 doi:10.20944/preprints202210.0086.v1

Article

Retweet Prediction Based on Heterogeneous Data Sources: The Combination of Text and Multilayer Network Features

Ana Meštrović [1,2]*, Milan Petrović [1,2] and Slobodan Beliga [1,2]

[1] Faculty for Informatics and Digital Technologies, University of Rijeka, 51000 Rijeka, Croatia
[2] Center for Artificial Intelligence and Cybersecurity, University of Rijeka, 51000 Rijeka, Croatia
* Correspondence: amestrovic@uniri.hr; Tel.: +385-51-584-718

Abstract: Retweet prediction is an important task in the context of various problems such as information dissemination analysis, automatic fake news detection, social media monitoring, etc. In this study, we explore the retweet prediction based on heterogeneous data sources. To classify the tweet according to the amount of retweets, we combine features extracted from the multilayer network and text. More specifically, we introduce a multilayer framework that proposes a multilayer network representation of Twitter. This formalism captures different users' actions and complex relationships, as well as other key properties of communication on Twitter. We select a set of local network measures from each layer and construct a set of multilayer network features. In addition, we adopt a BERT-based language model, namely Cro-CoV-cseBERT, to capture the high-level semantics and structure of tweets as a set of text features. We then train six machine learning algorithms (ML): Random Forest, Multilayer Perceptron, Light Gradient Boosting Machine, Category Embedding Model, Neural Oblivious Decision Ensembles, and an Attentive Interpretable Tabular Learning Model for the retweet prediction task. We compare the performance of all six algorithms in three different setups: with text features only, with multilayer network features only, and with both feature sets. We evaluate all setups in terms of standard evaluation measures. For this task, we first prepare an empirical dataset of 199,431 tweets in Croatian posted between January 1, 2020 and May 31, 2021. Our results indicate that the prediction model performs better by integrating multilayer network features with text features than using only one set of features.

Keywords: retweet prediction; multilayer network; natural language processing, text features, multilayer network, Twitter data

1. Introduction

Nowadays, social media platforms and online social networks have become an important source of information. They can serve as an important communication platform in a real-world crisis, emergency, or disasters [1,2]. Users tend to rely on online communication platforms for getting information, publishing posts that reflect their interests, views and activities [3]. At the same time, users can express their opinion on other posts via different forms of feedback such as reposts, quotes, mentions, replies, likes, etc. All these activities affect information spreading on social media [4]. In the last two decades, online social networks have increased the spread of information, but also of misinformation and disinformation, which can lead to an infodemic as a negative side effect [5,6]. Therefore, exploring patterns of information spreading on social networks is significant research in the context of disinformation and misinformation detection.

The primary motivation for this research is the analysis of crisis-related communication on social networks. As social media platforms such as Twitter, Facebook, Instagram, and Weibo play an increasingly important role, people are more likely to use them for information during the global crisis. This may even influence the course of the global crises particularly in the context of epidemics, climate change migration crises, economic crises or wars. For example, the outbreak of the COVID-19 disease caused a significant

 doi:10.20944/preprints202210.0086.v1

increase in social media usage among the public and it seriously affected the public's understanding of the COVID-19 risk [7]. In some countries there were many negative attitudes toward vaccines and anti-pandemic measures promoted on social networks [8]. Therefore, information spreading analysis during the global crisis is of great importance as one step of social media monitoring (infoveillance). Twitter is one of the largest social networks with around 330 million monthly active users [9]. Consequently, it is one of the most studied social networks. Recently, especially for the monitoring and tracking different aspects of healthcare information and public disease [8,10,11]. Among all the user behavior in social media, the retweet is considered one of the primary functions for spreading information on Twitter [12,13]. There is a large number of studies that deal with the prediction of information spreading on Twitter and other social networks. Many complex factors may influence the patterns of information spreading. Thus, different studies propose different sets of features for retweet prediction. Previous methods studied the problem using various linguistic features, the personal information of users, or network properties [12].

In some studies authors combined heterogeneous sets of features. For example in [14] the authors proposed a combination of content features with temporal and topological network features for retweet prediction. Another combination of heterogeneous data sources is proposed in [15] where Suh et al. analyzed two sets of features: content features (URLs, hashtags, mentions) and contextual features (number of followers and followees, the age of the account, the number of favorite tweets, and frequency of tweets).

However, there are properties that have not been fully explored in the task of retweet prediction. One less-studied approach is to use multilayer network properties as features, especially in combination with other features from heterogeneous data sources. The multilayer network is a formalism that captures various sorts of relationships over network data [16,17]. Used in the context of a social network, a multilayer network can represent different actions within the social network such as follow, share, quote, mention or reply as the separate network layer. Since each action has a different impact on information spreading, in this way it is possible to make a fine grade differentiation between layers and to include all this information as a predictor of retweeting. We have already shown that a multilayer network structure is fundamentally more expressive than individual layers in the examples of modeling a multilayer language network [18] and multidimensional knowledge network [19]. In [20] the authors use multilayer network features for disinformation detection in US and Italian news spreading on Twitter.

Inspired by these results we decided to employ multilayer network features in the more general task of information spreading prediction. However, in our approach, we construct a different multilayer network of Twitter and select different network measures to construct a multilayer network set of features. In addition, we combine multilayer network features with text features. To the best of our knowledge, this is the first attempt to use this set of multilayer network features in the task of retweet prediction and the first attempt to combine multilayer network features with text features.

We formalized our approach by introducing a multilayer framework for the representation of key elements of the communications on social networks. The main aim of this study is to explore the potential of the multilayer network measures as the set of features in the task of retweet prediction. Additionally, we investigate if the multilayer network features combined with text features perform better than just one set of features. Therefore, this study explores how message features extracted from heterogeneous data sources may affect tweet spreading in terms of retweeting.

Multilayer network features are extracted from the multilayer model of the social network in which a message is spreading. For the purpose of retweeting prediction we construct a multilayer network with four layers representing actions of following, mentioning, replying, and a layer of tweets and select several network measures from each layer. The text features are represented as a low dimensional vector (embedding) that captures its semantic and structure. More specifically, we adopt a BERT-based language

 doi:10.20944/preprints202210.0086.v1

model, namely Cro-CoV-cseBERT [8] for tweets' representation as embeddings, which we use as the set of text features.

We model the prediction problem as the binary classification task where one class contains tweets with just one retweet and another class contains tweets with more than one retweet. Next, we explore the performance of different feature sets by performing an extensive set of experiments in which we train six machine learning algorithms in three different setups: (i) classification based on text features, (ii) classification based on multilayer network features, and (iii) classification based on text and multilayer network features. More precisely, we train classifiers: Random Forest (RF), Multilayer Perceptron (MLP), Light Gradient Boosting Machine (LGBM), Category Embedding Model (CEM), Neural Oblivious Decision Ensembles (NODE), and Attentive Interpretable Tabular Learning (TabNet model). We evaluate the performance of trained classifiers on three different sets of features in terms of standard evaluation measures: accuracy, precision, recall and F1-score on the large dataset of tweets. For this purpose, we prepare an empirical dataset of 199,431 tweets in the Croatian language posted during the pandemic period between January 1, 2020 and May 31, 2021.

Our main research question is to whether the use of a multilayer network features and combination of features from heterogeneous data sources yields better results in terms of classification evaluation measures over text features. Additionally, we are interested in understanding which of the above features are most effective in the classification task and we analysed that using SHAP approach.

To summarize, the main contributions of this study are as follows:

1. We propose a multilayer framework as a formalism for the representation of communication on online social network.
2. We introduce a multilayer network representation of Twitter and select a set of measures from each layer to be extracted and combined with the metadata as the set of multilayer network features.
3. We perform experiments on a dataset of tweets using separately text features and multilayer network features and its combination and evaluate the performance for six machine learning classifiers.

The rest of the paper is organized as follows. Section 2 discusses some of the existing research work in the prediction of retweeting. Section 3 describes datasets, machine learning classifiers and the methods utilized in this study. Section 4 presents the results and analysis of our proposed approach. Section 5 discusses the proposed approach. Finally, Section 6 concludes our work.

2. Related Work

Information spreading analysis and the retweet prediction task have been carefully studied in a large number of research papers. There are many different ways to approach the problem of retweet prediction. It can be modeled as the binary or multiclass classification problem in which classes are defined according to the number of retweets, and the model should predict the class of a given tweet, as well as the regression/prediction problem in which the model should predict the number of retweets for a given tweet. There are also some other approaches such as prediction a p value, which is the probability of a retweet of the given tweet by the given retweeter [21], retweet time prediction [13] or the prediction of the size of retweet cascade size as in [22].

In the domain of complex networks this task is usually described as the link prediction in the network of retweeting. From the broader perspective, spreading patterns have been studied in many fields ranging from disease spreading [23,24] to information spreading in social networks [25].

In all these approaches one of the major research questions is related to the exploration of the properties that may affect the spreading. There are many possible factors that influence the information spreading in a social network, ranging from linguistic features,

 doi:10.20944/preprints202210.0086.v1

the personal information of users such as user profiles, user post history, user following relationships, or network properties.

Feature engineering and the selection of appropriate feature sets is an important step in all classification tasks. However, some studies have examined the potential of using deep neural networks to avoid the manual construction of features (such as [12] in which the authors proposed attention-based deep neural networks in the task of retweet prediction), it is still worth examining different possibilities in the construction sets of features, especially the combination of features from heterogeneous data sources. In this experiment, we combine the neural network approach for text feature extraction and the manual selection of features from the multilayer network.

In the following subsections we are focused on: **(i)** the research studies that involve features from the heterogeneous data sources, and **(ii)** on research with the multilayer network approach.

2.1. Retweet Prediction Based on Heterogeneous Data Sources

Suh et al. [15] examined two classes of features that might affect the retweetability of tweets: (i) the content features that include whether the tweet contains URLs, hashtags, and mentions and (ii) the contextual features that include the number of followers and followees, the age of the account, the number of favorite tweets, and the number and frequency of tweets. The results show that among content features, URLs and hashtags have a strong influence in retweeting. Among contextual features, the number of followers and followees as well as the age of the account seem to affect retweetability.

Similarly, in [14] the authors studied the effect of the message content in the task of the prediction of spreading a meme/idea. They analyzed the contribution and the limitations of the various feature sets on the information spreading. According to their results, it seems that a combination of content features with temporal and topological network features minimizes prediction error.

Fridaus et al. [3] analyzed the impact of the users' behaviors on retweet activities based on three aspects: topic preference, emotion, and personality. They proposed two types of retweet prediction models, one uses classification algorithms, and the other uses matrix factorization algorithms. The experimental results showed that in terms of the F1-score, the proposed classification models based on user behavior-related features provided a 5%-9% improvement over baseline models and the matrix factorization model showed a 4%-6% improvement over the baseline.

In [13] the authors proposed a novel Deep Fusion of Multimodal Features (DFMF) method for retweet time prediction. The method combines text features and node features in a way that it constructs a word embedding layer to learn the semantics of a tweet and a node embedding layer to learn social relationships within the network. The evaluation results demonstrated that the proposed method is more accurate in predicting the retweet time and can achieve as much as an 11.25 % performance improvement on the recall accuracy compared to Logistic Regression (LR) and Support Vector Machine (SVM).

Dai et al. [26] improved the SVM model for the prediction of user forwarding behavior of hot topics. The prediction of user retweeting behavior is based on combining three different data sources: user interest tags, user history behavior, and external factors influence.

Similarly, Ma et al. [27] explored features from different sources related to the hot topics discussed by the users' followees proposing a novel masked self-attentive model to perform retweet prediction. They incorporated the posting histories of users with an external memory and utilized a hierarchical attention mechanism to construct the users' interests. The obtained results of a dataset collected from Twitter show that the proposed method can achieve a better performance than state-of-the-art methods.

In addition to retweets, heterogeneous feature sources have also been successfully used to predict buzz tweets. Amitani et al. in [28], in their study on the classification of "buzz" tweets, examine the trends in social media and propose a classification method to

 doi:10.20944/preprints202210.0086.v1

study the factors that cause the buzz phenomenon on Twitter. This phenomenon can be understood as an explosion of popularity within a short period of time. The authors note that it is difficult to determine the causes of the buzz phenomenon based solely on the texts posted on Twitter. However, they developed a multitask neural network using both image and text extracted features as input and buzz class (buzz or non-buzz) and number of "likes" and "retweets" as output. The text features of the tweets were extracted using the pre-trained BERT model, and the image features were obtained from pre-trained models such as VGG16. The results of the experiments showed that the correct response rate for predicting buzz classes with the proposed method using both text and image features was higher than using the features alone [28].

There are many possibilities in combining heterogeneous data sources for feature extraction. As we claim in the introduction section, we expect multilayer network measures will have great potential in the prediction of retweeting.

2.2. Multilayer Approach in Social Networks Representation

Pierri et al. [20] modeled Twitter as a multilayer network including four layers: mention, reply, retweet and quote layer. They applied multilayer network measures as features for the detection of disinformation in US and Italian news spreading over Twitter. According to their results, a simple Logistic Regression model is able to classify disinformation vs mainstream networks with high accuracy (AUROC up to 94%).

Arenas et al. [29] modeled various kinds of interactions (specifically, retweeting, mentioning, and replying) as separate layers aiming to characterize interactions in online social networks during exceptional events that cause a large number of tweets (such as the discovery of the Higgs boson). They showed that a multilayer approach can reveal the presence of statistical regularities across different events, suggesting that there are some universal properties of online social networks during exceptional events.

In [30] the authors proposed a method based on the multilayer approach - capable of identifying influencers in online social networks. The layers represent users, items, and keywords, along with the intra-layer interactions among the actors of the same layer.

Magnani and Rossi proposed a model for the representation of multilayer networks and applied this model to two online social networks [31]. Their results confirmed that considering a multilayer network model allows us to extract results that do not correspond completely to the ones that can be obtained from each network layer separately.

In [32] the authors explored two online platforms Twitter and Foursquare analyzing the geo-social properties of links. They represented two platforms as a composite multilayer online social network, where each platform represents a layer in the network. According to their results, using the multilayer approach it is possible to successfully predict links across social networking services.

It is worth mentioning that in [33–35] the authors investigated the spreading patterns in multilayer networks, however, they did not apply machine learning algorithms, their approaches were based on diffusion modeling.

All of these studies of the multilayer network-based approach in modeling social networks are valuable and they have proved the potential of multilayer networks. In our research we adopted some of the ideas, however, we have modeled Twitter differently and utilized different network measures from all previous approaches.

3. Materials and Methods

3.1. Multilayer Framework Definition

Here we introduce a multilayer framework, a formalism that captures key properties of message spreading in social network. The model is based on the multilayer network that we use to represent online social network. Within the multilayer framework, we aggregate multilayer social network with a set of metadata corresponding to text messages published on social media.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 October 2022 doi:10.20944/preprints202210.0086.v1

First, we formalise a general framework that can be used in various tasks of analysis of messages in social media. This model can then be further adapted to a specific task. In this study, we use the proposed framework for tweet representation in the task of retweet prediction. More specifically, based on the multilayer framework, we extract multilayer network features and text features and use these features to train six ML models.

According to [16] a multilayer network is defined as a pair:

$$\mathcal{M} = (\mathcal{G}, \mathcal{C}) \tag{1}$$

where

$$\mathcal{G} = \{G^\alpha, \alpha \in \{1, \ldots, m\}\} \tag{2}$$

is a family of networks (graphs) $G^\alpha = (V^\alpha, E^\alpha)$ called network layers of $\mathcal{M}$ and $\mathcal{C} = E^{\alpha\beta} \subseteq V^\alpha \times V^\beta; \alpha, \beta \in \{1, \ldots, m\}, \alpha \neq \beta$ is the set of interconnections between nodes of different layers G^α and G^β where $\alpha \neq \beta$.

Similar to work presented in [6] layers are annotated as numbers from the set $\{1, \ldots, m\}$, where m is the number of layers. Same as one layer networks, multilayered networks can be directed or undirected, weighted or unweighted. Note that communication in social networks is best described using weighted and directed multilayer network.

Next, we introduce and consider a set T of metadata related to text messages posted on social network. Generally, set T includes all messaging metadata that is available, however, the concrete metadata represented within the framework may vary depending on the task. In the case of Twitter and retweet prediction task, this metadata includes information such as the number of retweets, quotes, mentions, etc. In the context of network analysis, these vectors may be attributes of nodes that represent messages. Finally, the multilayer framework is defined as a tuple:

$$\mathcal{MF} = (\mathcal{M}, T). \tag{3}$$

3.2. Twitter Communication Represented Using Multilayer Network

Given the framework $\mathcal{MF}$, we model Twitter data into four layers, thus $m = 4$. Each layer represent one aspect of communication on Twitter as follows. $G^1 = (V^1, E^1)$ is a tweets layer where Twitter messages are nodes and two nodes i and j are connected with the directed link if message i and j have at least three words and/or hashtag in common. The connection is established according to the timeline; from the first tweet to the second tweet. The weight represents the number of common words/hashtags. $G^2 = (V^2, E^2)$ is the follower layer where Twitter users are nodes. Two nodes i and j are connected with the directed link if user j follow user i. All weights are set to 1 since this layer is an unweighted network. $G^3 = (V^3, E^3)$ is a reply layer where Twitter users are nodes and two nodes i and j are connected with the directed link if user j replies to user i. $G^4 = (V^4, E^4)$ is a mention layer where Twitter users are nodes and two nodes i and j are connected with the directed link if user j mentions user i. The weight represents the number of mentions. Illustration of this model is represented in Figure 1.

Further explanations and details of the multiplex shown in Figure 1, the connections between the nodes of the same or different layers and the weights can be found in [6].

3.3. Multilayer Network Features

Next we select a set of local network measures: degree (in/out), strength (in/out), eigenvector centrality (in/out), Katz centrality (in/out), average clustering coefficient and number of communities.

In general local network measures are based on the number of node links, node position within the network and relationship with other nodes. These are centrality measures and they help in identification of the most influential individuals (nodes) in the network. These measures can give an insight into how nodes communicate with each other, which

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 October 2022 doi:10.20944/preprints202210.0086.v1

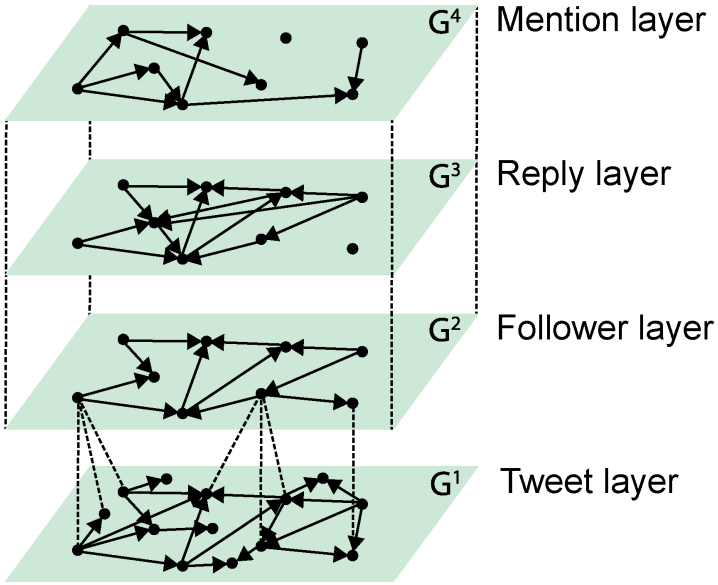

Figure 1. Twitter represented via multilayer network. The image is taken from [6] and adapted to the experiment of this study. Communication on Twitter captured via four layers of Multilayer network: G1 - Tweet layer, G2 - Follower layer, G3 - Reply layer and G4 - Mention layer. Note that more users' actions could be a represented as more separate layers (i.e. Retweet layer and/or Quote layer), however in the case of retweet prediction, these layers are related with the prediction value.

nodes are the most popular (hubs), how close are nodes with each other, and which nodes controls the network (in terms of information flow). In the context of retweeting prediction, node centrality measures can exhibit the nodes with the largest potential to be retweeted. It is important to emphasize that the appropriate usage of centrality measures depends on the understanding of the type of links in the network and network flow [36].

Degree centrality of a node is the measure that takes into account total number of links incident with a node. In the context of Twitter network, degree centrality can be interpreted as node with the largest number of followers or friends. But if we capture more that one layer than, degree centrality may also indicate the node with the largest number of mentions or replies. Higher degree implies popularity, and higher possibility to gain information that is flowing through the network. According to [37] for a node i and the number of its links to other nodes k_i, degree centrality is usually normalized by dividing it by the maximum possible degree $N1$:

$$dc_i = \frac{k_i}{N-1}. \tag{4}$$

In weighted networks a weighted degree is refereed to as **node strength**. Strength of a node i is defined as the sum of all weights attached to links belonging to this node [37]:

$$s_i = \sum_{j \in \Pi(i)} w_{ij}, \tag{5}$$

where $\Pi(i)$ denotes set of neighbouring nodes of a node i.

Eigenvector centrality is introduced by Bonacich [38]. It takes into account the centrality of the adjacent nodes. It can be interpreted as a measure of influence of a node in a network. A high eigenvector score means that a node is connected to many nodes who themselves have high scores. Relative scores are assigned to all nodes in the network based on the concept that connections to high-scoring nodes contribute more to the score than

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 October 2022 doi:10.20944/preprints202210.0086.v1

equal connections to low-scoring nodes. For the node i and constant λ centrality ce_i of node i is defined as [38]:

$$ce_i = \frac{1}{\lambda} \sum_{j \in \Pi(i)} ce_j. \tag{6}$$

Eigenvector centrality computes the centrality for a node based on the centrality of its neighbors. The eigenvector centrality for node i is the i-th element of the vector x defined by the equation:

$$Ax = \lambda x \tag{7}$$

where A is the adjacency matrix of graph G with eigenvalues λ. There is a unique solution x, all of whose entries are positive, if is the largest eigenvalue of the adjacency matrix [37].

For directed graphs equation 6 calculates the "left" eigenvector centrality which corresponds to the in-edges in the graph. For calculating out-edges eigenvector centrality it is necessary to reverse the graph G.

Katz centrality introduced by Leo Katz [39] calculates topological centrality that helps to discover the relative influence of each node on the network. It is a generalization of the eigenvector centrality. Katz centrality computes the centrality for a node based on the centrality of its neighbors. The general equation for calculating Katz centrality for node i is [39]:

$$kc_i = \alpha \sum_{j \in \Pi(i)} kc_j + \beta, \tag{8}$$

where parameter β controls the initial centrality and $\alpha < \frac{1}{\lambda_{max}}$.

Katz centrality computes the relative influence of a node within a network by measuring the number of the immediate neighbors (first degree nodes) and also all other nodes in the network that connect to the node under consideration through these immediate neighbors. For directed graphs it is possible to calculate in- and out- Katz centrality by taking into account that equation 8 can find "left" eigenvectors which corresponds to the in-edges in the graph. For out-edges Katz centrality it is necessary to use the reverse graph G.

Clustering coefficient of a node measures how well are neighbors interconnected and quantifies if they are becoming a clique. The local clustering coefficient is calculated as the proportion of links between the nodes within its neighborhood divided by the number of links that could possibly exist between them. Real-world networks (and in particular social networks) have on average higher clustering coefficient than random networks (when comparing networks of the same size). The clustering coefficient of a node i is defined as [40]:

$$C_i = \frac{e_{ij}}{k_i(k_i - 1)}, \tag{9}$$

where e_{ij} represents the number of pairs of neighbours of a node i that are connected.

For each layer we compute a set of network features separately and quantify different aspects of the information spreading process. Based on these centrality measures we have 9 features for each layer which makes 36 features in total.

In addition, we integrate network measures with the Twitter network metadata from $\mathcal{MF}$. We incorporate metadata from the Twitter network and use the following information as additional vector features for each tweet: number of user followers, number of user friends, number of mentions, number of hashtags, number of user statuses, indicator whether tweet contains a URL, indicator whether tweet contains media, indicator whether tweet contains COVID-19 related keywords, etc. We add some auxiliary variables such as whether the user is in the follower network, etc. Overall, 13 features are extracted from the set T of Twitter metadata.

The result is 49-dimensional vector as the representation of tweet extracted from the multilayer framework $\mathcal{MF}$.

 doi:10.20944/preprints202210.0086.v1

3.4. Text Features

When we are faced with the problem of natural language processing, the choice of an appropriate language model that will be useful in solving the given problem is certainly the development of a new sophisticated model or the choice of an existing language model that includes, i.e. takes into account, semantic, syntactic and other linguistic features of the text. The seminal work of [41] contributed to the emergence of numerous variants of text representation models in terms of low-dimensional vectors in continuous space-embeddings, where embeddings allow semantically related linguistic units to be represented with similar vector representations. As described in [8] the first generation was characterised by shallow language models, such as Word2Vec [41], Doc2Vec [42], GloVe [43] and fastText [44]. They have some shortcomings, such as static embeddings in which multiple concepts (i.e., different meanings of the same entity, polysemy) are not represented by different embedding vectors, or poor performance in new domains. Due to such shortcomings, the next generation of deep language models have been developed, namely ELMo [45], GPT/GPT-2 [46], GPT-3 [47] and BERT [48]. They replace static embeddings with contextualized representations and successfully solve the mentioned shortcomings. Moreover, they enable learning of context- and task- independent representations which yielded an improvement in performance on various NLP tasks [49,50].

To represent tweets in this study, we used the Cro-CoV-cseBERT language model from [8]. Cro-CoV-cseBERT is based on CroSloEngualBERT [51], a trilingual language model that was pre-trained on a large volume of texts from online news articles in Croatian, Slovenian and English, and additionally fine-tuned on a large corpus of texts related to COVID-19 in Croatian (dataset Cro-CoV-Texts). Cro-CoV-Texts contains 186,738 news articles and 500,504 user comments related to COVID-19 published on Croatian online news portals, as well as 28,208 COVID-19 tweets in Croatian (excluding tweets from the Senti-Cro-CoV-Tweets dataset) [8]. All texts from the dataset used for fine-tuning, were preprocessed following the same procedure as proposed in [52], which includes: replacing usernames, replacing urls, and translating emojis to ASCII code.

3.5. Classification Models

Here we describe six ML models that we trained for binary classification of tweets in our research.

Random forest, (RF) is well-known for taking care of data imbalances in different classes [53,54] especially for large datasets [55].

Multilayer perceptron, (MLP) is another relatively simple model that can be used to perform classification [56].

Light gradient boosting machine, (LGBM) classifier is based on decision trees to increases the efficiency of the model and reduces memory usage. It is described in [57].

Category Embedding Model, (CEM) is the basic model is pretty simple with the pretty simple architecture - a Feed Forward Network with the Categorical Features passed through an learnable embedding layer. It is similar to MLP, but with learned embeddings for category variables.

Neural oblivious decision ensembles, (NODE) for deep learning on tabular data is a model presented in ICLR 2020 [58]. According to the authors have beaten well-tuned gradient boosting models on many datasets. It uses a neural equivalent of oblivious trees (the kind of trees catboost uses) as the basic building blocks of the architecture.

Attentive interpretable tabular learning, (TabNet) is another model coming out of Google Research which uses sparse attention in multiple steps of decision making to model the output [59].

3.6. Data Collection and Experiment Setup

To perform the experiments, data were collected from the social network Twitter. The data were collected automatically using a pipeline for continuous collection of tweets over a long period of time, with the data structure organized so that there are records of users

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 October 2022 doi:10.20944/preprints202210.0086.v1

and their friends, their followers, and all their posts (i.e., published tweets) for a given period of time. The data collection pipeline is organized in such a way that it first collects accounts whose location is in Croatia, and then it collects all their friends and followers, as well as the published tweets of all the previously mentioned profiles.

The collected Twitter dataset (*Cro-Tweets2021*) captures tweets posted in the Croatian language during the period between January 1, 2020 and May 31, 2021. The data were collected using *tweepy* [60], a Python library for accessing the Twitter API.

After preprocessing the tweets and removing tweets without retweet, the final dataset consists of 199,431 tweets. After collecting and cleaning the data, we constructed the corresponding multilayer network $\mathcal{M}$ and multilayer framework $\mathcal{MF}$. Calculation of network measures was performed in Python package *NetworkX* [61].

Next we extracted the multilayer network features and text features. Before feature selection we performed detailed analysis of the features sets including mutual information analysis. The results of the feature analysis are available at https://github.com/InfoCoV/Multi-Cro-CoV-cseBERT/blob/main/notebooks/exploration/features_analysis.ipynb. The whole procedure of collecting and analysing tweets is described in Figure 2.

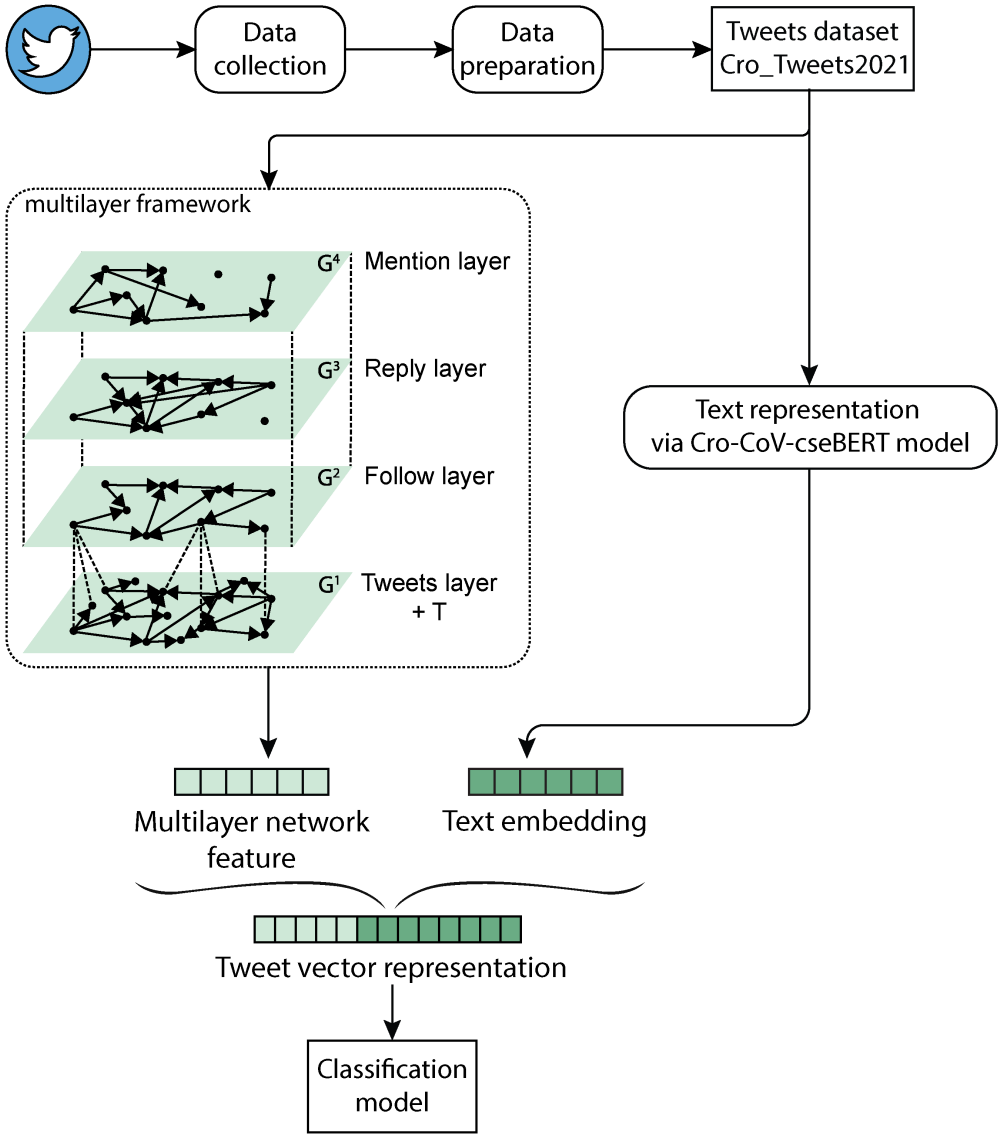

Figure 2. Tweets processing procedure. Detailed steps of tweets processing

 doi:10.20944/preprints202210.0086.v1

The *Cro-Tweets2021* dataset is publicly available at https://github.com/InfoCoV/InfoCoV/blob/main/Cro-Vect-Twitter.csv?fbclid=IwAR0m1Ahk6Jui200DQozGp4eeLa7n8AaBaf53ROLmMOUsSYCMaAvS2LTfwuc.

In the nest step we train six ML classifiers in the task of binary retweet classification: random forest, multilayer perceptron, light gradient boosting machine, category embedding model, neural oblivious decision ensembles and attentive interpretable tabular learning model. For the purpose of training the classification models, we split initial set of tweets, T into training, validation and test sets with an 80:10:10 ratio. It is important to mention that we split the tweets according to the time stamps of tweet.

After training and testing all classifiers, we perform the SHAP analysis [62] to identify the features that have the most impact on the classification.

4. Results

In this section we present the comparison results of the performance of six trained models using three different sets of features. These are features from the text, features from the network and their combination. We trained Random Forest (RF), Multilayer Perceptron (MLP), light gradient boosting machine (LGBM), Category Embedding Model (CEM), Neural Oblivious Decision Ensembles (NODE), and Attentive Interpretable Tabular Learning (TabNet). The evaluation was performed in terms of standard machine learning classification metric: accuracy (Acc), precision (P), recall (R) and $F1$-score ($F1$). Model performance was measured in a macro-averaged setting to ensure equal care for all classes.

Based on the results presented in Table 1, several important observations can be highlighted.

The first observation suggests that classifiers regularly achieve better results on network features than on text features in terms of all considered performance measures (Acc, P, R and $F1$).

Another observation concerns combined features (the union of text and network features), which provide classifiers with even more fruitful ground for inducing classification models. With respect to the standard measure of accuracy (Acc), the classifiers induced from the combined features show a meaningful improvement over those induced from the text features, ranging from 3.9 to up to 7.7%, while with respect to the $F1-score$ this progress ranges from 3.7 to up to 8.2%. Considering the features from the network, we also find that the performance improvement, which favors combined features over network features, is at most 1.4% for Acc and at most 1.7% for $F1-score$. There are also exceptions: for the LGBM classifier, performance remains the same whether features from the network or a combination of features are used, and the exception is the RF classifier, where combined features do not improve performance. In short, the observation based on the results suggests that the features from the network complement the text features well, and in such a combined set achieve better classification performance.

Considering only the most fruitful results obtained with set of combined features, in terms of $F1$-score, CEM is the most successful classification model with 67.9%, while TabNet is the worst with 66.6%. The MLP and NODE models perform well compared to the CEM model, as their performance is only one percentage point lower.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 October 2022 doi:10.20944/preprints202210.0086.v1

	Features	Acc	P	R	F1
RF	Text	0.600	0.608	0.606	0.599
	Network	**0.673**	**0.675**	**0.675**	**0.673**
	Combined	0.671	0.672	0.673	0.671
MLP	Text	0.631	0.632	0.632	0.631
	Network	0.667	0.666	0.662	0.662
	Combined	**0.679**	**0.678**	**0.678**	**0.678**
LGBM	Text	0.600	0.621	0.611	0.595
	Network	0.677	0.680	0.680	**0.677**
	Combined	**0.677**	**0.681**	**0.681**	**0.677**
CEM	Text	0.625	0.629	0.629	0.625
	Network	0.669	0.668	0.666	0.666
	Combined	**0.680**	**0.679**	**0.679**	**0.679**
NODE	Text	0.615	0.624	0.621	0.613
	Network	0.667	0.667	0.661	0.661
	Combined	**0.681**	**0.679**	**0.678**	**0.678**
TabNet	Text	0.630	0.630	0.630	0.629
	Network	0.663	0.662	0.658	0.659
	Combined	**0.669**	**0.667**	**0.666**	**0.666**

Table 1. Comparison of results for six trained models in combination with three different set of features.

Feature analysis

SHapley Additive exPlanations (SHAP), introduced in [62], is used to show the contribution or importance of each feature to the prediction of the model. SHAP values anlysis, in this case for Random Forest model, was performed on a sample of 1k examples from the test set. The absolute SHAP value indicates how much a single feature affected the prediction.

In order to understand the importance or contribution of the features for the whole dataset, the bee swarm plot is illustrated in Figure 3. In this plot, the features are ordered by their effect on the prediction in such a way that the most important feature is listed on the top, and the rest of the list is sorted in descending order. The features importance is determined according to SHAP values which are calculated with a unified framework for interpreting predictions [62] and presented simply with the mean average value for each feature. Features are sorted by the sum of the SHAP value magnitudes across all samples. Besides, plot also illustrates how higher and lower values of the feature affect the outcome. Small dots on the plot represent a single observation. The horizontal axis represents the SHAP value, while the color of the dot shows whether this observation has a higher (red) or lower value (blue) compared to other observations.

The features listed in Figure 3 are in order of global importance, with the first feature being the most important and the last being the least important. For the most important feature - *log1p_followers_count* - it is found to have a very high positive contribution when its values are high, and a very low negative contribution when its values are low. The same applies to the variable *entities.media*, which is second in the order of feature importance. For the third most important feature (*log1p_statuses_count*) high values of the variable were found to make a high negative contribution to prediction, while low values made a high positive contribution. Such conclusions can also be drawn from the plot for all other features. Moreover, it can be seen that some features such as *tweets_keywords_3_out_strength* hardly (or not at all) contribute to prediction, regardless of whether their values are high or low. It is interesting to note how some properties of the network are reflected in features

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 October 2022 doi:10.20944/preprints202210.0086.v1

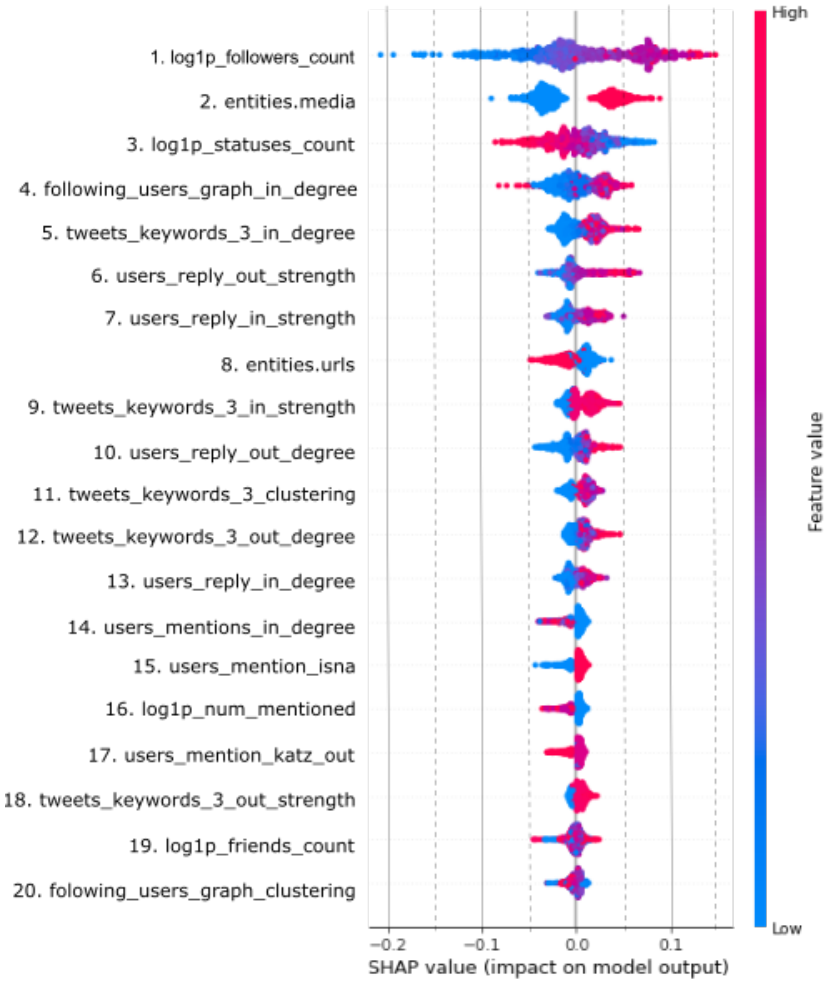

Figure 3. SHAP values analysis on beeswarm summary plot illustrating impact on model output.

that have a stronger impact than other features that reflect other properties of the network. The most important feature is number of followers (*1. log1p_followers_count*), and the first most important feature from the group of centrality measures is follower in-degree (*4. fallowing_users_graph_in_degree*). It is fair to say that the concept of followers plays an important role in the selection of contributing features. Apart from that, keywords are also a superior contributing feature, especially in the form of features resulting from centrality measures in/out-degree, in-strength and clustering coefficient (in Figure 3 those are the features 5, 9, 11 i 12). A detail to note is that in-degree centrality of keywords has a greater impact than out-degree. In terms of layers/graph types, the replay network has spawned a larger number of features with valuable impact than the mention layer/graph. Last but not least, network metadata also makes a satisfactory contribution to the retweet prediction, for example number of followers (follower count), number of changed statuses of the user (statuses count), presence of media in the tweet (entities media) or presence of url in the tweet (entities urls) are important features that are positioned at the top of the list.

5. Discussion

In this study, several aspects related to the retweet prediction task are investigated. The two main objectives were to explore the potential of multilayer representation of the social network for the retweet prediction and to analyse the possibilities of retweet prediction based on heterogeneous data sources.

Overall, multilayer network features perform better than text features for all six trained algorithms. According to that, we can confirm that multilayer network representation of

 doi:10.20944/preprints202210.0086.v1

Twitter have the great potential for retweet prediction. These findings are inline with results of the study [20] in which Pierri et al. have shown that multilayer network features perform better in the task of disinformation classification on Twitter. Although, this study modelled Twitter differently from here proposed $\mathcal{MF}$ and used different network measures, all these results indicate that multilayer approach in the task of retweet prediction is worth of further examination.

Furthermore, according to the results presented in the previous section we can conclude that the combination of multilayer network features with text features in general perform better than only one set of feature in the task of retweet amount prediction. We have to emphasize that the combination of features only slightly outperforms the multilayer network features, however this is the consistent for five from six models (only the random Forest algorithm have better performance in the case of multilayer network features). The potential of combination of features from heterogeneous data sources has been considered in several studies before [3,13–15,26,27] and it has been shown that combination of features is better than one features. Specifically, Suh et al. [15] also examined content and contextual features in the task of retweet prediction, but in much less extent than our study. To the best of our knowledge the combination of these two sets of features have not been examined already.

The further feature analysis performed by SHAP, indicates that among network measures, in-degree calculated from the follower network layer and in-degree calculated from the twitter network layer has the major impact to the model. Besides that, as expected, higher number of followers and some other metadata such as presence of media or url have a positive influence on retweeting.

Another important aspect of this research is the comparison of the performance of six different ML algorithms in the task of retweet prediction. We identify the CEM model as the one with the best performance according to all used evaluation measures in all three feature sets scenarios, while the overall lowest performance is achieved in the case of TabNet model. Again, it has to be emphasised that differences across all algorithms are not so significant. The only significant difference is in the performance of models using only text features in comparison to multilayer network set of features which seems to be significantly better for multilayer network features (as well as for combined features) for all six models. This is again indicator that multiyear network features have great potential in analysis of information spreading.

This research is an extension of our previous studies of online communication on social media during the COVID-19 pandemic. In [63] we compared the retweeting of COVID-19 related tweets and tweets that are not related to COVID-19. Our findings indicate that tweets that almost 60% of tweets related to COVID-19 belong to the high-spreadable class; while less than 40% of non COVID tweets belong to this high spreadable class. This suggests that tweet content may have a high impact to the retweeting (spreadability), especially during the global crisis, such as COVID-19 pandemic. In another study [64], we explored the potential of graph neural networks (GNNs) in the task of prediction if the user would tweet about COVID19 or not.

This research have several limitations that we plan to address in the future work. First, our results are not directly comparable to other studies, because we modelled the task of retweet prediction as the binary classification task into two classes: (i) class of tweets with only one retweet and (ii) class of tweets with more than one retweet. This way we try to predict if the amount of retweets would be poor or not, but we did not take into account tweets that are not retweeted at all. We decided to discard all tweets with no retweets because there are too many reasons why the tweet is not retweeted and this may negatively affect the prediction. We assumed that the prediction models would perform better if we concentrated only on the dataset of retweeted tweets in this first step. In addition we used this setup because, the structure of the dataset of Tweets, this way we ensured a balanced classes. However, in the future research we plan to include tweets with no retweet into the prediction task. Another limitation is that we used only one dataset of tweets to compare

 doi:10.20944/preprints202210.0086.v1

the performance of features and ML models. However, this dataset is a representative sample of tweets in the Croatian language posted during the pandemic years 2020 and 2021 and our intention was to analyse the crisis-related communication in the Croatia during the COVID-19 pandemic period. That is the reason why we trained and compared ML models on this specific dataset of Tweets.

6. Conclusions

In this paper we introduce a multilayer framework formalism for representation of online communication on social media. We utilized this formalism for feature extraction from heterogeneous data sources: multilayer network and text message. We performed detailed analysis of possible features and a combination of network and multilayer features in the task of binary classification of tweets according to the amount of retweeting.

The main focus of this research is to compare the performance of different sets of features and its combination. In addition, we evaluated six different ML classification models: random forest, multilayer perceptron, light gradient boosting machine, category embedding model, neural oblivious decision ensembles and attentive interpretable tabular learning model.

According to the overall results, solely multilayer network features performed significantly better than solely text features for all six algorithms. Overall, our results indicate that the structural features of Twitter represented as the multilayer network might be effectively exploited in the retweeting prediction task.

The combination of both feature sets has the best performance in the case of all classification models, except the random forest. We identify that the category embedding model with the has the best performance according to the F1-score which is 0.679. However, this result is only slightly better than results of other algorithms, and we can conclude that all six algorithms has similar performance in the task of retweet classification. Additionally we explored the impact of different features using SHAP analysis and determine that number of followers in the network, presence of media, number of changed user statues, in-degree on the follower network layer, in-degree on the twitter network layer features has the major impact to the model. Thus, we believe that our multilayer network-based approach provides useful insights to the future development of a system for prediction information spreading on social media.

The proposed approach can be further extended in the several directions and we have a plenty of plans for the future work. Firstly, we plan to test more multilayer network measures as predictors and also to explore the potential of deep learning automatic feature extraction from the multilayer network in the task of retweet prediction. Secondly, we plan to extend the multilayer framework model with the dynamic aspect (in the sense that we capture the dynamics of users' actions) and to use three sets of features for prediction of retweeting and information spreading on social media in general. Thirdly, we plan to utilize graph neural networks for link prediction.

Author Contributions: Conceptualization, A.M.; data curation, M.P.; formal analysis, S.B.; funding acquisition, A.M.; investigation, A.M., M.P. and S.B.; methodology: A.M. and S.B.; software, M.P. and S.B.; supervision, A.M.; visualization, validation, A.M. and S.B.; writing – original draft, A.M. and S.B.; writing – review editing, A.M. and S.B. All authors have read and agreed to the published version of the manuscript.

Funding: This work has been supported in part by the Croatian Science Foundation under the project IP-CORONA-04-2061, "Multilayer Framework for the Information Spreading Characterization in Social Media during the COVID-19 Crisis" (InfoCoV), and by University of Rijeka project number uniri-drustv-18-38.

Institutional Review Board Statement: Not applicable.

Data Availability Statement: The data presented and used in this study (*Cro-Tweets2021* dataset) are openly available at https://github.com/InfoCoV/InfoCoV/blob/main/Cro-Vect-Twitter.csv?fbclid=IwAR0m1Ahk6Jui200DQozGp4eeLa7n8AaBaf53ROLmMOUsSYCMaAvS2LTfwuc.

Acknowledgments: We would like to thank Velebit AI, especially Mladen Fernežir for leading the implementation of the classifiers.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

Acc	Accuracy
ASCII	American Standard Code for Information Interchanges
AUROC	Area Under the Receiver Operating Characteristics
BERT	Bidirectional Encoder Representations from Transformers
CEM	Category Embedding Model
COVID-19	Corona Virus Disease-19
ELMo	Embeddings from Language Models
F1	F1-score
GloVe	Global Vectors for Words Representations
GNN	Graph Neural Networks
GPT	Generative Pre-trained Transformer
LGBM	Light Gradient Boosting Machine
ML	Machine Learning
MLP	Multilayer Perceptron
NLP	Natural Language Processing
NODE	Neural Oblivious Decision Ensembles
P	Precision
R	Recall
RF	Random Forest
SHAP	SHapley Additive exPlanations
TabNet	Attentive Interpretable Tabular Learning

References

1. Carlos Cuello-Garcia, Giordano Pérez-Gaxiola, L.v.A. Social media can have an impact on how we manage and investigate the COVID-19 pandemic. *J Clin Epidemiol.* **2020**, *127*, 198–201. https://doi.org/10.1016/j.jclinepi.2020.06.028.
2. Bunker, D. Who do you trust? The digital destruction of shared situational awareness and the COVID-19 infodemic. *International Journal of Information Management* **2020**, *55*, 102201. Impact of COVID-19 Pandemic on Information Management Research and Practice: Editorial Perspectives, https://doi.org/https://doi.org/10.1016/j.ijinfomgt.2020.102201.
3. Firdaus, S.N.; Ding, C.; Sadeghian, A. Retweet Prediction based on Topic, Emotion and Personality. *Online Social Networks and Media* **2021**, *25*, 100165.
4. Wang, J.; Yang, Y. Tweet retweet prediction based on deep multitask learning. *Neural Processing Letters* **2022**, *54*, 523–536.
5. Eysenbach, G. Infodemiology: The epidemiology of (mis) information. *The American journal of medicine* **2002**, *113*, 763–765.
6. Petrović, M.; Levnajić, Z.; Meštrović, A. Analysis of the COVID-19 Communication on Twitter via Multilayer Network. In Proceedings of the The 2nd International Symposium on Automation, Information and Computing (ISAIC 2021), 2022, Vol. December 9-11, 2022.
7. Malecki, K.M.; Keating, J.A.; Safdar, N. Crisis communication and public perception of COVID-19 risk in the era of social media. *Clinical Infectious Diseases* **2021**, *72*, 697–702.
8. Babić, K.; Petrović, M.; Beliga, S.; Martinčić-Ipšić, S.; Matešić, M.; Meštrović, A. Characterisation of COVID-19-related tweets in the Croatian language: framework based on the Cro-CoV-cseBERT model. *Applied Sciences* **2021**, *11*, 10442.
9. Jay, A. FinancesOnline. https://financesonline.com/number-of-twitter-users/, 2021. Accessed: 2022-07-01.
10. Kuang, S.; Davison, B.D. Learning Word Embeddings with Chi-Square Weights for Healthcare Tweet Classification. *Applied Sciences* **2017**, *7*. https://doi.org/10.3390/app7080846.
11. Singh, C.; Imam, T.; Wibowo, S.; Grandhi, S. A Deep Learning Approach for Sentiment Analysis of COVID-19 Reviews. *Applied Sciences* **2022**, *12*. https://doi.org/10.3390/app12083709.
12. Zhang, Q.; Gong, Y.; Wu, J.; Huang, H.; Huang, X. Retweet prediction with attention-based deep neural network. In Proceedings of the Proceedings of the 25th ACM international on conference on information and knowledge management, 2016, pp. 75–84.
13. Yin, H.; Yang, S.; Song, X.; Liu, W.; Li, J. Deep fusion of multimodal features for social media retweet time prediction. *World Wide Web* **2021**, *24*, 1027–1044.
14. Tsur, O.; Rappoport, A. What's in a hashtag? Content based prediction of the spread of ideas in microblogging communities. In Proceedings of the Proceedings of the fifth ACM international conference on Web search and data mining, 2012, pp. 643–652.

15. Suh, B.; Hong, L.; Pirolli, P.; Chi, E.H. Want to be retweeted? large scale analytics on factors impacting retweet in twitter network. In Proceedings of the 2010 IEEE second international conference on social computing. IEEE, 2010, pp. 177–184.
16. Boccaletti, S.; Bianconi, G.; Criado, R.; Del Genio, C.I.; Gómez-Gardenes, J.; Romance, M.; Sendina-Nadal, I.; Wang, Z.; Zanin, M. The structure and dynamics of multilayer networks. *Physics reports* **2014**, *544*, 1–122.
17. Kivelä, M.; Arenas, A.; Barthelemy, M.; Gleeson, J.P.; Moreno, Y.; Porter, M.A. Multilayer networks. *Journal of complex networks* **2014**, *2*, 203–271.
18. Martinčić-Ipšić, S.; Margan, D.; Meštrović, A. Multilayer network of language: A unified framework for structural analysis of linguistic subsystems. *Physica A: Statistical Mechanics and its Applications* **2016**, *457*, 117–128.
19. Vukić, D.; Martinčić-Ipšić, S.; Meštrović, A. Structural analysis of factual, conceptual, procedural, and metacognitive knowledge in a multidimensional knowledge network. *Complexity* **2020**, *2020*.
20. Pierri, F.; Piccardi, C.; Ceri, S. A multi-layer approach to disinformation detection in US and Italian news spreading on Twitter. *EPJ Data Science* **2020**, *9*, 35.
21. Zaman, T.R.; Herbrich, R.; Van Gael, J.; Stern, D. Predicting information spreading in twitter. In Proceedings of the Workshop on computational social science and the wisdom of crowds, nips. Citeseer, 2010, Vol. 104, pp. 17599–601.
22. Kupavskii, A.; Ostroumova, L.; Umnov, A.; Usachev, S.; Serdyukov, P.; Gusev, G.; Kustarev, A. Prediction of retweet cascade size over time. In Proceedings of the Proceedings of the 21st ACM international conference on Information and knowledge management, 2012, pp. 2335–2338.
23. Moreno, Y.; Pastor-Satorras, R.; Vespignani, A. Epidemic outbreaks in complex heterogeneous networks. *The European Physical Journal B-Condensed Matter and Complex Systems* **2002**, *26*, 521–529.
24. Yang, R.; Wang, B.H.; Ren, J.; Bai, W.J.; Shi, Z.W.; Wang, W.X.; Zhou, T. Epidemic spreading on heterogeneous networks with identical infectivity. *Physics Letters A* **2007**, *364*, 189–193.
25. Ikeda, K.; Okada, Y.; Toriumi, F.; Sakaki, T.; Kazama, K.; Noda, I.; Shinoda, K.; Suwa, H.; Kurihara, S. Multi-agent information diffusion model for twitter. In Proceedings of the 2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT). IEEE, 2014, Vol. 1, pp. 21–26.
26. Dai, T.; Xiao, Y.; Liang, X.; Li, Q.; Li, T. ICS-SVM: A user retweet prediction method for hot topics based on improved SVM. *Digital Communications and Networks* **2022**, *8*, 186–193.
27. Ma, R.; Hu, X.; Zhang, Q.; Huang, X.; Jiang, Y.G. Hot topic-aware retweet prediction with masked self-attentive model. In Proceedings of the Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval, 2019, pp. 525–534.
28. Amitani, R.; Matsumoto, K.; Yoshida, M.; Kita, K. Buzz Tweet Classification Based on Text and Image Features of Tweets Using Multi-Task Learning. *Applied Sciences* **2021**, *11*. https://doi.org/10.3390/app112210567.
29. Omodei, E.; De Domenico, M.D.; Arenas, A. Characterizing interactions in online social networks during exceptional events. *Frontiers in Physics* **2015**, *3*, 59.
30. Oro, E.; Pizzuti, C.; Procopio, N.; Ruffolo, M. Detecting topic authoritative social media users: a multilayer network approach. *IEEE Transactions on Multimedia* **2017**, *20*, 1195–1208.
31. Magnani, M.; Rossi, L. The ml-model for multi-layer social networks. In Proceedings of the 2011 International Conference on Advances in Social Networks Analysis and Mining. IEEE, 2011, pp. 5–12.
32. Hristova, D.; Noulas, A.; Brown, C.; Musolesi, M.; Mascolo, C. A multilayer approach to multiplexity and link prediction in online geo-social networks. *EPJ Data Science* **2016**, *5*, 24.
33. Perc, M. Diffusion dynamics and information spreading in multilayer networks: An overview. *The European Physical Journal Special Topics* **2019**, *228*, 2351–2355.
34. De Domenico, M.; Granell, C.; Porter, M.A.; Arenas, A. The physics of spreading processes in multilayer networks. *Nature Physics* **2016**, *12*, 901–906.
35. Bródka, P.; Musial, K.; Jankowski, J. Interacting spreading processes in multilayer networks: a systematic review. *IEEE Access* **2020**, *8*, 10316–10341.
36. Matas, N. Comparing Network Centrality Measures as Tools for Identifying Key Concepts in Complex Networks: A Case of Wikipedia. *Journal of Digital Information Management* **2017**, *15*.
37. Newman, M. *Networks*; Oxford university press, 2018.
38. Bonacich, P. Power and centrality: A family of measures. *American journal of sociology* **1987**, *92*, 1170–1182.
39. Katz, L. A new status index derived from sociometric analysis. *Psychometrika* **1953**, *18*, 39–43.
40. Opsahl, T.; Panzarasa, P. Clustering in weighted networks. *Social Networks* **2009**, *31*, 155–163. https://doi.org/https://doi.org/10.1016/j.socnet.2009.02.002.
41. Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.S.; Dean, J. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems* **2013**, *26*.
42. Le, Q.; Mikolov, T. Distributed representations of sentences and documents. In Proceedings of the International conference on machine learning. PMLR, 2014, pp. 1188–1196.
43. Pennington, J.; Socher, R.; Manning, C.D. Glove: Global vectors for word representation. In Proceedings of the Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 2014, pp. 1532–1543.

44. Bojanowski, P.; Grave, E.; Joulin, A.; Mikolov, T. Enriching word vectors with subword information. *Transactions of the association for computational linguistics* **2017**, *5*, 135–146.
45. Peters, M.E.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; Zettlemoyer, L. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365* **2018**.
46. Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I.; et al. Language models are unsupervised multitask learners. *OpenAI blog* **2019**, *1*, 9.
47. Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language models are few-shot learners. *Advances in neural information processing systems* **2020**, *33*, 1877–1901.
48. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* **2018**.
49. Ethayarajh, K. How contextual are contextualized word representations? comparing the geometry of BERT, ELMo, and GPT-2 embeddings. *arXiv preprint arXiv:1909.00512* **2019**.
50. Babić, K.; Martinčić-Ipšić, S.; Meštrović, A. Survey of neural text representation models. *Information* **2020**, *11*, 511.
51. Ulčar, M.; Robnik-Šikonja, M. Finest bert and crosloengual bert. In Proceedings of the International Conference on Text, Speech, and Dialogue. Springer, 2020, pp. 104–111.
52. Müller, M.; Salathé, M.; Kummervold, P.E. Covid-twitter-bert: A natural language processing model to analyse covid-19 content on twitter. *arXiv preprint arXiv:2005.07503* **2020**.
53. Khoshgoftaar, T.M.; Golawala, M.; Van Hulse, J. An empirical study of learning from imbalanced data using random forest. In Proceedings of the 19th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2007). IEEE, 2007, Vol. 2, pp. 310–317.
54. Liu, X.Y.; Wu, J.; Zhou, Z.H. Exploratory undersampling for class-imbalance learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* **2008**, *39*, 539–550.
55. Dietterich, T.G. Ensemble methods in machine learning. In Proceedings of the International workshop on multiple classifier systems. Springer, 2000, pp. 1–15.
56. Ruck, D.W.; Rogers, S.K.; Kabrisky, M. Feature selection using a multilayer perceptron. *Journal of Neural Network Computing* **1990**, *2*, 40–48.
57. Kumar, S.; Mallik, A.; Panda, B. Link prediction in complex networks using node centrality and light gradient boosting machine. *World Wide Web* **2022**, pp. 1–27.
58. Popov, S.; Morozov, S.; Babenko, A. Neural oblivious decision ensembles for deep learning on tabular data. *arXiv preprint arXiv:1909.06312* **2019**.
59. Arık, S.O.; Pfister, T. Tabnet: Attentive interpretable tabular learning. In Proceedings of the AAAI, 2021, Vol. 35, pp. 6679–6687.
60. Roesslein, J. tweepy Documentation. *Online] http://tweepy. readthedocs. io/en/v3* **2009**, *5*.
61. Hagberg, A.A.; Schult, D.A.; Swart, P.J. Exploring network structure, dynamics, and function using NetworkX. In Proceedings of the Proceedings of the 7th Python in Science Conference; , 2008; pp. 11 – 15.
62. Lundberg, S.M.; Lee, S.I. A unified approach to interpreting model predictions. *Advances in neural information processing systems* **2017**, *30*.
63. Babić, K.; Petrović, M.; Beliga, S.; Martinčić-Ipšić, S.; Pranjić, M.; Meštrović, A. Prediction of COVID-19 related information spreading on Twitter. In Proceedings of the 2021 44th International Convention on Information, Communication and Electronic Technology (MIPRO). IEEE, 2021, pp. 395–399.
64. Petrović, M.; Hrelja, A.; Meštrović, A. Prediction of COVID-19 tweeting: classification based on graph neural networks. In Proceedings of the 2022 45th Jubilee International Convention on Information, Communication and Electronic Technology (MIPRO). IEEE, 2022, pp. 307–311.