Detecting Online Community Practices with Large Language Models: A Case Study of Pro-Ukrainian Publics on Twitter

Anonymous ACL submission

Abstract

Communities on social media display distinct patterns of linguistic expression and behaviour, collectively referred to as practices. These practices can be traced in textual exchanges, and reflect the intentions, knowledge, values, and norms of users and communities. This paper introduces a comprehensive methodological workflow for computational identification of such practices within social media texts. By focusing on supporters of Ukraine during the Russia-Ukraine war in (1) the activist collective NAFO and (2) the Eurovision Twitter community, we present a gold-standard data set capturing their unique practices. Using this corpus, we perform practice prediction experiments with both open-source baseline models and OpenAI's large language models (LLMs). Our results demonstrate that closed-source models, especially GPT-4, achieve superior performance, particularly with prompts that incorporate salient features of practices, or utilize Chain-of-Thought prompting. This study provides a detailed error analysis and offers valuable insights into improving the precision of practice identification, thereby supporting context-sensitive moderation and advancing the understanding of online community dynamics.[1]

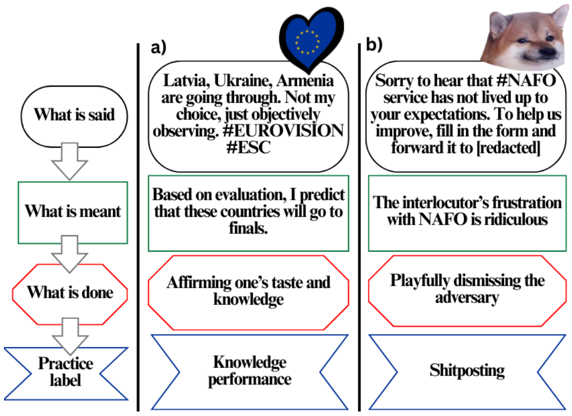

Figure 1: Analytical schema for identifying practices adopted from Gherardi (2012). Images represent communities analysed in this study – a) fans of the Eurovision Song contest known for their active use of Twitter, and b) NAFO, recognized for its efforts in debunking Russian propaganda on Twitter, characterized by avatars featuring Shiba Inu dogs.

1 Introduction

Online communities on platforms like Twitter[2] display distinctive and sustained patterns of behaviour and action, often referred to as *practices* (Mendes et al., 2023; Highfield, 2016; Meraz and Papacharissi, 2013), that are directed towards a goal and shaped by the socio-political context and affordances of digital platforms. Practices are significant because they reflect the values and beliefs of communities that engage in them (Trillò et al., 2022). For instance, consider `Knowledge performance`, or the practice of performative sharing of one's deeper than average knowledge of an issue. For fans of Eurovision song contest, this would involve sharing obscure facts related to the history or background of the contest or making predictions about its results, assuring one's taste, and affirming the value of pleasurable experiences. Not all online practices are as innocuous as `Knowledge performance`. Some, like creation of memes, may perpetuate and amplify racism (Matamoros-Fernández, 2017), while this practice can also be used to debunk disinformation. Prediction of practices at scale can enable context-sensitive approaches to facilitation of healthy networked environments (Seering, 2020). It can also help identify prevalence of practices, correlations between practices and other variables of interest, changes in practices over time or in response to external events. As the NLP community strives to improve LLMs' performance in tasks accounting for social context (Choi et al., 2023) and potentially

[1] Code available at: [insert github link]
[2] Now rebranded as X

harness them for interventions in online communities (Fraser et al., 2023; Bose et al., 2023), it is crucial to understand how LLMs handle predicting forms of community-specific sustained action as expressed in language.

Identifying practices in texts of online communities at scale addresses this need. It is also a necessary step towards *practice mapping*, which involves using computational and qualitative methods to investigate communities' patterns of language use and non-language-centered actions (such as sharing of URLs or interactions with other users).

While there exist frameworks for identification of practices through ethnographic, survey, or discourse analysis approaches (Gherardi, 2012; Trillò et al., 2022; Mendes et al., 2023), inferring them at scale in voluminous and ever-evolving digital trace data is a complex task. Nuanced identification of practices requires expertise in a community's vernacular, values, and context they operate within — a challenge preventing from easy crowdsourcing of such task and producing gold-standard data sets of practices for a community of interest large enough to support fine-tuning approaches. In addition, there is no consensus on how to best represent such expertise in a form of an in-context learning prompt in a way that could help the model "locate" (Reynolds and McDonell, 2021) the task of identifying instances based on a complex sociological notion.

In this work, in both codebook preparation and construction of Chain-of-Thought (COT) in-context learning prompts (Wei et al., 2023), we build upon an analytical schema for qualitative identification of practices in discourses of professional communities which separates practice identification into several steps – first examining the utterance, then identifying its meaning followed by inferring the intention and action behind it (Gherardi, 2012) (see Figure 1). We examine two online communities that support Ukraine following Russia's 2022 full-scale invasion through distinct forms of sustained action — NAFO, who engage in crowdfunding for the Ukrainian defence and debunk Russian propaganda through humorous or offensive posts, and fans of the Eurovision song contest, whose active Twitter community engaged with Ukraine's performers calling attention to the war during the 2022 and 2023 competition. To account for the difference in the social meaning between the two communities, we experiment with injecting into prompts salient features from our annotation codebook. To sum up, in this work, we propose a novel framework for using texts of social media posts to identify practices in online communities and present:

- Conceptualisation of an idea of practice as a unit of analysis that can be identified through text classification.
- A methodological workflow for constructing a gold-standard data set of practices from social media, tested on two online communities.
- Prompt-based text classification experiments utilising large language models in a zero- and few-shot setting to identify practices at scale.
- A set of human-annotator consistent prompts and Chain-of-Thought prompts that reflect the analytical schema for the qualitative identification of practices and improve the macro-averaged F1 score by 12.66% on average.
- An in-depth error analysis of the best-performing setting to assist in the future identification of practices from text data at scale.

2 Background

Speech acts and social meaning The view of linguistic utterances as accomplishing an action is captured in the notion of a speech, or an "illocutionary", act (Austin, 1962; Searle, 1968) – a performance of an action following a set of rules that ensure that the interlocutor understands the intention behind the utterance. This idea has informed numerous studies detecting intention and action in texts (Stolcke et al., 2000; Lampert et al., 2006; Carvalho and Cohen, 2006) and performing goal-oriented dialogue modelling (Young et al., 2010; Wen et al., 2017; Louvan and Magnini, 2020).

Works identifying speech acts in social media data have achieved this goal through transformer-based classifies (Saha et al., 2019, 2020) trained on expert-led semi-automated lexicons (Zhang et al., 2011; Vosoughi and Roy, 2016). These studies aimed to develop a generic classification approach, disregarding variation in online speech acts resulting from the authors' belonging to online communities or topical publics (Bruns, 2023). In contrast, our paper disaggregates online data sets prior to classification to ensure social meaning (Eckert, 2000, 2008) is preserved when identifying action in online communities.

Computational linguists have examined social meaning, or variation in language driven by community membership (Paris et al., 2012; Nguyen et al., 2021; Lucy and Bamman, 2021). However, only a handful of studies (Chancellor et al., 2018; August et al., 2020) considered how this variation reflects and produces norms and values. Our operationalisation of practice as a unit of analysis for text classifiers ensures community values and goals are captured in output of such classifiers. Importantly, by adopting the notion of practice – a pattern of action common among users sharing similar values and goals, it avoids the misconception (Bruns, 2019) of online communities as segregated homophilous "echo chambers" where members share an opinion on a topic (Mehta and Goldwasser, 2024). The two communities examined in this study have a common interest in Russia's invasion of Ukraine and take the side of the invaded country, but they achieve their goal of supporting Ukraine in distinct ways. As evident from practices like `Arguing` or `Shitposting` (see Section 5 for details), they are also aware of the opposing side and actively engage with them.

Practices of online communities Drawing from theories including Austin (1962) and Searle (1969), a body of work known as "practice turn" (Nicolini, 2012) understands practice as an activity sustained over time by a group of people in interaction with each other and material environment, oriented towards an object and grounded in norms and values. With language seen as an important form of situated action (Nicolini, 2012), practice turn scholars (Gherardi, 2012) have developed frameworks for identification of practices through close reading of texts associated with specific communities.

Several recent works produced typologies of social media practices through qualitative textual, survey, and interview analyses (Trillò et al., 2022; Mendes et al., 2023). However, to our knowledge, a more scalable approach has yet to be developed. In the field of computational linguistics, a handful of studies hinted at the idea of practice (Del Tredici and Fernández, 2017; Lucy and Bamman, 2021). Further engagement with this notion could enable a more comprehensive examination of variance in both language as action and values that guide such action. However, an objective of identifying practices in online corpora is a complex one. As we observe in section 5, it may implicitly incorporate a number of tasks, such as detection of stance, intent, presence of humour or sarcasm, and more.

In-context learning In-context learning with pre-trained Large Language Models (LLMs) has proven effective in these underlying tasks (Brown et al., 2020; Chowdhery et al., 2022), including processing texts from social media platforms (Roy et al., 2022; Sharma et al., 2023; Plaza-del arco et al., 2023; Zhu et al., 2023; Törnberg, 2023). Extremely large textual corpora, upon which LLMs are trained, contain conversations from social media platforms (Chowdhery et al., 2022; Brown et al., 2020), along with other texts that should allow the models to reproduce human knowledge.

The capability of LLMs to work with small annotated data sets is important for studies investigating practices of online communities. This capacity could help compare practices of multiple communities or one community across a prolonged period of time without needing to create training data sets of a size sufficient for fine-tuning (domain adaptation) for each community or period of interest. Despite this potential, capabilities of LLMs to perform complex reasoning tasks, such as context-sensitive classification, vary greatly depending on the prompt design (Loya et al., 2023). Studies have demonstrated effectiveness of prompts which include class features relevant to classification task (Bohra et al., 2023) or provide intermediate reasoning steps in a form of Chain-of-Thought prompting (Wei et al., 2023; Madaan et al., 2023). Our study builds upon these approaches to create prompts that 1) replicate the codebook proven most effective with human coders and 2) incorporate analytical reasoning utilised during the codebook construction.

While our focus is on achieving a reliable classification of practices in online communities, this approach can be applied to other NLP studies leveraging LLMs for tasks sensitive to social, political, and group context, such as frame prediction (Khanehzar et al., 2021; Frermann et al., 2023), identification of harmful online phenomena (ElSherief et al., 2021; Aich and Parde, 2022), and, more broadly, studies aiming to leverage LLMs for scaling up efforts of human annotators using prompt-based approaches (Munnangi et al., 2024; Sainz et al., 2023).

3 Practice Corpus

Data collection and preparation We developed our approach through examination of two online communities: (1) the North Atlantic Fella Organ-

Practice	NAFO	ESC	Practice	NAFO	ESC
L1-Advocacy	2.4	2.6	L{1,2}-Self-promotion	2.66	2.7
L1-Boosting	2.93		L1 - Shitposting	7.1	
L1-Charity		3	L2-Arguing	5.77	3.9
L1-Community imagining		2.9	L{2,3}-Audiencing	3.64	22.5
L1-Denouncing		3.1	L2-Betting		3.8
L1-Expressing solidarity	2.84	4.9	L2-Community work	12.95	
L1-Fundraising	2.84		L2-Expressing emotions		2.7
L1-Membership requests	2.75		L2-Play	5.68	
L1-Meme creation	3.02		L{3,2}-Knowledge performance	7.1	6.5
L1-Mobilising	8.96		L3-Not applicable	26.89	22.1
L1-News curation	2.66	19.3			

Table 1: Proportion of posts per practice in the gold standard data set. Total number of posts is 1127 for NAFO and 1000 for ESC. Priority for NAFO listed first in curly brackets where different between the case studies. Empty cells indicate practices not applicable to a case study.

isation (NAFO), a self-mobilised collective who debunk Russian propaganda and disinformation on Twitter; and (2) Twitter audiences of the Eurovision Song Contest (ESC)[3]. The communities either emerged in response to the invasion, like NAFO, or have many members sympathetic to Ukraine, like the ESC audience. Their selection was guided by our overarching interest in how support towards individuals and communities outside of one's nation is expressed via a global medium like social media.

We collected 4,079,694 tweets for the NAFO case study and a combined total of 585,129 tweets for the ESC in 2022 and 2023 through a keyword-based approach, querying Twitter Academic API. To maintain our research focus, we filtered out tweets produced by users opposing NAFO or Ukraine (details in Table 4, Appendix B). For irrelevant tweets and tweets unsupportive of Ukraine that were not captured by the upstream filtering, we established a category "Not applicable" which was included in the construction of the gold-standard data set and all experiments. Finally, we discarded retweets and tweets with fewer than three tokens after excluding hashtags, @-mentions, and URLs.

Codebook construction To capture practices in the collected data, we followed the analytical schema illustrated in Figure 1, with the first author examining 1400 randomly sampled tweets to produce a list of communities' practices (see Table 1). To account for the contextual specificity, the first author interviewed 27 community members, utilising an approach where an interviewee scrolls back through one's timeline while explaining motivations behind their posts (Robards and Lincoln, 2017) (see Appendix C). Following the initial codebook review, we began annotation and iteratively refined the codebook, similar to the approach by Card et al. (2015). Specifically, we introduced practice `Markers`, `Prioritisation schema`, and `Exclusion` criteria, collectively referred to as MPE in the following. By markers, we refer to conventionalised signals, including thematic or stylistic choices, which are specific to linguistic expression by members of a community (Bauman, 2000; Eckert, 2000, 2008), serving as a form of social meaning (Nguyen et al., 2021). Prioritisation schema was set up for posts that could be interpreted as multiple practices, with practices less common or most aligned with the research interest of the study treated as the highest priority (L1), and more common practices as the lowest priority (L3). Exclusion criteria were introduced to account for markers that could be misleading or ambiguous for coders. We arrived at the final version of the codebooks after three rounds of annotation.

Annotation and results The annotation task was performed by two of the study's authors, who labeled a combined random sample of 1900 tweets across five rounds. The decision to use domain experts for the annotation task was motivated by the importance of the expertise on online communities and context of Russia's war on Ukraine to facilitate interpretation of users' practices. The coders achieved maximum intercoder reliability, calculated as Krippendorff's alpha (Krippendorff, 2019), in the last round with 0.73 (mean of 0.68) for ESC and 0.77 (mean of 0.6) for NAFO case study. The two coders discussed labels upon which they disagreed in reconciliation meetings following each run until achieving a consensus. After completing the coding procedure, to obtain a minimum of 25 samples per each practice, an additional 227 tweets were sampled using keyword-based approach by

[3] See Appendix A for details on the two communities.

the first coder and validated by the second coder.

The labeled data set (Table 1) reveals an imbalanced class distribution across both case studies, with lower-priority categories (`Not applicable`, `Audiencing`) being the most frequent. Some initial insights could be gained from the labeled data set. For example, the `Charity` practice only appeared in the 2023 ESC data set, indicating that earlier into Russia's invasion, charitable causes were less likely to utilise the song contest as an opportunity for visibility.

4 Practice Prediction

Predicting practices automatically and with high quality would open new possibilities for understanding online action and its implications. Extending beyond semantic meaning (Fried et al., 2023), results of practice prediction can provide insights for better regulation of online activities. However, human annotation of large amounts of text data and its quality control is costly and time-consuming (Grosman et al., 2020). In the case of the proposed methodological workflow, high level of familiarity with the contextual and vernacular specificity of the community under investigation is also crucial for correct identification of practices, complicating the potential crowdsourcing of the annotation task.

4.1 Practice prediction tasks

Following previous studies on in-context learning with LLMs (Roy et al., 2022; Lu et al., 2022), we design our experiments as a text classification problem. We first experiment with injection into LLM prompts salient features of practices, represented as practice markers, prioritisation schema, and exclusion criteria (MPE). This prompt design is motivated by 21%[4] increase in the intercoder reliability of human annotators following introduction of MPE features in the codebook. Capturing thematic, stylistic, and other choices specific to the community under study, MPE prompts serve as a succinct way of expressing social meaning (Nguyen et al., 2021). This approach also echoes the work of Bohra et al. (2023), who developed a method for enhancing prompts for classification tasks with salient features of each class. While their approach is positioned as a substitute for demonstration examples, we also test how MPE performs in conjunction with practice examples.

[4]Average across two case studies

In addition, to investigate whether, in line with previous studies (Madaan et al., 2023), providing intermediate analytical steps can enhance a model's understanding of the prediction task, we also experiment with Chain-of-Thought (COT) prompts reflecting the schema used for our initial identification of practices during the codebook construction stage (Figure 1). Specifically, we design the prompt where each practice is first illustrated by a sample tweet, followed by two reasoning steps indicating its meaning and the intention and action behind it, and concluded with the practice label. We compare these results with prompts that only feature one-sentence practice descriptions.

Practice Description (PD) prompts consist of a short description of the community and its respective practices (Appendix E.3.1). It instructs the model to assign a single practice label to each tweet. Using a one-pass approach (Roy et al., 2022), we provide labels and definitions for all practices in one prompt. We then provide the model with a shuffled set of tweets for labeling. For *K=1* and *K=2* settings, for each practice we include in the prompt one or two examples of tweets, randomly selected from the training set.

PD+MPE prompts utilise the prompts consistent with the final version of the codebook constructed for human annotators (see Appendix D for codebooks, E.3.2 for prompts). The salient features of practices (MPE) are presented to the model as lists and short sentences following the practice description. The example below illustrates a part of the prompt for `Expressing solidarity` practice in the ESC case study.

> Expressing solidarity: L1. Tweets with only explicit and strong statements of support towards or solidarity with Ukraine with no other intent. Markers: "Slava Ukraini", "Glory to Ukraine", #StandWithUkraine. Exclusion criteria: "Let's go, Ukraine", "Congratulations, Ukraine", "Ukraine win" and similar cheers that may be meant for the performers should be labeled as "Audiencing"

PD+COT prompts utilise Chain-of-Thought (COT) prompts that replicate the analytical schema the first author utilised in the process of identifying practices in tweets (Figure 1) during the first step of codebook construction. In addition to the one-sentence practice descriptions, for each practice we include the tweet text ("what is said"), two analytical steps explaining "what is meant" and "what is done" by the tweet, followed by the expected label (see Appendix E.3.3 for prompts):

Tweet: How about some Ukrainian whiskey to pair with Eurovision? Other products available as well, and all proceeds will be donated to demining initiatives [URL] Let's think step by step: 1) The tweet advertises merchandise with profits supportting a pro-Ukrainian cause. 2) It engages in a form of aid towards Ukrainians suffering from Russia's war. Answer: Charity

4.2 Experimental Setup

Data set For each case study, we split the Practice Corpus into a test set (40% of all data) and a training set (60% of all data). We train all models using 5-folds cross-validation[5].

Baseline models We compare the performance of our proposed in-context learning prompts tested with GPT-3.5 and GPT-4[6], against several baselines. These include Random and Majority-class baseline, Linear Support Vector Machine (SVM) and Weighted-SVM with inverse class frequency and unigram features. We also compare our results with a prompt-free alternative to few-shot text classification with LLMs – a fine-tuning framework for sentence transformers SetFit (Tunstall et al., 2022). We test SetFit with two sentence-transformer models – MPNET (Song et al., 2020b) and DistilRoBERTA (Sanh et al., 2020). We test SetFit with one, two, and eight demonstration samples for each case study and model. The primary motivation for selecting these baselines is to explore open-source alternatives to OpenAI's LLMs that can reliably perform classification with a small amount of labeled data.

4.3 Results

Table 2 shows the practice prediction results for baselines and GPT models using practice description prompts. We assess the models based on their ability to accurately predict the practice label assigned to tweets, reporting macro-averaged F1 scores as mean and standard deviation across five folds (for precision and recall, refer to Appendix E.4). All tested models significantly outperform the Random and Majority baselines.

The SVM and Weighted-SVM models do not display promising results, only achieving F1 score of 60 or higher (detailed breakdown in Table 14, Appendix E.4.2) with practices where users consistently rely on set hashtags and accounts they mention – like NAFO's `Mobilising` which primarily included short tweets with hashtags used by the community for the purposes of calling each other's attention.

[5]For Random and Majority baselines, we utilise scikit-learn's (Pedregosa et al., 2011) dummy classifier and perform 1000 runs and 1 run respectively.

[6]GPT-3.5-turbo-instruct, GPT-4-1106-preview

	Setup	NAFO	ESC
	Random	06.11 (1.2)	07.63 (1.50)
	Majority	02.54	03.01
SVM	Linear	20.28 (1.17)	23.71 (2.57)
	Weighted	13.26 (1.97)	23.71 (2.22)
SetFit	MP(K=1)	10.41 (2.59)	10.55 (4.64)
	MP(K=2)	16.40 (1.96)	18.03 (5.72)
	MP(K=8)	25.67 (3.88)	32.13 (3.6)
	DR(K=1)	05.61 (1.84)	06.44 (2.16)
	DR(K=2)	10.18 (3.11)	13.48 (4.54)
	DR(K=8)	10.13 (3.12)	22.08 (8.19)
PD	GPT3.5(K=0)	**39.31 (1.85)**	**38.01 (2.24)**
	GPT3.5(K=1)	35.99 (2.63)	36.27 (4.21)
	GPT3.5(K=2)	21.95 (2.65)	12.15 (3.34)
	GPT4(K=0)	**47.65 (1.77)**	**49.33 (2.59)**
	GPT4(K=1)	46.62 (2.11)	49.24 (3.29)
	GPT4(K=2)	45.23 (2.30)	49.14 (2.41)

Table 2: Practice prediction results (macro-averaged F1 with standard deviation across five folds in brackets) for baseline models and practice description (PD) prompts. MP and DR stand for MPNET and DistilRoBERTA, respectively. K indicates the number of demonstration samples.

Transformer models fine-tuned on a small number of demonstration samples using SetFit framework display similar tendencies to SVMs, particularly struggling with practices where a correct identification involves the inference of an intent, such as `Self-promotion` or `Knowledge performance`, as well as `Fundraising` or `Expressing solidarity`. Despite this, increasing the number of demonstration samples from one or two to eight per practice category led to considerable improvement in F1 score with the transformer models.

Conversely, in line with previous studies (Reynolds and McDonell, 2021; Madaan et al., 2023), for in-context learning with practice description prompts, increasing the number of demonstration samples did not lead to a significant improvement. Overall, practice description prompts with both GPT-3.5 and GPT-4 largely outperform baselines, especially in the zero-shot setting. This result indicates that the extensive pre-training of these models may already to an extent equip them for handling a complex task of practice prediction, without the need for additional fine-tuning.

Building on the initial findings from Table 2, we delve into the effects of integrating practice descriptions with COT and MPE. As Table 3 (and tables 10 and 11 in the appendix) illustrate, including a succinct, expertly curated representation of the

Setup	NAFO	ESC
PD	46.62 (2.11)	49.24 (3.29)
PD+MPE	52.39 (2.39)†	53.33 (2.98)
PD+COT	51.96 (1.38)†	53.87 (2.59)†
PD+COT+MPE	**56.88** (2.06)†	**58.71** (5.15)†

Table 3: Comparison of practice description (PD) performance with the addition of MPE and COT prompts in the K=1 setting with GPT-4. Results are presented as macro-averaged F1 and standard deviation across five folds. A dagger indicates a statistically significant increase according to paired t-test calculated at $p \leq 0.05$.

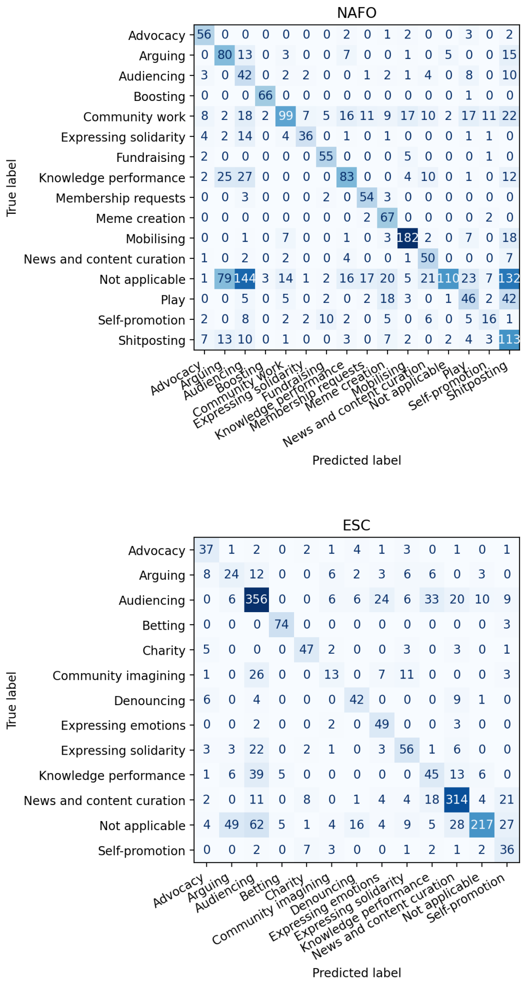

Figure 2: Confusion matrices for in-context learning with PD+COT+MPE prompts.

community's distinct use of language (PD+MPE prompts) increases the performance of the GPT-4 model. In addition, breaking down the task of practice prediction into analytical steps similar to those used by the human annotators upon initial identification of practices for codebook construction – combining practice description with Chain-of-Thought, PD+COT prompts – significantly improves GPT-4's performance.[7] Finally, we observe the best results with PD+COT+MPE prompt. We hypothesise that this type of prompt offers a more detailed description of the practice and the process for finding it that helps the pre-trained model to "locate" the category in the learned space (Reynolds and McDonell, 2021).

5 Discussion

Despite this potential, our results demonstrate that predicting a patterned intention and action behind online utterances with a limited number of samples is a difficult task for pre-trained large language models. In addition, even the best-performing setting (PD+COT+MPE prompt) fails to successfully predict a number of practices most closely aligned with our overarching research interest in communities' unique expression of support towards Ukraine.

We examine confusion matrices (Figure 2) and identify two categories of interest for each case study where the PD+COT+MPE prompting does not result in satisfactory performance. For NAFO, these are two of the most distinctive practices through which the collective combats Russian propaganda: `Shitposting`, or use of humorous or offensive posts to derail online discussions, and `Arguing`, or debating opponents with logic and facts. For the ESC case study, we select `Expressing solidarity` and `Community imagining`[8] due to their relevance for understanding how Russia's war on Ukraine altered practices of fans and their sense of belonging to the European community.

To seek insights that could improve practice prediction task results, we examine 450 false positive and 185 false negative tweets for these categories and identify prominent causes of errors.

Humour and sarcasm As observed in previous studies (Jentzsch and Kersting, 2023), humour and sarcasm presented challenges for the model with the PD+COT+MPE prompt. This was especially relevant for NAFO's practice of `Shitposting` which largely relies on jokes. Like in the example below, 53.85% of false negatives for this practice

[7] Due to budget constraints and length of our COT prompts, we only test COT prompt with the GPT-4 model.

[8] This practice refers to acts of discursively aligning oneself with a community, most often a nation state (Anderson, 1991). In the context of Eurovision (Sandvoss, 2008), this may involve publicly rooting for a performer representing one's country because they are "our own", apologising for lack of votes from one's country towards another country and so on.

were likely due to the model's failure to identify humour or sarcasm.

> You used to blame Ukraine's leaders, but look at you backpedaling. I can't see you riding a bike!

Overlapping practices According to our analysis, co-occurrence of multiple practices in one post was the most frequent cause of misclassification, accounting for 24.41% of errors overall. We observed it most prominently in false negative samples for Eurovision's `Expressing solidarity` practice (41.46% of misclassified samples). As in the example below, expressions of solidarity towards Ukraine co-occurred with speaking on behalf of the user's national community, expressing emotions, or engaging in `Audiencing` (live commentary).

> Amazingly done to Ukraine from the UK! You deserve to win. We're excited for you! #ESC2022 #WeStandWithUkraine.

We acknowledge that errors of this type may be inevitable, as studies indicate that even when investigated qualitatively, practices do not have easily identifiable boundaries and there exist overlaps between them (Gherardi, 2019; Gherardi and Nicolini, 2000). Due to this, human coders in our study also experienced difficulties with assigning one practice label per post. One potential avenue for the resolution of this issue could be treating practice prediction as a multi-label classification problem.

Misidentified stance of the author As elaborated in Section 3, one of the categories in our task involved identification of tweets by users supporting Russia and tweets unrelated to the war. We observed that PD+COT+MPE prompt was not always effective in identifying a pro-Russian stance in tweets. We attributed 24.52% of false positive samples for NAFO's `Shitposting` category to instances where the collective's adversaries deployed offensive language or logic to attack NAFO. While the expected label in this scenario was `Not applicable`, the model would classify such tweets as `Shitposting` or `Arguing`.

> You keep changing the subject – you are not good at this NAFO thing, fatty.

As a potential future solution to this issue, studies may introduce an upstream task of stance detection prior to classification of practices of a subset of users of interest, or incorporate this subtask in a form of a step in a COT prompt (Wei et al., 2023).

6 Conclusion

This paper proposes a systematic and scalable approach to associating the use of language in online texts with user practices as sustained patterns of behaviour shaped by sociopolitical and platform contexts. It provides a first empirically-driven systematic overview of practices on social media during Russia's war on Ukraine and presents a methodological workflow that can be applied by a wider range of studies aiming at identifying intention and action in communities of users.

The study advances our understanding of the potential of LLMs to make associations between utterances and online community practices. We demonstrate that even with a limited amount of gold-standard data, OpenAI's models, specifically GPT-4, are promising tools worth exploring. In addition, we show that representing the task of practice identification as a series of steps, and adding salient features as well as prioritisation and exclusion criteria to prompts, improves the performance of OpenAI's models.

Despite this promising results, these models still struggle with identifying sarcastic and humorous utterances as well as stance of the speaker in addition to the practice(s) they engage in. To address this, future studies may benefit from exploring approaches where identification of stance or sarcasm is treated as a separate task from practice prediction. Our error analysis also confirmed claims made in theoretical literature around overlap between practices of communities. To address this challenge, approaching practice prediction as a multi-label problem should be tested. Our hope is that computational linguistics and NLP communities continue to explore the practice prediction problem, enabling social scientists through insights and tools for scalable and efficient identification of user practices as manifested through language and beyond.

Limitations

We identify several limitations and shortcomings in our study as potential areas for future work. Our data set focuses on two case studies, connected by the overarching topic of Russia's war on Ukraine. The war has been a subject of varying interest from multiple communities across the world, while the two data sets were collected using only English-language keywords and contain predominantly English-language data.

The analysed communities of NAFO and ESC are also to an extent active on Discord, Reddit, TikTok, and other platforms, but our study is limited to Twitter data, which prevented us from exploring platform impact on the communities' practices. In addition, at the time of writing, Twitter Academic API, which we had utilised for data collection, is no longer freely available. This prevents future replication and longitudinal research on the communities of interest.

Our gold-standard data set is limited to one overarching topic, and is of a relatively small size. Our annotator agreement, while acceptable for studies examining human communication (Song et al., 2020a), can be improved. Our case study is of political nature, and there exists a risk of misuse of our modelling approach, as interpretations or applications of the model's outputs could be leveraged in ways that were not intended, influencing public perception or policy in an unanticipated manner.

Furthermore, while we utilise open-source baselines, in this study we focused on the performance of pre-trained OpenAI models. Such models are trained on data up to a specific cut-off date. For GPT-3.5, the date is September 2021 which is prior to Russia's February 2022 full-scale invasion. This lack of more up-to-date data may have impacted the results of the experiments outlined in this study. In addition, due to closed-source nature of Open AI models, potential changes to newer iterations may impact replicability of our results. We encourage future studies to work towards both improving practice detection with LLMs and achieving this through open-source models.

Ethics statement

Findings presented in this paper utilise text-based data collected in late 2022-early 2023 via Twitter Academic API. In accordance with the ethics clearance for this project, we have not requested consent from users who authored the texts, considering the risk they may be exposed to due to the research as minimal and the impracticality of contacting users with requests for consent. Despite this, depending on the national origin of a user, public engagement with the issue of Russia's war on Ukraine may put them at risk of persecution or social sanctions. To prevent re-identification and protect the privacy of our participants, we are only reporting on patterns emerging in collective practices as opposed to detailed descriptions of individual behaviour. While we utilised the original text of tweets during all experiments, we paraphrased or redacted it to prevent re-identification of the posts' authors in the appendices of the paper.

References

Ankit Aich and Natalie Parde. 2022. Telling a lie: Analyzing the language of information and misinformation during global health events. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4135–4141. European Language Resources Association.

Benedict R. O'G Anderson. 1991. *Imagined communities: reflections on the origin and spread of nationalism*, rev. and extended ed edition. Verso.

Tal August, Dallas Card, Gary Hsieh, Noah A. Smith, and Katharina Reinecke. 2020. Explain like I am a scientist: The linguistic barriers of entry to r/science. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, pages 1–12. Association for Computing Machinery.

J. L. Austin. 1962. *How to Do Things with Words: The William James Lectures Delivered in Harvard University in 1955*. Oxford University Press UK.

Richard Bauman. 2000. Language, identity, performance. *Pragmatics*, 10(1):1–5. Publisher: John Benjamins.

Arth Bohra, Govert Verkes, Artem Harutyunyan, Pascal Weinberger, and Giovanni Campagna. 2023. BYOC: Personalized few-shot classification with co-authored class descriptions. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13999–14015. Association for Computational Linguistics.

Olga Boichak and Andrew Hoskins. 2022. My war: Participation in warfare. *Digital War*, 3(1):1–8.

Ritwik Bose, Ian Perera, and Bonnie Dorr. 2023. Detoxifying online discourse: A guided response generation approach for reducing toxicity in user-generated text. In *Proceedings of the First Workshop on Social Influence in Conversations (SICon 2023)*, pages 9–14. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Axel Bruns. 2019. *Are filter bubbles real?* Digital futures. Polity Press.

Axel Bruns. 2023. From “the” public sphere to a network of publics: Towards an empirically founded model of contemporary public communication spaces. *Communication Theory*, 33(2):70–81.

Jean Burgess and Ariadna Matamoros-Fernández. 2016. Mapping sociocultural controversies across digital media platforms: One week of #gamergate on Twitter, YouTube, and Tumblr. *Communication Research and Practice*, 2(1):79–96.

Dallas Card, Amber Boydstun, Justin H Gross, Philip Resnik, and Noah A Smith. 2015. The media frames corpus: Annotations of frames across issues. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 438–444.

Vitor Carvalho and William Cohen. 2006. Improving “email speech acts” analysis via n-gram selection. In *Proceedings of the Analyzing Conversations in Text and Speech*, pages 35–41. Association for Computational Linguistics.

Stevie Chancellor, Andrea Hu, and Munmun De Choudhury. 2018. Norms matter: Contrasting social support around behavior change in online weight loss communities. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI ’18, pages 1–14. Association for Computing Machinery.

Minje Choi, Jiaxin Pei, Sagar Kumar, Chang Shu, and David Jurgens. 2023. Do LLMs understand social knowledge? evaluating the sociability of large language models with SocKET benchmark. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11370–11403. Association for Computational Linguistics.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. PaLM: Scaling language modeling with pathways.

Marco Del Tredici and Raquel Fernández. 2017. Semantic variation in online communities of practice. In *Proceedings of the 12th International Conference on Computational Semantics (IWCS) — Long papers*.

Penelope Eckert. 2000. *Linguistic variation as social practice: The linguistic construction of identity in Belten High*. Number 27 in Language in society. Blackwell Publishers.

Penelope Eckert. 2008. Variation and the indexical field. *Journal of Sociolinguistics*, 12(4):453–476. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-9841.2008.00374.x.

Mai ElSherief, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, and Diyi Yang. 2021. Latent hatred: A benchmark for understanding implicit hate speech. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 345–363, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Kathleen Fraser, Svetlana Kiritchenko, Isar Nejadgholi, and Anna Kerkhof. 2023. What makes a good counter-stereotype? Evaluating strategies for automated responses to stereotypical text. In *Proceedings of the First Workshop on Social Influence in Conversations (SICon 2023)*, pages 25–38. Association for Computational Linguistics.

Lea Frermann, Jiatong Li, Shima Khanehzar, and Gosia Mikolajczak. 2023. Conflicts, villains, resolutions: Towards models of narrative media framing. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8712–8732, Toronto, Canada. Association for Computational Linguistics.

Daniel Fried, Nicholas Tomlin, Jennifer Hu, Roma Patel, and Aida Nematzadeh. 2023. Pragmatics in language grounding: Phenomena, tasks, and modeling approaches. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12619–12640. Association for Computational Linguistics.

Silvia Gherardi. 2012. *How to Conduct a Practice-based Study: Problems and Methods*. Edward Elgar.

Silvia Gherardi. 2019. *How to conduct a practice-based study: problems and methods*, second edition. Edward Elgar Publishing.

Silvia Gherardi and Davide Nicolini. 2000. The organizational learning of safety in communities of practice. *Journal of Management Inquiry*, 9(1):7–18. Publisher: SAGE Publications Inc.

Timothy Graham and Jay Daniel Thompson. 2022. Russian government accounts are using a Twitter loophole to spread disinformation. Accessed: 2024-03-22.

Jonatas S. Grosman, Pedro H. T. Furtado, Ariane M. B. Rodrigues, Guilherme G. Schardong, Simone D. J. Barbosa, and Hélio C. V. Lopes. 2020. Eras: Improving the quality control in the annotation process for natural language processing tasks. *Information Systems*, 93:101553.

Tim Highfield. 2016. *Social Media and Everyday Politics*, 1st edition edition. Polity.

Tim Highfield, Stephen Harrington, and Axel Bruns. 2013. Twitter as a technology for audiencing and fandom. *Information, Communication & Society*, 16(3):315–339.

Sophie Jentzsch and Kristian Kersting. 2023. ChatGPT is fun, but it is not funny! humor is still challenging large language models. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 325–340. Association for Computational Linguistics.

Yuval Katz and Limor Shifman. 2017. Making sense? The structure and meanings of digital memetic nonsense. *Information, Communication & Society*, 20(6):825–842. Publisher: Routledge _eprint: https://doi.org/10.1080/1369118X.2017.1291702.

Shima Khanehzar, Trevor Cohn, Gosia Mikolajczak, Andrew Turpin, and Lea Frermann. 2021. Framing unpacked: A semi-supervised interpretable multi-view model of media frames. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2154–2166, Online. Association for Computational Linguistics.

Klaus Krippendorff. 2019. *Content Analysis: An Introduction to Its Methodology*. SAGE Publications, Inc.

Andrew Lampert, Robert Dale, and Cécile Paris. 2006. Classifying speech acts using verbal response modes. In *Proceedings of the 2006 Australasian Language Technology Workshop (ALTW2006)*, pages 34–41.

Samuel Louvan and Bernardo Magnini. 2020. Recent neural methods on slot filling and intent classification for task-oriented dialogue systems: A survey. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 480–496. International Committee on Computational Linguistics.

Manikanta Loya, Divya Sinha, and Richard Futrell. 2023. Exploring the sensitivity of LLMs' decision-making capabilities: Insights from prompt variations and hyperparameters. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3711–3716. Association for Computational Linguistics.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8086–8098. Association for Computational Linguistics.

Li Lucy and David Bamman. 2021. Characterizing english variation across social media communities with BERT. *Transactions of the Association for Computational Linguistics*, 9:538–556.

Aman Madaan, Katherine Hermann, and Amir Yazdanbakhsh. 2023. What makes Chain-of-Thought prompting effective? A counterfactual study. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1448–1535. Association for Computational Linguistics.

Ariadna Matamoros-Fernández. 2017. Platformed racism: The mediation and circulation of an Australian race-based controversy on Twitter, Facebook and YouTube. *Information, Communication & Society*, 20(6):930–946. Publisher: Routledge _eprint: https://doi.org/10.1080/1369118X.2017.1293130.

Sean McEwan. 2017. Nation of shitposters: Ironic engagement with the Facebook posts of Shannon Noll as reconfiguration of an Australian national identity. *PLATFORM: Journal of Media & Communication*, 8(2).

Nikhil Mehta and Dan Goldwasser. 2024. Using RL to identify divisive perspectives improves LLMs abilities to identify communities on social media.

Kaitlynn Mendes, William Hollingshead, Charlotte Nau, Jinman Zhang, and Anabel Quan-Haase. 2023. The evolution of #MeToo: A comparative analysis of vernacular practices over time and across languages. *Social Media + Society*, 9(3):20563051231196692. Publisher: SAGE Publications Ltd.

Sharon Meraz and Zizi Papacharissi. 2013. Networked gatekeeping and networked framing on #egypt. *The International Journal of Press/Politics*, 18(2):138–166. Publisher: SAGE Publications Inc.

Kateryna Minkina. 2022. Who are the NAFO fellas? The army of cartoon dogs fighting russian propaganda.

Monica Munnangi, Sergey Feldman, Byron C Wallace, Silvio Amir, Tom Hope, and Aakanksha Naik. 2024. On-the-fly definition augmentation of LLMs for biomedical NER.

Dong Nguyen, Laura Rosseel, and Jack Grieve. 2021. On learning and representing social meaning in NLP: A sociolinguistic perspective. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 603–612. Association for Computational Linguistics.

Davide Nicolini. 2012. *Practice theory, work, and organization: an introduction*, first edition edition. Oxford University Press.

OECD. 2008. *Local Economic and Employment Development (LEED) Local Development Benefits from Staging Global Events*. OECD Publishing.

Cecile Paris, Paul Thomas, and Stephen Wan. 2012. Differences in language and style between two social media communities. *Proceedings of the International AAAI Conference on Web and Social Media*, 6(1):539–542. Number: 1.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12(85):2825–2830.

Flor Miriam Plaza-del arco, Debora Nozza, and Dirk Hovy. 2023. Respectful or toxic? Using zero-shot learning with language models to detect hate speech. In *The 7th Workshop on Online Abuse and Harms (WOAH)*, pages 60–68. Association for Computational Linguistics.

Laria Reynolds and Kyle McDonell. 2021. Prompt programming for large language models: Beyond the few-shot paradigm. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI EA '21, pages 1–7. Association for Computing Machinery.

Brady Robards and Siân Lincoln. 2017. Uncovering longitudinal life narratives: Scrolling back on Facebook. *Qualitative Research*, 17(6):715–730.

Shamik Roy, Nishanth Sridhar Nakshatri, and Dan Goldwasser. 2022. Towards few-shot identification of morality frames using in-context learning. In *Proceedings of the Fifth Workshop on Natural Language Processing and Computational Social Science (NLP+CSS)*, pages 183–196. Association for Computational Linguistics.

Tulika Saha, Aditya Prakash Patra, Sriparna Saha, and Pushpak Bhattacharyya. 2020. A transformer based approach for identification of tweet acts. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. ISSN: 2161-4407.

Tulika Saha, Sriparna Saha, and Pushpak Bhattacharyya. 2019. Tweet act classification : A deep learning based classifier for recognizing speech acts in twitter. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. ISSN: 2161-4407.

Oscar Sainz, Iker García-Ferrero, Rodrigo Agerri, Oier Lopez de Lacalle, German Rigau, and Eneko Agirre. 2023. GoLLIE: Annotation guidelines improve zero-shot information-extraction.

Cornel Sandvoss. 2008. On the couch with Europe: The Eurovision Song Contest, the European Broadcast Union and belonging on the Old Continent. *Popular Communication*, 6(3):190–207.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter.

John R. Searle. 1968. Austin on locutionary and illocutionary acts. *The Philosophical Review*, 77(4):405–424. Publisher: [Duke University Press, Philosophical Review].

John R. Searle. 1969. *Speech Acts: An Essay in the Philosophy of Language*, 1st edition edition. Cambridge University Press.

Joseph Seering. 2020. Reconsidering self-moderation: The role of research in supporting community-based models for online content moderation. *Proceedings of the ACM on Human-Computer Interaction*, 4:107:1–107:28.

Arushi Sharma, Abhibha Gupta, and Maneesh Bilalpur. 2023. Argumentative stance prediction: An exploratory study on multimodality and few-shot learning. In *Proceedings of the 10th Workshop on Argument Mining*, pages 167–174. Association for Computational Linguistics.

Benjamin Shultz. 2023. In the spotlight: The Russian government's use of official Twitter accounts to influence discussions about its war in Ukraine. In *Proceedings of the 2nd ACM International Workshop on Multimedia AI against Disinformation*, MAD '23, pages 45–51. Association for Computing Machinery.

Hyunjin Song, Petro Tolochko, Jakob-Moritz Eberl, Olga Eisele, Esther Greussing, Tobias Heidenreich, Fabienne Lind, Sebastian Galyga, and Hajo G. Boomgaarden. 2020a. In validations we trust? The impact of imperfect human annotations as a gold standard on the quality of validation of automated content analysis. *Political Communication*. Publisher: Routledge.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020b. MPNet: Masked and permuted pre-training for language understanding.

Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics*, 26(3):339–374. Place: Cambridge, MA Publisher: MIT Press.

S.J.J. Tedjamulia, D.L. Dean, D.R. Olsen, and C.C. Albrecht. 2005. Motivating content contributions to online communities: Toward a more comprehensive theory. In *Proceedings of the 38th Annual Hawaii International Conference on System Sciences*, pages 193b–193b. ISSN: 1530-1605.

Tommaso Trillò, Blake Hallinan, and Limor Shifman. 2022. A typology of social media rituals. *Journal of Computer-Mediated Communication*, 27(4):zmac011.

Lewis Tunstall, Nils Reimers, Unso Eun Seo Jo, Luke Bates, Daniel Korat, Moshe Wasserblat, and Oren Pereg. 2022. Efficient few-shot learning without prompts.

Petter Törnberg. 2023. ChatGPT-4 outperforms experts and crowd workers in annotating political twitter messages with zero-shot learning.

Soroush Vosoughi and Deb Roy. 2016. Tweet acts: A speech act classifier for Twitter. *Proceedings of the International AAAI Conference on Web and Social Media*, 10(1):711–714. Number: 1.

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.

Tsung-Hsien Wen, Yishu Miao, Phil Blunsom, and Steve Young. 2017. Latent intention dialogue models. In *Proceedings of the 34th International Conference on Machine Learning*, pages 3732–3741. PMLR. ISSN: 2640-3498.

Steve Young, Milica Gašić, Simon Keizer, François Mairesse, Jost Schatzmann, Blaise Thomson, and Kai Yu. 2010. The hidden information state model: A practical framework for POMDP-based spoken dialogue management. *Computer Speech & Language*, 24(2):150–174.

Renxian Zhang, Dehong Gao, and Wenjie Li. 2011. What are tweeters doing: Recognizing speech acts in Twitter. In *Proceedings of the 5th AAAI Conference on Analyzing Microtext*, AAAIWS'11-05, pages 86–91. AAAI Press.

Yiming Zhu, Peixian Zhang, Ehsan-Ul Haq, Pan Hui, and Gareth Tyson. 2023. Can ChatGPT reproduce human-generated labels? A study of social computing tasks.

A Description of case studies

Case study 1: NAFO emerged through their shared efforts to counter Russian propaganda and disinformation on Twitter and to gather funds to support Ukraine (Boichak and Hoskins, 2022). The collective uses humour, sarcasm, and seemingly nonsensical and repetitive texts (Katz and Shifman, 2017) to debunk Russian propaganda and disrupt narratives of prominent pro-Russian accounts on Twitter. They also often engage in what scholars (McEwan, 2017) and communities highly active online refer to as "shitposting" – or using ironic, aggressive, or poor-quality content to derail a discussion or provoke opponents to break Twitter's Terms of Service. The below interaction illustrates the centrality of humour, community-specific vernacular, and dedicated hashtags and keywords for mobilisation and community-building in NAFO. It involves the author calling out an instance of Russian propaganda, inviting other users to engage with it and respond with memes, "shitposting", or debunking.

> Call to #NafoAticle5 - a highest-order nonsense has been pronounced and needs to be handled by the #Fellas @user @user @user @user @user tag all fellas who can help [link]

The expression "nonsense pronounced" refers to NAFO's interaction with Russian ambassador to Austria, Mikhail Ulyanov, who, as other Russian embassies and officials on Twitter, played a prominent role in spreading disinformation on the platform (Graham and Thompson, 2022; Shultz, 2023). Originally used by the ambassador to insult one of the NAFO members, the phrase was reclaimed by the collective and featured prominently in their exchanges. This example also illustrates the importance of contextual knowledge for interpretation of texts produced by NAFO's members – an additional challenge for computational detection of their discursive practices.

Case study 2: Eurovision Song Contest (ESC) is an annual singing competition, in which countries from the European continent, Australia, and beyond, are represented by one 3-minute musical performance, with the winner decided through a combination of a jury and a popular vote. Organised by the European Broadcasting Union since 1956, Eurovision is the most watched non-sporting event in the world (OECD, 2008). The ESC data set, while sharing a thematic connection through the focus on Russia's 2022 full-scale invasion of Ukraine, represents a different form of an online community – one emerging every year around May to discuss the preparation, the two semi-finals, and the finals of the contest. Audiences from around the world tweet about the contest as they watch the televised broadcast, making their tweets visible to other audience members through the event-wide (e.g., #esc) and country-specific (e.g., #SBSEurovision for Australia) hashtags. While ESC fandom on Twitter is centered around the broadcast, previous scholarship (Highfield et al., 2013) identified distinct fan practices such as `Audiencing`, or public performance of being a part of the Eurovision audience through live commentary on the performances. The tweet below is an example of `Audiencing`, where the user speaks of their favourite performances, referring to them by country names:

> Ok, it's Ukraine or Czech for me. But Netherlands, Romania, and Portugal are worth a mention. #Eurovision

In 2022, Russia was banned from performing in the contest, and the Ukrainian folk hip-hop band Kalush Orchestra won with a record-breaking number of points received from the voting public. Kalush used their performances as an opportunity to call the audiences' attention to the plight of the Ukrainian military and civilians trapped inside the Azovstal steel plant in Mariupol – a risky move as performers are banned from political statements according to the rules of the event. While during peacetime, Ukraine as the winner would be hosting the following year's competition, in 2023, the UK hosted on Ukraine's behalf. The 2023 contestants form Ukraine, Tvorchi, used the spotlight to promote humanitarian initiatives and call attention to Russia's shelling of their hometown, Ternopil. In this way, the 2022 and 2023 Eurovision presents an opportunity to explore the interconnection between global entertainment spectacles and political activism online.

B Tweet selection

Table 4 presents the data sets used in this study. We queried Twitter Academic API using keywords that could allow us to identify users (1) engaged with NAFO through mentioning it in their posts, and (2)

Query	Timeframe	Total Tweets	Filtered Tweets
`NAFO`	2022/05/01 – 2023/05/01	4,079,694	1,315,982
`(Eurovision OR #esc) (Ukraine OR Kalush OR UKR OR [Ukrainian flag emoji])`	2022/04/10 – 2022/06/10	444,455	125,569
`(Eurovision OR #esc) (Ukraine OR Tvorchi OR UKR OR [Ukrainian flag emoji])`	2023/04/09 – 2023/06/09	140,674	38,504

Table 4: Summary of tweet data collected. This table presents the queries used to collect tweets, the timeframe for each query, the total number of tweets retrieved, and the number of tweets remaining after filtering.

engaged with Ukraine's performance at the 2022 and 2023 Eurovision Song Contest through mentioning the event together with a reference to Ukraine or the two performers representing the country – Kalush Orchestra, a folk hip-hop collective who won the contest in 2022, and Tvorchi, an electronic music duo who placed 6th in 2023. For Eurovision, the period of collection was set as a month before and a month after the competition date for each year. For NAFO, we began the collection in May 2022 – the month when the movement emerged (Minkina, 2022).

We filtered out tweets that were likely to contain posts supporting Russia, and not Ukraine, in the full-scale invasion. To do so, building on issue mapping, a methodology for studying online communities through their engagement with issues involving disagreement (Burgess and Matamoros-Fernández, 2016), we constructed three retweet networks and conducted a close reading of posts by central and random nodes in each cluster. This allowed us to identify some users who produced posts out of scope of our study and discard them from further analysis. The final number of tweets in each data set is presented in **Filtered Tweets** column of Table 4.

C Interviews

In this section of the appendix, we provide the interview guide utilised for 27 semi-structured online interviews conducted as a part of this project. We recruited interview participants using a combination of purposive and random stratified sampling. For the latter, we separated users by their contribution to the overall volume of tweets in our data sets using a 1/9/90 distribution (Tedjamulia et al., 2005).

We separated the interviews into three parts – general questions about their social media use followed by a scroll back (Robards and Lincoln, 2017) section where either the interviewer or the participants shared their screen and scrolled through the interviewee's timeline of Twitter posts, and closing questions. To prompt participant reflections on patterns in Twitter activity in relation to Russia's full-scale invasion of Ukraine, we asked them about memorable posts, motivations behind them and the extent of coordination or collaboration with other users in the first part of the interviews. Similarly, questions from the scroll back section allowed us to gauge the regularity of certain types of posts over others. We did not explicitly prompt users to name practices they engaged in, and did not introduce them to the theoretical construct of "practice". Despite this, especially with NAFO case study, participants themselves named and provided definitions for a number of practices, such as `Shitposting`, `Bonking`, or `Boosting`. For example, one participant explained:

> The whole, you know, the putting of terrible memes under the Russian embassy and, you know, pro-Russian accounts instead of arguing because it's impossible to, it's ridiculous to argue with these people. Some do actually but it's ridiculous. I mean, it's like talking to a wall. It's really, it's a total waste of energy, but people still do it. But the whole, you know, insulting the ambassadors and things like that. That's what we call shitposting.

C.1 Interview guide

C.1.1 Indicative interview questions (General)

1. How did you first learn about Russia's invasion of Ukraine (Russia's war on Ukraine)?
2. Where do you obtain information about the invasion?
3. What digital media platforms or other outlets do you use to share information about the war?

4. Tell me about memorable posts that you have made in relation to Russia's invasion of Ukraine.
5. What did you pay attention to when making those posts?
6. Who is your intended audience?
7. Do you coordinate your posts with someone?
8. Do you have a connection to Ukraine?

C.1.2 Social media scroll back questions

Explain to the participant that you have pre-selected some of their posts and give them a choice for you to share the screen first or for them to share their Twitter timeline and scroll back to some posts that were important or meaningful to them. If you were the one to share your screen and show participants pre-selected posts, ask them about any other posts they remember. Feel free to let them scroll through their timeline. If the participants were the ones sharing their screen and did not touch upon some of the pre-selected tweets of interest, ask them if they could discuss some of the posts you selected. Record the video of the screen sharing process. Questions to ask about each post:

1. What happened on the day when you shared this post?
2. What drove you to make it?
3. What makes this post memorable or particularly effective?
4. What happened as a result of you posting it? Did it subvert or follow your expectations?

C.1.3 Closing questions

1. What would you like to see happen because of your posts?
2. What would you like to see happen with regards to Russia's invasion?
3. What will you be doing when the situation is resolved?

D Codebooks

D.1 Coding instructions

Do:

- Read the text of each tweet `tweet_text` column of the coding file, sheet labelled Tweets).
- If you do not have a working level of proficiency in the language of a tweet, utilise machine translation (DeepL or Google Translate).
- Assign one code from the dropdown of the `code` column.
- To make the assignment easier, consider possible codes in their order of priority – L1 > L2 > L3.
- If the text of the tweet cannot be interpreted as one of the practices in the dropdown, label it as Not Applicable.
- Use Common Examples and Markers to help you make judgement but prioritise the general description of the practice over presence or absence of the markers and examples listed in the codebook.

Do not:

- Inspect tweets in-situ using Twitter's keyword search or other approaches to understand the context of the utterance.

- Expand URLs included in the text of tweets.
- Evaluate the effectiveness or depth of user's commitment to the practice they are engaged in. Do focus on what their tweet is doing and do not base your judgement on how well or how genuinely the action is performed.

Special cases:

- If a tweet corresponds to more than one practices, try to establish the practice that in your opinion represents the intent of the author more strongly and assign them as codes. If this is not possible, label as Not Applicable.
- If a tweet corresponds to a practice from the available options, but the author clearly does not support Ukraine or Ukrainians, label it as Not Applicable.

D.2 Description of Practices

Practice	Priority	Description	Sample text (paraphrased)
Advocacy	L1	Reaching out to powerful actors to direct their course of action. **Markers**: at-mentioning Elon Musk or politicians	This will result in troll farms funded by malicious state actors like russia to become prolific and will make any efforts to correct the information they share impossible. This is a horrible decision @elonmusk! #NAFOfellas [link]
Arguing	L2	Trying to persuade an opponent. **Common examples:** pointing out falsity of information (debunking), misguidedness of their argument, or pointing out to a different perspective. **Markers:** tweet type: reply; providing factual evidence, "point".	@user What's your point after all, beside that you don't like being spammed with memes? Does that mean NAFO are bots if their clowning is not to your taste? This is not a valid argument to dismiss actions of people because you don't like them.
Audiencing	L2	News-related banter that does not entail knowledge sharing or deep commentary, rather an emotional or pleasurable experience of watching the events of the war together. **Markers:** "HIMARS O'Clock", "bavovna", "what [...] doin", military terminology	Here come the Riders of Ronan, this will be huge! #NAFO #RussiaIsLosing [GIF]
Boosting	L1	Short replies usually including the word "Boost" aimed to increase visibility of the content of someone else. **Markers:** User handles in the beginning of the tweet, "boost", URL.	@user @user @user @user Boost [GIF]
Community work	L2	Maintaining NAFO's cohesion, development, and growth through positive or supportive messages. **Common examples:** Appreciation of NAFO as a community, definition of its value and values, encouragement of other users to join, promises of mutual following, highlighting of prominent fellas, directing fellas to other potential communities, ideation around how the community can grow, NAFO-themed items. **Markers:** mentions of "the way", "fella". **Exclusion criteria:** targets of practice are other potential real or "imagined" communities, such as one's country, European Union etc. Calls to other NAFO members to engage in an activity should be coded as Mobilising.	@user Stay on this platform. Only children use Facebook. #NAFOfellas, boost him to the skies. [image]

Expressing solidarity	L1	Explicit statements of support towards or solidarity with Ukraine. **Markers:** "Slava Ukraini", "Glory to Ukraine", #StandWithUkraine.	#NAFO stands with Ukraine!
Fundraising	L1	Calls to donate money to a cause related to Ukraine. **Markers:** "donate", #RageDonate, names of weapons or military regiments (only in combination with donation markers).	Y'all, we are close! If all fellas made donations like [redacted], we would get it done today!
Knowledge performance	L3	Showcasing a deeper than average level of knowledge about the invasion or Twitter as a platform. **Markers:** "algorithm", military terms, political actors.	Watch this: he criticizes green efforts by the city of Budapest, while his boss imports russian energy with hands covered in Ukrainian blood. How ironic! @user
Membership requests	L1	Requesting a NAFO avatar – the accepted way of joining NAFO. **Markers:** "get a fella", #fellarequests, details around items to be depicted in the avatar, URLs.	@OfficialNAFO Would it be possible to make a fella based on Goose from Untitled Goose Game? [Link]
Meme creation	L1	Explicit tweets about meme making. **Markers**: use of a word "meme", "need", "forge", "make". **Exclusion criteria**: tweets using memes for a purpose – either to annoy someone (Shitposting) or for enjoyment (Play)	@user We should make a remake of this with NAFO dogs! #squadGoals [GIF]
Mobilising	L1	Directing or spurring action of other members of the collective. **Common examples**: pointing to a target of shitposting or a poll. **Markers:** #article5, #NAFOarticle5, #NAFOfellas, #NAFOexpansion, #NAFOfella, #NAFOhelp in combination to statements like "Check this out".	@user You're so clueless it disgusts me! #NAFO #NAFOfellas Have a look at this!!!
News curation	L1	Sharing of news and information. **Markers:** names of places or politicians, URLs, "says" or other verbs in Present Simple, "interview", news headline writing style.	From the ISW newest report on Ukraine: "Russian authorities continue to forcibly deport Ukrainian children from occupied Ukraine to Russia". #UkraineStolenChildren #NAFO
Play	L2	Having fun without a practical purpose. **Common examples:** Explicit jokes, memes, fantasies around NAFO. **Markers:** CIA, Bonk, Langley, Crimea Beach party, racoons, tractors. **Exclusion criteria:** Tweets with a clear adversarial target should be coded as Shitposting.	Put your hands together for Bonkenstein playing their rock classic Bonk Frei Vatnik [image]
Sarcasm	L2	Using words that likely imply the opposite of their literal meaning. **Common examples:** arguments with actors critical of NAFO or supporters of Russia.	@user Is this how liberation of Russian speakers look like? #ukraine #nafo [Link]
Self-promotion	L1	Highlighting one's own efforts or achievements as a NAFO fella. **Common examples:** stories of having successfully removed an actor from the platform or being blocked by a prominent pro-Russian account. **Exclusion criteria:** if the tweet starts with an account handle of a prominent pro-Russian account, code as Shitposting. **Markers:** "bonked", vatnik, Medvedev, Zakharova, Jason Hinckle, or other famous pro-Russian account.	@user Stayed up past midnight to bonk a few local vatniks. #SlavaUkraine
Shitposting	L1	Posting humorous, silly, offensive, or off-topic content to highlight flaws of propaganda / argument and to provoke an adversary to break the platform's ToS. **Markers:** Tweet type: reply, to Russian embassies, Ambassador Ulyanov, Kim Dot Com, Andrew Korybko, [redacted], Langley, CIA handlers, nonsense pronounced, copium.	This is a call from [redacted] #NAFO Twitter headquarters. We approved your application for NAFO Twitter fellaship and the ownership of your account has been transferred to NAFO. If you see a dog meme, the transfer has been successful [image].

Not Applicable	L3	Any other tweet not fitting any of these categories, also includes practices of adversaries of NAFO.	@user When you do not have an argument, insult the opponent, it always helps (according to NAFO handbook).

Table 5: Practice descriptions for NAFO case study

Practice	Priority	Description	Sample text (paraphrased)
Advocacy	L1	Requesting assistance towards Ukraine targeting either powerful actors (e.g., Twitter accounts of politicians) or broader communities online or offline. **Markers:** #SaveMariupol #SaveAzovstal. **Common examples:** requests to vote for Ukraine in the contest.	Russian genocide is killing Ukrainians, we need your help to exfiltrate #azovstal defenders #savemariupol #saveazovstal #eurovision [image]
Arguing	L2	Trying to persuade an opposing actual or imagined audience. **Markers:** “you people”, “those who”, tweet type: reply (user handles at the beginning of the tweet).	if you are upset about Ukraine’s Eurovision win, get over it. it’s a song competition. not a big deal. some people saying crazy ass stuff rn on this site.
Audiencing	L3	Performing as an audience of the Eurovision. **Markers:** “love”, country names, performers’ names, any other references indicating that the author is watching the show as they tweet. **Common examples:** commenting on performances, personal top-N, excitement about the event starting or ending, jokes, playful commentary related to performances and the contest, messages congratulating winners or performers.	I love Ukraine’s performance. a bucket hat, a flute, rapping, mad trousers - what else you need? #Eurovision
Betting	L2	Requests to participate in a bet, results of bets. **Markers:** “bet”, at-mention of RequestABet, “odds”	#RequestABet Eurovision, Ukraine to win, Norway, UK, Serbia and Czech Republic to finish in the top 10. Any odds please?
Charity	L1	Highlighting past and future instances of help to Ukraine through a charitable cause or activity. Often would be undertaken as a part of PR by a company or organisation. **Common examples:** requests for donations, **Markers:** events supporting refugees, donation links.	Check out one of the projects during #Eurovision. Local and Ukrainian kids celebrated their important connection by creating kites and flying them together [link]
Community imagining	L1	Speaking to or about a collective “we” beyond the individual. Capturing a collective feeling or addressing an imagined community. **Markers:** geopolitical entities (countries, EU), when used not to denote performers, “us” meaning Eurovision fans, “this country”.	Beyond words. I shed tears yesterday watching #Eurovision rehearsals and the show tonight. So proud of how we really did this for Ukraine and stood with them. This is what a special relationship means, forget the US. Glory to Ukraine! [image]
Denouncing	L1	Criticising of Russia and other actors that advertently or inadvertently support Russia. **Markers:** expletives, explicit mentions of Russian atrocities in various parts of Ukraine, #RussiaIsATerroristState. **Exclusion criteria:** actors or actions not related to Russia’s war on Ukraine such as criticism of Eurovision for decisions unrelated to the war.	Ukraine were winners of 2022 Eurovision. As we speak, Russia continues terrorising the whole Ukrainian territory. Btw, did you know that Eurovision has been going for 67 years, but the Soviet Union only stood for 68 years. Which one of the two is still going strong? Jealousy and fear is all Russia has to offer.
Expressing emotions	L2	Explicit expressions of various emotions without other apparent intent.**Markers**: “crying”, “laughing”, extensive emotion-centric emoji.	#Eurovision I’m in tears. I love Ukraine so much.
Expressing solidarity	L1	Making statements of support for Ukraine **Markers:** “Slava Ukraini”, “Glory to Ukraine”, #StandWithUkraine. **Exclusion criteria**: “Let’s go, Ukraine” and similar cheers that may be meant for the performers.	@user I have never watched Eurovision before today, but I hope Ukraine wins. Stay strong, Europe is with Ukrainians.

Knowledge performance	L2	Showcasing a level of knowledge about Eurovision beyond an average audience member. **Markers:** trivia, references to previous years, “EBU”. **Common examples:** predictions or attempts to theorise the reasoning behind some actions of the EBU, strategies of performers.	It’s a trend today, but Ukraine was a little all over the place. Camera work was messy at the start. Note there was no blue and yellow prominent - seems like the EBU achieved their goal with that #Eurovision
News curation	L1	Sharing news and other forms of information. **Common examples:** articles by news media outlets, entertainment or tabloid news, Eurovision fan communities producing reports from the ground, including on the results of the vote, music recommendations and reviews. **Markers:** URL, “says” or other verbs in Present Simple, “interview”, news headline writing style	NATO deputy lauds Eurovision win, says song highlights Ukrainian bravery [link] #tech
Self-promotion	L2	Showcasing efforts or success of oneself or one’s in-group. **Common examples:** PR tweets, tweets of Eurovision participants themselves. **Markers:** URLs.	Had my hands full creating a #eurovision-themed German lesson. We’ll cover entries from several countries and will do a vote in the class. Find it on TES to use it [link] #germanteaching #MFLteaching
Not Applicable	L3	Any other tweet not fitting any of these categories, also includes practices of users who oppose Ukraine and/or its involvement in Eurovision.	Ukraine with their subpar soccer team looks likely to win UEFA. Ukraine also just won Eurovision over the best song. Isn’t that “nice”?

Table 6: Practice descriptions for ESC case study

E Experimental details

E.1 Computational resources

We conducted all our experiments on a consumer Windows laptop (3.0 GHz Intel Core i7-1185G7 with 16GB of RAM). Utilised Python packages included `scikit-learn 1.3.2`, `openai 1.30.1`, and `sentence-transformers 2.2.2`. We calculate computational costs for OpenAI models based on the current official pricing for the GPT-3.5-turbo-instruct ($0.0005 / 1K context tokens, $0.0015 / 1K output tokens) and GPT-4-1106-preview ($0.01 / 1K context tokens, $0.03 / 1K output tokens). Combined costs of GPT-3.5 experiments were 59.03 USD, while GPT-4 – 1452.52 USD.

E.2 Model hyperparameters

For both Support Vector Machine models and transformer baseline models used with the SetFit framework, we utilise the respective default hyperparameter settings (tables 7, 8). For OpenAI’s models, we utilise the temperature setting of 0, the frequency penalty of 0.5, and the presence penalty of 0.

Table 7: Hyperparameters for Support Vector Machines models

Setting	Linear	Weighted
Kernel	Linear	RBF
Regularisation	1.0	1.0
Class weight	None	Balanced

Table 8: Hyperparameters for Setfit Models

Setting	MPNet	DistilRoBERTa
Model	paraphrase-mpnet-base-v2	all-distilroberta-v1
Loss class	Cosine similarity	Cosine similarity
Batch size	16	16
Iterations	20	20

E.3 In-context prompts

E.3.1 Practice description (PD)

ESC

Task: You will be provided with a tweet, created by a member of an online community and categorize it based on the practice they are engaged in.

Definition: In this context, “practice” refers to the distinct ways of communicating or performing actions using language that are unique to the online community under study.

Community Description: You will be examining tweets from fans and audiences of the Eurovision Song Contest that are supportive of Ukraine during and around the time of the 2022 and 2023 contests.

Instructions: For a given tweet, assign the appropriate label based on the following practices or categories. In your response, return only one label from this list: [Advocacy, Arguing, Audiencing, Betting, Charity, Community imagining, Denouncing, Expressing emotions, Expressing solidarity, Knowledge performance, News and content curation, Self-promotion, Not applicable]

Descriptions of practices are below.

`Advocacy`: Tweets that address powerful actors (politicians, governments, international organisations, celebrities) or broader communities online or offline and try to direct their course of action towards helping Ukraine in the war or in the competition.
`Arguing`: Argumentative tweets by Ukraine supporters that try to persuade actual or imagined opponents and get them to support Ukraine.
`Audiencing`: Tweets that provide shallow, brief, or humorous real-time commentary on the performance of Ukraine and other countries in Eurovision.
`Betting`: Tweets that request to participate in a money-related bet, results of bets.
`Charity`: Tweets that highlight past and future instances of help to Ukraine through a charitable cause or activity. Often would be undertaken as a part of PR by a company or organisation.
`Community imagining`: Tweets in which the author speaks on behalf of their country, region of the world, or community, addressing people in same or other countries or communities, capturing or conveying a collective sentiment or opinion.
`Denouncing`: Tweets that criticise Russia and other actors that advertently or inadvertently support Russia.
`Expressing emotions`: Tweets with explicit mentions of various emotions without other apparent intent.
`Expressing solidarity`: Tweets with only explicit and strong statements of support towards or solidarity with Ukraine with no other intent.
`Knowledge performance`: Tweets in which the authors use their deep or broad knowledge about various aspects of the Eurovision Song Contest to evaluate performances in detail or make predictions about outcomes of the contest.
`News and content curation`: Tweets that share news, fan blogs, or similar content that reports on events of Eurovision or the Russia-Ukraine war.
`Self-promotion`: Tweets in which the author humbly brags about themselves or their company. This may include talking about creations, purchases, donations, content they produced, or other past or planned efforts or achievements.
`Not applicable`: If a tweet does not correspond to any of the specified practices or is not supportive of Ukraine and its performance in Eurovision, label it as “Not applicable”.
Input Tweet:

NAFO

Task: You will be provided with a tweet, created by a member of an online community and categorize it based on the practice they are engaged in.

Definition: In this context, “practice” refers to the distinct ways of communicating or performing actions using language that are unique to the online community under study.

Community Description:You will be examining tweets from members of an online self-mobilized collective called “NAFO”, which focuses on countering Russian propaganda about the war in Ukraine. They achieve this through the use of humor or factual information.

Instructions:For a given tweet, try and assign the appropriate label based on the following practices or categories. In your response, return only one label from this list: [Advocacy, Arguing, Audiencing, Boosting, Community work, Expressing solidarity, Fundraising, Knowledge performance, Membership requests, Meme creation, Mobilising, News and content curation, Play, Self-promotion, Shitposting, Not applicable]

Descriptions of practices are below.

`Advocacy`: Tweets that address powerful actors (politicians, governments, international organisations, celebrities) and try to direct their course of action.

`Arguing`: Argumentative tweets that try to persuade an opponent and get them to support Ukraine.

`Audiencing`: Tweets that provide shallow, brief, and opinionated commentary on events of the war or situation on Twitter.

`Boosting`: Short replies usually including the word "Boost" aimed to increase visibility of the content of someone else.

`Community work`: Tweets that maintain NAFO's camaraderie, cohesion, development, and growth through positive, supportive, or celebratory messages about the movement, recruitment of new members or correcting behaviour of existing members.

`Expressing solidarity`: Tweets with only explicit and strong statements of support towards or solidarity with Ukraine with no other intent.

`Fundraising`: Tweets that call to donate money to a cause related to Ukraine.

`Knowledge performance`: Tweets that showcase the speaker's deep or broad knowledge about the invasion or Twitter as a platform, or make predictions.

`Membership requests`: Tweets that request or provide users with a NAFO avatar – the accepted way of joining NAFO.

`Meme creation`: Explicit tweets about meme making.

`Mobilising`: Tweets that direct or spur action (such as retweeting, sharing of information, responding to a poll, or engaging with a target) of other members of NAFO.

`News and content curation`: Tweets that repost news articles and other reports about the war or NAFO.

`Play`: Humorous tweets that do not have a practical purpose, aside from having fun.

`Self-promotion`: Any tweet in which the user speaks about themselves in the first person, putting an emphasis on their future or past deeds as a NAFO member.

`Shitposting`: Tweets that contain humorous, unrealistic, silly, offensive, or off-topic content to highlight flaws of propaganda or argument and annoy an adversary.

`Not applicable`: If a tweet does not correspond to any of the specified practices or is not supportive of Ukraine and NAFO, label it as "Not applicable".

Input Tweet:

E.3.2 PD+MPE

ESC

{Task, practice definition, community description, and instructions from the PD prompt}

`Advocacy`: L1. {PD} Markers: #SaveMariupol #SaveAzovstal or similar hashtags. Common examples: requests to political leaders or the European Broadcasting Union (EBU) to be more supportive of Ukraine or its representatives, requests to vote for Ukraine in the contest.

`Arguing`: L2. {PD} Markers: "those who", "you people" or similar.

`Audiencing`: L3. {PD} Markers: "love", country names, performers' names, any other references indicating that the author is watching the show as they tweet. Common examples: brief commenting on performances, personal top-10, messages of excitement about the event starting or ending, jokes, playful commentary related to performances and the contest, messages congratulating winners or performers. Exclusion criteria: For more detailed commentary on performances or predictions of results, label as "Knowledge performance".

`Betting`: L2. {PD} Markers: "bet", RequestABet, "odds". Exclusion criteria: Figurative use of the word "bet".

`Charity`: L1. {PD} Common examples: requests for donations. Markers: mentions of events supporting refugees, URLs for donations.

`Community imagining`: L1. {PD} Markers: geopolitical entities (countries, EU), when used not to denote performers, "us", "we", "ours" meaning Eurovision fans, compatriots or performers representing

one's country, "this country". Common examples: Expressions of gratitude from Ukrainians for support, apologies for not voting for a country enough, celebration of a win by a performer from one's own country.
`Denouncing`: L1. {PD} Markers: expletives, explicit mentions of Russian atrocities or attacks on various parts of Ukraine. Exclusion criteria: actors or actions not related to Russia's war on Ukraine such as criticism of Eurovision for decisions unrelated to the war.
`Expressing emotions`: L2. {PD} Markers: "crying", "laughing", extensive emotion-centric emoji.
`Expressing solidarity`: L1. {PD} Markers: "Slava Ukraini", "Glory to Ukraine", #StandWithUkraine. Exclusion criteria: "Let's go, Ukraine", "Congratulations, Ukraine", "Ukraine win" and similar cheers that may be meant for the performers should be labeled as "Audiencing".
`Knowledge performance`: L2. {PD} Markers: Trivia about participants, references to performances in previous years, "EBU". Common examples: predictions or attempts to theorise the reasoning behind some actions of the EBU, strategies of performers etc.
`News and content curation`: L1. {PD} Markers: URL, "says" or other verbs in Present Simple, "interview", news headline writing style.
`Self-promotion`: L2. {PD} Markers: Mentions of or URLs pointing to creations, purchases, donations, content user themselves produced, or other past or planned efforts or achievements.
`Not applicable`: L3. {PD} Markers: Statements suggesting Ukraine's win is predictable because the war, voting is rigged; spam tweets featuring hashtags related to trending topics other than Eurovision or the war.
Input Tweet:

NAFO

{Task, practice definition, community description, and instructions from the PD prompt}

`Advocacy`: L1. {PD} Markers: @-mentions of Elon Musk, politicians, UN or similar entities, mentions of weapons to be donated to Ukraine (ATACMS, Taurus, Leopards). Exclusion criteria: Code tweets about the action NAFO should take as "Community work"; code tweets that do not target powerful entities via hashtags or account handles as "Arguing"; code tweets asking to donate funds as "fundraising".
`Arguing`: L2. {PD} Markers: tweet type is reply; "you", "point", "facts", "example", "evidence". Exclusion criteria: Code detailed factual or historical information as "Knowledge performance". Code tweets in which users exchange comments without disagreement as "Audiencing".
`Audiencing`: L2. {PD} Markers: "HIMARS O'Clock", "bavovna", military terminology, "what airdefence doin", "Russia is losing". Exclusion criteria: Code detailed commentary or predictions as "Knowledge performance".
`Boosting`: L1. {PD} Markers: User handles in the beginning of the tweet, "boost", URLs.
`Community work`: L2. {PD} Markers: Mentions of "the way" or phrases "This is the way", "fella", "NAFO-themed", "NAFO expansion", "movement", "team". Exclusion criteria: Calls to other NAFO members to engage in an activity together should be coded as Mobilising.
`Expressing solidarity`: L3. {PD} Markers: "Slava Ukraini", "Glory to Ukraine", #StandWithUkraine, "Russian warship". Exclusion criteria: If the tweet contains another form of action or practice, such as putting an emphasis on the goodness of the speaker (Self-promotion) or requesting to become a member of NAFO (Membership requests), prioritise the other codes. Tweets that express solidarity towards NAFO should be coded as "Community work".
`Fundraising`: L1. {PD} Markers: "Donate", "kibble", "feed the wolves", #RageDonate, (only in combination with donation markers) names of weapons, equipment, or military regiments.
`Knowledge performance`: L3. {PD} Markers: "algorithm", "I", "mine", military terms, political actors, historical facts or facts about Twitter or other users, condescending tone. Exclusion criteria: Code tweets that put emphasis on how the interlocutor is wrong as "Arguing".
`Membership requests`: L1. {PD} Markers: "get a fella", #fellarequests, "ready", details around items to be depicted in the avatar, URLs. Exclusion criteria: Code tweets that suggest someone should join NAFO as "Community work".

`Meme creation`: L2. {PD} Markers: #FellaRequests, use of a word “meme”, “need”, “forge”, “make”, “template”. Exclusion criteria: tweets using memes for a purpose – either to annoy someone (code as Shitposting) or for enjoyment (code as Play), word “meme” featured in news about NAFO (code as “News and content curation”).
`Mobilising`: L1. {PD} Markers: #article5 or #NAFOarticle5, #NAFOfellas #NAFOfella #NAFOhelp in combination to statements like “Check this out”, “retweet”, “RT”, “you know what to do”. Exclusion criteria: If an activity entails donation of money or goods, label as “Fundraising”.
`News and content curation`: L1. {PD} Markers: Names of places or politicians, URLs, “says” or other verbs in Present Simple, “interview”, news headline writing style.
`Play`: L2. {PD} Markers: “CIA”, “bonk”, “Langley”, “Crimea Beach party”, “racoons”, “tractors”. Exclusion criteria: Code tweets with a clear adversarial target as Shitposting.
`Self-promotion`: L1. {PD} Markers: first-person point of view (“I did”, “I am”, “I would”, “my favourite”), “bonked”, “vatnik”, Medvedev, Zakharova, Jason Hinckle.
`Shitposting`: L1. {PD} Markers: Tweet type: replies to Russian embassies, Ambassador Ulyanov, Kim Dot Com, Andrew Korybko, words like “[redacted]”, “Langley”, “CIA handlers”, “nonsense pronounced”. Exclusion criteria: If a tweet appears like Shitposting but is dismissive of NAFO, code as “Not applicable”.
`Not applicable`: L3. {PD} Markers: slurs or insults targetting NAFO, complaints about NAFO.
Input Tweet:

E.3.3 PD+COT

ESC

{Task, practice definition, community description, instructions, and practice descriptions from the PD prompt}

Here are a few examples of tweets with their assigned practice and reasoning behind it.
Tweet: {tweet} Let’s think step by step: 1) The author means that immediate assistance is needed for Ukrainian Mariupol defenders. 2) It advocates for saving Mariupol and those defending it from the Russian invasion. Answer: `Advocacy`
Tweet: {tweet} Let’s think step by step: 1) The author means that while it’s a controversial opinion and Ukraine deserves to host Eurovision, it is not a good idea to do so currently. 2) They present arguments for why their opinion is correct. Answer: `Arguing`
Tweet: {tweet} 1) The author means that either Spain or Ukraine will win this year. 2) It provides a brief commentary on the Eurovision performances. Answer: `Audiencing`
Tweet: {tweet} Let’s think step by step: 1) This tweet speaks about authors’ predicted Eurovision results. 2) It makes a bet by mentioning a betting-related account @RequestABet. Answer: `Betting`
Tweet: {tweet} Let’s think step by step: 1) The tweet advertises merchandise with profits supportting a pro-Ukrainian cause. 2) It engages in a form of aid towards Ukrainians suffering from Russia’s war. Answer: `Charity`
Tweet: {tweet} Let’s think step by step: 1) The author means that they wanted their country, the UK, to win, but acknowledge that Ukraine’s performance was also good. 2) They express a sense of national pride for the UK. Answer: `Community imagining`
Tweet: {tweet} Let’s think step by step: 1) This tweet states instances of Russia’s cruel war on Ukraine and oppressive domestic policies. 2) It is criticizing these actions. Answer: `Denouncing`
Tweet: {tweet} Let’s think step by step: 1) The author speaks of Russia’s attack on Ukraine and that they are empathetic towards Ukraine. 3) They express continuous support for Ukraine in the war. Answer: `Expressing solidarity`
Tweet: {tweet} Let’s think step by step: 1) This tweet means that a part of Eurovision broadcast made them emotional. 2) Its main intent is to express the author’s emotions. Answer: `Expressing emotions`
Tweet: {tweet} Let’s think step by step: 1) The author is making a prediction about Ukraine winning a Eurovision. 2) The tweet’s main intent is to showcase author’s deep or broad knowledge of Eurovision. Answer: `Knowledge performance`

Tweet: {tweet} Let's think step by step: 1) This tweet speaks about the author's own accomplishment. 2) Its main emphasis is on promoting a piece of content made by the author as they share a link to it. Answer: `Self-promotion`

Tweet: {tweet} Let's think step by step: 1) This tweet is a short, factual sentence about the song contest and its background. 2) It is a form of news content which includes a URL likely pointing to the article. Answer: `News and content curation`

Tweet: {tweet} Let's think step by step: 1) The tweet is claiming Ukraine won because of political reasons. 4) The tweet is not supportive of Ukraine. Practice: `Not applicable`

Input Tweet:

NAFO

{Task, practice definition, community description, instructions, and practice descriptions from the PD prompt}

Here are a few examples of tweets with their assigned practice and reasoning behind it.

Tweet: {tweet} Let's think step by step: 1) The author means that to win, Ukraine needs to have an advantage in weapons, and that the Western leaders need to send Ukraine those weapons (Leopard tanks). 2) It advocates for providing Ukraine with weapons. Answer: `Advocacy`

Tweet: {tweet} Let's think step by step: 1) The author means that their opponent is wrong about who the author is and why they support Ukraine. 2) They present arguments in favour of supporting Ukraine. Answer: `Arguing`

Tweet: {tweet} Let's think step by step: 1) The author is briefly commenting on a news piece about the war, likely referring to Russia's military failure. 2) They are engaged in discussing the events of the war together with others. Answer: `Audiencing`

Tweet: {tweet} Let's think step by step: 1) The author means that they support the cause or content of the tweet they are replying to, as well as Ukraine and NAFO. 2) It attempts to increase visibility of the original tweet as tweets with more replies are more likely to get recommended by the Twitter algorithm. Answer: `Boosting`

Tweet: {tweet} Let's think step by step: 1) The tweet refers to an accomplishment of NAFO and suggests the collective's members need to continue their important efforts. 2) It celebrates the collective, encourages members to continue being a part of it, and creates a sense of community. Answer: `Community work`

Tweet: {tweet} Let's think step by step: 1) The author means that they will always support Ukraine and believe in the country winning in the war. 2) They pay respect to Ukraine. Answer: `Expressing solidarity`

Tweet: {tweet} Let's think step by step: 1) The author speaks about a fundraiser for someone in the Ukrainian military. 2) They are encouraging others to donate and spread the fundraiser further. Answer: `Fundraising`

Tweet:{tweet} Let's think step by step: 1) The author means that a certain development on Twitter is due to the activity of pro-Russian and other actors. 2) The tweet's main intent is to showcase author's deep or broad knowledge of the information environment of Twitter during the war. Answer: `Knowledge performance`

Tweet: {tweet} Let's think step by step: 1) The author means that they would like to have a NAFO avatar created featuring certain attributes. 2) The tweet's main intent is to request membership in NAFO. Answer: `Membership requests`

Tweet: {tweet} Let's think step by step: 1) The author means that NAFO should create a template for memes inspired by a film "Red Notice". 2) The tweet's main intent is to support meme creation efforts of NAFO. Answer: `Meme creation`

Tweet: {tweet} 1) The author means that NAFO should pay attention to a tweet by a potential pro-Russian actor. 2) The tweet's main intent is to make as many NAFO members as possible to engage with a pro-Russian user and counter Russian propaganda. Answer: `Mobilising`

		NAFO			ESC		
		F	P	R	F	P	R
	Random	6.11(1.2)	6.18(1.3)	6.18(1.3)	7.63(1.5)	7.69(1.6)	7.71(1.5)
	Majority	2.54	1.60	6.25	3.01	1.87	7.69
SVM	Linear	20.28(1.17)	33.84(2.96)	19.13(1.38)	23.71(2.57)	44.9(2.62)	23.25(1.97)
	Weighted	13.26(1.97)	28.09(4.1)	14.39(1.57)	23.71(2.22)	44.9(4.75)	23.25(1.72)
SetFit	MPNET(K=1)	10.41(2.59)	15.08(4.73)	13.12(2.65)	10.55(4.64)	11.88(5.26)	14.75(6.20)
	MPNET(K=2)	16.4(1.96)	22.66(6.85)	18.21(3.73)	18.03(5.72)	20.75(8.4)	21.14(5.12)
	MPNET(K=8)	25.67(3.88)	33.61(9.16)	26.08(3.58)	32.13(3.6)	39.9(5.21)	31.3(3.63)
	DistilRoBERTA(K=1)	5.61(1.84)	7.91(2.76)	9.01(2.12)	6.44(2.16)	5.79(2.60)	11.82(3.75)
	DistilRoBERTA(K=2)	10.18(3.11)	12.02(2.29)	13.11(3.22)	13.48(4.54)	19.14(6.94)	15.78(2.95)
	DistilRoBERTA(K=8)	10.13(3.12)	12.03(2.28)	12.41(3.12)	22.08(8.19)	26.26(13.35)	23.14(6.49)

Table 9: Detailed practice prediction results for baseline models. We report macro-averaged F1, prediction and recall, with standard deviation in brackets. Results are averaged across five folds for SVM and SetFit models and across 1000 runs for the Random baseline. We only repeat a run with the Majority classifier once.

Tweet: {tweet} Let's think step by step: 1) This tweet is a short, factual sentence about the events of Russia's war on Ukraine. 2) It is a form of news content which includes a URL likely pointing to the article. Answer: `News and content curation`
Tweet: {tweet} Let's think step by step: 1) The author is pointing out a reseblance of an image to Nazgul, a character from Lord of the Rings. 2) They are playfully engaging with others through popular culture references. Answer: `Play`
Tweet: {tweet} Let's think step by step: 1) This tweet speaks about the author's own accomplishment of writing a thread that attracted online and media attention. 2) Its main emphasis is on celebrating the author's achievement as an effective supporter of Ukraine. Answer: `Self-promotion`
Tweet: {tweet} Let's think step by step: 1) This tweet shares an image that portrays Putin as female. 2) It uses crude humour to mock Putin and derail Russian propaganda efforts. Answer: `Shitposting`
Tweet: {tweet} Let's think step by step: 1) The tweet is accusing NAFO of hypocricy. 4) The tweet is not supportive of NAFO as it tries to portray NAFO in bad light. Answer: `Not applicable`
Input Tweet:

E.3.4 Few shot (PD, MPE)

Few shot PD and MPE prompts were constructed by appending the following instruction and demonstration tweets to the above prompts:

> Here are a few examples of tweets with their assigned practice.
>
> Tweet: {tweet} Practice: {practice}
>
> Tweet: {tweet} Practice: {practice}
>
> ...

E.4 Detailed results of experiments

E.4.1 Overview of macro-averaged F1, precision, and recall

For all models used in this study, we report macro-average F1, precision, and recall metrics in tables 9 (baseline models) and 10 (OpenAI models).

E.4.2 Per-class results for all models

To provide a detailed view of model performance and acknowledge the label skew in the ground truth data, we also outline class-wise metrics for all models in tables 11, 12, 13, and 14.

For our experiments with variations of the in-context learning prompt with GPT-4 model and one demonstration sample per class (compared in Table 11), we observe that adding either COT reasoning steps or MPE features improves the F1 score for all categories, in comparison to practice description prompts.

		NAFO			ESC		
		F	P	R	F	P	R
GPT3.5 (PD)	K=0	39.31(1.85)	41.37(3.09)	41.61(1.37)	38.01(2.24)	41.12(1.61)	47.03(2.7)
	K=1	35.99(2.63)	51.66(2.55)	33.06(2.82)	36.27(4.21)	48.56(2.69)	37.99(5.9)
	K=2	21.95(2.65)	49.5(3.34)	19.77(2.17)	12.15(3.34)	43.86(9.8)	13.42(2.48)
GPT3.5 (PD+MPE)	K=0	**43.39(2.28)**†	**45.52(2.96)**†	**48.35(2.1)**†	**40.31(2.44)**	**45.27(2.78)**†	43.16(3.03)
	K=1	**38.48(2.68)**	**57.67(4.92)**	**34.16(3.2)**	**38.35(5.02)**	**51.11(5.02)**	36.69(5.16)
	K=2	**27.32(5.18)**†	**56.99(4.8)**	**23.47(3.9)**	**16.26(7.2)**	43.52(17.63)	**16.23(5.18)**
GPT4 (PD)	K=0	47.65(1.77)	48.52(1.2)	55.05(1.37)	49.33(2.59)	51.09(2.79)	56.73(3.17)
	K=1	46.62(2.11)	47.24(1.77)	53.24(2.19)	49.24(3.29)	47.62(2.54)	56.78(4.67)
	K=2	45.23(2.3)	46.58(5.2)	50.18(2.6)	49.14(2.41)	50.31(3.01)	54.33(2.91)
GPT4 (PD+MPE)	K=0	**53.54(1.24)**†	**52.68(1.05)**†	**62.38(1.85)**†	**56.06(5.07)**†	**57.41(4.93)**†	**59.93(4.69)**
	K=1	**52.39(2.39)**†	**52.69(1.91)**†	**57.52(2.24)**†	**53.33(2.98)**	**52.71(3.15)**†	**60.56(4.51)**
	K=2	**51.31(2.54)**†	**53.54(2.78)**†	**57.2(3.31)**†	**54.44(5.74)**†	**55.3(4.83)**	**57.39(6.16)**
PD+COT	K=1	**51.96(1.38)**†	**55.10(1.17)**†	**58.60(0.49)**†	**53.87(2.59)**†	**53.68(2.08)**†	**61.30(3.21)**†
PD+COT+MPE	K=1	**56.88(2.06)**†	**58.60(2.66)**†	**64.15(1.80)**†	**58.71(5.15)**†	**57.84(5.02)**†	**62.89(4.93)**†

Table 10: Detailed results for experiments with OpenAI models. We compare the performance of the base practice description prompts with MPE and COT prompts. We report macro-averaged F1, precision, and recall. Bold font indicates an increase with MPE or COT prompt in comparison to the same setting with base prompts. The dagger indicates statistically significant results of paired t-test calculated at $p \leq 0.05$ when comparing the base and MPE or COT prompt result.

COT prompts appear to be more successful with categories where meaning and intention is hard to infer from the tweet text without additional contextual information. This was the case with ESC's `Denouncing` practice and `Knowledge performance` practice in both case studies, where it was particularly important for the model to infer the intention of "humbly bragging" about one's knowledge or strongly criticising Russia as the invading country.

In contrast, the MPE prompt demonstrated significant increase for practices `Expressing solidarity` (increase from 39.75 to 56.42) and `Community work` (improvement from 36.67 to 45.84) in the NAFO data set. Both of these practices rely on community or task specific-vernacular which was inluded in the form of markers. For example, `Community work` uses words like "fellas" used to address the members of the collective and phrases like "This is the way" to communicate the movement's values – we hypothesise that the MPE prompt helped highlight such instances in the test data.

MPE also outperformed COT in the "Not applicable" category, where the model was expected to identify practices of users supporting Russia. While we did not anticipate COT will perform significantly worse with this category, it is conceivable that constructing a prompt emphasising intention and action leads to the model "forgetting" to incorporate an implicit stance detection task.

As stated in Section 5, we encourage future studies to make the stance detection task an explicit part of the COT prompt. Alternatively, as Table 11 demonstrates, combining COT and MPE prompts may lead to improvement in the results of the practice prediction task.

	Practice	**PD**	**PD+MPE**	**PD+COT**	**PD+COT+MPE**
NAFO	**macro-averaged F1 (All)**	46.62(2.11)	52.39(2.39)	51.96(1.38)	**56.88(2.06)**
	Advocacy	70.18(9.97)	73.58(3.03)	**76.08(4.67)**	73.42(5.49)
	Arguing	39.41(7.05)	44.12(4.00)	40.92(5.15)	**48.91(5.83)**
	Audiencing	13.74(6.93)	22.23(7.31)	18.02(4.13)	**23.18(5.66)**
	Boosting	91.16(2.34)	94.62(3.34)	**95.73(4.40)**	95.47(4.93)
	Community work	36.67(6.26)	45.84(10.23)	39.90(4.80)	**49.75(5.47)**
	Expressing solidarity	39.75(8.77)	56.42(10.72)	48.14(7.49)	**63.66(12.37)**
	Fundraising	76.13(4.17)	73.14(7.20)	77.70(8.80)	**79.21(7.92)**
	Knowledge performance	47.34(9.72)	47.89(5.91)	51.79(6.18)	**55.77(6.50)**
	Membership requests	58.09(5.98)	63.76(10.06)	70.56(12.66)	**72.19(6.93)**
	Meme creation	49.37(7.53)	54.82(13.10)	63.51(11.15)	**64.46(4.26)**
	Mobilising	75.30(5.09)	76.85(3.72)	79.56(3.15)	**81.93(2.64)**
	News and content curation	42.67(7.21)	42.58(11.05)	57.68(10.30)	**58.72(9.77)**
	Play	20.66(7.10)	34.17(2.78)	32.07(9.87)	**37.99(8.14)**
	Self-promotion	20.79(10.66)	23.14(7.07)	24.69(7.01)	**32.69(9.93)**
	Shitposting	34.56(6.76)	36.66(4.27)	37.28(5.69)	**42.03(7.06)**
	Not applicable	30.10(12.22)	**48.38(6.00)**	17.79(5.54)	30.77(2.78)
ESC	**macro-averaged F1 (All)**	49.24(3.29)	53.33(2.98)	53.87(2.59)	**58.71(5.15)**
	Advocacy	50.61(11.87)	55.81(15.90)	60.45(11.65)	**61.32(6.26)**
	Arguing	28.43(16.29)	24.11(9.51)	27.08(7.89)	**29.95(9.89)**
	Audiencing	44.14(3.30)	59.77(9.75)	61.01(1.13)	**70.17(2.80)**
	Betting	89.06(5.60)	90.50(3.68)	**92.33(5.04)**	91.65(4.23)
	Charity	67.97(6.19)	**73.88(6.57)**	73.73(3.96)	73.23(7.35)
	Community imagining	23.12(7.83)	24.82(10.86)	**26.07(11.64)**	24.32(15.33)
	Denouncing	53.42(9.69)	48.89(10.39)	63.38(6.92)	**63.90(12.92)**
	Expressing emotions	44.68(8.99)	52.71(9.86)	41.74(8.17)	**64.30(11.61)**
	Expressing solidarity	40.35(9.54)	44.95(12.02)	49.98(7.52)	**57.37(12.77)**
	Knowledge performance	30.49(7.55)	32.05(12.78)	37.83(9.18)	**38.50(13.97)**
	News and content curation	75.84(5.90)	77.24(3.53)	**80.48(2.11)**	80.02(2.80)
	Self-promotion	36.88(20.73)	**45.66(18.18)**	39.82(16.49)	44.18(18.28)
	Not applicable	55.21(5.73)	62.89(9.69)	46.42(3.50)	**64.35(2.09)**

Table 11: Per-class comparison of GPT-4's performance in *K=1* setting with PD (Practice Descrtiption), PD+MPE (Markers, Priority, Exclusion criteria), PD+COT (Chain-of-Thought), and PD+COT+MPE in-context learning prompts. We report a mean F1 score for each class across five folds, and a macro-averaged F1 score for all categories, with standard deviation in brackets. Bold font indicates the highest score for the specific practice.

	Practice	K0(PD)	K0(PD+MPE)	K1(PD)	K1(PD+MPE)	K2(PD)	K2(PD+MPE)
NAFO	**macro-averaged F1 (All)**	47.65(1.77)	53.54(1.24)	46.62(2.11)	52.39(2.39)	45.23(3.47)	51.31(2.54)
	Advocacy	69.22(2.62)	71.17(7.35)	70.18(9.97)	73.58(3.03)	64.01(16.14)	67.85(12.26)
	Arguing	36.78(1.13)	44.29(5.08)	39.41(7.05)	44.12(4.00)	31.79(3.22)	34.13(6.10)
	Audiencing	14.22(2.67)	19.32(5.92)	13.74(6.93)	22.23(7.31)	10.84(3.14)	18.40(8.84)
	Boosting	88.81(6.15)	74.90(8.73)	91.16(2.34)	94.62(3.34)	92.45(5.62)	96.95(3.24)
	Community work	41.25(4.53)	48.15(5.70)	36.67(6.26)	45.84(10.23)	39.92(6.99)	46.61(8.28)
	Expressing solidarity	53.02(12.90)	63.38(7.73)	39.75(8.77)	56.42(10.72)	32.67(10.12)	49.03(15.74)
	Fundraising	73.28(5.06)	75.72(5.69)	76.13(4.17)	73.14(7.20)	80.89(7.48)	75.59(4.94)
	Knowledge performance	47.97(6.65)	52.01(6.20)	47.34(9.72)	47.89(5.91)	39.29(4.14)	43.41(6.70)
	Membership requests	60.43(12.70)	63.92(5.32)	58.09(5.98)	63.76(10.06)	66.89(14.60)	75.26(11.22)
	Meme creation	57.34(5.20)	62.94(2.03)	49.37(7.53)	54.82(13.10)	48.19(7.38)	58.07(8.51)
	Mobilising	73.67(4.13)	80.17(3.70)	75.30(5.09)	76.85(3.72)	68.80(12.68)	77.42(2.89)
	News and content curation	51.74(8.14)	50.29(7.82)	42.67(7.21)	42.58(11.05)	37.05(9.13)	45.93(16.90)
	Play	37.71(3.41)	44.77(5.07)	20.66(7.10)	34.17(2.78)	19.93(6.86)	35.39(6.70)
	Self-promotion	13.49(9.08)	27.24(11.90)	20.79(10.66)	23.14(7.07)	16.05(11.58)	19.62(9.50)
	Shitposting	32.44(5.01)	41.48(6.42)	34.56(6.76)	36.66(4.27)	35.35(2.93)	42.26(5.97)
	Not applicable	10.94(2.61)	36.83(1.66)	30.10(12.22)	48.38(6.00)	39.53(10.64)	35.06(10.52)
ESC	**macro-averaged F1 (All)**	49.33(2.59)	56.06(5.07)	49.24(3.29)	53.33(2.98)	49.14(2.41)	54.44(5.74)
	Advocacy	53.45(12.19)	59.58(7.56)	50.61(11.87)	55.81(15.90)	45.63(13.48)	55.93(14.00)
	Arguing	25.87(13.74)	28.85(13.46)	28.43(16.29)	24.11(9.51)	17.94(9.18)	16.83(10.85)
	Audiencing	63.14(2.92)	71.25(1.62)	44.14(3.30)	59.77(9.75)	50.04(11.09)	65.46(0.99)
	Betting	90.94(2.81)	90.46(2.42)	89.06(5.60)	90.50(3.68)	87.40(5.06)	84.86(5.56)
	Charity	69.94(5.50)	72.81(6.32)	67.97(6.19)	73.88(6.57)	69.43(5.65)	69.75(7.11)
	Community imagining	12.25(4.01)	25.07(14.57)	23.12(7.83)	24.82(10.86)	18.21(1.28)	21.57(14.18)
	Denouncing	50.58(7.21)	52.16(9.85)	53.42(9.69)	48.89(10.39)	56.43(9.89)	64.35(10.24)
	Expressing emotions	36.48(5.69)	60.68(9.46)	44.68(8.99)	52.71(9.86)	60.53(2.73)	61.02(9.85)
	Expressing solidarity	42.21(10.64)	46.40(14.35)	40.35(9.54)	44.95(12.02)	44.11(7.31)	48.21(6.69)
	Knowledge performance	24.68(7.10)	35.00(13.97)	30.49(7.55)	32.05(12.78)	22.86(6.54)	33.15(8.28)
	News and content curation	81.07(3.10)	81.20(3.76)	75.84(5.90)	77.24(3.53)	58.30(6.50)	72.62(1.90)
	Self-promotion	38.50(14.17)	40.77(15.34)	36.88(20.73)	45.66(18.18)	48.82(14.08)	47.17(19.55)
	Not applicable	52.14(1.87)	64.50(3.61)	55.21(5.73)	62.89(9.69)	59.17(6.03)	66.80(3.02)

Table 12: Per-class results for OpenAI's **GPT-4 model**. We report a mean F1 score for each class across five folds, and a macro-averaged F1 score for all categories, with standard deviation in brackets. K indicates the number of demonstration samples.

	Practice	**K0(PD)**	**K0(PD+MPE)**	**K1(PD)**	**K1(PD+MPE)**	**K2(PD)**	**K2(PD+MPE)**
NAFO	**macro-averaged F1 (All)**	39.31(1.85)	43.39(2.28)	35.99(2.63)	38.48(2.68)	21.95(2.65)	27.32(5.18)
	Advocacy	55.31(6.20)	60.97(8.99)	52.81(15.35)	67.52(7.38)	42.41(12.16)	67.31(10.37)
	Arguing	28.89(4.13)	27.19(2.92)	11.77(6.71)	8.42(10.16)	2.42(5.42)	0.00(0.00)
	Audiencing	0.00(0.00)	11.06(8.09)	0.00(0.00)	2.86(6.39)	0.00(0.00)	2.86(6.39)
	Boosting	86.50(8.60)	87.79(4.35)	77.48(11.03)	68.63(6.53)	55.58(6.59)	42.54(14.68)
	Community work	28.31(3.60)	32.27(5.51)	22.51(6.67)	36.97(9.98)	3.61(3.48)	10.88(7.01)
	Expressing solidarity	53.01(11.94)	48.98(10.06)	38.27(7.59)	39.40(16.41)	10.42(14.38)	10.91(11.28)
	Fundraising	70.40(4.40)	75.46(8.03)	71.15(6.75)	67.76(9.00)	58.33(22.90)	56.75(10.51)
	Knowledge performance	2.13(2.94)	12.08(2.66)	2.13(2.94)	5.28(3.29)	2.11(2.92)	2.32(3.18)
	Membership requests	22.92(5.97)	39.94(5.53)	49.87(9.63)	63.02(16.24)	17.20(4.02)	33.01(25.32)
	Meme creation	58.58(11.95)	65.93(12.21)	56.34(21.60)	66.29(8.27)	24.69(14.29)	50.81(21.86)
	Mobilising	62.46(3.16)	68.80(4.63)	61.18(7.88)	68.55(6.77)	49.47(28.49)	68.56(5.17)
	News and content curation	38.58(8.18)	44.18(7.12)	41.35(9.94)	41.73(4.55)	29.30(12.81)	34.33(12.75)
	Play	33.82(1.53)	36.81(8.26)	19.28(11.44)	17.43(3.61)	7.30(4.77)	9.29(6.54)
	Self-promotion	16.06(6.54)	14.70(4.82)	5.08(7.05)	5.54(8.18)	0.00(0.00)	0.00(0.00)
	Shitposting	27.63(3.75)	29.30(4.55)	17.15(8.18)	5.55(4.33)	3.02(4.36)	1.11(2.48)
	Not applicable	44.35(2.61)	38.78(3.25)	49.48(0.93)	50.68(1.62)	45.38(1.80)	46.39(1.42)
ESC	**macro-averaged F1 (All)**	38.01(2.24)	40.31(2.44)	36.27(4.21)	38.35(5.02)	12.15(3.34)	16.26(7.20)
	Advocacy	44.58(13.50)	55.31(17.11)	48.79(12.85)	49.10(13.22)	20.28(9.26)	19.71(21.05)
	Arguing	12.75(5.81)	13.53(10.99)	10.07(8.84)	2.22(4.97)	0.00(0.00)	2.50(5.59)
	Audiencing	30.99(3.72)	55.97(3.22)	12.87(5.41)	59.74(5.17)	2.74(3.17)	40.64(14.67)
	Betting	88.40(8.04)	74.61(12.97)	87.34(7.61)	78.20(13.10)	30.64(19.82)	35.21(19.26)
	Charity	65.30(7.92)	59.82(5.54)	52.46(13.95)	56.01(11.58)	25.42(17.51)	27.31(14.85)
	Community imagining	0.00(0.00)	0.00(0.00)	3.48(3.20)	0.00(0.00)	0.00(0.00)	2.50(5.59)
	Denouncing	52.86(13.73)	60.03(9.11)	39.77(9.69)	39.67(7.69)	8.41(7.72)	3.08(6.88)
	Expressing emotions	49.45(3.19)	69.71(4.23)	62.69(8.87)	54.10(32.29)	16.13(13.05)	13.50(15.42)
	Expressing solidarity	34.56(4.87)	38.91(7.24)	35.38(11.85)	34.68(14.68)	7.84(4.44)	12.02(11.00)
	Knowledge performance	14.25(5.35)	8.08(9.01)	17.60(7.50)	11.26(12.81)	4.06(5.89)	8.54(9.46)
	News and content curation	35.87(7.83)	30.40(6.96)	26.47(12.35)	31.51(7.03)	2.88(3.04)	6.34(10.32)
	Self-promotion	25.33(16.94)	11.48(12.96)	30.82(19.71)	30.79(19.51)	2.86(6.39)	0.00(0.00)
	Not applicable	39.82(4.11)	46.13(0.55)	43.79(6.05)	51.23(4.60)	36.65(1.90)	39.98(1.95)

Table 13: Per-class results for OpenAI's **GPT-3.5** model. We report a mean F1 score for each class across five folds, and a macro-averaged F1 score for all categories, with standard deviation in brackets. K indicates the number of demonstration samples.

	Practice	SVM-L	SVM-W	MPNet-K2	RoBERTa-K2	MPNet-K8	RoBERTa-K8
NAFO	**macro-averaged F1 (All)**	20.28(1.17)	13.26(1.97)	16.40(1.96)	10.18(3.11)	25.67(3.88)	10.13(3.12)
	Advocacy	3.08(6.88)	5.43(7.54)	0.00(0.00)	0.00(0.00)	14.71(22.21)	0.00(0.00)
	Arguing	0.00(0.00)	10.76(7.66)	2.05(4.59)	0.00(0.00)	12.80(7.60)	0.00(0.00)
	Audiencing	0.00(0.00)	2.50(5.59)	1.90(4.26)	0.00(0.00)	6.59(7.67)	0.00(0.00)
	Boosting	68.39(8.75)	43.85(10.96)	16.62(30.65)	15.71(35.14)	22.15(32.51)	15.71(35.14)
	Community work	27.65(7.08)	20.30(4.58)	17.91(10.31)	11.94(13.47)	27.70(6.12)	13.76(12.70)
	Expressing solidarity	24.86(9.82)	2.86(6.39)	12.15(12.60)	11.37(15.95)	16.27(12.70)	6.67(14.91)
	Fundraising	0.00(0.00)	0.00(0.00)	36.08(29.71)	10.00(22.36)	48.81(33.09)	0.00(0.00)
	Knowledge performance	2.31(3.22)	15.23(12.77)	14.61(12.23)	4.21(9.42)	23.66(12.75)	12.13(17.86)
	Membership requests	30.74(11.71)	4.00(8.94)	28.67(30.26)	3.33(7.45)	42.10(34.68)	3.33(7.45)
	Meme creation	46.41(18.43)	7.37(6.76)	24.62(33.77)	18.00(26.83)	41.21(39.04)	18.00(26.83)
	Mobilising	72.42(6.91)	68.35(4.85)	57.03(14.60)	39.82(26.89)	67.41(5.15)	45.88(29.23)
	News and content curation	0.00(0.00)	0.00(0.00)	6.45(14.43)	0.00(0.00)	10.48(10.38)	0.00(0.00)
	Play	0.00(0.00)	3.53(7.89)	0.61(1.36)	0.87(1.94)	13.93(6.60)	0.87(1.94)
	Self-promotion	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	8.58(9.92)	0.00(0.00)
	Shitposting	0.00(0.00)	5.39(7.41)	4.40(6.10)	3.33(7.45)	7.48(6.49)	3.33(7.45)
	Not applicable	48.57(1.54)	22.56(22.39)	39.27(7.77)	44.25(3.11)	46.80(7.25)	42.40(4.89)
ESC	**macro-averaged F1 (All)**	29.51(2.57)	23.71(2.22)	18.03(5.72)	13.48(4.54)	32.13(3.61)	22.08(8.19)
	Advocacy	39.84(9.07)	32.35(11.87)	4.85(10.84)	6.15(13.76)	24.81(15.94)	15.90(22.24)
	Arguing	0.00(0.00)	2.86(6.39)	9.67(10.94)	5.22(11.67)	18.48(18.49)	2.11(4.71)
	Audiencing	53.07(3.28)	36.70(18.01)	44.07(4.82)	41.40(12.58)	55.16(6.66)	56.11(5.27)
	Betting	81.42(4.02)	77.70(5.40)	33.77(34.85)	31.76(29.44)	42.75(38.52)	30.84(42.42)
	Charity	23.13(9.70)	5.69(7.98)	20.89(33.83)	8.57(19.17)	52.41(23.76)	32.69(20.14)
	Community imagining	6.67(9.43)	0.00(0.00)	0.00(0.00)	0.00(0.00)	16.50(9.70)	0.00(0.00)
	Denouncing	10.64(10.35)	7.56(10.86)	13.52(13.65)	0.00(0.00)	18.37(24.33)	4.71(10.52)
	Expressing emotions	16.82(11.29)	7.33(10.11)	17.75(30.57)	11.43(25.56)	14.54(19.53)	10.14(15.56)
	Expressing solidarity	19.35(7.91)	15.06(13.10)	1.86(4.16)	0.00(0.00)	11.23(5.01)	3.08(4.22)
	Knowledge performance	0.00(0.00)	20.81(5.14)	4.07(9.10)	1.33(2.98)	25.65(5.66)	5.78(5.53)
	News and content curation	67.40(1.61)	58.00(9.22)	49.70(7.92)	48.63(9.18)	69.23(5.69)	69.68(4.02)
	Self-promotion	13.08(21.69)	7.08(9.83)	8.03(12.91)	0.00(0.00)	15.73(14.45)	7.21(10.61)
	Not applicable	52.21(5.52)	37.15(17.53)	26.20(17.51)	20.69(13.43)	52.89(7.20)	48.84(6.32)

Table 14: Per-class results for **SVM and SetFit baselines**. We report a mean F1 score for each class across five folds, and a macro-averaged F1 score for all categories, with standard deviation in brackets. K indicates the number of demonstration samples.