Assessing the Feasibility of Predicting Social Network Users' Behavior Using Linguistic and Social Cues

Stephanie Anneris MALVICINI [a,b] Nina PARDAL [c]
Laura ALONSO ALEMANY [d] Maria Vanina MARTINEZ [a]
Gerardo Ignacio SIMARI [b,e]

[a] *Artificial Intelligence Research Institute (IIIA-CSIC), Barcelona, Spain*
[b] *DCIC, Universidad Nacional del Sur (UNS), Argentina*
[c] *University of Huddersfield, UK*
[d] *FAMAF, Universidad Nacional de Córdoba, Argentina*
[e] *ICIC (UNS–CONICET), Argentina*

Abstract. Social media polarization is increasingly shaping everyday life, driving divisions in beliefs, relationships, and communities. As societal groups grow more divided, trust diminishes and cooperation becomes harder, making polarization a pressing issue today that requires both social and technical solutions. While the need to intervene in order to mitigate polarization is often clear, it is first essential to establish methods for identifying these phenomena. One way of tackling that problem is focusing on the development of computational tools that help in modelling and predicting online users' behavior and interactions within a social network environment. In this study, we propose various such models based on different machine learning techniques and study how social cues and the linguistic features of the conversations in which they are involved, impact their behavior. Concretely, we tested two approaches for predicting the linguistic features in future messages, and explored both classical machine learning and large language models. Our preliminary experiments, conducted with real-world data, yielded interesting results that will be useful for the design of future studies.

Keywords. AI Applications, Natural Language Processing, Agents and Multi-agent Systems

1. Introduction

As established in the Global Risks Report 2025 [1], polarization within societies further hardens views and affects policy-making, while also contributing to intensify misinformation and disinformation. False or misleading content is complicating the geopolitical environment as a mechanism for foreign entities to affect voter intentions, create doubt among the general public worldwide about what is happening in conflict zones, and tarnish the image of products or services from another country. Misinformation, disinformation and societal polarization am-

plify the other leading risks, from state-based armed conflict to extreme weather events. Both political and societal polarization skew narratives and distort facts, contributing to low and declining trust in media.

To mitigate polarization, and before considering intervention strategies, we need to be able to detect such phenomena. This is not an easy task due to the existence of multiple definitions and interrelated factors: group polarization [2,3], political polarization [4,5], opinion polarization [6], social science polarization definition [7,8,9], interactional polarization [10], among others. Also, the metrics and methodologies used to measure polarization are diverse, encompassing techniques such as sentiment analysis, opinion formation models, network metrics, and interactions' examination. Considering that there is limited research on the linguistic side of the materialization of polarization, and inspired by the work of [11], as well as following the idea presented in [12] on linguistic markers of polarization, our goal is to build a model capable of answering the following question: *Is it possible to build a model that, given a user's past behavior and the current context, can predict with a level of confidence the linguistic cues of the user's response?* Here, we explore both machine learning (ML) and large language models (LLMs) to explore our research question, while developing representations for conversations and users from social networks, focusing on the available linguistic cues.

2. Literature Review

Modeling human behavior computationally is a longstanding goal in computational sociology. Agent-based modeling (ABM) has been the dominant method for simulating social behavior over the past four decades, linking micro-level actions with macro-level societal trends and capturing interdependencies in human behavior [13]. This work focuses on ABM applied to social media user behavior. Most existing simulations use simple reactive agents, often emphasizing whether users communicate and how information propagates. For instance, [14] proposes a probabilistic framework modeling users as stochastic processes to predict responses to advocacy posts. In [15], a latent topic model is presented incorporating non-uniform attention to simulate opinion diffusion, demonstrating that psychosocial models better describe user behavior. ATMiner [16] predicts user opinion, sentiment, and action on controversial topics to understand online behavior.

The work in [17] uses X (formerly Twitter) data to construct virtual communities, distinguishing between strong and weak links to measure influence, correlating it with lexical diversity, sentiment polarity, and social graph variables. Similarly, [18] models networks as directed graphs and integrates sociological and psychological user types for realistic simulations. The DUAPM model [19] predicts user activity on micro-blogging platforms using personal data and interactions. Hate speech detection is studied in [20], showing that including conversational context improves accuracy, though results vary across protected characteristics. [21] proposes a proactive NLP-based approach to detect antisocial behavior for early intervention, while [22] predicts retweets using user profiles based on topic preferences, emotions, and personality, and [11] predicts user reactions to social feeds using personality traits and social cues. Our work builds on

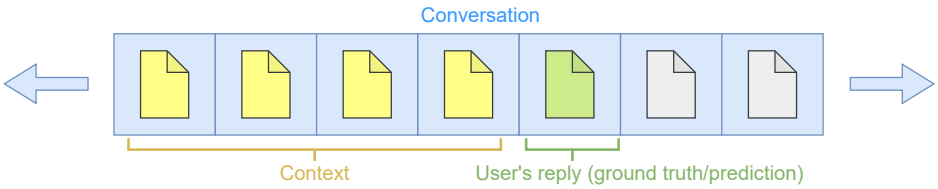

Figure 1. Context and prediction task.

the latter, extending context representation to include linguistic features, emotions, and communicative intent. [23] uses linguistic features for author profiling and micro-targeting in political contexts, aiding crisis communication strategies for public institutions. For polarization analysis, [12] combines machine learning with log-likelihood analysis to detect linguistic markers of polarization, focusing on ethos-based attacks or support; [24] further provides annotated corpora for pathos-based emotion in polarized social media discourse.

3. Predicting Users' Reactions to their Social Media Feeds

We aim to model user response given their past behavior and current conversational context. The context consists of prior messages in a thread (cf. Figure 1). Given a context, we derive linguistic and user-based features described below.

3.1. Linguistic Features in Messages

For each context, we compute features from annotated linguistic elements across messages. Many features involve averages of ratio and count values, normalized and discretized into intervals: $[0, 25)$, $[25, 50)$, $[50, 75)$, $[75, 100]$.

(C1) Moral Foundation Distribution (MFD): Based on the Moral Foundations Dictionary[1], we group categories into *vice* and *virtue*. We compute normalized averages of their word ratios and counts, producing four interval-based features.

(C2) Polarization Words: We use a polarization dictionary [25] to compute the average and ratio of polarizing terms, applying the same normalization process.

(C3) Abusive Words: Using abusive language annotations [26], we derive intervalized features for average ratio and count.

(C4) Valence: Based on standardized valence scores [27], we classify words as positive or negative and compute their average ratio and amount across messages, resulting in four interval features.

(C5–C6) Emotion Features: Using Ekman's emotion taxonomy [28], we extract the predominant emotion per context and its distribution across seven categories, each discretized into intervals.

(C7–C8) Sentiment Features: We use [29] to derive the predominant sentiment (positive, negative, neutral) and the interval-based distribution across categories.

[1] https://moralfoundations.org/other-materials

3.2. User-Based Features

The following features capture stable traits and context-specific behavior derived from user history.

(U1) Personality Type: Using the Symanto Big Five API [2], we assign each user a binary score (high/low) across five traits: Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism, yielding $2^5 = 32$ combinations.

(U2) Psychographics: Using Symanto's psycholinguistic profiling[3], we identify 32 additional user types by analyzing three key dimensions:

Emotionality: High if rationality score exceeds 0.5, low otherwise.
Communication Style: Binary labels for *self-revealing* and *fact-oriented*.
Communication Needs: Binary labels for *information-seeking* and *action-seeking*.

While Big-5 reflects stable personality traits, Symanto psychographics capture dynamic dataset- and topic-behavior.

3.3. Prediction Task

We define classification tasks aimed at predicting the linguistic cues of the user's response within the context. Each linguistic feature becomes a separate classification problem—favoring accuracy and parallelizability. Both multiclass and binary setups are considered, with interval bins and feature transformations as follows:

(P1–P2) Avg. Abusive / Polarization: Intervalized average ratio and amount values (as in C2/C3); for binary, thresholds are at 50%.

(P3) Predominant MFD: Compare virtue vs. vice intervals (amount/ratio). Multiclass: virtue>vice, equal, vice>virtue. Binary merges virtue≥vice into one class.

(P4) Predominant Sentiment: Multiclass as in (C7); binary merges positive and neutral vs. negative.

(P5) Predominant Emotion Super-Set: Multiclass as in (C5); Binary split: *neutral, joy, surprise* vs. *fear, anger, disgust, sadness*.

(P6) Predominant Valence: As with (P3), comparing positive vs. negative intervals (amount/ratio), with binary reducing to two classes.

A detailed description of the prediction classes for both binary and multiclass scenarios is provided in the supplementary material (Appendix A).

4. Experiments and Results

We conducted the following two experiments. First, we trained classical machine learning models for the prediction task. In the second one, we use LLMs to try to assess the feasibility and complexity of the prediction task itself.

[2] https://www.symanto.com/nlp-tools/nlp-api/emotion-text-analysis
[3] https://www.symanto.com/wp-content/uploads/2021/12/Symantos-Psychographics.pdf

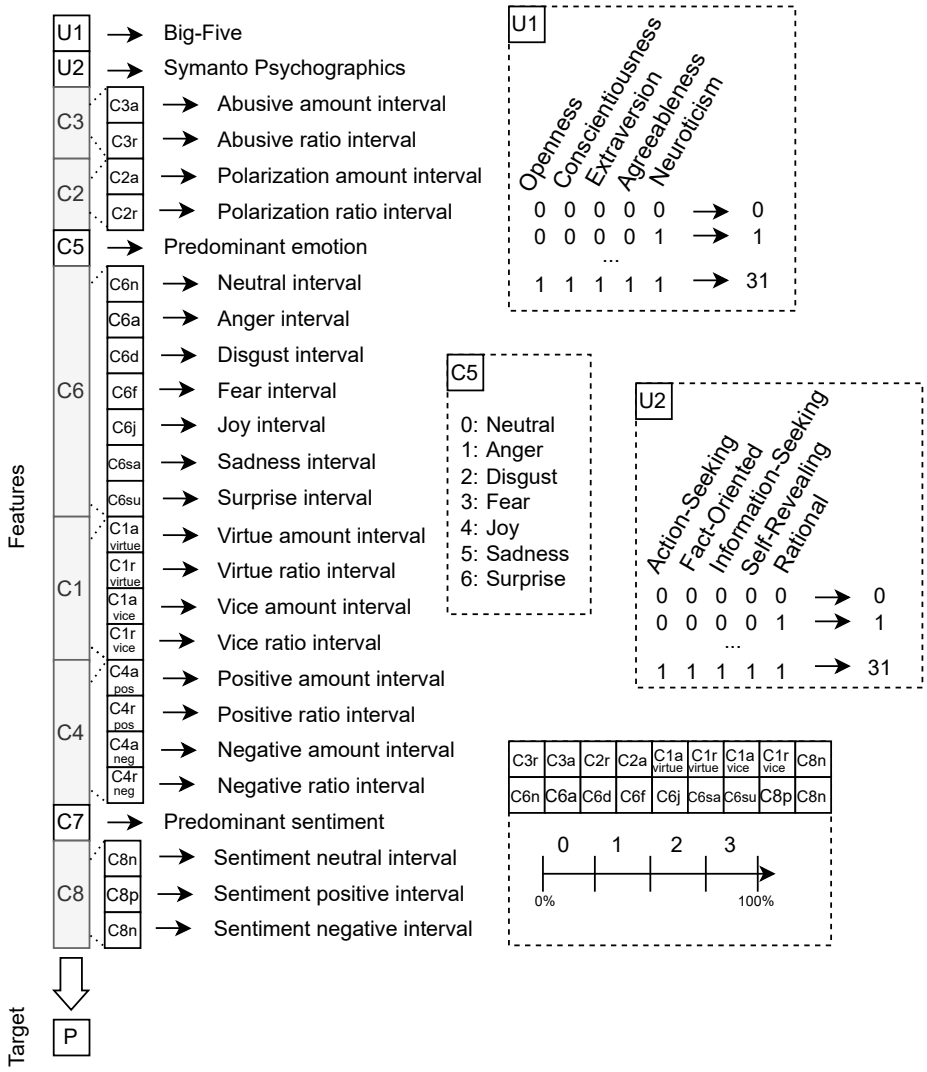

Figure 2. Summary of features in the context.

4.1. Using Classical Machine Learning

The first set of experiments is designed to predict the features described in Section 3.3 using classical machine learning models, exploring various approaches for calculating linguistic features (Section 3.1).

Dataset. The dataset for this study, related to climate change, is built from three files. The topic was determined by the dataset we had available, but our approach can be applied to any topic[4].

[4]Dataset can be made available upon request.

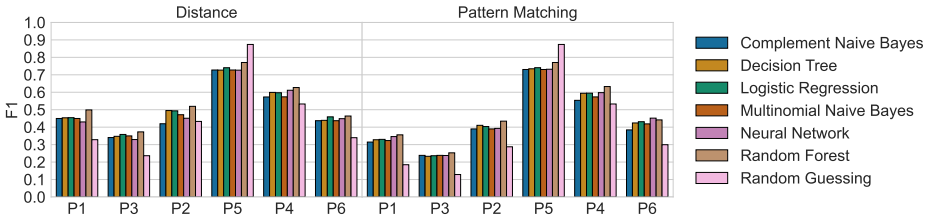

Figure 3. F1 Scores for the ML experiments, for both Distance and Pattern Matching setups.

- **Influencer's File**: 301 tweets by 9 influential figures on climate change.
- **Response's File**: 212,220 tweets from 106,383 unique users in response to the influencers, representing a subset of responses.
- **Profile's File**: 383,121 tweets from 425 users, not specifically related to climate change, providing a broader context.

To analyze user personalities and communication styles, we only keep users with at least 10 tweets, reducing the dataset to 2,309 users. Tweets are then grouped by user to calculate the relevant features. To address class imbalance, we used imbalanced-learn's RandomUnderSampler[5] to balance the binary classification tasks by randomly reducing the majority class. The dataset is split 90% for training and 10% for testing.

Word Counting. We used two approaches to calculate the linguistic features (Section 3.1): pattern matching and word distance using embeddings. Ratios are computed dividing the count of relevant words by the total tweet word count.

- **Pattern Matching**: We use the SnowballStemmer from NLTK [30] to preprocess the data, matching words in the tweet to those in a dictionary. The word count increases for each match.
- **Word Distance** (Embeddings): Tweets and dictionary words are tokenized and stop words, punctuation, and URLs are removed. We use FastText embeddings (crawl-300d-2M-subword) [31] to calculate cosine similarity between word vectors, setting a threshold of 0.6. If a word is within this threshold of a dictionary word, it counts towards the total.

Models. We trained several ML classifiers, fine-tuning hyperparameters using RandomizedSearchCV [32] to optimize performance. A detailed list of the models and their corresponding hyperparameters can be found in Appendix B.

Results. Binary classification outcomes were obtained following the methodology described in Section 3.3. Figure 3 shows F1 score results. To see the complete performance metrics—including precision and recall—along with F1 score differentials relative to a Random Guessing baseline, are detailed in Appendix C.

The F1 score serves as the main evaluation metric, although its behavior was not consistent with expectations. Notably, the **Predominant Sentiment** and **Emotion** prediction tasks yielded the highest F1 scores, although the emotion task performs poorly compared to the baseline. In contrast, the **Abusive** and

[5]Documentation available here

MFD tasks exhibited substantially lower F1 scores. Among the evaluated models, the **Random Forest** classifier achieved the best performance in the vast majority of the cases. Although the **distance-based** approach generally surpassed **pattern-matching** in absolute terms, the relative improvement over the Random Guessing baseline was greater for the pattern-matching method. Precision results follow similar trends, with **Predominant Emotion** and **Sentiment** showing the highest precision. **Abusive** and **MFD** predictions remain weak. Overall, precision values were low, indicating a high rate of false positives across models.

Recall results are more consistent across tasks, with **Abusive** and **MFD** tasks showing the greatest improvements over the baseline, while **Emotion** underperforms. Recall surpasses the baseline for most tasks, suggesting that the models capture many actual positive instances but misclassify neutral or positive content.

While the distance-based method offered marginal improvements in overall metrics, it may introduce undesired effects, particularly due to the loss of informative detail in embedding representations. This is especially problematic when working with clean or limited datasets. Furthermore, the study's reliance on a single dataset underscores the difficulty of acquiring optimal data and highlights the importance of developing or curating datasets with targeted characteristics. Finally, the obtained results made us also ponder the feasibility of the performing the proposed tasks automatically. A way to try to assess such feasibility would be to compare against human performance. However, given the lack of resources to perform such experiment, in the next section we decided to use LLMs as a simple way of understanding the difficulty of the proposed taks.

4.2. Using Large Language Models

To assess the feasibility of our prediction tasks, we designed and conducted a series of experiments using Large Language Models (LLMs), as follows:

Dataset. The same dataset as before, but split into training and test sets (80/20) to prevent label leakage, ensuring no conversation appears in both sets.

Linguistic Features. Word counting is based on pattern matching, applied to both context and predictions. The prediction labeling follows the methodology in Section 3.3, with LLM-specific adaptations:

- (P1)(P2) **Avg. Abusive, Avg. Polarization**: Binary features are converted to "High" if either the amount or ratio of relevant words is in the [50, 100] range; otherwise, "Low".
- (P3)(P6) **Predominant Moral Foundation Category, Predominant Valence**: We use word counts to identify the dominant category. If no clear majority exists, the label is "Equal".
- (P4)(P5) **Predominant Sentiment, Predominant Emotion**: Multiclass labels assigned based on the dominant category name, replacing numerical values.

Model. We use Llama-3.2-1B-Instruct[6], an instruction-tuned, multilingual, open-source LLM by Meta, released in September 2024. Its optimized transformer architecture is trained on a mix of public data. We disabled sampling

[6] https://huggingface.co/meta-llama/Llama-3.2-1B-Instruct

(do_sample=False) to reduce response randomness and focus on the most probable outputs. Llama was chosen for its privacy-friendly, cost-effective local deployment and suitability for instruction-following tasks.

Prompts. Prompts are based on the machine learning input vector (Figure 2), following recommended prompting practices [7]. To represent user personality and conversation dynamics effectively, we include the last five tweets of each conversation, leveraging findings from [11], which show improved performance with recent content. This aligns with cognitive insights from [14] and [15] regarding user attention and recency bias. Both zero-shot and few-shot (one example per class) prompt formats are tested. The complete prompt templates are provided in Appendix D.

Results. Evaluation of few-shot and zero-shot language model predictions across various classification tasks shows that providing examples in the prompt significantly improves performance, particularly for the **Abusive** and **Polarization** tasks. In the **zero-shot** setting, models often default to predicting a single class, leading to poor overall metrics. **Few-shot** prompting helps the model recognize and differentiate between multiple classes, resulting in improved F1, recall, and precision scores. **Emotion** and **Sentiment** classification benefit from few-shot prompting, though performance remains uneven—especially in Emotion, where certain categories like "Disgust" are rarely predicted due to class imbalance in the dataset. For Sentiment, few-shot prompting increases both correct and incorrect predictions of the "Negative" class, while "Positive" remains underrepresented across both approaches. **MFD** predictions improve with prompting, expanding from two to all three classes, though with trade-offs in class-level accuracy. **Valence** predictions show mixed results; while class coverage increases with few-shot prompting, accuracy for certain labels decreases. Table 1 presents the precision, recall, and F1 scores for both the zero-shot and few-shot experiments. Appendix D contains a complete table including the weighted metric values and the confusion matrices.

These findings highlight two key considerations. First, if the objective is to detect harmful behavior for timely intervention, models must minimize false positives—favoring few-shot approaches. Second, when the goal is to predict future user behavior (e.g., after an intervention), accurate prediction across all classes becomes critical, and current model performance remains insufficient. Additionally, the prompt design heavily influences model behavior, indicating a need for further exploration of prompt structure, format, and granularity. The dataset's limitations, particularly the lack of balanced class representation in Emotion tasks, further underscore the need for improved or augmented data.

5. Conclusions and Future Work

The results of our experiments show that building a model of an agent such that we can answer the question *Given past behavior and the current context, how do*

[7] https://aws.amazon.com/es/blogs/machine-learning/best-prompting-practices-for-using-meta-llama-3-with-amazon-sagemaker-jumpstart/

Table 1. LLM Experiment: Comparison of Zero Shot vs. Few Shot performance

		Zero Shot		Few Shot			Zero Shot		Few Shot	
	Task	**Micro**	**Macro**	**Micro**	**Macro**	**Task**	**Micro**	**Macro**	**Micro**	**Macro**
P		0.0	N/A	0.2350	N/A		0.3915	0.4377	0.4829	0.4251
R	Abus.	0.0	N/A	0.7483	N/A	Sent.	0.3915	0.3571	0.4829	0.3775
F1		0.0	N/A	0.3577	N/A		0.3915	0.2734	0.4829	0.3469
P		0.0	N/A	0.3841	N/A		0.2541	0.2912	0.4337	0.3410
R	Polar.	0.0	N/A	0.8187	N/A	MFD	0.2541	0.3283	0.4337	0.3395
F1		0.0	N/A	0.5228	N/A		0.2541	0.1611	0.4337	0.3283
P		0.0868	0.1235	0.2672	0.1471		0.3558	0.2410	0.3439	0.3628
R	Emo.	0.0868	0.1365	0.2672	0.1636	Val.	0.3558	0.3727	0.3439	0.3666
F1		0.0868	0.0300	0.2672	0.0969		0.3558	0.2872	0.3439	0.2952

we expect a given user to respond?, is a difficult task, especially when the goal is to obtain as the agent's output something as fine-grained as linguistic cues. Given this preliminary evaluation, we cannot yet obtain a definitive answer to our research question, since our results are mixed. Therefore, we plan to develop a more general generative agent approach, generating the response as text and analyzing linguistic features afterwards. Additional information such as considering the weight of each tweet, whether it is by relationship between the users, or the recency of the tweets, could be tried in order to better guide the learning process.

Acknowledgments. Partially funded by MCIN/AEI (CHIST-ERA iTrust) – PCI2022-135010-2, PID2022-139835NB-C21, PIE 2023-5AT010 CSIC, Universidad Nacional del Sur (UNS) – PGI 24/ZN057, and ANPCyT – PICT-2018-0475 (PRH-2014-0007). The authors thank Cristian Cozar for his help with tasks involving setting up this line of research.

References

[1] Elsner M, Atkinson G, Zahidi S. The Global Risks Report 2025; 2025. Available from: `https://www.weforum.org/publications/global-risks-report-2025/`.

[2] Myers DG, Lamm H. The group polarization phenomenon. Psychological bulletin. 1976;83(4):602.

[3] Hogg MA, Turner JC, Davidson B. Polarized norms and social frames of reference: A test of the self-categorization theory of group polarization. Basic and Applied Social Psychology. 1990;11(1):77-100.

[4] Harel TO, Maoz I, Halperin E. A conflict within a conflict: intragroup ideological polarization and intergroup intractable conflict. Current Opinion in Behavioral Sciences. 2020;34:52-7.

[5] Demszky D, Garg N, Voigt R, Zou J, Gentzkow M, Shapiro J, et al. Analyzing Polarization in Social Media: Method and Application to Tweets on 21 Mass Shootings. In: Proceedings of NAACL-HLT; 2019. p. 2970-3005.

[6] Diaz GA, Urribarri D, Ganuza ML, Chesñevar C, Estevez E, Maguitman A. Polviz: Assessing Opinion Polarization in Social Media through Visual Analytics and Argumentation. In: Proc. ICTPEG; 2024. p. 337-47.

[7] Isenberg DJ. Group polarization: A critical review and meta-analysis. Journal of personality and social psychology. 1986;50(6):1141.

[8] Sunstein CR. The law of group polarization. University of Chicago Law School, John M Olin Law & Economics Working Paper. 1999;(91).

[9] Bliuc AM, Bouguettaya A, Felise KD. Online intergroup polarization across political fault lines: An integrative review. Frontiers in Psychology. 2021;12:641215.

[10] Tyagi A, Uyheng J, Carley KM. Affective polarization in online climate change discourse on Twitter. In: Proc. ASONAM. IEEE; 2020. p. 443-7.
[11] Gallo FR, Simari GI, Martínez VM, Falappa MA. Predicting user reactions to Twitter feed content based on personality type and social cues. FGCS. 2020;110:918-30.
[12] Gajewska E, Budzyńska K. Digital polarisation: Analysis of 'us' versus 'them' rhetoric in public discussions on social media. In: Proc. PP-RAI; 2024. .
[13] Brugière A, Nguyen-Ngoc D, Drogoul A. Handling Multiple Levels in Agent-Based Models of Complex Socio-Environmental Systems: A Comprehensive Review. Frontiers in Applied Mathematics and Statistics. 2022;8:1020353.
[14] Hogg T, Lerman K, Smith LM. Stochastic models predict user behavior in social media. arXiv preprint arXiv:13082705. 2013.
[15] Kang JH, Lerman K, Getoor L. LA-LDA: a limited attention topic model for social recommendation. In: Proc. SBP-BRiMS. Springer; 2013. p. 211-20.
[16] Gao H, Mahmud J, Chen J, Nichols J, Zhou M. Modeling user attitude toward controversial topics in online social media. In: Proc. ICWSM. vol. 8; 2014. p. 121-30.
[17] Lahuerta-Otero E, Cordero-Gutiérrez R. Looking for the perfect tweet. The use of data mining techniques to find influencers on twitter. Computers in human behavior. 2016;64:575-83.
[18] Rodermund SC, Lorig F, Berndt JO, Timm IJ. An agent architecture for simulating communication dynamics in social media. In: Proc. MATES. Springer; 2017. p. 19-37.
[19] Yang P, Yang G, Liu J, Qi J, Yang Y, Wang X, et al. DUAPM: An effective dynamic micro-blogging user activity prediction model towards cyber-physical-social systems. IEEE Transactions on Industrial Informatics. 2019;16(8):5317-26.
[20] Pérez JM, Luque FM, Zayat D, Kondratzky M, Moro A, Serrati PS, et al. Assessing the impact of contextual information in hate speech detection. IEEE Access. 2023;11:30575-90.
[21] Singh R, Subramani S, Du J, Zhang Y, Wang H, Miao Y, et al. Antisocial Behavior Identification from Twitter Feeds Using Traditional Machine Learning Algorithms and Deep Learning. EAI Endorsed Transactions on Scalable Information Systems. 2023;10(4).
[22] Firdaus SN, Ding C, Sadeghian A. Retweet prediction based on topic, emotion and personality. Online Social Networks and Media. 2021;25:100165.
[23] García-Díaz JA, Colomo-Palacios R, Valencia-García R. Psychographic traits identification based on political ideology: An author analysis study on spanish politicians' tweets posted in 2020. FGCS. 2022;130:59-74.
[24] Gajewska E, Budzynska K, Konat B, Koszowy M, Kiljan K, Uberna M, et al. Ethos and Pathos in Online Group Discussions: Corpora for Polarisation Issues in Social Media. arXiv preprint arXiv:240404889. 2024.
[25] Simchon A, Brady WJ, Van Bavel JJ. Troll and divide: the language of online polarization. PNAS Nexus. 2022;1(1):pgac019.
[26] Wiegand M, Ruppenhofer J, Schmidt A, Greenberg C. Inducing a Lexicon of Abusive Words – A Feature-Based Approach. In: Proc. ACL: Human Language Technologies. Berlin: Association for Computational Linguistics; 2018. p. 1046-56.
[27] Warriner AB, Kuperman V, Brysbaert M. Norms of valence, arousal, and dominance for 13,915 English lemmas. BEHAVIOR RESEARCH METHODS. 2013;45(4):1191-207.
[28] Ekman P. Expression and the nature of emotion. Approaches to emotion. 1984;3(19):344.
[29] Pérez JM, Rajngewerc M, Giudici JC, Furman DA, Luque F, Alemany LA, et al.. pysentimiento: A Python Toolkit for Opinion Mining and Social NLP tasks; 2024. Available from: `https://arxiv.org/abs/2106.09462`.
[30] Bird S, Loper E, Klein E. Natural Language Processing with Python. O'Reilly Media Inc.; 2009.
[31] Mikolov T, Grave E, Bojanowski P, Puhrsch C, Joulin A. Advances in Pre-Training Distributed Word Representations. In: Proc. LREC; 2018. .
[32] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research. 2011;12:2825-30.