applied sciences

MDPI

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 spreading analysis, automatic fake news detection, social media monitoring, etc. In this study, we explore retweet prediction based on heterogeneous data sources. In order to classify a tweet according to the number of retweets, we combine features extracted from the multilayer network and text. More specifically, we introduce a multilayer framework for the multilayer network representation of Twitter. This formalism captures different users' actions and complex relationships, as well as other key properties of communication on Twitter. Next, we select a set of local network measures from each layer and construct a set of multilayer network features. We also 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 trained six machine learning (ML) algorithms: 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 compared 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 evaluated all the setups in terms of standard evaluation measures. For this task, we first prepared an empirical dataset of 199,431 tweets in Croatian posted between 1 January 2020 and 31 May 2021. Our results indicate that the prediction model performs better by integrating multilayer network features with text features than by using only one set of features.

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

Citation: Meštrović, A.; Petrović, M.; Beliga, S. Retweet Prediction Based on Heterogeneous Data Sources: The Combination of Text and Multilayer Network Features. *Appl. Sci.* **2022**, *12*, 11216. https://doi.org/10.3390/app122111216

Academic Editor: Christos Bouras

Received: 23 September 2022
Accepted: 3 November 2022
Published: 5 November 2022

1. Introduction

Nowadays, social media platforms and online social networks have become an important source of information. Users tend to rely on online communication platforms for getting information, publishing posts that reflect their interests, views, and activities [1]. At the same time, users can express their opinions about other posts via different forms of feedback, such as reposts, quotes, mentions, replies, likes, etc. All these activities affect information spreading on social media [2]. In the last two decades, online social networks have increased the spread of information, but also misinformation and disinformation, which can lead to an infodemic and other negative side effects [3,4]. Therefore, exploring patterns of information spreading on social networks is a significant aspect of research in the domain of disinformation detection and infodemic. The primary motivation for this research was the analysis of crisis-related communication on social networks. However, the proposed approach can be applied in other domains for tasks related to retweet prediction.

Because social media platforms such as Twitter, Facebook, Instagram, and Weibo play an increasingly important role, people are more likely to turn to them in search of information during a global crisis. They can serve as an essential communication platform in real-world crises, emergencies, or disasters [5,6]. Social media may even influence the

course of 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 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 a global crisis is of great importance as one of the steps 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, it has frequently been used for monitoring and tracking different aspects of healthcare information and public disease [8,10,11]. Among all the user behavior on social media, retweets are considered to be one of the primary ways of 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 by using various linguistic features, personal information of users, or network properties [12].

In some recent studies, authors combined heterogeneous data sources: information content, network structure, dynamics of spreading, information metadata, and other properties that can be referred to as heterogeneous data sources [1,13–17]. However, there are properties that have not been fully explored in the task of retweet prediction. One less-studied approach is the use of 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 [18,19]. 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. Because 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 predictors of retweeting. We have already shown that a multilayer network structure is fundamentally more expressive than individual layers in examples of modeling a multilayer language network [20] and multidimensional knowledge network [21]. In [22], 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 time that anyone has attempted to use this set of multilayer network features in the task of retweet prediction and the first time anyone has attempted to combine multilayer network features with text features. We formalized our approach by introducing a multilayer framework for the representation of key elements of communications on social networks.

The main objective 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 whether 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 within which a message is spreading. For the purpose of retweeting prediction, we propose and construct a multilayer network representation with four layers representing actions of following, mentioning, replying, and a layer of tweets, and select several network measures from each layer. Text features are represented as a low-dimensional vector (embedding) that captures its semantics and structure. More specifically, we adopt a

BERT-based language model, namely Cro-CoV-cseBERT [8] for representation of tweets as embeddings, which we use as a set of text features.

We model the prediction problem as a binary classification task in which the first class contains tweets with just one retweet and the second class contains tweets with more than one retweet. Next, we explore the performance of different feature sets by conducting an extensive set of experiments in which we train six machine learning models 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 trained the following 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 evaluated the performance of trained classifiers on three different sets of features in terms of standard evaluation measures: accuracy, precision, recall, and F1 score on a large dataset of tweets. For this purpose, we prepared an empirical dataset of 199,431 tweets in the Croatian language posted during the pandemic period between 1 January 2020 and 31 May 2021.

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

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

1. We propose a multilayer framework formalism for Twitter representation based on multilayer network and select a set of measures from each layer to be extracted and combined with the metadata as the set of multilayer network features.
2. We conducted a set of experiments on a dataset of tweets using separate text features and multilayer network features and their combination and evaluated the performance of six machine learning classifiers.
3. We performed an analysis of feature importance to determine the impact of two sets of features for the task of retweet prediction and studied various multilayer network features chosen by using SHAP.

The rest of the paper is organized as follows. Section 2 discusses some of the existing research 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 extensively studied by a number of researchers. There are many different ways to approach the problem of retweet prediction. It can be modelled as a 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 a regression/prediction problem in which the model should predict the number of retweets for a given tweet [23]. 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 user [24] or users' retweet behaviour prediction [15], retweet time prediction [13] or the prediction of the size of retweet cascade size as in [25]. In the domain of complex networks, this task is usually described as the link prediction in the network of retweeting. From a broader perspective, spreading patterns have been studied in many fields ranging from disease spreading [26,27] to information spreading on social networks [28].

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

personal information of users such as user profiles, user post history, user following relationships, network properties, etc. There are studies that try to predict retweeting based solely on text properties [29,30] or based on topic and emotion extracted from text [31]. On the other hand, many authors use only network properties to predict retweeting [32,33].

Some recent studies show the advantages of combining features from different, heterogeneous sources: information content, network structure, dynamics of spreading, information metadata, and other properties that can be referred to as heterogeneous data sources [1,13]. There are still some combinations of heterogeneous data sources worth exploring as predictors of spreading. One such approach is combining multilayer network features with text features proposed in this study. In the following subsections we give an overview of two research domains related to the proposed approach: (i) research studies that combine features from heterogeneous data sources in the task of retweet prediction, and (ii) research of the multilayer network approach.

2.1. Retweet Prediction Based on Heterogeneous Data Sources

Usually, in approaches based on heterogeneous data sources authors combine features extracted from the message text and from the social network. There have been attempts to combine content features (extracted from text) and contextual features (extracted from networks) even in some earlier research studies. However, the proposed feature sets are relatively simple and include only partial information extracted from the text and networks. Thus, Suh et al. [34] 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 revealed that, when it comes to content features, URLs and hashtags have a strong influence on retweeting. When it comes to contextual features, the number of followers and followees, as well as the age of the account, seem to affect retweetability. Similarly, in [35] the authors studied the effect of the message content in the task of the spreading prediction of ideas. 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.

In one more recent study, Sharma and Gupta [14] explored the impact of different numerical features extracted from tweet content and network in the task of retweet prediction. They proposed three features from the author's profile that can capture the behavioral pattern of the user: author total activity, author activity per year, and author tweets per year. They performed their experiment by using a large dataset (Scott, Jason, and Sketch the Cow. "Archiveteam-Twitter-Stream-2018-08: Free Download, Borrow, and Streaming". Internet Archive, Archive Team: The Twitter Stream Grab: https://archive.org/details/archiveteam-twitter-stream-2018-08, accessed on 20 September 2022) containing 100 million random tweets from the online twitter archive of August 2018. Their results showed that the proposed model has better accuracy when user features are combined with tweet content features. Yin et al. [13] combined text and network features and proposed a novel deep fusion of multimodal features (DFMF) method for retweet time prediction. Their 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 proposed method is evaluated on the real-world Twitter dataset (UDI-Twitter Crawl-Aug20123: https://wiki.illinois.edu/wiki/display/forward/Dataset-UDI-TwitterCrawl-Aug2012, accessed on 20 September 2022) that contains 284 million following relationships, 3 million user profiles, and 50 million tweets. The evaluation results showed that the proposed method was 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).

The properties used for retweet prediction may be extracted from the author of the tweet (author-centered prediction) or from the user who will retweet (user-centered prediction). Thus, some recent studies combine features extracted from author, tweet text, and user. For example, Fu et al. [15] proposed a prediction model MDF-RP (multidimensional feature-based retweeting prediction) that combines features extracted from three different sources: author, tweet, and user in order to predict if the user will retweet the given tweet message. The evaluation of the proposed model was performed on the dataset of messages crawled from Weibo (social network in China, similar to Twitter) from 1 June 2018 to 31 July 2018 with 3352 users and 316,829 tweets (the dataset is available upon request). Their results based on different classifiers showed that the performances of MDF-RP outperformed the basic features in terms of precision, recall, and F1 score.

On the other hand, Fridaus et al. performed only user-centered prediction [1]. They 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 of which uses classification algorithms, and the other matrix factorization algorithms. For their experiment, authors collected a dataset of tweets (dataset of Twitter IDs: https://github.com/snadiaf/Twitter-Data, accessed on 20 September 2022). The evaluation 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. Similarly, Ma et al. [17] 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 results obtained on a dataset collected from Twitter with a total of 411,054 users and 36,807,681 tweets showed that the proposed method can perform better than state-of-the-art methods. Dai et al. [16] improved the SVM model for the prediction of user forwarding behavior of hot topics which was also based on user features. The prediction of user retweeting behavior is based on combining three different data sources: user interest tags, user history behavior, and external factors influence. For the purpose of the experiment, the collected dataset of tweets was divided into a training dataset and a test dataset (the training and test datasets are available at: https://www.sciencedirect.com/topics/engineering/test-dataset, accessed on 20 September 2022). In addition to retweet prediction, heterogeneous feature sources have also been successfully used to predict buzz tweets. Amitani et al. [36], in their study on the classification of "buzz" tweets, examine the trends on social media and propose a classification method to 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 texts posted on Twitter. However, they developed a multitask neural network by 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 by 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 when using the text or image features alone [36].

2.2. Multilayer Network-Based Approach in Social Networks Representation

As can be seen from the previous section, there are many possibilities in combining heterogeneous data sources for feature engineering in the task of retweet prediction, and almost all of them include some features extracted from the social network, involving author-centered properties, user-centered properties or a combination of both. However, there is still a dearth of research studies that represent a social network as a multilayer network which can capture different users' actions (such as reply, quote, mention, and

follow) as separate layers. As we claim in the introduction section, network properties extracted from different layers provide more detailed insight into online communication, and based on this it is possible to better predict the action of retweeting. Thus, we expect that multilayer network measures combined with text features have great potential in the prediction of retweeting.

One such study carried out by Pierri et al. [22] modeled Twitter as a multilayer network including four layers: mention, reply, retweet, and quote layer. They applied multilayer network representation of Twitter in the task of disinformation detection in US and Italian news spreading over Twitter. More precisely, they used multilayer network measures for classifying news articles pertaining to disinformation vs. mainstream news by solely inspecting their diffusion mechanisms on Twitter. They trained a logistic regression model to classify disinformation vs. mainstream networks on two large-scale datasets of diffusion cascades (tweets for United States and Italy). The proposed approach has a high accuracy (AUROC up to 94%) in the task of disinformation detection and suggests that a similar approach based on multilayer networks might be possible for the task of information-spreading prediction in general.

There are some other studies that used multilayer networks to model different aspects of online social networks communication. Thus, Arenas et al. [37] modelled 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 [38], the authors proposed a method based on a multilayer approach-capable of identifying influencers on online social networks. The layers represent users, items, and keywords, along with the intralayer 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 [39]. 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 [40], the authors explored two online platforms, Twitter and Foursquare, analyzing the geosocial properties of links. They represented the two platforms as a composite multilayer online social network, wherein each platform represents a layer in the network. According to their results, by using the multilayer approach it is possible to successfully predict links across social networking services. It is also worth mentioning that in [41–43] the authors investigated the spreading patterns in multilayer networks. However, they did not apply machine learning algorithms, and 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 proven the potential of multilayer networks when it comes to capturing important properties of online communication. In our research, we adopted some of these ideas. However, we have modeled Twitter differently and utilized network measures that differ from all previous approaches.

3. Materials and Methods

3.1. Multilayer Framework Definition

In this section we introduce a multilayer framework, as a formalism that can capture various aspects of message spreading on Twitter. This is an extension of our previous work published in [4], in which we proposed a communication multilayer framework for representing communication on social media. Here, we define a framework based on the multilayer network that we use to represent different users' actions on Twitter. Within the multilayer framework, we aggregate the multilayer social network with a set of metadata corresponding to text messages published on social media. Next we select set of network

measures from each layer and define a set of multilayer network features used to train six ML models for retweet prediction.

According to [18], a multilayer network is defined as a pair:

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

where

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

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

Similar to work presented in [4], layers are annotated as numbers from the set $\{1, \ldots, m\}$, where m is the number of layers. Like one-layer networks, multilayered networks can be directed or undirected, weighted or unweighted. Note that communication in social networks is best described by using a weighted and directed multilayer network.

Next, we introduce and consider a set T of metadata related to text messages posted on social networks. 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 the 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 by Using Multilayer Network

Given the framework $\mathcal{MF}$ defined according to the (3), we model a Twitter network as the multilayer framework consists of four layers, thus $m = 4$. Each layer represents one aspect of communication on Twitter as follows.

The first layer is the tweet layer, $G^1 = (V^1, E^1)$, in which nodes represent Twitter messages and two nodes i and j are connected with the directed link if message i and j have at least three words and/or hashtags in common. The direction of the link is defined according to the timeline; from the first tweet to the second tweet. The link weight is defined as the number of common words/hashtags. The second layer is the follower layer, $G^2 = (V^2, E^2)$, in which nodes represent Twitter users and two nodes i and j are connected with the directed link if user j follows user i. This is an unweighted network; however, the whole network is weighted, and thus all weights in this layer are set to 1. The third layer is the reply layer, $G^3 = (V^3, E^3)$, in which nodes represent Twitter users and two nodes i and j are connected with the directed link if user j replies to user i. The weight is defined as the number of replies. The fourth layer is is a mention layer, $G^4 = (V^4, E^4)$, in which nodes represent Twitter users and two nodes i and j are connected with the directed link if user j mentions user i. The weight is defined as the number of mentions. Next, we define a set of interconnections between nodes of different layers. The first layer of tweets (tweet layer) is connected with the second layer (follower layer) in the way that there is a directed link from the user (follower) to every tweet that he/she posts. The reply layer is connected with the tweet layer in the way that there is a directed link from the user to the tweet if the user replies to this tweet. Analogously, the mention layer is connected with the tweet layer in the way that there is a directed link from the user to the tweet if the user mentions this tweet. The rest of the layers, G^2, G^3 and G^4 represent a multiplex network in which interconnections are established between the same nodes. A multiplex network is a special case of the multilayer network in which interlayer links can only connect nodes that represent the same node in different layers.

The model of Twitter represented as the multilayer network is illustrated in Figure 1. The figure is taken from [4] and adapted to the experiment of this study.

Further explanations and details of the multiplex, which is illustrated in Figure 1, the connections between the nodes of the same or different layers, and the weights can be found in [4].

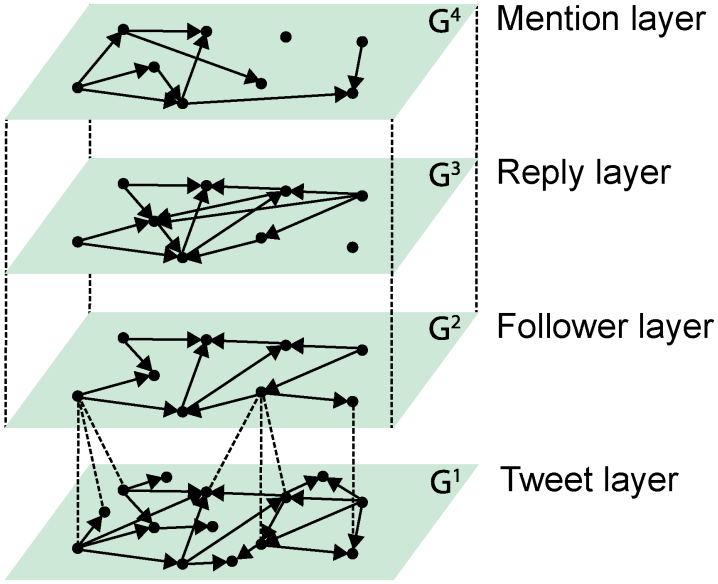

Figure 1. Twitter represented via multilayer network. Communication on Twitter captured via four layers of multilayer network: G^1, tweet layer; G^2, follower layer; G^3, reply layer; and G^4, mention layer. The interconnections between nodes of the tweet layer and the follower layer are established if the user (node) from G^2 posts a tweet (node) from G^1. Other interlayer links are not represented in this figure due to the better visibility; however, G^2, G^3 and G^4 are connected as multiplex network as explained in the text above.

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 nodes are the most popular (hubs), how close are nodes with each other, and which nodes control 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 [44].

Degree centrality of a node is the measure that takes into account the 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. However, if we capture more than one layer, degree centrality may also indicate the node with the largest number of mentions or replies. A higher degree implies popularity, and a higher possibility to gain information that is flowing through the network. According to [45], 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 $N - 1$:

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

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 [45]:

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

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

Eigenvector centrality is introduced by Bonacich [46]. 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 that 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 equal connections to low-scoring nodes. For the node i and constant λ centrality ce_i of node i is defined as [46]:

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

Eigenvector centrality computes the centrality for a node based on the centrality of its neighbors. The eigenvector centrality for node i is the ith element of the vector x defined by the equation [45]

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

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 λ is the largest eigenvalue of the adjacency matrix [45].

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 [47] 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 [47]:

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

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 correspond to the in-edges in the graph. For out-edge 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 [48]

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

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 five centrality measures, we calculate values for in- and out-centrality measure (except clustering coefficient which is undirected) according to the equations defined in (4)–(6), (8), (9). As a result, we have nine 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 a 49-dimensional vector as the representation of tweets extracted from the multilayer framework $\mathcal{MF}$.

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 certainly involves the development of a new sophisticated model or the choice of an existing language model that includes, e.g., semantic, syntactic and other linguistic features of the text. The seminal work of [49] 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 [49], Doc2Vec [50], GloVe [51], and fastText [52]. 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 [53], GPT/GPT-2 [54], GPT-3 [55], and BERT [56]. 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 [57,58].

To represent tweets in this study, we used the Cro-CoV-cseBERT language model from [8]. Cro-CoV-cseBERT is based on CroSloEngualBERT [59], 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].

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 [60,61], especially for large datasets [62].

Multilayer perceptron (MLP) is another relatively simple model that can be used to perform classification [63].

The 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 [64].

The category-embedding model (CEM) is a basic model that is relatively simple with relatively simple architecture—a feed-forward network with categorical features passed

through a learnable embedding layer. It is similar to MLP, but with learned embeddings for category variables.

The neural oblivious decision ensembles (NODE) for deep learning on tabular data is a model presented in ICLR 2020 [65]. According to the authors, it has beaten well-tuned gradient-boosting models on many datasets. It uses a neural equivalent of oblivious trees (the kind of trees that catboost uses) as the basic building blocks of the architecture.

Attentive interpretable tabular learning (TabNet) is another model coming out of Google Research that uses sparse attention in multiple steps of decision making to model the output [66].

3.6. Data Collection and Experiment Setup

To perform the experiments, data were collected from the social network Twitter. The data were collected automatically by using a pipeline for a continuous collection of tweets over a long period of time, with the data structure organized so that there are records of users 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 located 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 1 January 2020 and 31 May 2021. The data were collected by using tweepy [67], 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. Next, we performed cleaning and processing of tweets following the same procedure as proposed in [68]. This includes several steps including: replacing usernames, replacing URLs, and translating emojis to ASCII code.

In the next step, we constructed the corresponding multilayer network $\mathcal{M}$ and multilayer framework $\mathcal{MF}$. Calculation of network measures was performed in a Python package *NetworkX* [69]. Then, we extracted the multilayer network features and text features. Before feature selection, we performed a detailed analysis of the features sets (The results of the feature analysis are available at: https://github.com/InfoCoV/Multi-Cro-CoV-cseBERT/blob/main/notebooks/exploration/features_analysis.ipynb, accessed on 20 September 2022) including mutual information analysis. The whole procedure of collecting and analysing tweets is described in Figure 2, and *Cro-Tweets2021* dataset (https://github.com/InfoCoV/InfoCoV/blob/main/Cro-Vect-Twitter.csv?fbclid=IwAR0m1Ahk6Jui200DQozGp4eeLa7n8AaBaf53ROLmMOUsSYCMaAvS2LTfwuc, accessed on 20 September 2022) is publicly available.

In the next, 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 the 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 tweets.

After training and testing all classifiers, we perform the SHAP analysis [70] to identify the features that have the most impact on the classification.

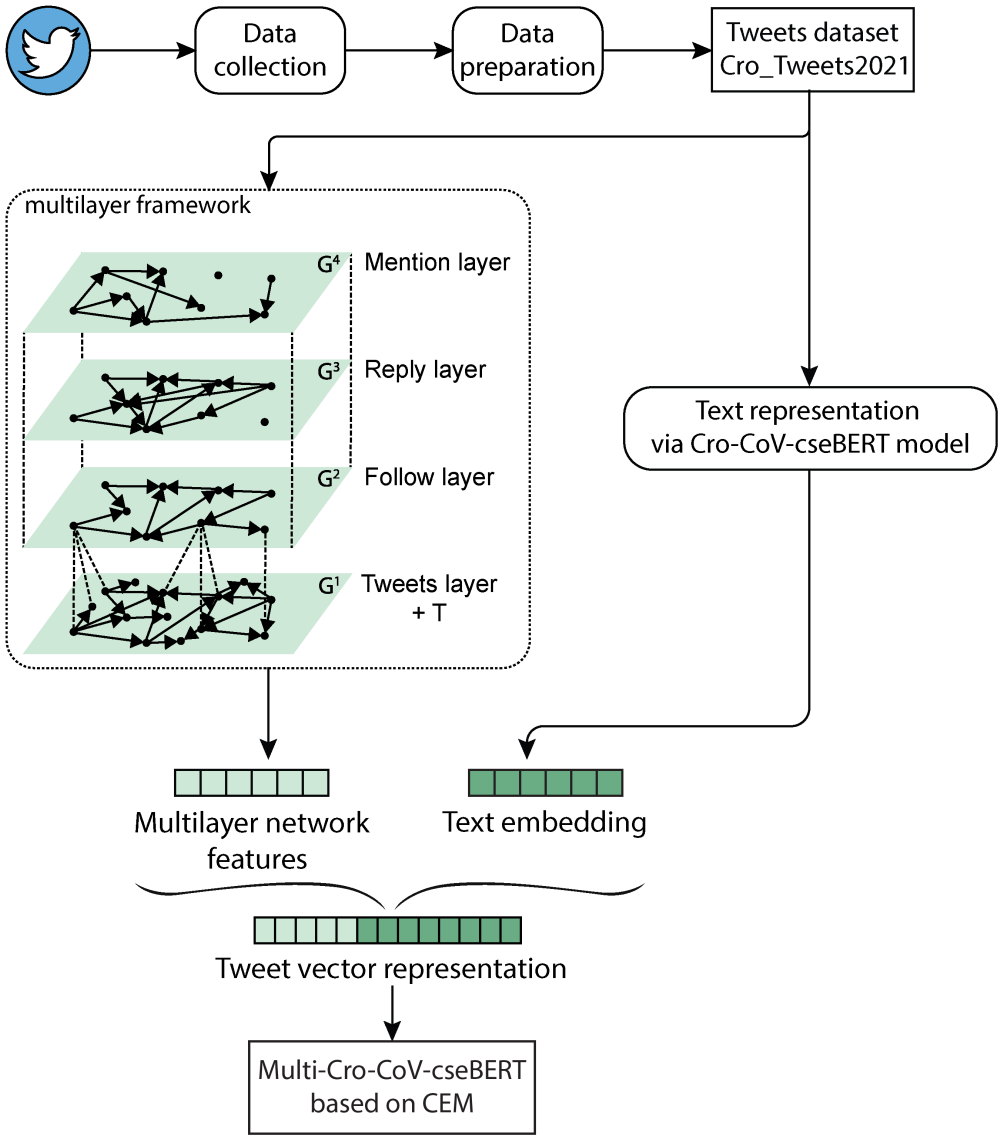

Figure 2. Tweets processing procedure.

4. Results

In this section, we present the comparison results of the performance of six trained models by using three different sets of features. These are features from the text, features from the multilayer network, and their combination. In the next step, we perform SHAP values analysis.

4.1. Evaluation Results

We trained and compared six ML models, namely RF, MLP, LGBM, CEM, NODE, and TabNet. The evaluation was performed in terms of standard machine learning classification metric such as: accuracy (*Acc*), precision (*P*), recall (*R*), and *F*1-score (*F*1). 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.

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

	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**

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 *F*1).

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%, whereas with respect to the *F*1 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 *F*1 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 are obtained with a set of combined features, in terms of *F*1 score, CEM is the most successful classification model with 67.9%. The lowest performance is achieved for the TabNet model with 66.6%. The MLP and NODE models perform well compared to the CEM model, as their performance is only one percentage point lower.

Based on these results, we decided to integrate the CEM model as the part of the Multi-Cro-CoV-cseBERT model for retweet prediction based on multilayer network and text features. We will further use this model for information-spreading analysis in the domain related to COVID-19 pandemic.

4.2. Feature Analysis

Shapley additive explanations (SHAP), introduced in [70], is used to show the contribution or importance of each feature to the prediction of the model. SHAP values analysis, in the case of the RF model, was performed on a sample of 1000 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 [70] 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. In addition, the 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, whereas the color of the dot shows whether this observation has a higher (red) or lower value (blue) compared to other observations.

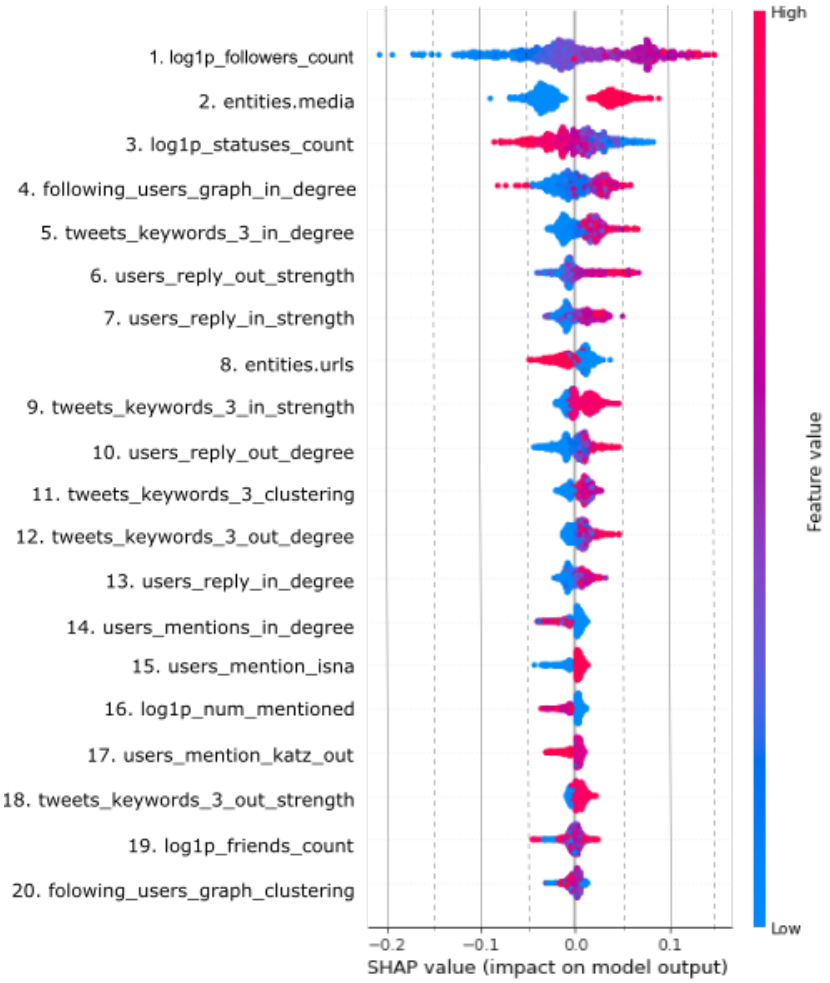

Figure 3. SHAP values analysis on bee swarm summary plot illustrating impact on model output.

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. The most important feature—*log1p_followers_count*—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, whereas 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 do not) 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

that have a stronger impact than other features that reflect other properties of the network. The most important feature is the number of followers (*1. log1p_followers_count*), and the 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 make 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.

Compared with other similar studies, our results are in line with the findings of Suh et al. [34]. They also examined the content and contextual features in the task of retweet prediction, but to a much lesser extent than our study. In general, their findings suggest that retweetability has a very close relationship with the social network context of the authors and the informational content and value contained in tweets, similar to our results. In particular, they showed that among content features, URLs and hashtags have strong potential in the prediction of retweeting. Furthermore, among contextual features which are related to the network, the number of followers and followees as well as the age of the account affect retweetability. In our results, the number of followers also has a high correlation with retweeting. However, according to our SHAP analysis, the presence of media has a much stronger impact than the presence of URLs. Next, the in-degree measures of the follower layer and the tweet layer are ranked as high, as well as the impact of the in-strength and out-strength centrality measures of the retweet layer. This suggests that multilayer network measures have the potential for retweet prediction which has not been shown before.

5. Discussion on Retweet Prediction Based on Heterogeneous Features

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, the multilayer network features perform better than text features for all six trained models. According to that, we can confirm that multilayer network representation of Twitter has great potential for retweet prediction. These findings are in line with results of the study [22] in which Pierri et al. have shown that multilayer network features perform better in the task of disinformation classification on Twitter. Although this study modeled Twitter differently than the $\mathcal{MF}$ method proposed here and used different network measures, all these results indicate that multilayer approach in the task of retweet prediction is worthy 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 performs better than only one set of features in the task of predicting the number of retweets. We have to emphasize that the combination of features only slightly outperforms the multilayer network features. However, this is consistent for five of six models (only the RF algorithm has 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 [1,13–17,34,35], and it has been shown that the combination of features is better than one set of features. Specifically, in [15] authors showed that models based on multidimensional features extracted from author, tweet, and user outperform models based

on the standard set of features. Our approach also combines features extracted from tweet with data related to author and user, but in a different way.

To the best of our knowledge, the combination of multilayer network measures and text embedding as features for information-spreading prediction has not been examined previously. The feature analysis performed by the SHAP approach indicates that among multilayer network measures, the in-degree calculated from the follower layer and the in-degree calculated from the tweet layer have highest impact on the model. Furthermore, the impact of the in-strength and out-strength centrality measures of the retweet layer are also high. Besides that, as expected, the number of followers and some other metadata, such as the presence of media and URLs, have a positive influence on retweeting. These findings suggest that except for the standard measures, multilayer network measures may be valuable in retweet-prediction models.

Here we need to emphasize that feature engineering and the selection of appropriate feature sets is an important step in all classification tasks, although some studies have examined the potential of using deep neural networks to avoid the manual construction of features. For example, in [12] authors proposed attention-based deep neural networks in the task of retweet prediction, and in [71] the authors applied graph representation learning to extract the structural attributes of the ego network and predict user retweet behavior. However, it is still worth examining different possibilities in the construction sets of features, especially the combination of features from heterogeneous data sources. In this way, it is possible to detect which data sources have higher influence to information spreading, and in the next step we can include these sources as an input into a deep neural network. In this context, the next research direction of the proposed approach is to perform joint representation learning from heterogeneous data sources: multilayer network and text.

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 set scenarios, whereas 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 that use only text features in comparison to a 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 an indicator that multiyear network features have great potential in the 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 [72], we compared the retweeting of COVID-19-related tweets and tweets that are not related to COVID-19. Our findings indicate that nearly 60% of tweets related to COVID-19 belong to the high-spreadable class, whereas less than 40% of non-COVID-19 tweets belong to this high-spreadable class. This suggests that tweet content may have a high impact on retweeting (spreadability), especially during a global crisis, such as the COVID-19 pandemic. In another study [73], we explored the potential of graph neural networks (GNNs) in the task of prediction if the user would tweet about COVID-19 or not. By using the proposed multi-Cro-CoV-cseBERT model for retweet prediction, we will further analyse the information-spreading patterns in the domain of COVID-19-related communication on Twitter.

This research has several limitations that we plan to address in 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. In this way, we try to predict whether 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 this way we ensured balanced classes of the dataset. However, in 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 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 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 and Future Work

In this paper, we introduce a multilayer framework formalism for the representation of online communication on social media. We utilized this formalism for feature extraction from heterogeneous data sources: multilayer networks and text messages. We performed a 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, exclusively multilayer network features performed significantly better than exclusively text-based 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 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 have similar performance in the task of retweet classification. Additionally, we explored the impact of different features by using SHAP analysis and determine that the number of followers in the network, the presence of media, the number of changed user statuses, the in-degree on the follower network layer, and the in-degree on the Twitter network layer features have major impacts on the model. Thus, we believe that our multilayer network-based approach provides useful insights into the future development of a system for predicting information spreading on social media.

The proposed approach can be further extended in the several directions, and we have several of plans for future work. First, 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 (accessed on 20 September 2022).

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. Firdaus, S.N.; Ding, C.; Sadeghian, A. Retweet Prediction based on Topic, Emotion and Personality. *Online Soc. Netw. Media* **2021**, *25*, 100165. [CrossRef]
2. Wang, J.; Yang, Y. Tweet retweet prediction based on deep multitask learning. *Neural Process. Lett.* **2022**, *54*, 523–536. [CrossRef]
3. Eysenbach, G. Infodemiology: The epidemiology of (mis) information. *Am. J. Med.* **2002**, *113*, 763–765. [CrossRef]
4. Petrović, M.; Levnajić, Z.; Meštrović, A. Analysis of the COVID-19 Communication on Twitter via Multilayer Network. In Proceedings of the 2nd International Symposium on Automation, Information and Computing (ISAIC 2021), Beijing, China, 3–6 December 2022.
5. Cuello-Garcia, C.; Pérez-Gaxiola, G.; van Amelsvoort, L. Social media can have an impact on how we manage and investigate the COVID-19 pandemic. *J. Clin. Epidemiol.* **2020**, *127*, 198–201. [CrossRef] [PubMed]
6. Bunker, D. Who do you trust? The digital destruction of shared situational awareness and the COVID-19 infodemic. *Int. J. Inf. Manag.* **2020**, *55*, 102201. doi: 10.1016/j.ijinfomgt.2020.102201. [CrossRef] [PubMed]
7. Malecki, K.M.; Keating, J.A.; Safdar, N. Crisis communication and public perception of COVID-19 risk in the era of social media. *Clin. Infect. Dis.* **2021**, *72*, 697–702. [CrossRef]
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. *Appl. Sci.* **2021**, *11*, 10442. [CrossRef]
9. Jay, A. FinancesOnline. Available online: https://financesonline.com/number-of-twitter-users/ (accessed on 1 July 2022).
10. Kuang, S.; Davison, B.D. Learning Word Embeddings with Chi-Square Weights for Healthcare Tweet Classification. *Appl. Sci.* **2017**, *7*, 846. [CrossRef]
11. Singh, C.; Imam, T.; Wibowo, S.; Grandhi, S. A Deep Learning Approach for Sentiment Analysis of COVID-19 Reviews. *Appl. Sci.* **2022**, *12*, 3709. [CrossRef]
12. Zhang, Q.; Gong, Y.; Wu, J.; Huang, H.; Huang, X. Retweet prediction with attention-based deep neural network. In Proceedings of the 25th ACM International on Conference on Information and Knowledge Management, Indianapolis, IN, USA, 24–28 October 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. [CrossRef]
14. Sharma, S.; Gupta, V. Role of twitter user profile features in retweet prediction for big data streams. *Multimed. Tools Appl.* **2022**, *81*, 27309–27338. [CrossRef] [PubMed]
15. Fu, X.; Cheng, S.; Zhao, L.; Lv, J. Retweet Prediction Based on Multidimensional Features. *Wirel. Commun. Mob. Comput.* **2022**, *2022*, 1863568. [CrossRef]
16. Dai, T.; Xiao, Y.; Liang, X.; Li, Q.; Li, T. ICS-SVM: A user retweet prediction method for hot topics based on improved SVM. *Digit. Commun. Netw.* **2022**, *8*, 186–193. [CrossRef]
17. 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 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, Paris, France, 21–25 July 2019; pp. 525–534.
18. 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. *Phys. Rep.* **2014**, *544*, 1–122. [CrossRef]
19. Kivelä, M.; Arenas, A.; Barthelemy, M.; Gleeson, J.P.; Moreno, Y.; Porter, M.A. Multilayer networks. *J. Complex Netw.* **2014**, *2*, 203–271. [CrossRef]
20. Martinčić-Ipšić, S.; Margan, D.; Meštrović, A. Multilayer network of language: A unified framework for structural analysis of linguistic subsystems. *Phys. A Stat. Mech. Its Appl.* **2016**, *457*, 117–128. [CrossRef]
21. Vukić, D.; Martinčić-Ipšić, S.; Meštrović, A. Structural analysis of factual, conceptual, procedural, and metacognitive knowledge in a multidimensional knowledge network. *Complexity* **2020**, *2020*, 9407162. [CrossRef]
22. Pierri, F.; Piccardi, C.; Ceri, S. A multi-layer approach to disinformation detection in US and Italian news spreading on Twitter. *EPJ Data Sci.* **2020**, *9*, 35. [CrossRef]
23. Nesi, P.; Pantaleo, G.; Paoli, I.; Zaza, I. Assessing the reTweet proneness of tweets: Predictive models for retweeting. *Multimed. Tools Appl.* **2018**, *77*, 26371–26396. [CrossRef]
24. 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, Sierra Nevada, Spain, 17 December 2010; Volume 104, pp. 17599–17601.
25. 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 21st ACM International Conference on Information and Knowledge Management, Maui, HI, USA, 29 October–2 November 2012; pp. 2335–2338.
26. Moreno, Y.; Pastor-Satorras, R.; Vespignani, A. Epidemic outbreaks in complex heterogeneous networks. *Eur. Phys. J. Condens. Matter Complex Syst.* **2002**, *26*, 521–529. [CrossRef]
27. 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. *Phys. Lett. A* **2007**, *364*, 189–193. [CrossRef]
28. 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), Warsaw, Poland, 11–14 August 2014; Volume 1, pp. 21–26.
29. Daga, I.; Gupta, A.; Vardhan, R.; Mukherjee, P. Prediction of likes and retweets using text information retrieval. *Procedia Comput. Sci.* **2020**, *168*, 123–128. [CrossRef]
30. Kushwaha, A.K.; Kar, A.K.; Ilavarasan, P.V. Predicting retweet class using deep learning. In *Trends in Deep Learning Methodologies*; Elsevier: Amsterdam, The Netherlands, 2021; pp. 89–112.
31. Firdaus, S.N.; Ding, C.; Sadeghian, A. Topic specific emotion detection for retweet prediction. *Int. J. Mach. Learn. Cybern.* **2019**, *10*, 2071–2083. [CrossRef]
32. Pierri, F.; Piccardi, C.; Ceri, S. Topology comparison of Twitter diffusion networks effectively reveals misleading information. *Sci. Rep.* **2020**, *10*, 1–9. [CrossRef] [PubMed]
33. Miao, W.Y.; Fang, B.; Meng, L.Q. Retweet Prediction within Communities on SNS Based on Social Network Analysis. *J. Comput.* **2018**, *29*, 147–160.
34. 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, Minneapolis, MN, USA, 20–22 August 2010; pp. 177–184.
35. Tsur, O.; Rappoport, A. What's in a hashtag? Content based prediction of the spread of ideas in microblogging communities. In Proceedings of the Fifth ACM International Conference on Web Search and Data Mining, Seattle, WA, USA, 8–12 February 2012; pp. 643–652.
36. Amitani, R.; Matsumoto, K.; Yoshida, M.; Kita, K. Buzz Tweet Classification Based on Text and Image Features of Tweets Using Multi-Task Learning. *Appl. Sci.* **2021**, *11*, 567. [CrossRef]
37. Omodei, E.; De Domenico, M.D.; Arenas, A. Characterizing interactions in online social networks during exceptional events. *Front. Phys.* **2015**, *3*, 59. [CrossRef]
38. Oro, E.; Pizzuti, C.; Procopio, N.; Ruffolo, M. Detecting topic authoritative social media users: A multilayer network approach. *IEEE Trans. Multimed.* **2017**, *20*, 1195–1208. [CrossRef]

39. 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, Kaohsiung, Taiwan, 25–27 July 2011; pp. 5–12.
40. 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 Sci.* **2016**, *5*, 24. [CrossRef]
41. Perc, M. Diffusion dynamics and information spreading in multilayer networks: An overview. *Eur. Phys. J. Spec. Top.* **2019**, *228*, 2351–2355. [CrossRef]
42. De Domenico, M.; Granell, C.; Porter, M.A.; Arenas, A. The physics of spreading processes in multilayer networks. *Nat. Phys.* **2016**, *12*, 901–906. [CrossRef]
43. Bródka, P.; Musial, K.; Jankowski, J. Interacting spreading processes in multilayer networks: A systematic review. *IEEE Access* **2020**, *8*, 10316–10341. [CrossRef]
44. Matas, N. Comparing Network Centrality Measures as Tools for Identifying Key Concepts in Complex Networks: A Case of Wikipedia. *J. Digit. Inf. Manag.* **2017**, *15*, 203–213. [CrossRef]
45. Newman, M. *Networks*; Oxford University Press: Oxford, UK, 2018.
46. Bonacich, P. Power and centrality: A family of measures. *Am. J. Sociol.* **1987**, *92*, 1170–1182. [CrossRef]
47. Katz, L. A new status index derived from sociometric analysis. *Psychometrika* **1953**, *18*, 39–43. [CrossRef]
48. Opsahl, T.; Panzarasa, P. Clustering in weighted networks. *Soc. Netw.* **2009**, *31*, 155–163. [CrossRef]
49. Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.S.; Dean, J. Distributed representations of words and phrases and their compositionality. *Adv. Neural Inf. Process. Syst.* **2013**, *26*, 3111–3119.
50. Le, Q.; Mikolov, T. Distributed representations of sentences and documents. In Proceedings of the International Conference on Machine Learning, PMLR, Bejing, China, 22–24 June 2014; pp. 1188–1196.
51. Pennington, J.; Socher, R.; Manning, C.D. Glove: Global vectors for word representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 25–29 October 2014; pp. 1532–1543.
52. Bojanowski, P.; Grave, E.; Joulin, A.; Mikolov, T. Enriching word vectors with subword information. *Trans. Assoc. Comput. Linguist.* **2017**, *5*, 135–146. [CrossRef]
53. Peters, M.E.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; Zettlemoyer, L. Deep contextualized word representations. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), New Orleans, LA, USA, 1–6 June 2018; ACL: New Orleans, LA, USA, 2018; pp. 2227–2237. [CrossRef]
54. Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I. Language models are unsupervised multitask learners. *OpenAI Blog* **2019**, *1*, 9.
55. 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. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 1877–1901.
56. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Minneapolis, MN, USA, 2–7 June 2019; ACL: Minneapolis, MN, USA, 2019; pp. 4171–4186. [CrossRef]
57. Ethayarajh, K. How Contextual are Contextualized Word Representations? Comparing the Geometry of BERT, ELMo, and GPT-2 Embeddings. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, 3–7 November 2019; ACL: Hong Kong, China, 2019; pp. 55–65. [CrossRef]
58. Babić, K.; Martinčić-Ipšić, S.; Meštrović, A. Survey of neural text representation models. *Information* **2020**, *11*, 511. [CrossRef]
59. Ulčar, M.; Robnik-Šikonja, M. Finest bert and crosloengual bert. In Proceedings of the International Conference on Text, Speech, and Dialogue, Brno, Czech Republic, 8–11 September 2020; Springer: Berlin/Heidelberg, Germany, 2020; pp. 104–111.
60. 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), Patras, Greece, 29–31 October 2007; Volume 2, pp. 310–317.
61. Liu, X.Y.; Wu, J.; Zhou, Z.H. Exploratory undersampling for class-imbalance learning. *IEEE Trans. Syst. Man Cybern. Part B (Cybernetics)* **2008**, *39*, 539–550.
62. Dietterich, T.G. Ensemble methods in machine learning. In Proceedings of the International Workshop on Multiple Classifier Systems, Cagliari, Italy, 21–23 June 2000; Springer: Berlin/Heidelberg, Germany, 2000; pp. 1–15.
63. Ruck, D.W.; Rogers, S.K.; Kabrisky, M. Feature selection using a multilayer perceptron. *J. Neural Netw. Comput.* **1990**, *2*, 40–48.
64. Kumar, S.; Mallik, A.; Panda, B. Link prediction in complex networks using node centrality and light gradient boosting machine. *World Wide Web* **2022**, *25*, 2487–2513. [CrossRef]
65. Popov, S.; Morozov, S.; Babenko, A. Neural Oblivious Decision Ensembles for Deep Learning on Tabular Data. In Proceedings of the International Conference on Learning Representations, Addis Ababa, Ethiopia, 26–30 April 2020.
66. Arık, S.O.; Pfister, T. Tabnet: Attentive interpretable tabular learning. In Proceedings of the AAAI, Online, 2–9 February 2021; Volume 35, pp. 6679–6687.
67. Roesslein, J. Tweepy Documentation. 2009. Available online: http://tweepy.readthedocs.io/en/v3 (accessed on 10 September 2022).

68. Müller, M.; Salathé, M.; Kummervold, P.E. Covid-twitter-bert: A natural language processing model to analyse covid-19 content on twitter. *arXiv* **2020**, arXiv:2005.07503.
69. Hagberg, A.A.; Schult, D.A.; Swart, P.J. Exploring network structure, dynamics, and function using NetworkX. In Proceedings of the 7th Python in Science Conference, Pasadena, CA, USA, 19–24 August 2008; pp. 11–15.
70. Lundberg, S.M.; Lee, S.I. A unified approach to interpreting model predictions. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 4768–4777.
71. Guo, H.; Yang, L.; Liu, Z. UserRBPM: User Retweet Behavior Prediction with Graph Representation Learning. In Proceedings of the International Conference on Mobile Multimedia Communications, Okayama, Japan, 12–16 December 2021; Springer: Berlin/Heidelberg, Germany, 2021; pp. 613–632.
72. 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), Opatija, Croatia, 27 September–1 October 2021; pp. 395–399.
73. 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), Opatija, Croatia, 23–27 May 2022; pp. 307–311.