrics to perform the evaluation.

Meaning preservation: To evaluate meaning preservation in the rewritten tweets across different styles (formal, casual, factual), we use Sentence-BERT (SBERT), which was designed to generate sentence embeddings which capture semantic intent. We find this a suitable metric for the performance of the rewriting techniques, so as to allow for the surface-level changes in vocabulary, while preserving the intended meaning. Specifically, we use the model distiluse-base-multilingual-cased, which is suitable for Dutch language datasets. The embeddings for the original tweet and the rewrite are compared using cosine similarity, with higher scores indicate better meaning preservation.

Predicted engagement: The objective of rewriting the tweets is garnering better audience engagement in the form of likes, retweets, quotes and comments. Hence, in addition to evaluating meaning preservation, we also incorporate an engagement prediction model to assess the potential engagement on Twitter (through likes, retweets, comments and quotes). For simplicity, we predict an overall score for the overall engagement and not specifically for each of the engagement indicators. As explained in Section 3.2, our prediction model is based on TwHIN-BERT (twhin-bert-base) (Zhang et al., 2023) and trained on 10,000 Dutch language tweets and the engagement they received. We provide each of the retweets as an input to the BERT-based model and predict the engagement as 'high' or 'low'. We evaluate the performance of a model as the percentage of rewritten tweets which were 'low' engagement originally and predicted as 'high' in the rewritten form.

3.5 Human Evaluation

A randomly chosen sample of 150 pairs of original and rewritten tweets were annotated by two native Dutch speakers for two paramaters: style adherence and content preservation. The evaluation was done on a likert scale of 1-5, with 1 being the lowest score and 5 being the highest. The key instructions given to the annotators were: 1) **Style Adherence** : Does the rewritten tweet adhere to the specified style ? 2) **Content Preservation** : Does the rewritten tweet preserve the content(meaning) of the original tweet? The annotation scores and comments are analysed to further understand the performance and possible limitations of the rewriting methods. The detailed instructions given to the annotators are provided in Appendix A.

4 Results

4.1 Comparison of Rewriting Techniques

The results of our experiments are presented in Table 2, where we evaluate the rewrites of a given tweet in three distinct styles: casual, formal, and factual. As described in Section 3.3, for each of the three language models (GPT-3.5, GPT-4o, and Mistral-7B), we applied six prompting techniques : zero-shot, few-shot, and feedback approaches, with or without persona integration in the prompt. We assessed the performance of the generated rewrites using three metrics described in Section 3.4: n-gram overlap, as represented by ROUGE-L; semantic similarity, as measured by S-BERT to evaluate meaning preservation; and the potential social media engagement, predicted by a BERT-based engagement prediction model (predEngagement). Each metric provided unique insights into the performance of the style rewriting techniques. For the sake of brevity, we are only showing the scores by the best performing prompting technique for each LLM and text style.

The results show that GPT-4o outperformed GPT-3.5 and Mistral-7B across all metrics. In particular, GPT-4o outperformed the other models in semantic similarity especially in the factual and formal styles, demonstrating its capacity to preserve meaning while altering the tone. On the other hand, the engagement prediction model shows interesting insights: casual rewrites consistently received higher predicted engagement scores across all models and prompting techniques, with the few-shot with persona method yielding the highest engagement predictions. This suggests that integrating user personas and few-shot learning significantly enhances the engagement potential, particularly in the casual style.

4.2 Results of Human Evaluation

The human coders scored the rewritten tweets on a scale of 1(the lowest) to 5 (the highest) in terms of style adherence and content preservation (annotators' agreement Krippendorff's alpha 0.618 for style adherence and 0.640 for content preservation). The average score received for style adherence was 4.135 and that for content preservation was 4.335. The coders' average scores for the three language styles are given in Appendix A.

From the average scores in 150 tweets, casual style rewrites appeared to score lower in style adherence compared to other styles, and formal style tended to score higher in terms of content preservation. As a further step to analyze the differences in style adherence and content preservation in the three different language styles, statistical significance tests were performed. ANOVA test revealed overall significant differences between language styles for style adherence ($p = 0.0006$) and content preservation ($p = 0.023$). As a follow-up test for pairwise differences, Tukey's HSD test indicated that casual style had significantly lower style adherence scores compared to formal and factual styles. For content preservation, pairwise comparisons (Tukey's HSD test) did not identify any significant differences between specific style pairs. This suggests that style adherence is significantly more challenging in casual style rewrites compared to other two styles, whereas the pairwise differences between langauge styles for content preservation were not statistically significant.

The coders' observations also provide qualitative insights about the rewrites. The coders observed that the rewriting focused on vocabulary and much less on structural changes (e.g. sentence structure, passive/active voice). The casual style rewriting also often included a small phrase at the end (e.g. "how cool is that?, "what a mess!"), which could be repetitive for usage in news tweets. Rewriting in factual style was also found to shorten the original text, in comparison to rewrites in other styles.

5 Discussion

The results provide insights into rewriting a given tweet according to the specified language styles.

We see that ROUGE-L score is lowest when rewriting tweet into casual style. At the same time, rewriting into casual style reported higher in the predictive engagement score. We also observe that while comparing the semantic similarity of the original tweets with the rewritten tweets (results from GPT-4o with feedbackPersona), the formal and factual style rewrites score higher than casual style rewrites. This indicates the need to further check to what extent the intended meaning could be lost while rewriting into a specified language style.

In this section, we present a critical reflection of our research. We use pre-trained large language models for the rewriting of the tweets with the objective of increasing audience engagement. The GPT models were chosen given their superior performance in text-based tasks. However, usage of such proprietary and closed source models may not be feasible for mid- or small-sized newsrooms. Considering the practical challenges and constraints of implementing a tweet rewriting system for a newsroom, it would be beneficial to also include more smaller sized, open-source models and evaluate their performance in the same task.

Another limitation lies in the evaluation metrics chosen. ROUGE-L and S-BERT based semantic similarity provide useful insights into structural and semantic preservation between the original and rewritten text, there could be several stylistic nuances which might not have been captured. Additionally, while the twHIN-BERT based models have proven performance in tweet engagement prediction, the prediction of potential engagement is still speculative. While the prediction using this model might be a reasonable approximation, the measurement of engagement could still be validated involving real-life audience.

The manual analysis of the rewritten tweets also point to certain limitations of using automated methods for tweet rewriting. As mentioned in Section 4, the coders noted that the rewrites often included removing or replacing certain words : thus the changes were vocabulary-based, rather than structural. For rewrites in the casual style, modifications frequently involved appending a brief phrase at the end of the text. For example, the tweet, *Terwijl de oorlog in Oekraïne voortduurt, opent een Russische prijsvechter zijn deuren in België* (English : While the war in Ukraine continues, a Russian prize fighter opens its doors in Belgium) is rewritten in casual style to *Terwijl de oorlog in Oekraïne doorgaat, opent een Russische budgetwinkel gewoon zijn deuren in België. Apart toch?* (English : While the war in Ukraine continues, a Russian budget store opens its doors in Belgium. Strange, right?)
Despite the high scores for style adherence and meaning preservation, such patterns, if occurring frequently, might render the rewrites repetitive, especially considering its applicability for news tweets.
Ethical concerns should also be carefully examined in this research. We find that rewriting tweets in a casual format increases the predicted engagement - but such a model might inadvertently prioritize sensationalism at the expense of journalistic values. A casual tone might not be suitable for certain topics, irrespective of the predicted engagement. For some topics, maintaining the neutrality and briefness of the tone is of importance. As observed by the human coders, an automated system might repeatedly use the same pattern (such as adding a conversational tag at the end in the case of casual rewrites), which might not be ideal for usage in news headlines. This system has been designed to provide helpful information to the journalist and is not meant as a replacement for human expertise; whether to use the rewritten tweets is completely up to the discretion of the journalist.

6 Conclusion

This research explored the research gap of rewriting news text on social media to align with specified language styles. We carried out experiments to evaluate to what extent the potential engagement received by a tweet can be changed using different prompting strategies and LLMs. In our evaluation, we considered three language styles of news tweets: formal, casual and factual, three LLMs: GPT-3.5-turbo-0125, GPT-4o, Mistral-7B and six different prompting strategies: zero-shot/few-shot/feedback, persona/no persona, for the evaluation of the rewritten tweets. We evaluated the performance of the techniques using semantic similarity calculated using S-BERT embeddings and potential engagement class predicted by a twHIN-BERT based model. Our research indicates superior performance of GPT-4o models in all linguistic styles, especially when combined with persona prompting and feedback on engagement (feedbackPersona technique in Table 2). Human evaluation of the rewritten tweets overall indicates the high adherence to the specified style and preservation of content in the

rewritten tweets; however, it also identifies specific patterns in rewriting (such as including a conversational phrase at the end for casual rewrites or making the text shorter for factual rewrites), which indicates scope for future improvement.

Acknowledgements

This work is supported by the Volkswagen Foundation under the "Artificial Intelligence and the Society of the Future" program (grant number: 9B390).

References

Kholoud Khalil Aldous, Jisun An, and Bernard J Jansen. 2019. View, like, comment, post: Analyzing user engagement by topic at 4 levels across 5 social media platforms for 53 news organizations. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 13, pages 47–57.

Theo Araujo and Toni GLA van der Meer. 2020. News values on social media: Exploring what drives peaks in user activity about organizations on twitter. *Journalism*, 21(5):633–651.

Marcello Barbella and Genoveffa Tortora. 2022. Rouge metric evaluation for text summarization techniques. *Available at SSRN 4120317*.

Britta C Brugman, Christian Burgers, Camiel J Beukeboom, and Elly A Konijn. 2022. Humor in satirical news headlines: Analyzing humor form and content, and their relations with audience engagement. *Mass Communication and Society*, pages 1–28.

Biraj Dahal, Sathish AP Kumar, and Zhenlong Li. 2019. Topic modeling and sentiment analysis of global climate change tweets. *Social network analysis and mining*, 9:1–20.

Yijiang River Dong, Tiancheng Hu, and Nigel Collier. 2024. Can llm be a personalized judge? *arXiv preprint arXiv:2406.11657*.

Christiane Eilders. 2006. News factors and news decisions. theoretical and methodological advances in germany.

Zhichao Fang, Rodrigo Costas, and Paul Wouters. 2022. User engagement with scholarly tweets of scientific papers: A large-scale and cross-disciplinary analysis. *Scientometrics*, 127(8):4523–4546.

Salvatore Giorgi, Sharath Chandra Guntuku, McKenzie Himelein-Wachowiak, Amy Kwarteng, Sy Hwang, Muhammad Rahman, and Brenda Curtis. 2022. Twitter corpus of the# blacklivesmatter movement and counter protests: 2013 to 2021. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 16, pages 1228–1235.

Dimitris C Gkikas, Katerina Tzafilkou, Prokopis K Theodoridis, Aristogiannis Garmpis, and Marios C Gkikas. 2022. How do text characteristics impact user engagement in social media posts: Modeling content readability, length, and hashtags number in facebook. *International Journal of Information Management Data Insights*, 2(1):100067.

Chunjia Han, Mu Yang, and Athena Piterou. 2021. Do news media and citizens have the same agenda on covid-19? an empirical comparison of twitter posts. *Technological Forecasting and Social Change*, 169:120849.

Xu Han, Xingyu Gu, and Shuai Peng. 2019. Analysis of tweet form's effect on users' engagement on twitter. *Cogent Business & Management*.

Tony Harcup and Deirdre O'neill. 2017. What is news? news values revisited (again). *Journalism studies*, 18(12):1470–1488.

Sui He. 2024. Prompting chatgpt for translation: A comparative analysis of translation brief and persona prompts. *arXiv preprint arXiv:2403.00127*.

Ahmad Hany Hossny, Lewis Mitchell, Nick Lothian, and Grant Osborne. 2020. Feature selection methods for event detection in twitter: a text mining approach. *Social Network Analysis and Mining*, 10:1–15.

Kristina Janét, Othello Richards, and Asheley R Landrum. 2022. Headline format influences evaluation of, but not engagement with, environmental news. *Journalism Practice*, 16(1):35–55.

Junseok Kim, Nakyeong Yang, and Kyomin Jung. 2024. Persona is a double-edged sword: Enhancing the zero-shot reasoning by ensembling the role-playing and neutral prompts. *arXiv preprint arXiv:2408.08631*.

Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. 2010. What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, pages 591–600.

Kenza Lamot, Tim Kreutz, and Michaël Opgenhaffen. 2022. "we rewrote this title": How news headlines are remediated on facebook and how this affects engagement. *Social Media+ Society*, 8(3):20563051221114827.

Guillaume Lample, Sandeep Subramanian, Eric Smith, Ludovic Denoyer, Marc'Aurelio Ranzato, and Y-Lan Boureau. 2019. Multiple-attribute text rewriting. In *International Conference on Learning Representations*.

Nikola Ljubešić, Ilia Markov, Darrja Fišer, and Walter Daelemans. 2020. The lilah emotion lexicon of croatian, dutch and slovene. In *Proceedings of the Third Workshop on Computational Modeling of People's Opinions, Personality, and Emotion's in Social Media. Barcelona, Spain (Online), ACL, pp. 153–157, December, 2020*, pages 1–5.

Mounica Maddela, Fernando Alva-Manchego, and Wei Xu. 2020. Controllable text simplification with explicit paraphrasing. *arXiv preprint arXiv:2010.11004*.

Soo Jung Moon and Patrick Hadley. 2014. Routinizing a new technology in the newsroom: Twitter as a news source in mainstream media. *Journal of Broadcasting & Electronic Media*, 58(2):289–305.

Nic Newman, Richard Fletcher, Kirsten Eddy, Craig T Robertson, and Rasmus Kleis Nielsen. 2023. Digital news report 2023.

Kole Norberg, Husni Almoubayyed, Stephen E Fancsali, Logan De Ley, Kyle Weldon, April Murphy, and Steve Ritter. 2023. Rewriting math word problems with large language models. *Grantee Submission*.

Ning Xin Nyow and Hui Na Chua. 2019. Detecting fake news with tweets' properties. In *2019 IEEE conference on application, information and network security (AINS)*, pages 24–29. IEEE.

Heather L O'Brien. 2011. Exploring user engagement in online news interactions. *Proceedings of the American society for information science and technology*, 48(1):1–10.

Sejoon Oh, Gaurav Verma, and Srijan Kumar. 2024. Adversarial text rewriting for text-aware recommender systems. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 1804–1814.

Carlos Olea, Holly Tucker, Jessica Phelan, Cameron Pattison, Shen Zhang, Maxwell Lieb, and J White. 2024. Evaluating persona prompting for question answering tasks. In *Proceedings of th e 10th international conference on artificial intelligence and soft computing, Sydney, Australia*.

Orestis Papakyriakopoulos and Ellen Goodman. 2022. The impact of twitter labels on misinformation spread and user engagement: Lessons from trump's election tweets. In *Proceedings of the ACM Web Conference 2022*, pages 2541–2551.

Kunwoo Park, Haewoon Kwak, Jisun An, and Sanjay Chawla. 2021. How-to present news on social media: A causal analysis of editing news headlines for boosting user engagement. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 15, pages 491–502.

Jiashu Pu, Ling Cheng, Lu Fan, Tangjie Lv, and Rongsheng Zhang. 2023. Just adjust one prompt: Enhancing in-context dialogue scoring via constructing the optimal subgraph of demonstrations and prompts. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9472–9496.

Nazanin Salehabadi, Anne Groggel, Mohit Singhal, Sayak Saha Roy, and Shirin Nilizadeh. 2022. User engagement and the toxicity of tweets. *arXiv e-prints*, pages arXiv–2211.

Lei Shu, Liangchen Luo, Jayakumar Hoskere, Yun Zhu, Yinxiao Liu, Simon Tong, Jindong Chen, and Lei Meng. 2024. Rewritelm: An instruction-tuned large language model for text rewriting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18970–18980.

Catherine E Slavik, Charlotte Buttle, Shelby L Sturrock, J Connor Darlington, and Niko Yiannakoulias. 2021. Examining tweet content and engagement of canadian public health agencies and decision makers during covid-19: mixed methods analysis. *Journal of Medical Internet Research*, 23(3):e24883.

Jaebong Son, Hyung Koo Lee, Sung Jin, and Jintae Lee. 2019. Content features of tweets for effective communication during disasters: A media synchronicity theory perspective. *International Journal of Information Management*, 45:56–68.

Tony Sun, Kellie Webster, Apu Shah, William Yang Wang, and Melvin Johnson. 2021. They, them, theirs: Rewriting with gender-neutral english. *arXiv preprint arXiv:2102.06788*.

Martina Toshevska and Sonja Gievska. 2021. A review of text style transfer using deep learning. *IEEE Transactions on Artificial Intelligence*, 3(5):669–684.

Damian Trilling, Petro Tolochko, and Björn Burscher. 2017. From newsworthiness to shareworthiness: How to predict news sharing based on article characteristics. *Journalism & mass communication quarterly*, 94(1):38–60.

Pawan Kumar Verma, Prateek Agrawal, Ivone Amorim, and Radu Prodan. 2021. Welfake: word embedding over linguistic features for fake news detection. *IEEE Transactions on Computational Social Systems*, 8(4):881–893.

Kasper Welbers and Michaël Opgenhaffen. 2019. Presenting news on social media: Media logic in the communication style of newspapers on facebook. *Digital journalism*, 7(1):45–62.

Qiongkai Xu, Lizhen Qu, Zeyu Gao, and Gholamreza Haffari. 2020. Personal information leakage detection in conversations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6567–6580.

Xinyang Zhang, Yury Malkov, Omar Florez, Serim Park, Brian McWilliams, Jiawei Han, and Ahmed El-Kishky. 2023. Twhin-bert: A socially-enriched pre-trained language model for multilingual tweet representations at twitter. In *Proceedings of the 29th ACM SIGKDD conference on knowledge discovery and data mining*, pages 5597–5607.

A Appendix A

Annotation Instructions for Evaluating Rewritten Tweets

RewriteStyle	Style Adherence		Content Preservation	
	coder-1	coder-2	coder-1	coder-2
formal	4.85	3.90	4.46	4.81
factual	4.79	3.79	4.08	4.38
casual	4.10	3.58	4.02	4.55

Table 3: Average scores of human annotation of the rewritten tweets in a scale of 1 to 5, for style adherence and content preservation

Thank you for participating in this annotation task. Your role is to evaluate rewritten tweets based on their adherence to a specified language style and their ability to preserve the meaning of the original tweet. The data provided to you consists of three inputs:
Original Tweet: The tweet in its original form.
Language Style: The style in which the tweet has been rewritten (Formal, Casual, or Factual).
Rewritten Tweet: The tweet rewritten in the specified language style.
For each rewritten tweet, you will evaluate two aspects on a scale of 1 (lowest) to 5 (highest):
1. Style Adherence This measure evaluates how well the rewritten tweet adheres to the specified language style. Below are explanations for each style:
Formal: The language should be formal, featuring possibly higher complexity and a more advanced vocabulary. There should be no use of informal or conversational language.
Casual: The language should be casual and conversational, aiming to relate to the audience. Informal expressions and a friendly tone are acceptable.
Factual: The language should be factual, objective, and concise. The rewritten tweet should avoid opinions, perspectives, or unnecessary details, focusing solely on conveying the core information in a brief manner.
Rating Guidelines for Style Adherence:
5: The rewritten tweet fully adheres to the specified style with no deviations.
4: The rewritten tweet mostly adheres to the specified style, with minor deviations.
3: The rewritten tweet somewhat adheres to the specified style, with some deviations.
2: The rewritten tweet poorly adheres to the specified style, with several deviations.
1: The rewritten tweet does not adhere to the specified style at all.
2. Content Preservation This measure evaluates how well the rewritten tweet preserves the meaning (content) of the original tweet. The rewritten tweet should retain the message and details of the original tweet, even if the language style has changed.
Rating Guidelines for Content Preservation:
5: The rewritten tweet completely preserves the meaning of the original tweet with no loss of content.
4: The rewritten tweet mostly preserves the meaning of the original tweet, with only minor omissions or alterations.
3: The rewritten tweet preserves the meaning of the original tweet to some extent, but there are noticeable omissions or alterations.
2: The rewritten tweet poorly preserves the meaning of the original tweet, with significant omissions or alterations.
1: The rewritten tweet does not preserve the meaning of the original tweet at all.

Steps for Annotations
Read the Inputs: Carefully read the Original Tweet, the specified Language Style, and the Rewritten Tweet.
Assess Style Adherence: Based on the specified style (Formal, Casual, or Factual), rate the rewritten tweet on a scale of 1 to 5 for style adherence.
Assess Content Preservation: Compare the rewritten tweet to the original tweet and rate the rewritten tweet on a scale of 1 to 5 for content preservation.

Annotation Scores The annotation scores given by the human coders for the rewritten tweets is given in Table 3.