Harnessing Personalization Methods to Identify and Predict Unreliable Information Spreader Behavior

Shaina Ashraf, Fabio Gruschka, Lucie Flek, Charles Welch
Conversational AI and Social Analytics (CAISA) Lab, University of Bonn
{sashraf,flek,cfwelch}@bit.uni-bonn.de

Abstract

Studies on detecting and understanding the spread of unreliable news on social media have identified key characteristic differences between reliable and unreliable posts. These differences in language use also vary in expression across individuals, making it important to consider personal factors in unreliable news detection. The application of personalization methods for this has been made possible by recent publication of datasets with user histories, though this area is still largely unexplored. In this paper we present approaches to represent social media users in order to improve performance on three tasks: (1) classification of unreliable news posts, (2) classification of unreliable news spreaders, and, (3) prediction of the spread of unreliable news. We compare the User2Vec method from previous work to two other approaches; a learnable user embedding layer trained with the downstream task, and a representation derived from an authorship attribution classifier. We demonstrate that the implemented strategies substantially improve classification performance over state-of-the-art and provide initial results on the task of unreliable news prediction.

1 Introduction

The distribution of information and news over the internet has enabled the uncontrolled spread of unreliable news and calls for the development of new social norms of careful information evaluation and sharing. Algorithms decide the newsfeed for their users and the widespread propagation of unreliable news has led to the need of automated means of detecting such information. Much research has addressed this issue with a variety of corpora containing different types of unreliable news, however few corpora exist which contain a longitudinal component of the individuals who spread unreliable news.

Studies have analyzed the language used when unreliable news is spread, finding differences in social and self-referencing words, denial, complaints, generalizing terms, lower cognitive complexity, less exclusive words, and more negative emotion and action words (Sharma et al., 2019; de Oliveira et al., 2021). Naturally, the way these expressions are formed varies across individuals, making it important to model users to improve detection. Initial work has begun to apply such methods, though the application of personalization methods for this task is still largely unexplored (Sakketou et al., 2022; Mu and Aletras, 2020).

In this work, we show that unreliable news can be more accurately detected when using personalization. Personalization has different meanings across literature in natural language processing (Flek, 2020) but in this work it refers to the process of building personalized representations of users in order to better model their behaviors. Our contributions are (1) state-of-the-art results on the FACTOID and Twitter datasets for detecting unreliable news spreaders by improving user embeddings, (2) an exploration of the task of predicting when unreliable news will be spread, showing improvements over the best model from previous work, and (3) a comparison of the performance of recent personalization methods for both tasks.

2 Related Work

Previous work uses neural methods to combine text-based features, such as those from statements related to news data Karimi et al. (2018). Liu and Wu (2018) use RNN and CNN-based methods to build propagation paths for detecting misinformation at the early stages of propagation. Shu et al. (2019) propose a tri-relationship embedding framework to model relationships among publishers, news stories, and social media users for fake news detection. Karadzhov et al. (2017) introduced a framework for fully-automatic fact checking using external

Proceedings of the The 8th Workshop on Online Abuse and Harms (WOAH), pages 146–158
June 20, 2024 ©2024 Association for Computational Linguistics

sources. They use a deep neural network with LSTM text encoding, semantic kernels and task-specific embeddings that are combined to encode a claim together with portions of possibly relevant text from the web. Cui et al. (2019) propose an explainable fake news detection system, DEFEND, which considers users' comments to explain if news is fake or real. Nguyen et al. (2020) propose a fake news detection method that uses a graph learning framework to represent social contexts. Ghanem et al. (2021) propose FakeFlow model, to enhance fake news detection by analyzing the flow of affective information, such as emotions, sentiment, and hyperbolic language, within texts. By segmenting input texts into smaller units, FakeFlow effectively models the interactions between topical and affective terms, thereby improving its ability to identify fake news articles. Duan et al. (2020) extracted linguistic and sentiment features from users' tweet. Also the presence of emojis, hashtags and political bias has been taken into account for prediction. (Khilji et al., 2023) captured contextual information of user by exploring personalization methods based on user metadata and credibility features for debunking misinformation

Researchers are also examining cognitive factors influencing people's ability to distinguish fake news (Pennycook and Rand, 2019). Data-driven studies analyzing bots' participation in social media discussion (Howard and Kollanyi, 2016), user reactions to reliable/unreliable news posts (Glenski et al., 2018a), and demographic characteristics of users propagating unreliable news sources (Glenski et al., 2018b), are also integral to our understanding of the problem space.

In the exploration of penalization techniques for the identification and prediction of misinformation spreaders, the work of (Plepi et al., 2023; Plepi and Flek, 2021) presents the importance of incorporating user-specific context alongside conversation text and have achieved significant results in both their sarcasm detection and perception classification tasks. (Salemi et al., 2023) also showcases the significant benefits of integration personalization techniques into large language models through extensive experimentation, including zero-shot and fine-tuned setups. Similarly, Lian et al. (2022) proposes an innovative incremental user embedding model that dynamically integrates recent user interactions into accumulated history vectors, utilizing a transformer encoder for personalized text classification.

Sakketou et al. (2022) introduced the misinformation spreader dataset, FACTOID, that captures long-term context of users' historical posts. They provide initial findings on the dataset, which serve as a baseline for our experiments. The user histories allow us to address a new temporal task of predicting when someone will spread misinformation. These histories are categorized across several contentious topics, offering a comprehensive view of misinformation spread on Reddit. These categories include general political debate, SARS-CoV-2 (COVID-19), gender rights, climate change, vaccinations, abortion, gun rights, and debates about 5G technology. Each category encapsulates discussions from multiple subreddits, encompassing a variety of stances and biases. The dataset's breadth across these topics allows for a broader understanding of misinformation trends and the development of strategies to anticipate.

Mu and Aletras (2020) predict, using only language information, whether a social media user will propagate news items from unreliable or reliable sources before they share any news items. Unreliable users have a history of sharing content from unreliable sources at least three times, while reliable users only share content from trustworthy sources. They define a binary classification task and train a machine learning model on a dataset of user histories leading up to their first news repost, labeled as either reliable or unreliable. Comparatively, our study expands on this approach. While they use data up until the first news item is shared, our work includes news items within a user's history. We compare their best performing method to ours, as described in §3.4.

3 Methodology

In this section, we discuss the approaches for the different setups for personalized representations in our work. We use static word representations from GloVe pretrained on the respective dataset as input for the most of our methods. To facilitate comparisons with previous work, we also explored Word2Vec representations that were pretrained using both datasets. This allowed us to investigate whether our results benefit from leveraging global word-word co-occurrence statistics and the linear substructures within the word vector space. With these word representations we are able to learn personalized user embeddings. We further discuss the task setup and definitions.

3.1 Definitions

In the Twitter dataset, users are classified as *reliable* or *unreliable* based on their sharing habits. Mu and Aletras (2020) define unreliable sources to be propaganda, clickbait, conspiracy theories, or satire. In the FACTOID dataset, misinformation is defined to encompass various forms of politically oriented false or misleading news. This includes unintentionally misleading news, deliberately deceptive disinformation, politically skewed hyperpartisan news, and humorously false satirical news (Sakketou et al., 2022).

Ruffo et al. (2023) provide a detailed description and taxonomy of information types. The two datasets we study both cover misinformation, disinformation, as the news may be intentionally or unintentionally spread, as well as malinformation, which includes things like propaganda and is spread with a malicious intent. We adopt the term *unreliable* to refer to these types of information propagated by online users.

3.2 Task Definitions

We address three tasks, the first of which classifies users, and two that classify individual posts, as visualized in Figure 1.

Unreliable News Spreader Detection We classify if a given user is a spreader of unreliable news or not. Each user u^i is associated with a posting history H^i, as in (Sakketou et al., 2022).

Unreliable News Post Classification For the classification of unreliable news posts, we want to predict $y^i_j \in \{\text{unreliable}, \text{information}\}$ with the pretrained embeddings $\mathcal{E}^i_j$ and the post history.

Unreliable News Post Prediction For the prediction of unreliable news posts, we want to predict $y^i_j \in \{\text{unreliable}, \text{information}\}$ only with the pretrained or task embeddings $\mathcal{E}^i_j$.

3.3 Splitting User Data

When we are classifying users as unreliable news spreaders, we use all data for that user, as in previous work. However, when we are classifying posts, we need to use only posts that precede a post that we want to classify. To do this, we split users into *artificial users* at points in time delimited by the number of preceding posts and experiment with different limits to the number of preceding posts.

We partition the post history of each user u^i into chunks of size X and create an artificial user a^i_j for each chunk. The j-th artificial user for real user i is defined as $a^i_j \in \mathcal{A} = \{a^1_1, \ldots, a^1_{M_1}, \ldots, a^N_1, \ldots, a^N_{M_N}\}$ where $M_i = \lceil \frac{L^i}{X} \rceil$ represents the number of artificial users created, and each user u^i, with a length of post history denoted by L, is split into segments of size X.

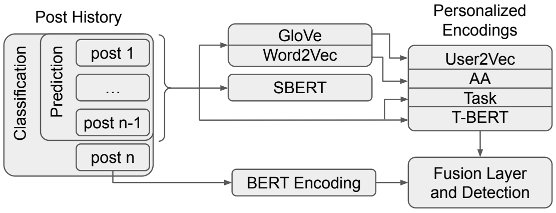

Figure 1: Visualization of task setup for prediction and classification tasks. The fusion and detection box represents a final layer of our neural model, which assigns a label corresponding to the task type.

For each post history chunk, h^i_j, we take the first $X - 1$ posts and reserve the label of the X-th post as classification target. After that we drop all a^i_j with $|h^i_j| < 20$ to compute the initial user representation for $\mathcal{E}^i_j$ based on their corresponding historical posts.

3.4 User Representations

User2Vec Amir et al. (2016) presented User2Vec, which computes user embeddings from a corpus of their text. For the unreliable news spreader approach we calculate the embeddings $\mathcal{E}^i \in R^d$ of user u^i based on their corresponding historical posts $\mathcal{H}^i$. Computing the embeddings $\mathcal{E}^i_j$ requires pretrained word embeddings, which we compute both with word2vec and GloVe (Pennington et al., 2014; Mikolov et al., 2013).

Task Embeddings This approach uses an embedding layer initialized with Xavier initialization (Glorot and Bengio, 2010), which takes in a user ID and converts it into a vector representation in the forward pass. It is updated during training, so it is expected to encode signals of misinformation spreaders.

Authorship Attribution Much previous work has addressed authorship attribution (AA), the task of classifying, from a predetermined set of authors, which author wrote a given text (Stamatatos, 2009). Recent personalization work has looked into deriving user representations from authorship attribution classifiers (Plepi et al., 2022a; Welch et al., 2022). We use SBERT to encode all posts (Reimers and Gurevych, 2019) and use the resulting vectors for classification by passing them through a feed-forward layer with input size 768. We calculate performance on the validation set with the embeddings before the classification layer ($d = 400$) for each post for a user and average these to get the resulting AA embedding. This is in contrast to previously mentioned methods that use the distribution of predictions or probabilities, which have a dimension size equal to the number of users. This model achieves an accuracy of 1.5% which is 170x better than chance for the FACTOID dataset and 0.5% on the Twitter dataset (175x better than chance).

Combined We perform ablations using each combination of two of the above methods, and for using all three at the same time.

T-BERT Mu and Aletras (2020) presented a truncated version of the BERT (T-BERT) which takes initial 512 words pieces from the text of each user as input. We also followed the same approach in all three of our tasks. For post classification and prediction tasks, we computed user contextualized T-BERT embeddings by taking the recent 512 tokens from each user and concatenate them with each post before passing to model.

4 Datasets

Our study leverages two pre-existing datasets, FACTOID (Sakketou et al., 2022) and a Twitter dataset (Mu and Aletras, 2020). Initially, we considered other datasets, including CMU-MisCov19 (Memon and Carley, 2020), and data from the PAN shared tasks (Rangel et al., 2020), however they were not suitable for our experimentation as they only provide Tweet IDs or labels for authors not for tweets and some have missing information for users, lacking content for the user personalization techniques.

FACTOID consists of 4,150 users with 3.4M posts. We use the balanced user split from their paper, which consists of 1,086 unreliable news spreaders and an equal amount of real information spreaders for 2,172 in total. A user is annotated as a unreliable news spreader if they have at least two posts with unreliable news links. We split the data into train/test to balance the number of spreaders.

We consider posts unreliable news if they have one or more unreliable news links. When splitting to create artificial users as described in §3.3, we vary the number of context posts, using 50, 100, and 200 posts per user, resulting in 12.8k, 12.5k, and 11.6k artificial users respectively. We then balance the post-level data to have an equal number of real and unreliable news posts, resulting in 19,654 total. Posts contain 119 tokens on average (σ=206). Other datasets designed for identifying unreliable news spreaders only include binary labels for the user-level. To obtain pretrained embeddings with unsupervised learning algorithms we use data from users history, most of which is unlabeled (see Table 1).

Twitter provides all necessary information including user labels and IDs, which enabled us to recompile the posting history of each user. Unfortunately, not all tweets were available for us to crawl, resulting in only 3.5K users whereas the original dataset had 6.2K users. The dataset has 2.6M posts, with an approximate distribution of 40:60 between users circulating unreliable news and other information sharers. Posts contain 25 tokens on average (σ=18). The corpus was recrawled in Plepi et al. (2022b) and further details on collection can be found in their paper. Given that this dataset indicated negligible social interaction among its users, our focus was predominantly on the personalization techniques (rather than the temporal graphs they explored).Users who shared at least three unreliable links were labeled as misinformation spreaders. Note that this is different from the FACTOID dataset, as we wanted to be consistent with both original works.

	FACTOID	Twitter
Total Posts	3,354,450	2,626,176
Total Users	4,150	3,541
Unreliable Spreaders	1,086	1,455
Reliable Spreaders	3,064	2,086
Unreliable Posts	9,835	1,521,415
Reliable Posts	70,168	1,104,761

Table 1: Comparison of datasets and label distributions.

5 Experiments

To evaluate the performance of the unreliable news spreader detection models, we use 5-fold cross validation, for consistency with previous work. We compare the proposed personalized embeddings with several previous models for the unreliable news detection methods. For post-level tasks we show results after 10 iterations with 20 epochs each and learning rate of $1e - 5$. For post-level tasks we encode posts with BERT (Devlin et al., 2019)

Model		F1 Score	
Sakketou et al. (2022)		0.61	
T-BERT		0.58	
T-BERT+U2V-GloVe		0.59	
	U2V-GloVe	**U2V-W2V**	**AA**
RF	0.71	0.60	0.74
Ridge	0.73	0.67	0.67
LR	0.71	0.63	0.64
SVM	**0.75**	0.63	0.69

Table 2: Unreliable News Spreader Detection results on the balanced FACTOID dataset using the logistic regression (LR), ridge regression (Ridge), support vector machine (SVM) and random forest (RF) classifiers compared to previous work and our combined model. Reported values are the F_1- scores over a 5-fold Cross Validation. Bold denotes the best overall performance on the task.

before concatenating user representations. We compare to a *Random* method, which is a model with a random vector as input and concatenated to BERT. We also compare to the best model from Mu and Aletras (2020), T-BERT. We did not compare to the graph-based methods used in Sakketou et al. (2022). They found that the graph-based method on Reddit achieved 0.3% higher F1 than the User2Vec random forest method. We find that the construction of the Reddit graph also is unlikely to signify interaction between users as many users reply to posts without responding to other comments and without knowing other users. Due to these reasons and the high model complexity of the graph attention network, we did not use this model for our tasks.

5.1 Setup & Parameters

To obtain User2Vec features we use the parameters mentioned in Amir et al. (2016). For the vector size parameter we adjust GloVe and Word2Vec to the same dimension $d = 400$ based on manual tuning.

5.2 Results

For comparison with previous work, we provide results for the unreliable news spreader detection task in a similar format and using mostly the same classifiers as previous work. For results at the post-level we report results as a distribution over 10 runs.

Unreliable News Spreader Detection The results for the unreliable news spreader detection on the Factoid and Twitter datasets are shown in Table 2 and Table 3 respectively. In Table 2, the best model from Sakketou et al. (2022) is our baseline at 0.61 F1, which uses a User2Vec (U2V) model trained on the Google News Corpus using word2vec (W2V). We compared this setup to one where the word embeddings are pretrained on in-domain data using their corpus with both word2vec (U2V-W2V) and GloVe (U2V-GloVe). Note that the User2Vec method is initialized with static embeddings only so contextualized embeddings from large pretrained language models are incompatible with this approach. We used the same classic machine learning classifiers (i.e. random forest, logistic regression, support vector machines) for the sake of comparison. We also compared to the best performing method from (Mu and Aletras, 2020) (T-BERT).

Model		F1 Score
T-BERT		0.51
T-BERT+U2V-GloVe		0.65
	U2V-GloVe	**AA**
RF	0.62	0.70
Ridge	0.70	0.76
LR	0.75	**0.82**
SVM	0.70	0.76

Table 3: Unreliable News Spreader Detection results on the balanced Twitter dataset using the logistic regression (LR), ridge regression (Ridge), support vector machine (SVM) and random forest (RF) classifiers compared to previous work and our combined model. Reported values are the F_1- scores over a 5-fold Cross Validation. Bold denotes the best overall performance on the task.

We included one more model based on T-BERT but with the U2V-GloVe vectors concatenated to the input before being passed to a final classification layer. We found that this improved performance on the FACTOID dataset, but only slightly over the T-BERT baseline. Simpler classification models with high quality user embeddings learned through the authorship attribution and User2Vec methods outperformed the language model approach, which we attribute to their training method, which takes all of a users previous data into account when learning a representative vector, whereas BERT can only encode a limited history.

In Table 3, the results are evaluated on the Twitter data by following the same models and embedding methods used in the FACTOID dataset to assess their performance in detecting unreliable news spreaders. Here, we did not include the word2vec approaches, as they performed poorly

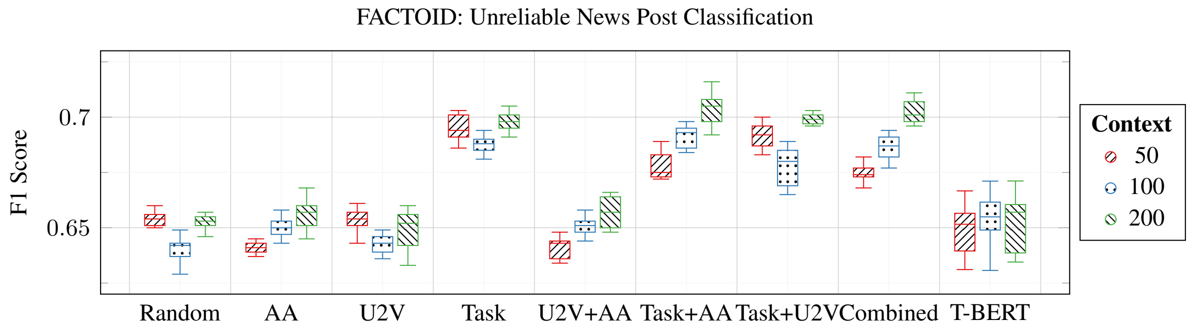

Figure 2: Distributions of F_1-scores for personalization methods and combinations while varying the number of context posts (p) or tokens (t) for the task of classifying unreliable news posts.

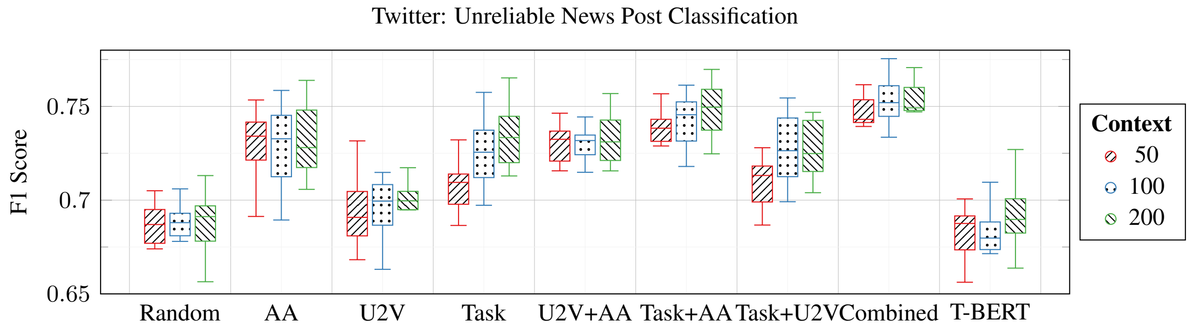

Figure 3: Distributions of F_1-scores on the Twitter dataset for personalization methods and combinations while varying the number of context posts for the task of classifying unreliable news posts.

on the other task compared to GloVe (which includes Sakketou et al. (2022)). Interestingly, the highest performance with 82% F_1 is achieved by the model trained on authorship attribution embeddings. Here the T-BERT with U2V-GloVe embeddings performed much higher than the T-BERT baseline, but still lower than the best U2V-GloVe and authorship attribution embedding approaches. For further experiments with the commonly used LIWC features, see Appendix A. Note that we do not compare to the task embedding method because it requires data from a user for both training and testing, while this task setup has separate users across the splits.

Unreliable News Post Classification Figure 2 shows the F_1 measure for the unreliable news detection task using FACTOID Dataset. Task embeddings in combination with the pretrained authorship attribution features achieve the best results with a median F_1 score of 72%. The worst score is obtained by the User2Vec approach with 65%. If we compare the different input sizes, the AA features benefit from having more data to train on. Other approaches considered individually seem not to learn better features with higher input sizes. The combinations follow this trend from the AA embeddings. The combination of all three seems negatively impacted by User2Vec. However, the influence is not statistically significant (Kruskal and Wallis, 1952).

Similarly, Figure 3 shows the results for unreliable news detection on Twitter. The combined approach using all user representations had the best performance with a median F_1 score 75%. It is interesting to note that, all approaches appear to learn better features with fewer users and bigger message chunks. Contrary to the FACTOID dataset, the authorship attribution approach performs better, as it did for the unreliable news spreader task, than the User2Vec embeddings. T-BERT performs relatively low on this task and not much higher than our random baseline. We believe that the lack of reproducibility of Twitter datasets in general could lead to such discrepancies.

Unreliable News Post Prediction Figure 4 shows results for unreliable news prediction for the FACTOID dataset. In this comparison, we see that authorship attribution features lose up to 16% F_1 with fewer users and more potentially irrelevant context. With a smaller context of 50, the difference is lower by 6% than in the classification task.

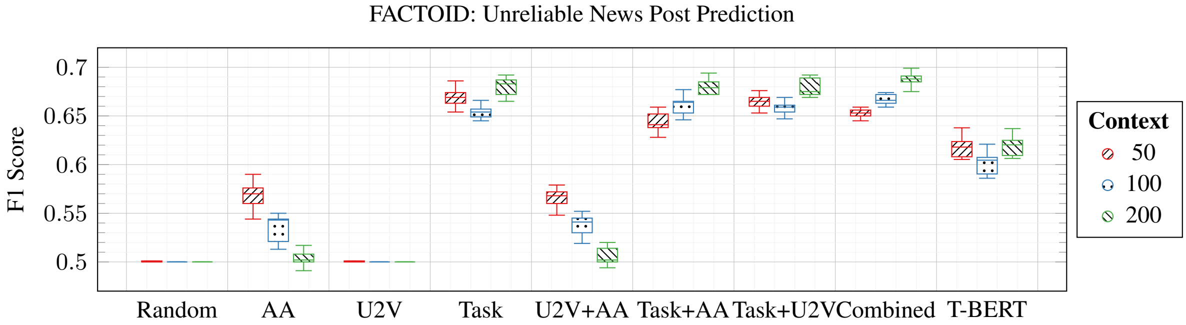

Figure 4: Distributions of F_1-scores for personalization methods and combinations while varying the number of context posts (p) or tokens (t) for the task of predicting unreliable news posts.

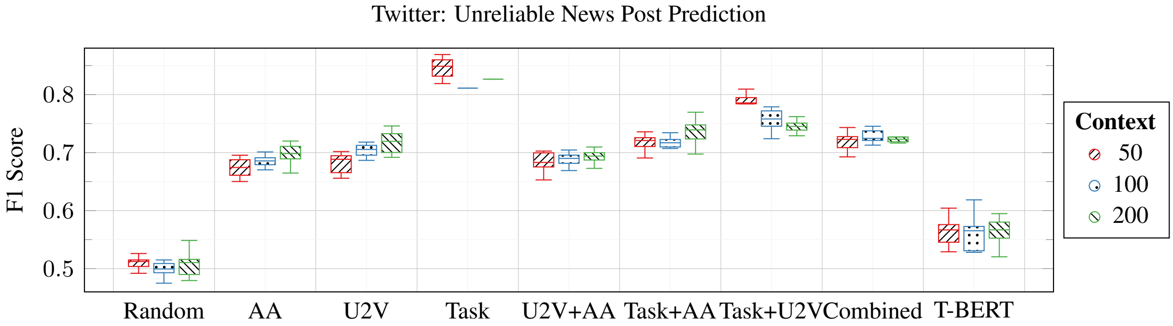

Figure 5: Distributions of F_1-scores on the Twitter dataset for personalization methods and combinations while varying the number of context posts for the task of predicting unreliable news posts.

User2Vec performs similar to chance and task embeddings remain high performing, not differing in the median ($p < 0.0003$). Combinations of personalization methods show similarly high performance. Here T-BERT shows competitive performance but still underperforms all of our methods that use task embeddings.

Similarly, Figure 5 displays the results of the unreliable news prediction task using the Twitter dataset. Although these methods rely only on user embeddings and omit post text, we can observe that the model is still learning high quality representations as the results are encouraging. The best score is obtained by task embeddings with median F_1 85%, combining task and User2Vec embeddings perform second best. We see competitive performance from the User2Vec embeddings whereas they performed randomly on the FACTOID dataset. The truncated BERT encodings caused the model to perform poorly, likely due to the fact that it does not seem to capture enough context for the prediction task. Interestingly, T-BERT performs better for the FACTOID dataset, and all of our methods outperform it on the Twitter dataset, leading to a new state-of-the-art for this task.

Linguistic Analysis In addition to our primary focus on comparing results of user personalization methods across two datasets, we explored linguistic characteristics of the spreaders' posts. Specifically, we looked at sentiment scores, which provide an indication of the emotional tone expressed in the content. These sentiment scores were computed using VADER (Valence Aware Dictionary and sEntiment Reasoner, Hutto and Gilbert (2014)), a lexicon and rule-based sentiment analysis tool specifically designed to gauge sentiments expressed in social media. Our analysis revealed that unreliable news spreaders exhibit significantly different sentiment scores compared to reliable news spreaders. We tested this observation using a two-sample t-test, which yielded a $p < 0.0001$. This provides strong statistical support for our observation: unreliable news spreaders indeed have a significantly different sentiment score than reliable news spreaders. Interestingly, our analysis also identified a negative correlation of -0.11 between the number of unreliable news posts and sentiment score. This suggests that as individuals disseminate more unreliable news, their sentiment score decreases, implying a less positive linguistic style among unreliable news spreaders as they become more active in the propagation of unreliable news. By examining selected instances, we observed a consistent pattern. Sentiment scores experienced a downward shift as individuals approached the posting of a unreliable news item.

We also looked at the correlation between the labels in the FACTOID dataset and the LIWC categories, similarly to Mu and Aletras (2020). However, we did not find significant correlations between the groups. On the Twitter dataset, they found correlations, for instance, between the use of power and analyitic words with unreliable news spreaders, and informal and netspeak language with reliable news spreaders. These differences could be due to the difference in writing styles between Reddit and Twitter users.

6 Discussion

The task of unreliable news post prediction could provide insight into the patterns of users who spread unreliable news which could help inform the design of social media policies or interventions to prevent such cases. We compared to the best method from previous work, T-BERT, which we found competitive with the embedding combinations for post prediction on the Twitter dataset but with lower scores for post classification and prediction on the FACTOID dataset. When a higher number of context posts were available, the embedding methods more consistently outperformed T-BERT. On the spreader detection task, we found that when we had high-quality user representations derived from other deep learning models, simple classifiers were able to achieve higher performance than the T-BERT baselines, which may introduce more noise and complexity than necessary.

Our results indicated that embedding performance varied depending on the dataset and the specific task at hand. For instance, User2Vec excelled at capturing long-term behavioral patterns, making it particularly effective for tasks where a user's historical behavior is a key factor. However, it may not have been as adept at capturing the nuances of individual posts or the specific contexts in which they were made. Authorship attribution focused on the unique linguistic style of users, making it effective for identifying unreliable news spreaders who have a consistent writing style, but less so for those who vary their writing style. These embeddings were particularly useful in post-classification, where they were concatenated with text to provide a more comprehensive representation. Task embeddings were updated during training, allowing them to adapt to the unique challenges posed by unreliable news detection. This adaptability was a key reason why they often outperformed other methods in our experiments. On the other hand, the combination of all user representations (U2V+AA+Task) showed the best performance on the Twitter dataset, suggesting that a multifaceted approach that leverages various aspects of user behavior and post characteristics can provide a more robust solution for unreliable news detection.

In summary, the effectiveness of each user representation strategy is highly dependent on the specific challenges posed by the task of unreliable news detection and the nature of the dataset. There's no one size fits all solution, and the optimal strategy may involve a combination of different user representations to capture the multifaceted nature of user behavior and unreliable news spread.

In a linguistic analysis, we identified that unreliable news spreaders tend to exhibit distinct sentiment scores that decrease as they circulate more unreliable news. However, no significant correlations were observed between LIWC categories and reliable/unreliable news spreaders as was found in previous work.

7 Conclusions

In this work, we systematically studied the application of recent personalization methods to three distinct yet interrelated tasks. These tasks included user-level detection of unreliable news spreaders, post-level classification of unreliable news, and predicting when unreliable news will be spread.

We found significant improvements in the task of detecting unreliable news spreaders at a user level when applying User2Vec embeddings learned with GloVe pretrained on in-domain data. This result indicates that a closer alignment with the domain of the data yields superior performance in identifying unreliable news agents. Moreover, for post-level tasks such as classifying unreliable news and predicting its propagation, we discovered that task embeddings learned jointly with the downstream task outperformed other personalization methods and previous work. Furthermore, our findings suggest that combining different personalization methods can further boost performance.

In addition to these primary findings, our exploration into linguistic characteristics yielded intriguing insights. We observed a significant difference in sentiment scores between unreliable news spreaders and reliable news spreaders, with unreliable news spreaders exhibiting a less emotive linguistic style. We also noticed a negative correlation between the number of unreliable news posts and sentiment scores, indicating a decline in sentiment as the frequency of these posts increased.

Future work could explore the integration of our approach with other forms of analysis, such as network analysis or more nuanced linguistic analysis, for a more comprehensive understanding of unreliable news dynamics. We release our code [1] and data split to facilitate further research in this vital field and support shared scientific goals.

Limitations

Previous work from Sheikh Ali et al. (2022); Sakketou et al. (2022) characterizes a user as a unreliable news spreader based on whether at least two unreliable news links were detected in their post history, while Mu and Aletras (2020) requires at least three posts. If we look inside the results of our model, it seems to classify users as unreliable news spreaders if at least one unreliable news link was detected. For example this post of a randomly selected user:

"`https://www.dailymail.co.uk/news/article-4364984/Ivanka-Trump-hit-claim-ripping-designs.html` is well in keeping of the Trump family trend of stealing ideas and claiming them as one's own."

This post contains an unreliable news link.[2] This user was classified as an unreliable news spreader but according to the definition of an unreliable news spreader, they are a reliable news spreader. Which leads to the question how many times a user should post about unreliable news in order to be considered as a unreliable news spreader? Although this threshold of two unreliable news posts is somewhat arbitrary and should be adjusted for the desired application, it serves to show the effectiveness of our approach.

Our methods look at the text of posts being shared on social media. The links shared by individuals contain additional multi-modal information. Often these links contain images or video. Our model does not take the link content into account and future work could improve model performance by modeling this information.

The datasets that we use were both labeled using curated lists of reliable and unreliable news sources. As such, it is possible that labels contain some noise, as reliable sources may sometimes have less reliable articles and vice versa. It is also possible that bias exists in the websites providing ground truth labels. As such, there is a risk that this could lead a trained model to incorrectly classify certain topics or populations. Relatedly, the previous work that created these datasets assumed that the sharing of a source was inherently an act of spreading unreliable news. A dataset that also contained the stance of the sharer toward the articles would allow for more nuance regarding what is shared, one may wish to separate those wishing to inform others of the unreliability of news from those who are promoting it.

Ethics Statement

If we develop language models for authorship attribution, they could be used to find other online accounts of a person, given posts on a single one of their accounts. This could potentially be used for user profiling and surveillance of target populations (Rangel Pardo et al., 2013). Furthermore, the identification of unreliable news spreaders must be carefully applied in practice, as people may be misclassified, leading to the suppression of speech for these individuals.

User-augmented classification efforts risk invoking harmful stereotyping, as the algorithm labels people as unreliable news spreaders or classifies users posts as unreliable news. These can be emphasized by the semblance of objectivity created by the use of a computer algorithm (Koolen and van Cranenburgh, 2017).

There are forms of bias that apply specifically in natural language processing research. For example, gender bias in a text such as the use of words or syntactic constructs that connote or imply an inclination or prejudice against one gender (Hitti et al., 2019). Machine learning algorithms trained in natural language processing tasks have exhibited various forms of systemic racial and gender biases. For example hate speech detection (Bolukbasi et al., 2016) or learned word embeddings (Park et al., 2018).

[1] Github:`https://github.com/caisa-lab/WOAH24-FakenewsSpreader`

[2] According to `https://mediabiasfactcheck.com/`

Acknowledgements

This work has been supported by the Federal Ministry of Education and Research of Germany (BMBF) as a part of the Junior AI Scientists program under the reference 01-S20060, and the state of North Rhine-Westphalia as part of the Lamarr Institute for Machine Learning. Any opinions, findings, conclusions, or recommendations in this material are those of the authors and do not necessarily reflect the views of the BMBF or Lamarr Institute.

References

Silvio Amir, Byron C. Wallace, Hao Lyu, Paula Carvalho, and Mário J. Silva. 2016. Modelling context with user embeddings for sarcasm detection in social media. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pages 167–177, Berlin, Germany. Association for Computational Linguistics.

Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam Tauman Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 4349–4357.

Limeng Cui, Kai Shu, Suhang Wang, Dongwon Lee, and Huan Liu. 2019. defend: A system for explainable fake news detection. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management, CIKM 2019, Beijing, China, November 3-7, 2019*, pages 2961–2964. ACM.

Nicollas R de Oliveira, Pedro S Pisa, Martin Andreoni Lopez, Dianne Scherly V de Medeiros, and Diogo MF Mattos. 2021. Identifying fake news on social networks based on natural language processing: trends and challenges. *Information*, 12(1):38.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Xinhuan Duan, Elham Naghizade, Damiano Spina, and Xiuzhen Zhang. 2020. Rmit at pan-clef 2020: Profiling fake news spreaders on twitter. In *CLEF 2020 Labs and Workshops, Notebook Papers*. CEUR Workshop Proceedings.

Lucie Flek. 2020. Returning the N to NLP: Towards contextually personalized classification models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7828–7838, Online. Association for Computational Linguistics.

Bilal Ghanem, Simone Paolo Ponzetto, Paolo Rosso, and Francisco Rangel. 2021. FakeFlow: Fake news detection by modeling the flow of affective information. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 679–689, Online. Association for Computational Linguistics.

Maria Glenski, Tim Weninger, and Svitlana Volkova. 2018a. Identifying and understanding user reactions to deceptive and trusted social news sources. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 176–181, Melbourne, Australia. Association for Computational Linguistics.

Maria Glenski, Tim Weninger, and Svitlana Volkova. 2018b. Propagation from deceptive news sources who shares, how much, how evenly, and how quickly? *IEEE Transactions on Computational Social Systems*, 5(4):1071–1082.

Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, volume 9 of *Proceedings of Machine Learning Research*, pages 249–256, Chia Laguna Resort, Sardinia, Italy. JMLR Workshop and Conference Proceedings, PMLR.

Yasmeen Hitti, Eunbee Jang, Ines Moreno, and Carolyne Pelletier. 2019. Proposed taxonomy for gender bias in text; a filtering methodology for the gender generalization subtype. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 8–17, Florence, Italy. Association for Computational Linguistics.

Philip N. Howard and Bence Kollanyi. 2016. Bots, #StrongerIn, and #Brexit: Computational Propaganda during the UK-EU Referendum. *CoRR*, abs/1606.06356.

Clayton Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media*, volume 8, pages 216–225. The AAAI Press.

Georgi Karadzhov, Preslav Nakov, Lluís Màrquez, Alberto Barrón-Cedeño, and Ivan Koychev. 2017. Fully automated fact checking using external sources. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 344–353, Varna, Bulgaria. INCOMA Ltd.

Hamid Karimi, Proteek Roy, Sari Saba-Sadiya, and Jiliang Tang. 2018. Multi-source multi-class fake news detection. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1546–1557, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Abdullah Faiz Ur Rahman Khilji, Anubhav Sachan, Divyansha Lachi, and et al. 2023. Can we debunk disinformation by leveraging speaker credibility and perplexity measures?

Corina Koolen and Andreas van Cranenburgh. 2017. These are not the stereotypes you are looking for: Bias and fairness in authorial gender attribution. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 12–22, Valencia, Spain. Association for Computational Linguistics.

William H Kruskal and W Allen Wallis. 1952. Use of ranks in one-criterion variance analysis. *Journal of the American statistical Association*, 47(260):583–621.

Ruixue Lian, Che-Wei Huang, Yuqing Tang, Qilong Gu, Chengyuan Ma, and Chenlei Guo. 2022. Incremental user embedding modeling for personalized text classification. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2022, Virtual and Singapore, 23-27 May 2022*, pages 7832–7836. IEEE.

Yang Liu and Yi-fang Brook Wu. 2018. Early detection of fake news on social media through propagation path classification with recurrent and convolutional networks. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 354–361. AAAI Press.

Shahan Ali Memon and Kathleen M Carley. 2020. CMU-MisCov19: a Novel Twitter dataset for characterizing COVID-19 misinformation. *Zenodo*.

Tomás Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pages 3111–3119.

Yida Mu and Nikolaos Aletras. 2020. Identifying twitter users who repost unreliable news sources with linguistic information. *PeerJ Computer Science*, 6:e325.

Van-Hoang Nguyen, Kazunari Sugiyama, Preslav Nakov, and Min-Yen Kan. 2020. FANG. In *Proceedings of the 29th ACM International Conference on Information &; Knowledge Management*. ACM.

Ji Ho Park, Jamin Shin, and Pascale Fung. 2018. Reducing gender bias in abusive language detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2799–2804, Brussels, Belgium. Association for Computational Linguistics.

James W. Pennebaker, Ryan Boyd, Kayla Jordan, and Kate Blackburn. 2015. *The development and psychometric properties of LIWC2015*. University of Texas at Austin.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Gordon Pennycook and David G. Rand. 2019. Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition*, 188:39–50. The Cognitive Science of Political Thought.

Joan Plepi, Magdalena Buski, and Lucie Flek. 2023. Personalized intended and perceived sarcasm detection on Twitter. In *Proceedings of the 3rd Workshop on Computational Linguistics for the Political and Social Sciences*, pages 8–18, Ingolstadt, Germany. Association for Computational Linguistics.

Joan Plepi and Lucie Flek. 2021. Perceived and intended sarcasm detection with graph attention networks. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4746–4753, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Joan Plepi, Béla Neuendorf, Lucie Flek, and Charles Welch. 2022a. Unifying data perspectivism and personalization: An application to social norms. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7391–7402, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Joan Plepi, Flora Sakketou, Henri-Jacques Geiss, and Lucie Flek. 2022b. Temporal graph analysis of misinformation spreaders in social media. In *Proceedings of TextGraphs-16: Graph-based Methods for Natural Language Processing*, pages 89–104, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Francisco Rangel, Anastasia Giachanou, Bilal Ghanem, and Paolo Rosso. 2020. Overview of the 8th Author Profiling Task at PAN 2020: Profiling Fake News Spreaders on Twitter. In *CLEF 2020 Labs and Workshops, Notebook Papers*. CEUR-WS.org.

Francisco Rangel Pardo, Paolo Rosso, Irina Chugur, Martin Potthast, Martin Trenkmann, Benno Stein, Ben Verhoeven, and Walter Daelemans. 2013. Overview of the 2nd author profiling task at pan 2014. *CEUR Workshop Proceedings*, 1180.

Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. 2017. Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2931–2937, Copenhagen, Denmark. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Giancarlo Ruffo, Alfonso Semeraro, Anastasia Giachanou, and Paolo Rosso. 2023. Studying fake news spreading, polarisation dynamics, and manipulation by bots: A tale of networks and language. *Computer science review*, 47:100531.

Flora Sakketou, Joan Plepi, Riccardo Cervero, Henri Jacques Geiss, Paolo Rosso, and Lucie Flek. 2022. FACTOID: A new dataset for identifying misinformation spreaders and political bias. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3231–3241, Marseille, France. European Language Resources Association.

Alireza Salemi, Sheshera Mysore, Michael Bendersky, and Hamed Zamani. 2023. Lamp: When large language models meet personalization. *ArXiv preprint*, abs/2304.11406.

Karishma Sharma, Feng Qian, He Jiang, Natali Ruchansky, Ming Zhang, and Yan Liu. 2019. Combating fake news: A survey on identification and mitigation techniques. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(3):1–42.

Zien Sheikh Ali, Abdulaziz Al-Ali, and Tamer Elsayed. 2022. Detecting users prone to spread fake news on Arabic Twitter. In *Proceedinsg of the 5th Workshop on Open-Source Arabic Corpora and Processing Tools with Shared Tasks on Qur'an QA and Fine-Grained Hate Speech Detection*, pages 12–22, Marseille, France. European Language Resources Association.

Kai Shu, Suhang Wang, and Huan Liu. 2018. Exploiting tri-relationship for fake news detection. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, volume abs/1712.07709. AAAI Press.

Kai Shu, Suhang Wang, and Huan Liu. 2019. Beyond news contents: The role of social context for fake news detection. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining, WSDM 2019, Melbourne, VIC, Australia, February 11-15, 2019*, pages 312–320. ACM.

Efstathios Stamatatos. 2009. A survey of modern authorship attribution methods. *Journal of the American Society for information Science and Technology*, 60(3):538–556.

Yla R Tausczik and James W Pennebaker. 2010. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology*, 29(1).

Charles Welch, Chenxi Gu, Jonathan K. Kummerfeld, Veronica Perez-Rosas, and Rada Mihalcea. 2022. Leveraging similar users for personalized language modeling with limited data. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1742–1752, Dublin, Ireland. Association for Computational Linguistics.

A Psycholinguistic Features for Misinformation Spreader Detection

Several previous papers have addressed the use of psycholinguistic features for the detection of misinformation spreaders (Rashkin et al., 2017; Shu et al., 2018). We decided to compare our approach to the use of such features using the commonly used lexicon, Linguistic Inquiry and Word count (LIWC; (Tausczik and Pennebaker, 2010; Pennebaker et al., 2015)). The lexicon provides a set of word categories for over 6k words, representing linguistic and psycholinguistic processes.

We construct a feature-vector using the lexicon by counting each word category and concatenating these into a single vector. We also experimented with a concatenation of the LIWC feature vector and the User2Vec representations. We provide results in Table 5. The methods for results that do not use LIWC are copied from §5 for comparison. We include only the GloVe results here, as they performed better than Word2Vec. We find that the LIWC features underperform the personalization methods, and even lower performance when combined with the User2Vec approach.

B Additional Training Details

We use the transformers HuggingFace model `bert-base-uncased`. The model has 12 layers, a hidden size of 768, 12 heads, and 110M parameters. It was trained on lower-cased English text. The non-BERT models run in a few minutes on a single CPU. The BERT models for the post-level tasks take 9-10 hours to run for one context size for 10 runs on an NVIDIA A100 GPU.

Model	**U2V**	**LIWC**	**LIWC+U2V**	**AA**	**Baseline**
RF	0.71	0.57	0.68	0.74	0.61
Ridge	0.73	0.64	0.71	0.67	-
LR	0.71	0.58	0.71	0.64	0.60
SVM	**0.75**	0.61	0.71	0.69	0.61

Table 4: Psycholinguistic feature comparison for unreliable news spreader detection results on the balanced FACTOID dataset using the logistic regression (LR), ridge regression (Ridge), support vector machine (SVM) and random forest (RF) classifiers. Reported values are the F_1- scores over a 5-fold Cross Validation. User2Vec approaches use GloVe embeddings for training.

Model	**U2V**	**LIWC**	**LIWC+U2V**	**AA**	**Baseline**
RF	0.62	0.65	0.74	0.70	-
Ridge	0.70	0.65	0.73	0.76	-
LR	0.75	0.65	0.74	**0.82**	-
SVM	0.70	0.63	0.71	0.76	-

Table 5: Psycholinguistic feature comparison for unreliable news spreader detection results on the balanced Twitter dataset using the logistic regression (LR), ridge regression (Ridge), support vector machine (SVM) and random forest (RF) classifiers. Reported values are the F_1-scores over a 5-fold Cross Validation. User2Vec approaches use GloVe embeddings for training.