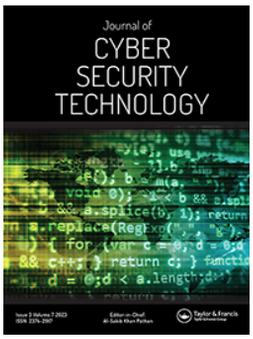

Journal of Cyber Security Technology

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/tsec20

Feature analysis of fake news: improving fake news detection in social media

Johnathan Leung, Dinusha Vatsalan & Nalin Arachchilage

To cite this article: Johnathan Leung, Dinusha Vatsalan & Nalin Arachchilage (2023): Feature analysis of fake news: improving fake news detection in social media, Journal of Cyber Security Technology, DOI: 10.1080/23742917.2023.2237206

To link to this article: https://doi.org/10.1080/23742917.2023.2237206

Published online: 24 Jul 2023.

JOURNAL OF CYBER SECURITY TECHNOLOGY
https://doi.org/10.1080/23742917.2023.2237206

Taylor & Francis
Taylor & Francis Group

Feature analysis of fake news: improving fake news detection in social media

Johnathan Leung[a], Dinusha Vatsalan[b] and Nalin Arachchilage[a]

[a]The School of Computer Science, The University of Auckland, Auckland, New Zealand; [b]The School of Computing, Macquarie University, Sydney, Australia

ABSTRACT

Fake news is a threat to society, and its spread can have real-world consequences in many situations. For example, attackers have weaponised fake news to influence user opinion by causing users to emotionally react to fake news. On the other hand, fake news can also be a threat to national democracy. Therefore, we investigate textual sentiment, visual sentiment, behavioural and metadata features that entice users, using a dataset of posts from Reddit and another from Twitter, that were already categorised into the labels of fake news and real news. First, we extract features, such as visual sentiment, textual sentiment, behavioural reactions, and metadata, and then analyse various features for fake news prediction. We then run a machine learning experiment to classify posts that help improve fake news detection in social media.

ARTICLE HISTORY
Received 5 April 2023
Accepted 8 July 2023

KEYWORDS
Fake news; feature selection; visual sentiment; user behaviour; machine learning

1. Introduction

Diffusion of fake news has become prevalent and poses a threat to society [1–3] as well as to democracy as it can negatively influence the users' trust in governmental institutions [4]. Major events such as the 2016 United States presidential election have involved fake news that proliferate on social media and influence voters [1]. More recently, in the COVID-19 pandemic, an "infodemic" of information that contradicts official advice related to vaccines and health measures have rapidly proliferated on social media, which caused people to ignore recommended health guidelines and ultimately threaten public health [5]. False rumours have also played a role in the reaction in other situations, such as terrorist attacks, by causing false information to propagate among the public and drown out contradicting facts [2,3].

On social media, the spread of fake news is a significant issue, and fake news spreads faster than real news, possibly due to the emotional reactions of readers and the novelty of fake news [6]. Fake news can be weaponised by attackers to

CONTACT Nalin Arachchilage nalin.arachchilage@gmail.com The School of Computer Science, The University of Auckland, Auckland, New Zealand

influence online user opinion, such as during the 2016 US and 2017 French elections [7].

Given that a major part of the spread of fake news is due to human behaviour, there is a clear need to investigate and detect the emotional and behavioural factors in fake news that drive humans to share fake posts [6]. In particular, the aspects of the particular content within the post, like attention-grabbing features of text and images, may be a characteristic of fake news which push users to trust the news [8]. A recent study by Vatsalan and Arachchilage [9] has investigated the importance of particular behavioural and emotional features in a fake news dataset. Though the study findings mentioned the importance of such features in the classification task, it did not measure the impact of including the different categories of features on the prediction accuracy. Furthermore, it does not study emotional aspects in other mediums, such as the visual medium.

Therefore, we aim to conduct a study of different categories of features for fake news detection, namely emotional data in images and texts, as well as metadata (e.g. the poster, the referenced URL) and behavioural data (e.g. post score, reply count). By doing so, we will be able to comprehensively identify the characteristics of fake news that entice users and evaluate the importance of different characteristics. In turn, this work aims to contribute to improving the performance of fake news detection. We use machine learning (ML) techniques to classify fake news using the identified influential or impactful features. We conduct an empirical study and present our findings using two real datasets.

The remainder of the paper is structured as follows: we discuss the related work, we present the methodology and then the experiment evaluation discussing the findings, and conclude the paper.

2. Related work

2.1. Fake news identification

Several works have been conducted for fake news detection using a variety of approaches [10]. One way is to look at the user behavioural aspects of fake vs real posts by directly analysing the user reactions. Ma et al. [11] analysed the propagation tree of posts of fake news to track the spread and engagement with the fake news, and found distinct characteristics about patterns of diffusion of rumours. Chen et al. [12] used deep learning to look at Twitter posts to gauge user reactions to rumours, and found specific features that indicate an emotional reaction to fake claims. However, they only consider the text-based analysis of user reactions.

The aspects of text posts can also be a factor that affects user behaviour regarding fake news. Kapusta et al. [13] extracted the features of text from real and fake news posts, specifically the count of sentences and words and the

sentiment rating, and compared them between the fake and real posts. Ajao et al. [14] also used text sentiment by combining it with word embeddings, reporting an improvement over previous approaches on a particular Twitter dataset.

Singh et al. [8] extracted language features and attributes from the visual sentiment mode in their model to predict if a news article was fake. The results showed that emotive visual features, primarily the presence of violence, and signs of manipulation (like overlaying text), and the primary image colour, differed between real and fake news. Specific features of text, such as pronouns and emotive words also were found to be important. Castillo et al. [15] used multiple mediums in their model to detect fake news from a Twitter dataset, including language features of the text, metadata from the user and Twitter information about the topic that the tweet belongs to.

Regarding fake news detection with a multimodal approach in general, Kirchknopf et al. [16] integrated text, images, and metadata in a multimodal framework on the Fakeddit dataset from Reddit. They used a neural network, which consisted of a Long Short-Term Memory (LSTM) network for text and a convolutional neural network for visual input, and additional inputs for metadata.

Vatsalan and Arachchilage [9] also utilised text sentiment, behavioural, and metadata features in social media fake news posts. Although previous studies used a subset of these features and/or investigated the importance of these features in the fake news classification task, they did not measure the impact of including the different features on the classifier accuracy and did not comprehensively evaluate and select appropriate features from a wide range of categories including sentiments in visual medium.

2.2. Visual sentiment analysis

Online posts are often accompanied by images in addition to text, and the visual content can play a part in emotionally enticing readers to fake content [8]. Therefore, visual sentiment analysis can be used in the task of fake news detection.

Multiple techniques have been developed to automate the task of visual sentiment detection. One method is to use descriptive features from the images, such as Borth's approach of using colour histograms, GIST descriptors, and visual Bag-of-Words to calculate the visual sentiment [17].

As an alternative method, convolutional neural network (CNN) has performed well on image-related tasks in ML, and hence they have seen use in image sentiment. For instance, Vadicamo et al. [18] trained multiple different models based on different architectures of CNNs on a Twitter dataset of images labelled with the sentiment polarity (which is derived from the text that accompanies the image). You and Luo

[19] have also used a CNN architecture for visual sentiment analysis.

3. Methodology

As illustrated in Figure 1, our methodology consists of several steps. First, we extract textual sentiment, visual sentiment, behavioural and metadata features from fake and real news in a dataset. Then, we analyse the data and evaluate the correlations between these set of features, as well as compare features between real and fake content. Based on the analysis, we identify important and impactful features for fake news detection, i.e. to distinguish between real and fake news. Then, we build a classifier using machine learning techniques, which will try to classify a post as fake or not, using these features as training data. In the following subsections, we will describe our methodology in detail.

3.1. Dataset

In our study, we use the following datasets.

3.1.1. Fakeddit [20]

This is a multimedia dataset consisting of posts from the Reddit social media website. Reddit is a website made of multiple "subreddits", or communities with a particular theme; for example, "nottheonion" contains stories which appear fake but are in fact true, and "photoshopbattles" hosts contests that let users post manipulated images. Some subreddits contain real content, such as news subreddits, but others contain content which are misleading in some way. Each post contains a title, optional image, and can link to some external content (such as a news article or some other website). Users can leave comments on the posts. Users can leave upvotes and downvotes on the posts and comments; the total score is the number of upvotes minus the number of downvotes. This

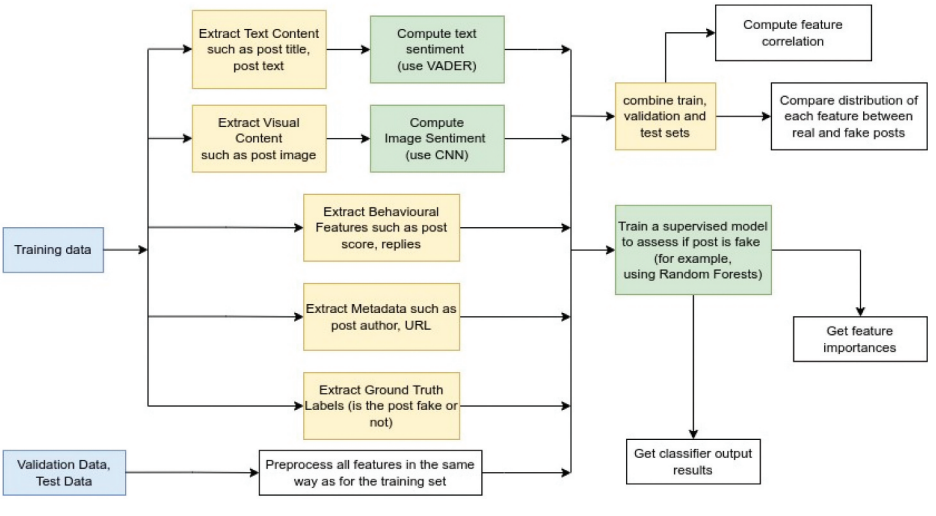

Figure 1. The methodology used in this study.

dataset provides a two-way fake/real label for each post, and a post is *pseudo-labelled* according to the theme of the *whole subreddit* that the post was created in.

We only take the dataset of posts which have text and images. Before filtering, we have 564,000 posts in the training set, 59342 in the validation set, 59319 in the test set. Within the data that we used for experiments, the training partition has 340,945 rows with both images and text, the testing partition has 35,997 and the validation partition has 35,597 rows. About 49% of posts are classified as real.

3.1.2. Twitter dataset [21]

This is a set of labelled Twitter posts for fake news detection, based on the MediaEval Twitter conspiracy detection dataset. This dataset contains fake and real posts on Twitter, and the content is related to 5 G technology and the COVID-19 pandemic. On Twitter, a post consists of text and can be accompanied by links and images, and users can 'favourite' and retweet (i.e., forward the post). Each post in the dataset is accompanied by a label of whether it is true, and they were manually annotated.

We did not do any filtering for this dataset. This dataset contains 1245 true posts and 479 fake posts.

3.2. Data types

Multiple data types are used in our study:

3.2.1. Textual emotional content

The emotions in posts can entice users, and in the context of fake news, this can lead users to wrongly believe it [9,14]. These aspects can target a user to fall for fake information and further spread it to others. Although the emotions are not readily available as data, they are present in text and the strength and direction of the post's emotions can be computed. Our hypothesis is that analysing the text in the data using sentiment analysis techniques and extracting the sentiments from texts in news/posts would help improve the classification/prediction model used for fake news identification.

3.2.2. Visual emotional content

News content can also be accompanied by visual data, such as a picture or video. The visual medium can convey additional information about the data. To attract people to fake news, an attacker can also attach visual content with strong emotional content to reinforce the emotional aspect of the post. For example, a shocking image can be used to accompany text with negative sentiment and trigger negative emotions in the user, and cause them to be trapped into the fake content. We extract the

image sentiment to attempt to further help improve the prediction model.

3.2.3. Behavioural features

These features correspond to the behavioural aspects of users with regard to the news/posts in social media. Behavioural features can be captured from the number of shares or re-tweets, number of likes and dislikes and number of comments to a news/post in social media. Behavioural features influence users in reading and sharing the news, as often users get influenced by the behaviour of other users (who could be known friends, unknown users, in the common circle of friends, or popular users) in social media. We hypothesise incorporating such behavioural features into the classification/prediction model for fake news identification can also help improve accuracy of the fake news prediction model.

3.2.4. Metadata

In general, this includes features such as the author, the linked URLs in the post text, the creation time. These features are potentially useful in fake news detection because readers can refer to these metadata to see if the post is from a trustworthy source or links to reliable websites.

3.2.5. Ground truth labels

Ground truth labels in the form of 'fake' and 'legitimate' labels are required in our study for the analysis tasks as well as for training the fake news classification and prediction models.

3.3. Models

We use several models in our study to extract the required features and to predict the fake news/posts.

3.3.1. Text sentiment analysis model

To study the emotions associated within text data of the news, we use the VADER (Valence Aware Dictionary and sEntiment Reasoner) sentiment analysis tool [22], which is specifically developed to extract sentiment polarity expressed through post text. It considers several aspects of sentiments, including the use of exclamation marks, capitalization, intensifying words (e.g. extremely), conjunctions (e.g. nevertheless), emojis, slang, acronyms and emoticons. It returns the positive, negative, and neutral sentiment scores, that indicate the proportions of text that fall in these categories (that are summed to 1.0), as well as the compound score, which is a weighted sum of lexicon ratings. (normalized between −1 and +1)

3.3.2. Image sentiment analysis model

We also utilise the emotions associated with the image. To extract the image emotional sentiment, we use the convolutional neural network developed by Vadicamo et al. [18] that is based on VGG19. The authors have published the weights of a pretrained model, so it can be directly used for inferring image sentiment for our study. It returns the positive, negative and neutral sentiment scores of images, which sum to 1.

3.3.3. Fake news prediction model

In addition to user sentiment/emotional features extracted using the above-mentioned models, we also extract user behavioural information, such as upvote score and upvote ratio of the post/news, and the number of comments to the post, as well as metadata of the news, like the author and the domain of the linked content. We use all these features to train a supervised fake news classification model, such as random forests, to utilise the extracted features in order to predict whether or not the news samples are ”fake”.

3.4. Steps

We conducted the following steps to predict fake news:

3.4.1. Feature analysis and selection

We first compare the sentiment scores of image, title (and comment) of fake and legitimate news posts calculated by the sentiment models to study the different ranges of sentiment scores and patterns in fake vs. legitimate news. This allows us to identify the most influential sentiment (negative, positive, or neutral) in fake news and sentiments conveyed through which medium (image, title, comments, or news content) contributes more to distinguish fake news from legitimate news. For example, fake news might have a title with strong sentiments to emotionally attract users.

We then analyse the correlations between the features of the posts. For example, sentiments within the post content may be linked to the user behavioural data, such as the post-voting score.

We next evaluate the feature importance scores measured by Gini index and rank the features accordingly to identify which features contribute more to the learning of fake news prediction model.

3.4.2. Prediction

Finally, we train a supervised model (e.g. random forest) on the selected features to predict the fake news/posts. We study how effective the prediction model trained on metadata, user sentiment and behavioural features is on the test dataset. We evaluate the prediction accuracy by comparing the predicted labels

('fake' and 'legitimate') by the classifier with the ground-truth labels using accuracy, precision, recall, and f1-measure.

4. Results

In this section, we present the experimental results of our study on two fake news datasets described in Section 3.1. We provide the results of feature analysis in Section 4.1 and the results of fake news prediction model in Section 4.2.

4.1. Feature analysis

We analyse the features and their importance and correlations with each other to validate our hypothesis of using such different features for improved fake news prediction model. For Fakeddit dataset, the features under consideration are:

- image neg/neu/pos:The three-way (negative, neutral, and positive) image sentiment scores as determined from the neural network model.
- text neg/neu/pos/comp:The negative, neutral, and positive text senti- ment scores as returned by the VADER sentiment analysis tool.
- score: The score which is calculated as the upvotes of the post minus the down- votes; this corresponds to behavioural feature, i.e. users' reactions to the post.
- upvote ratio: The ratio of votes that are upvotes; this corresponds to positive reactions of users to the posts.
- The number of comments made by users to the post.
- The author, which is identified by a unique username; this is metadata.
- The domain, which is the website domain of the linked image in the news post; this is also metadata.

For the Twitter dataset, we use the following features:

- image neg/neu/pos:The three-way image sentiment scores calculated using the CNN model.
- text neg/neu/pos/comp:These are the negative, neutral, positive, and compound scores of text sentiment, according to the VADER tool.
- retweet count, favourite count: User reaction statistics for a post, which are user behavioural features. user {verified, followers count, friends count, listed count, statuses count, age days, favourites count}: These correspond to statistics for the author's credibility and interactions in social media that come under "metadata" category. The listed count indicates the number of

public lists the user is part of; the statuses count indicates the number of tweets.

Among these features, the domain and author are categorical data. For the purpose of analysis, we generated another feature named domain proportion true as a feature that represents, out of the posts that have a corresponding domain, the proportion of posts that are labelled as real posts. This acts as a proxy for the categorical variable. For example, if a particular post has domain abc.com, and 3 out of 5 posts in the training set with domain abc.com are labelled as true, then domain proportion true is set to 0.6 for any post with domain abc.com. If domain proportion true is some other value for another category, like 0.1 for xyz.com, then this feature can distinguish between different categories. Similarly, author proportion true is created based on the author data for similar purposes.

4.1.1. Results on Fakeddit dataset

We first compare the features for real and fake news in Figure 2 to analyse the importance of these features for fake news detection. In the title, the neutral sentiment is lower for fake posts than real posts, the negative sentiment is higher, and the compound sentiment is also lower. This suggests that fake news may have more negative sentiments to make the headline more emotional to grab users.

We also can observe that the upvote ratio is higher for fake posts, suggesting that many users may have a more firm and acceptance reaction to the content

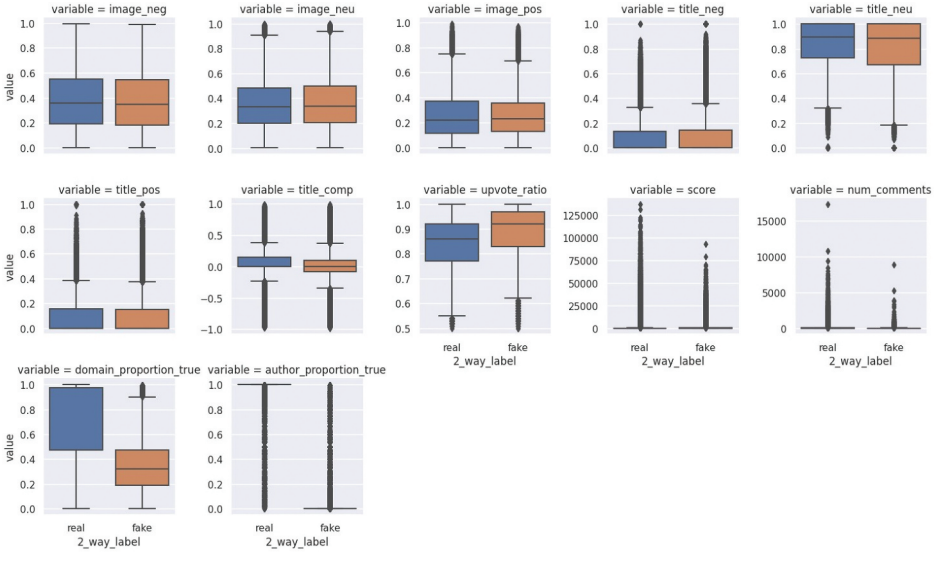

Figure 2. Fakeddit: Each box plot shows the distributions of values for each feature/variable compared between fake and real news posts.

of fake posts, as a higher proportion of upvotes means that the overall audience's judgement of the post is more approving. On the other hand, the number of comments and the post score are lower for fake content than for real content, which means that comparatively not many users react or comment to fake posts. However, among the user reactions, fake news tend to receive mostly up votes compared to real posts, i.e. more acceptance from users.

The sentiments expressed on images appear to be less strongly correlated to identifying whether a post is real or fake. However, the images appear to have lower positive sentiment scores for fake content than real content. The author proportion statistics show that the authors of real posts tend to post a high number of real content, and authors of fake posts conversely post a low amount of real content. A similar trend is found for the domain feature as well. This is significant as it shows that the post is more likely to be fake, if it is created by an author known for posting fake content or has a domain that is associated with fake contents.

Next, we analyse the correlation results as summarised in Figure 3. Some interesting and useful observations have been made through this analysis, as follows. Within the images, the neutral sentiment scores are negatively correlated with the negative sentiment scores and the positive sentiment scores, which is not surprising as images with higher negative or positive scores

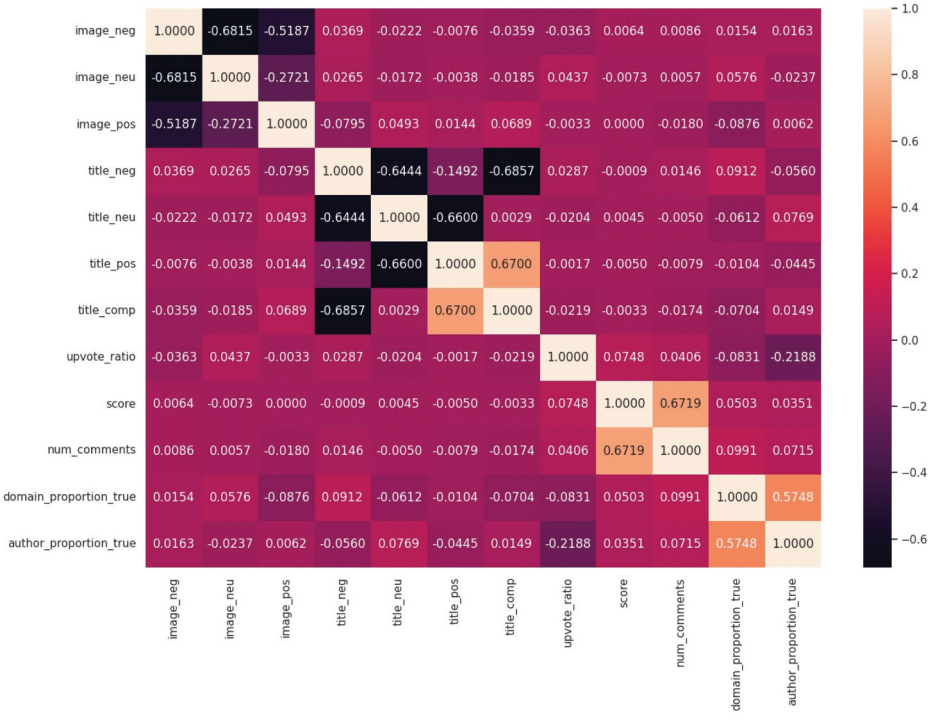

Figure 3. Correlation coefficients between the features for Fakeddit.

generally have lower neutral scores. Within the titles, we see that the negative sentiment scores are strongly negatively correlated with the neutral scores and the positive scores, which is expected as well as titles with negative sentiments are generally negatively correlated with positive and/or neutral sentiments in titles.

Within the behavioural features, we see a highly strong (linear) correlation between score and the number of comments. However, the upvote ratio is not as strongly correlated with either of these two features. This implies that the ”approval rate” of the post is not as strongly correlated with the number of people reacting to the post through votes and comments.

With regard to title and image sentiments, the negative sentiment of the title is negatively correlated with the positive sentiment in the images. This shows that the image sentiment is related in some way to the title text sentiment, and hence, in general, the emotions conveyed by the image are similar to that of the text, especially the title of the news.

We next analyse the feature importance for the prediction task on the Fakeddit dataset. The feature importance scores based on Gini importance metric are shown in Figure 5. In the prediction task, the domain/author (metadata category) seem to be the most important features for predicting if a post is fake or real. This is followed by behavioural features. To a smaller extent, the image sentiment and the text sentiment, also have a minor contribution to the classification task. This largely corroborates the significance of ”metadata” and ”behavioural” features in the classification task, as well as the importance of incorporating sentiment features for further enhancement of the prediction model.

4.1.2. Results on Twitter dataset

We now analyse and compare the features of real and fake news on the Twitter Mediaeval dataset; the results are presented in Figure 4. In the Twitter dataset, we see that the median compound text sentiment score is lower for fake news than that of the real posts, and the median negative text sentiment score is higher for fake posts and the median positive text sentiment score is lower. Overall, the positive and neutral sentiment scores in both images and texts seem to be lower for fake news than real news, while the negative sentiment scores in texts are higher for fake news than real news. This aligns with our previous observations from the Fakeddit dataset.

For the behavioural features, such as retweet count, favourites count and so on, most fake posts have a relatively low number of counts with a few outliers. We can see that the behavioural features have higher values for real posts in the datasets (at least at the high extremes), and we can also observe a similar trend for most of the author-related metadata features. In other words, it appears that real posts are posted by users with more followers, friends, favourite counts, and so on, as well

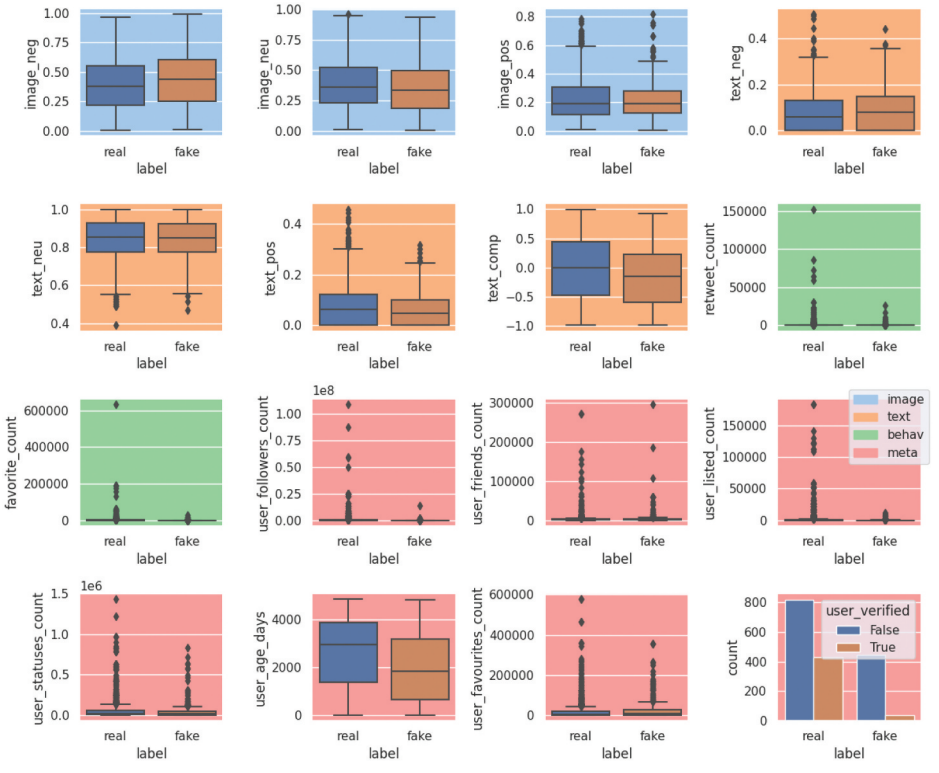

Figure 4. Comparison of features between fake and real posts of Twitter.

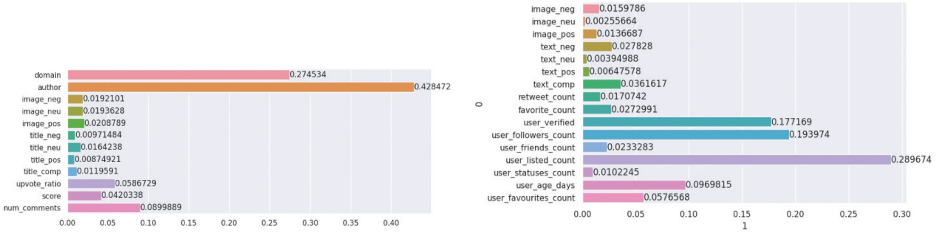

Figure 5. Feature importances for the Fakeddit (left) and Twitter (right) datasets.

as reacted by many users as compared to fake posts. We observe similar correlations between different categories of features in this dataset, as with Fakeddit dataset.

Similar findings for the feature importance scores on the Twitter dataset are provided in Figure 5 (right). Some metadata features, specifically the listed count, verified, and followers count, have high feature importances. This is followed by behavioural features (retweet count, favourite count) and text sentiment features (the negative and composite scores), and then the image sentiment scores.

4.2. Fake news classification results

4.2.1. Results on Fakeddit dataset

For evaluating the fake news classification model on the Fakeddit dataset, we use the training, test, and validation partitions provided by the authors of the dataset [20]. We used the random forest classification model to predict the fake news. The random forest model has hyperparameters to be fine-tuned, which we tuned using the validation set. For the categorical values, we encoded them using a "target encoding" method described in [23]. For each category, it encodes the values into numerical values as a mix of the posterior probability of the target given a category and the prior probability of the target across all training samples.

Table 1 shows the performance of the Random Forest classifier using the extracted features from Fakeddit dataset, with different combinations of features. From the results with only one category of features, the metadata features category produces the best results, showing that predicting the fake or real nature of posts using the user profile statistics and domain produces good prediction accuracy. The behavioural category has high prediction accuracy and combined with metadata it improves the prediction accuracy. This shows a high correlation between the fake or real nature of the posts and the users' behaviours or reactions to the posts. This implies that the user reactions are useful to distinguish the fake or true nature of posts.

When comparing the performance of the classification model with and without using sentiments in images, there are not always major improvements in prediction accuracy (for example, text + behav + meta compared to image + text + behav + meta has less than 0.1% improvement). However, there is a considerable improvement from using the sentiments in text to sentiments in image and text, of about a 1% improvement. This suggests that, compared to using text sentiment only, the image sentiment has additional information that can be more informative for detecting the fake posts. When metadata or

Table 1. Classifier results. Note that f1 score is related to the fake class.

name	accuracy	precision	recall	f1
image+text+behav+meta	0.8897	0.8915	0.8922	0.8918
text+behav+meta	0.8894	0.8902	0.8930	0.8916
behav+meta	0.8849	0.8838	0.8914	0.8876
image+behav+meta	0.8839	0.8828	0.8904	0.8866
text+meta	0.8310	0.8785	0.7755	0.8238
image+text+meta	0.8298	0.8744	0.7776	0.8232
meta	0.8100	0.8918	0.7136	0.7928
image+meta	0.8091	0.8817	0.7222	0.7940
image+text+behav	0.7361	0.7301	0.7649	0.7471
text+behav	0.7346	0.7272	0.7667	0.7464
image+behav	0.7167	0.7071	0.7579	0.7316
behav	0.7135	0.6984	0.7702	0.7326
image+text	0.5865	0.5889	0.6245	0.6062
text	0.5744	0.5585	0.7866	0.6532
image	0.5325	0.5349	0.6312	0.5791

behavioural data are added, these features are considered to be more influential than image sentiments. If we look at the contribution of sentiments in ”text”, we can see that in several cases, such as between ”meta” and ”text+meta”, and ”image” and ”image+text”, there are high improvements in the prediction accuracy (about 2% and 5.5%, respectively) when text is added in.

Looking at the top performing categories, the ”behav+meta” category appears to account for most of the prediction accuracy. Interestingly, all categories with ”meta” score high in precision (around 87 – 89%), which means that out of the posts that the model predicted to be”fake”, a high amount is actually classified correctly. However, this accuracy can be further improved by combining other features, such as user behavioural features, as validated by these results.

Also, within the single category of features, sentiments in text have the highest recall of about 80%, but does not have such a high precision. Arguably, recall has more cost than precision for fake news prediction problem, and this can be accomplished by incorporating sentiments in the texts of the posts/news.

4.2.2. Results on Twitter dataset

We ran a similar set of experiments on the Twitter dataset. We ran experiments with 5-fold stratified cross validation repeated 4 times (i.e. with 4-fold cross validation in the inner loop).

The results for the Twitter dataset are shown in Table 2. Note that the number of true posts is 72.2% in this dataset (that means a highly imbalanced dataset). Given that the mean accuracies are very similar for all categories, it suggests that the various features are not informative to a major degree for the fake news classification task, and so the classifier has chosen a naive strategy of predicting the majority class (true). This could be due to the fact of the class imbalance in

Table 2. Fake news prediction results for the unbalanced Twitter dataset; f1 score is for the fake class.

name	mean accuracy	mean f1	mean precision	mean recall	stddev accuracy
behav+meta	0.7198	0.1893	0.5806	0.1060	0.0127
image	0.7206	0.0100	0.2460	0.0016	0.0043
meta	0.7210	0.2170	0.5267	0.1034	0.0076
image+text	0.7210	0.0101	0.2639	0.0016	0.0032
text	0.7213	0.0344	0.2679	0.0037	0.0031
image+text+behav	0.7213	0.0248	0.2625	0.0026	0.0023
text+behav	0.7213	0.0365	0.4398	0.0089	0.0042
image+behav	0.7214	0.0331	0.5519	0.0063	0.0039
behav	0.7222	0.0000	nan	0.0000	0.0009
image+behav+meta	0.7226	0.2164	0.5254	0.1123	0.0137
text+behav+meta	0.7245	0.2239	0.5208	0.1378	0.0110
text+meta	0.7262	0.2639	0.5289	0.1691	0.0125
image+meta	0.7265	0.2378	0.5461	0.1103	0.0138
image+text+meta	0.7284	0.2447	0.5518	0.1436	0.0097
image+text+behav+meta	0.7288	0.2034	0.5006	0.1180	0.0113

Table 3. Fake news prediction results on the balanced Twitter dataset; f1 score is for the fake class. The 2nd –6th columns are the mean over all folds.

Feature categories	mean accuracy	mean f1	mean precision	mean recall	stddev accuracy
image	0.5150	0.4988	0.5170	0.4874	0.0409
text	0.5492	0.5417	0.5518	0.5363	0.0289
image+text	0.5539	0.5566	0.5545	0.5613	0.0357
image+behav	0.5615	0.5722	0.5586	0.5879	0.0293
behav	0.5693	0.6015	0.5599	0.6516	0.0289
image+text+behav	0.5793	0.6092	0.5692	0.6578	0.0335
text+behav	0.5838	0.6206	0.5699	0.6840	0.0292
image+meta	0.6699	0.6957	0.6455	0.7557	0.0314
image+text+behav+meta	0.6704	0.6926	0.6488	0.7445	0.0288

the Twitter data. These results show that not only features, but also training the model on a balanced dataset is essential for accurate classification.

We therefore conducted another set of experiments on a balanced Twitter dataset, where we sampled equal number of fake and real news posts, similar to [18]. In other words, we balanced the dataset by removing some real posts randomly so that the number of fake posts becomes equal to the number of real posts. The fake news prediction results on the balanced Twitter dataset are in Table 3.

In the Twitter-balanced dataset results, the accuracy with the ″meta″ features are higher than that of categories without ″meta″ in all cases. Furthermore, we see that ″image+text″ sentiment features improve the performance than when using ″image″ sentiment features only, which means that in this case the sentiments in texts would have a contribution to the prediction task. However, there is not a significant increase in the prediction results between using the sentiments from ″text″ only and sentiments from ″image+text″. Finally, when all the features are combined, the prediction model was able to achieve the highest prediction performance with regard to all metrics. This validates the importance of a combination of all categories of features that all these different features are important to improve the fake news prediction task.

5. Discussion

From our experimental results on the Fakeddit dataset, we can observe that the metadata (the post's associated domain and author features) category and the behavioural category are highly useful for classification, as given by their high feature importance scores and the prediction accuracy results with these feature categories. Combining sentiment-related features also has a marginal contribution to the accuracy of the prediction model when combined with these feature categories.

The Twitter dataset is highly imbalanced and hence limits the evaluation of fake news prediction. Class imbalance is one of the key factors that impact the success of the model learning in terms of accuracy and fairness

on the training dataset. This imbalance problem is similar to that in [18] Therefore, the results on the original Twitter dataset do not provide any important findings. However, the balanced Twitter dataset results are highly consistent with the Fakeddit dataset results. As for the Fakeddit dataset, users tend to have a pattern of posting multiple times in the same subreddit (and hence topic area), and therefore there is a pattern of what kind of posts a domain host (for example, a domain belonging to a reputable news organisation will generally host reliable content). This helps the prediction model to distinguish real news from fake news based on such metadata.

As for images, we do not see any significant impact of using sentiments conveyed in images on the fake news detection. However, this does not mean that the images do not play a part in attracting the user to the fake news. They indeed help improve the model in picking up the fake news posts to a certain extent. For example, the recall improves highly when sentiments from images used. It might be due to the fact that the sentiment analysis model used for images does not provide more accurate results. The CNN model used for image sentiment analysis is also a pre-trained model, which might not be accurate in different domains or datasets. Moreover, the CNN image sentiment analysis model may not be suitable for different types of images in our study, for example memes and screenshots which are prevalent in the data. This suggests that other particular aspects of images could be investigated further, as done in [8]. The text sentiment does have some correlation with the truth of the post, in both Fakeddit and Twitter datasets, and this reflects the results in [14] and [13].

In summary, different metadata, behavioural and sentiment features play an important role in improving the prediction performance in different aspects (precision, recall, accuracy, f1-score, combination or all of these metrics), and hence incorporating all the different features could help boost fake news detection.

6. Conclusion

This work investigated and analysed the role of multiple features in fake news, specifically focusing on text and visual sentiments, and various metadata and user behavioural data extracted from the posts. Our results reveal that, overall, the metadata and behavioural features are the most important features in fake news detection, and it appears that sentiment features could help improve the performance when combined with a certain set of features together. We also compared the features for real vs. fake news and investigated the differences in the distribution of the different features between fake and real news, and how they are correlated to each other. Our findings can be used to develop a comprehensive feature selection strategy for an effective and user-centric fake news detection model and its browser plugin for social media platforms.

Disclosure statement

No potential conflict of interest was reported by the author(s).

ORCID

Nalin Arachchilage http://orcid.org/0000-0002-0059-0376

References

[1] Bovet A, Makse HA. Influence of fake news in twitter during the 2016 US presiden- tial election. Nat Commun. 2019;10(1):7. doi: 10.1038/s41467-018-07761-2

[2] Gupta A, Lamba H, Kumaraguru P. $1.00 per RT #BostonMarathon #PrayFor- Boston: Analyzing fake content on Twitter. 2013 APWG eCrime Researchers Summit 17-18 September 2013 San Francisco, CA, USA, IEEE; 2013. p. 1–12. doi:10.1109/eCRS.2013.6805772

[3] Starbird K, Maddock J, Orand M, et al., *Rumors, false flags, and digital vigilantes: Misinformation on Twitter after the 2013 Boston Marathon bombing*, in *iConference*, Grandville, MI, USA. 2014, pp. 654–662.

[4] Pew Research Center, *Many Americans say made-up news is a critical problem that needs to be fixed* (2019). Available at accessed 2022 04 29. https://www.pewresearch.org/journalism/2019/06/05/many-americans-say-made-up-news-is-a-critical-problem-that-needs-to-be-fixed/

[5] Carrion-Alvarez D, Tijerina-Salina PX. Fake news in COVID-19: A perspective. Health Promot Perspect. 2020;10(4):290–291. doi: 10.34172/hpp.2020.44

[6] Vosoughi S, Roy D, Aral S. The spread of true and false news online. Science. 2018;359 (6380):1146–1151. doi: 10.1126/science.aap9559

[7] Lazer DMJ, Baum MA, Benkler Y, et al. The science of fake news. Science. 2018;359 (6380):1094–1096. doi: 10.1126/science.aao2998

[8] Singh VK, Ghosh I, Sonagara D. Detecting fake news stories via multimodal anal- ysis. J Assoc Inf Sci Technol. 2021;72(1):3–17. doi: 10.1002/asi.24359

[9] Vatsalan D, Arachchilage NAG, *Hold On! Your emotion and behaviour when falling for fake news in social media*, in *CHI'21 Workshop on Technologies to Support Critical Thinking in an Age of Misinformation*, May 8–13, 2021; Yokohama, Japan, virtual, 2021. https://www.preprints.org/manuscript/202105.0477/v1.

[10] Sharma K, Qian F, Jiang H, et al. Combating fake news. ACM Trans Intell Syst Technol. 2019;10(3):1–42. doi: 10.1145/3305260

[11] Ma J, Gao W, Wong KF, *Detect rumors in microblog posts using propagation structure via Kernel Learning*, in *Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Stroudsburg, USA. 2017, pp. 708–717.

[12] Chen T, Li X, Yin H, et al. Call attention to rumors: Deep attention based recurrent neural networks for early rumor detection. In Trends and Applications in Knowledge Discovery and Data Mining: PAKDD 2018 Workshops, BDASC, BDM, ML4Cyber, PAISI, DaMEMO, June 3, 2018; Melbourne, VIC, Australia, 2018; 40–52. Springer International Publishing.

[13] Kapusta J, Benko Ľ, Munk M. Fake news identification based on sentiment and frequency analysis. In: Learning and analytics in intelligent systems. Cham: Springer International Publishing; 2019. pp. 400–409. doi:10.1007/978-3-030-36778-7_44.

[14] Ajao O, Bhowmik D, Zargari S, *Sentiment aware fake news detection on online social networks*, in *IEEE (ICASSP)*, New York, USA. 2019, pp. 2507–2511.
[15] Castillo C, Mendoza M, Poblete B. Information credibility on twitter The 20th international conference on World wide web 28 March 2011, Hyderabad, India, New York, NY, USA: ACM Press; 2011; p. 675–684. doi:10.1145/1963405.1963500
[16] Kirchknopf A, Slijepcevic D, Zeppelzauer M. Multimodal detection of information disorder from social media. 2021 International Conference on Content-Based Multimedia Indexing (CBMI), 28-30 June 2021, Lille, France, IEEE; 2021; p. 1–4. 10.1109/CBMI50038.2021.9461898
[17] Borth D, Ji R, Chen T, et al., *Large-scale visual sentiment ontology and detectors using adjective noun pairs*, in *ACM international conference on Multimedia*, New York, USA. 2013, pp. 223–232.
[18] Vadicamo L, Carrara F, Cimino A, et al., *Cross-media learning for image sentiment analysis in the wild*, in *ICCVW*, New York, NY, USA. IEEE, 2017, pp. 308–317.
[19] You Q, Luo J, Jin H, et al. Robust image sentiment analysis using progressively trained and domain transferred deep networks. Proce AAAI Conf Artif Intell. 2015;29(1). doi: 10.1609/aaai.v29i1.9179
[20] Nakamura K, Levy S, Wang WY. R/Fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection. The Twelfth Language Resources and Evaluation Conference, May 2023, Marseille, France, European Language Resources Association, 2020; p. 6149–6157. https://aclanthology.org/2020.lrec-1.0
[21] Cheema GS, Hakimov S, Mu¨ller-Budack E, et al. On the role of images for analyzing claims in social media. CLEOPATRA Workshop 2021, 2021; Germany, Germany: Leibniz Universität Hannover (LUH); 2021; p. 1–15. doi: 10.5446/52942
[22] Hutto C, Gilbert E, *VADER: A parsimonious rule-based model for sentiment analysis of social media text*, Proceedings of the International AAAI Conference on Web and Social Media, Ann Arbor, Michigan, USA; 8, (2014; p. 216–225.
[23] Micci-Barreca D. A preprocessing scheme for high-cardinality categorical attributes in classification and prediction problems. ACM SIGKDD Explorations Newsl. 2001;3(1):27–32. doi: 10.1145/507533.507538