Assessing the combination of DistilBERT news representations and diffusion topological features to classify fake news

Carlos Abel Córdova Sáenz, Marcelo Dias, Karin Becker

Instituto de Informática – Universidade Federal do Rio Grande do Sul (UFRGS)
Porto Alegre – RS – Brazil

Abstract. Fake news (FN) have affected people's lives in unimaginable ways. The automatic classification of FN is a vital tool to prevent their dissemination and support fact-checking. Related work has shown that FN spread faster, deeper, and more broadly than truthful news on social media. Deep learning has produced state-of-the-art solutions in this field, mainly based on textual attributes. In this paper, we propose to combine compact representations of the textual news properties generated using DistilBERT, with topological metrics extracted from their propagation network in social media. Using a dataset related to politics and distinct learning algorithms, we extensively assessed the components of the proposed solution. Regarding the textual attributes, we reached results comparable to state-of-the-art solutions using only the news title and contents, which is useful for FN early detection. We assessed the influential topological metrics, and the effect of their combination with the news textual features. We also explored the use of ensembles. Our results were very promising, revealing the potential of the features proposed and the adoption of ensembles.

Categories and Subject Descriptors: Information Systems [**Data Mining**]: Data Streaming; Artificial Intelligence [**Machine Learning**]: Supervised Learning; Information Systems [**Web Applications**]: Social Network

Keywords: DistilBERT, fake news, fake news classification, topological features, ensembles

1. INTRODUCTION

The fake news phenomenon has increased in the last decade, affecting various aspects of everyday life, including politics, health, education, among others. Social networks play an active role in this context, as the same mechanisms for democratizing information are used to spread untruths. The effects of a rumor or fake news can be tragic, compromising democracy worldwide or affecting people's lives in unimaginable ways [Wang 2017].

Currently, there is no consensus on the concept of fake news [Zhou and Zafarani 2020], which can be defined broadly or strictly [Shu et al. 2017]. In the broad interpretation, news, statements, speeches or posts on social networks are considered to contain false information related to public figures and organizations. This aspect also includes works for the detection of rumors, satires and bots [Bondielli and Marcelloni 2019].

In the strict definition adopted by this work, Fake News (FN) refer to fake journalistic articles whose veracity can be verified and which were published intentionally to deceive the consumer of the news [Shu et al. 2017]. The concept emphasizes authenticity and intention, in addition to indicating that FN are similar to news that followed the journalistic protocol, making it difficult for their recipients to identify them.

Identifying misleading information is not easy for humans [Zhou and Zafarani 2020], and the harmful potential is so relevant that many fact-checking initiatives are being developed. Such initiatives are

This research is partially supported by CNPq (process: 131178/2020-2).

directed either by groups of the mainstream media individually (e.g., Washington Post and CNN in the United States - USA, and Folha de São Paulo and Estadão in Brazil) or in consortium (e.g., Project Comprova[1]), as well as by less influential journalistic groups (e.g., PolitiFact[2], Agência Lupa[3]). However, the quantity in which they are produced, the speed of their dissemination and the complexity in performing manual fact checking lead to the need for automatic mechanisms to combat fake news [Reis et al. 2019].

FN Detection is the task that aims to identify whether a news item is false or true. Works focused on the news classification task were developed using supervised machine learning approaches [Bondielli and Marcelloni 2019; Zhou and Zafarani 2020]. Such approaches are based on the training of classifiers using labeled data and are essentially differentiated by the learning algorithms used (shallow or deep learning), and by the features explored in the task, which are divided into features related to news itself, and the social context of the news spread [Shu et al. 2019].

Features extracted from the news (e.g., title, text and image) allow the early detection of false news, i.e., before it spreads, as they do not depend on the spread of the news on social networks. However, this approach usually limits solutions to the domain of training data used to construct predictive models. A study [Reis et al. 2019] argues that features related to the news source and the engagement generated in its propagation are the most discriminatory in the FN classification. A proposal for the classification of FN outlets based on the topology of the propagation network is presented in [Pierri et al. 2020]. Proposals for features representing the social context include news broadcast profiles on social networks [Shu et al. 2019], social behavior (e.g., likes) [Bauskar et al. 2019] and propagation patterns [Shu et al. 2020].

Related works that explore textual attributes and Deep Learning [Liao et al. 2021; Shu et al. 2019; Zhou et al. 2020] in FN classification have reported the best results. A new trend in natural language processing (NLP) is to create models by transfer learning from representations of encoded languages using massive amounts of data, such as BERT (Bidirectional Encoder Representations from Transformers) [Devlin et al. 2019]. Another opportunity is to verify if the topological approach to detect FN broadcasting vehicles proposed in [Pierri et al. 2020] can contribute to FN classification regardless of the source.

In this article, we explore the combination of textual content of the news and the topology of the news diffusion networks for FN classification. More specifically, we propose the use of DistilBERT [Sanh et al. 2019], a lighter version of BERT, to generate features that compactly represent the news. As social context, we propose to represent the properties of the diffusion network of each news item on Twitter by topological metrics, considering tweets, retweets and mentions. We developed experiments using a politics-related dataset available on FakeNewsNet[4] (FNN) [Shu et al. 2018], using different algorithms for supervised learning and stacking ensemble. We assessed the contribution of each type of feature separately and their combination.

This article is an extension of our previously presented work [Sáenz et al. 2020]. We have significantly evolved it by: a) extending the set of topological features used for classification and assessing their contribution for FN classification; b) extending the experimental settings, and analyzing in more detail the performance results; c) leveraging stacking ensembles to improve the classification of fake news; and d) updating the theoretical background and related work.

With regard to related work, our main contributions are:

- a solution for fake news classification that combines compact DistilBERT representations of textual

[1]https://projetocomprova.com.br/
[2]https://www.politifact.com/
[3]https://piaui.folha.uol.com.br/lupa/
[4]https://github.com/KaiDMML/FakeNewsNet

content of the news and topological metrics describing its diffusion network. The combination of classifiers in a stacking ensemble achieved the best results, comparable to the state-of-the-art solutions[Shu et al. 2019; Zhou et al. 2020; Liao et al. 2021].
- experiments based on the fine-tuning of language representation models (DistilBERT), which is still little explored in FN classification. Our results were promising, showing that the classification of FN based only on the title and content of the news achieves results close to the state-of-the-art [Shu et al. 2019; Zhou et al. 2020; Liao et al. 2021], which also consider the text of the propagation posts;
- evaluation of the contribution of topological features of propagation networks as representative attributes of social engagement, previously restricted to identifying communication vehicles that propagate false news [Pierri et al. 2020]. We improved our previous results [Sáenz et al. 2020] by considering additional topological features.
- an encompassing experimental setting to assess all components of the proposed approach.

The remaining of this article is structured as follows. Section 2 describes the theoretical background and Section 3 presents works related to the classification of FN. Section 4 describes the proposed combination of features extracted from the news and topological metrics of their dissemination for the classification of FN. Section 5 details the experiments performed. Section 6 presents the conclusions and points to future research.

2. THEORETICAL BACKGROUND

2.1 Social Network Analysis

Network Analysis consists on studying the properties and characteristics of networks (or graphs), which are composed by a set of nodes connected through links called edges. Social network analysis (SNA) is the process of investigating social structures through the use of networks and graph theory [Hansen et al. 2020]. SNA has been extensively deployed to understand various phenomena in internet social media networks (e.g., Twitter, Facebook), in applications such as Political Polarization analysis [Takikawa and Nagayoshi 2017], Fake News detection [Zhou and Zafarani 2019], Bots detection [Wang et al. 2016] or Community Detection [Leão et al. 2020], among others. Typically nodes represent social entities (e.g., people or actors in the network) connected by edges representing static (e.g., friendship , follower, subscriber) or dynamic relationships (e.g., respond, mention, like). For instance, diffusion networks are used to analyze how the social transmission of a behavior follows the social network of associations or interactions among individuals, since individuals who spend a lot of time together, or who interact more have more opportunity to learn from each other. In Twitter, for instance, diffusion models can be created by representing as nodes users who post tweets, and connecting by edges users who replied to, retweeted or mentioned them in each others' tweets [Pierri et al. 2020]. In this work, we will adopt this type of diffusion model to represent the interaction between users in the spreading of news.

Many insights about the nature and behavior of users in a social network structure can be derived from topological metrics describing the network [Costa et al. 2007], among them:

- Number of nodes and number of edges, which describe how large the network is.
- Nodes' degree (in-degree, out-degree), which in the context of diffusion networks represent the interaction between users (e.g., retweet, reply to, or mention).
- Average degree, which is the average of all nodes degrees, it is a global metric that represents how connected are the nodes in the network.
- Network density, which is a measure of how many actual connections between nodes exist, compared to the possible number of connections.

- Network diameter, which is the length of the longest shortest path between any two nodes in the network. It provides an intuition of how difficult it can be to reach a node from any other on the network.
- Network strongly and weakly connected components. Both types of components are directed subgraphs. In strongly connected components, all nodes in the same subgraph can reach each others (and be reachable from each other). A weakly connected component is one in which all components are connected by some path, ignoring direction. They both indicate the closed groups can be found in the network.
- Network's clustering coefficient, which represents the probability of finding sub-groups of highly connected nodes within the network.
- Network's k-core, which is the greatest number k of edges that every node in the graph could at least have without being empty. It is a measure of sparsity of the graph.

In this work we will explore social topological metrics to investigate whether they can contribute as discriminatory features for fake news classification.

2.2 Supervised Machine Learning

Supervised machine learning is the task of learning a function that maps an input to an output based on example input-output pairs. The resulting models can be used to predict the output of new, unknown data records. While classification algorithms deal with discrete labels, regression is a predictive modeling task that deals with continuous values. Traditional classification algorithms are SVM (support vector machine), Naive Bayes, variations on tree induction (e.g., C5, Random Forest), Logistic regression, among others [Murphy 2012].

An ensemble consists in combining different base (or weak) models with the purpose of a collective decision. It is a performance boosting technique used for different machine learning tasks. The premise is that each individual base model contributes with a different hypothesis space, and their aggregation in a final model reduces the computational cost of training a single model for a complex task [Zhang and Ma 2012]. There are different types of ensemble, such as bagging, boosting stacking and mixture of experts. In this work, we adopt the mixture-of-experts ensemble models [Polikar 2009], composed of different base classifiers, all of them trained on the same data, but by different algorithms (and/or parametrizations), as a means to generate variability. Then a second level classifier is used to assign weights for their combination according to some rule, such as voting majority, averaged probabilities, etc.

Deep learning has emerged as a powerful technique that allows computational models to learn representations of large sets of data using computing power. In a nutshell, deep learning uses a cascade of multiple layers of nonlinear processing units for feature extraction and transformation. The lower layers, closer to the data input, learn simpler features, while higher layers learn more complex features derived from lower layer ones [Zhang et al. 2018]. The scenario of deep neural networks for natural language processing has significantly changed with recent work on transfer learning based on language models pre-trained in an unsupervised manner on massive sets of data. BERT [Devlin et al. 2019] is a bidirectional model based on the transformer architecture, which replaces the sequential nature of recurrent networks with a much faster attention-based approach. A pre-trained BERT model can be fine-tuned with just one additional output layer to create state-of-the-art models for a wide range of tasks (e.g., sentiment analysis) without substantial task-specific architecture modifications, and with much smaller training sets. DistilBERT [Sanh et al. 2019] is based on knowledge distillation, which consists on training a more compact model to reproduce the behaviour of a larger one. This allows the fine-tuning to be based on much more compact language representations, which provides faster and lighter text processing, with a much lower computational cost, allowing experiments even with limited computational resources, such as those caused by the current pandemic situation. The experiments that compare BERT and DistilBERT reveal DistilBERT preserves 95% of the full BERT model, while using 66M parameters (instead of 110M).

In this work, we will leverage DistilBERT to generate compact representations of news, to be used as features. We will explore both traditional classification algorithms and ensembles to classify fake news based on properties of the news and topological features of their diffusion network.

3. RELATED WORK

The exploitation of FN in the context of elections, central to Donald Trump's victory in 2016 in the USA and followed in other countries, motivated a significant interest in the theme. Surveys like [Zhou et al. 2020; Bondielli and Marcelloni 2019; Shu et al. 2017] contribute with a conceptual framework and the compilation of important works in the area.

Much of the work relies on the use of supervised machine learning for FN classification, either through traditional or deep learning algorithms. According to [Shu et al. 2019], the two large groups of attributes used for FN Detection are features extracted from the news content or the social context. The first involves textual characteristics extracted from the headline or text of the news (e.g., n-grams), derived attributes (e.g., linguistic characteristics, emotions) or obtained from images published with the news. The second involves properties extracted from the user's profile, patterns of social interaction or news spread. A study [Reis et al. 2019] evaluates the contribution of different types of features to the classification of FN, concluding that all contribute in a discriminatory way, but that some may be more useful, among them, those extracted from the social engagement generated by news. Using 5 different algorithms, it reports F-measure results ranging from 0.75 to 0.81 in the tested datasets.

Supervised approaches require labeled data for training [Zhou and Zafarani 2020], and several efforts have focused on building data sets for this purpose, such as [Wang 2017; Shu et al. 2018]. The present work makes use of the Politifact dataset, one of those available in the FakeNewsNet (FNN) [Shu et al. 2018] repository. Unlike most datasets, Politifact includes not only properties textual contents related to the news, but also the interactions that resulting in its propagation into the Twitter social network. Thus, it allows the classification of FN using the news content, the social context or their combination. In order to respect Twitter's privacy policy, FNN provides a program that automates the download of news data (title, text, image URLs, etc.) and information related to the social context (tweets, retweets, user profiles, timelines of users, followers and followed by users who tweeted about the news).

It is possible to find more than a dozen published proposals using the Politifact dataset. We highlight the approaches dEFEND [Shu et al. 2019], with F-measure 0.92 using textual attributes extracted from news and posts, GCAL [Liao et al. 2021] with F-measure of 0.92 using the same previous features, but also users profiles, and SAFE [Zhou et al. 2020], which has the best result reported using only the news content (0,89).

dEFEND uses the textual content of the news and tweets that mentioned it. It proposes the use of encoders to extract features from this content and co-attention mechanisms (news and comments) in order to improve the classification performance and to select sentences that justify the classification performed (explainability). GCAL evolves dEFEND, by exploring user profiles, comments and news to generate an heterogeneous graph. Such a graph is processed by a neural network of graphs with attention mechanisms, which allow not only to perform a classification with a considerable F1 metric, but also to find, like its predecessor, sentences within the news item that can explain the prediction made. SAFE, on the other hand, explores the similarity between textual content (title and text) and images of the news using convolutional neural networks. These works show the relevant role of the use of Deep Learning applied to the textual content of the news.

To deal with the dependence on the content of the training corpus, other proposals on Politifact propose the use of the social context, such as user profile [Shu et al. 2019], reactions [Bauskar et al.

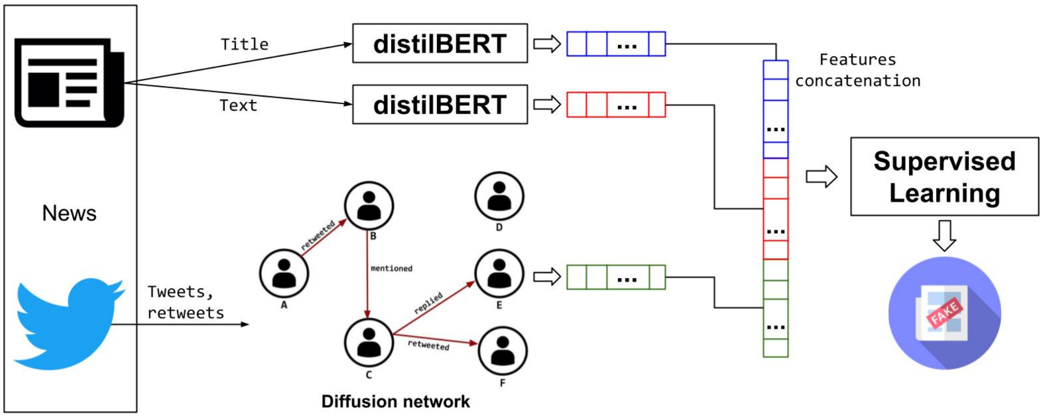

Fig. 1. Architecture proposed for the classification of Fake News

2019] and patterns of news spread [Shu et al. 2020]. These works confirmed the importance of features on social engagement in the news classification [Reis et al. 2019], but better performance is obtained when combined with the textual news features [Shu et al. 2019; Shu et al. 2020].

The analysis of social networks topology has been realized in different contexts to detect different patterns [da Fonseca Vieira et al. 2019]. A study [Pierri et al. 2020] shows promising results for automatic classification of communication vehicles as mainstream or of misinformation based exclusively on topological metrics of the diffusion network, which have the advantage of being more difficult to be simulated through robots. These same attributes are explored in the present work for the classification of news.

The present work differs from those related to FN classification by combining DistilBERT for processing the textual content of the news, with topological attributes extracted from the news dissemination network.

4. FEATURE EXTRACTION FROM NEWS AND DIFFUSION NETWORKS FOR FAKE NEWS CLASSIFICATION

The proposal for FN classification evaluated in this work combines a) news representations based on fine-tuning models of language representations using DistilBERT [Sanh et al. 2019], and b) characteristics of the social context using topological metrics of the networks used in the dissemination of fake news. Unlike [Shu et al. 2019; Zhou et al. 2020], we extract textual features only from the headline and/or text of the news. The innovative characteristic of our work is to represent the social context by metrics that characterize the topology of the social network used for their dissemination, i.e., tweets and retweets where the URL of the news is present. The inclusion of properties of the topology of the diffusion network enables to have a set of features independent of the news domain, and that is not easy to reproduce artificially using robots.

Figure 1 outlines the proposed approach. We combine the two types of features by concatenating the vectors representing each aspect in a single vector, used as input to a supervised classification algorithm. According to [Gadzicki et al. 2020], this approach of combining multi-modal features is referred to as early-maturing fusion.

4.1 Textual features of news

In our work, we used DistilBERT as an encoder to create a compact representation of the news. In this way, the headline and text of the news, which are originally unstructured data, are transformed into another structured representation: vectors of floating numbers that summarize the textual content.

In our experiments, we used both the headline, the news, and the combination of both, to verify the most relevant properties of the news in FN detection. Our proposal differs from [Zhou et al. 2020], which seeks local patterns of increasing complexity using convolutions, and from [Shu et al. 2019] which uses encoders that align news and associated posts.

We used the library *transformers*[5], which includes DistilBERT. Regarding the raw text of the headline and news content, we use functions for the tokenization, padding and masking. Then, we extracted the vector representation of them with DistilBERT in base and lowercase ('distilbert-base-uncased').

4.2 Diffusion network features

In article, we propose leveraging topological metrics from the diffusion networks that spread news (fake or true) in Twitter, more specifically, by retweeting, replying or mentioning users in these tweets. This approach was originally proposed for classifying news outlets as mainstream or misinformation.

Considering this purpose, for each news item, we construct a diffusion network using the tweets and retweets that include the respective news URL. In this diffusion network, the nodes represent users who (re)tweeted the news, responded to these tweets, or are mentioned in them. Pairs of nodes are connected by directed and unweighted edges, whenever the user represented by the origin has retweeted, answered or mentioned the destination node. An example is shown in Figure 1, where users A, B, C, D, E and F (re)tweeted a post that contains an URL representing a news i, fake or true. In this graph, user C was mentioned by B, responded to a tweet from E, and retweeted a tweet from F. User A retweeted a tweet from user B. Finally User D did not interact with other users in the context of this news broadcast. Thus, this network represents the way people interact to spread this news item on Twitter.

Once constructed the diffusion network for a given piece of news, we calculate a set of metrics that characterizes the respective topological properties. These metrics represent the complexity of the respective network, its propagation power and the strength of connection and cohesion among the participants. We seek to determine whether the way in which users interact with each other and form closed groups, can contribute to the detection of FN. The metrics to be calculated, demonstrated in [Pierri et al. 2020] to have managed to reach several cases in diffusion networks, such as when users within the network form groups, networks where there is no mono-directionality in the diffusion of news or networks where there is a single user who distributes the news among all others (broadcast). The metrics adopted in [Pierri et al. 2020], and experimented in our original work [Sáenz et al. 2020] are:

- Number of strongly connected components
- Size of the largest strongly connected component
- Number of weakly connected components
- Size of the largest weakly connected component
- Diameter of the largest weakly connected component
- Clustering coefficient
- K-Core number

In this article, we included four additional metrics that characterize the network as a whole and that were used in some other works [Zhou and Zafarani 2019; Shu et al. 2020], to determine if they can contribute to the improvement of the results:

- Number of nodes
- Number of edges

[5]https://huggingface.co/transformers/

Table I. Politifact news dataset in experiments

Type	# True news	# Fake News	Total
Complete original dataset	624	432	1056
News with textual content	505	385	890
News with textual content and diffusion network	304	355	659

- Density of the network
- Diameter of the network

To calculate these topological attributes, we first uploaded the news with their tweets and retweets to the graph-oriented database *neo4j*[6]. There, each news, tweets and retweets were represented as nodes of an heterogeneous network. The links in that network had different types and connected different types of nodes: news with tweets and tweets with retweets. We designed a query to construct and extract the diffusion networks for each news item, with the characteristics described before, from the network stored on *neo4j*. For each of the extracted networks, we imported them and calculated the metrics listed before using *networkx*[7].

5. EXPERIMENTS

5.1 Dataset

We used the news set *Politifact* available in the FNN repository. Using the program made available by the FNN to extract the data, we were able to collect 507 true and 385 false news out of a total of 1058 news available. To build the news dissemination network, the respective tweets and retweets were also collected where the news is referenced. However, some of the tweets/retweets could not be downloaded for various reasons (e.g., removal on Twitter). In order to avoid the reproduction of untrustworthy diffusion networks, in our experiments involving topologies we despise all the news with problems in the collection of tweets/retweets. The dataset with social context was limited to 304 true news and 355 false news, with their tweets and retweets.

Table I contrasts the amount of news from the repository and collected in each case.

5.2 Objectives and experiment setup

We developed a set of experiments with the following objectives:

(a) Determine the most influential textual properties of the news to the classification of FN, when compressed using DistilBERT: news title, text or both.
(b) Investigate which topological features of the news diffusion network add value to FN classification, and how to best represent them.
(c) Explore different ways to combine DistilBERT textual representations of news and topological features.
(d) Determine which classification algorithm, within a set of candidate algorithms, produces the best results.

In terms of algorithms, we adopted Logistic Regression (LR), Random Forest (RF), K-Nearest Neighbors (KNN), Support Vector Machines (SVM) and Naïve Bayes (NB) algorithms[8]. These algorithms were explored in the feature analysis developed in [Reis et al. 2019], except for Logistic

[6]https://neo4j.com/
[7]http://networkx.github.io/
[8]We also tried other algorithms, including multi-layer perceptron, which did not yielded good results, and thus are not reported in this article

Regression. We also experimented with mixture-of-experts ensembles, which combine predictive models trained using distinct algorithms, using voting and averaged probabilities to consolidate these results in a single, final prediction.

To evaluate the results, we used Precision, Recall and F1. For the aggregation of results considering both fake and true news classes, we used weighted average. We applied a *repeated cross-validation* strategy to train and test each algorithm and features configuration, with 10 repetitions and 10 folds ($r = 10$, $k - folds = 10$). All results reported in the remaining of this section refer to the average of the values obtained in these set of executions. In the remaining of this section we present these results using charts, but the tables in Annex A detail all average results, and the corresponding standard deviations.

We performed statistical tests to verify if there is a significant difference in the performance of sets of models [Demšar 2006]. We used the ANOVA statistical test to compare a set of distributions, and two-tailed Student T-Test for a pairwise comparison of models. In both statistical tests we used a confidence level of 0.05, and we adopted the null hypothesis that there was no significant difference between the results obtained from the compared models. The p-values obtained from these tests can also be found in Annex A.

All these experiments were developed in the *Python* environment, using the *scikit-learn*[9] library. We used the default parameters/hyperparameters of these algorithms as available in the library. Attempts to improve parametrization (including GridSearchCV[10] for SVM) did not yield better results.

5.3 Experiments with DistilBERT

The first experiment was carried out to compare the classification performance according to the textual properties of the news. Thus, all the news retrieved from the repository were used, i.e., 890 news. The algorithms were trained using as input: a) the vector corresponding only to the title, b) the vector corresponding only to the textual content of the news, and c) the two concatenated vectors.

The results in terms of weighted averaged F1 (W-F1) are presented in the chart displayed in Figure 5.3 (the detailed results are in Table VI in Annex A). We observed that the use of DistilBERT on the combination of title and text yields the best results in four out of the five algorithms. These differences are statistically significant compared to title only and text only, according to p-values presented in Table VII. The performance of the models trained using title only or text only are statistically comparable. The best performing algorithm was LR ($WF1 = 0.906$), followed by RF ($WF1 = 0.892$) and SVM ($WF1 = 0.878$) (Table VIII).

These results revealed a promising approach of minimum requirements for FN detection. First, in terms of absolute performance, we obtained a result close to state of the art represented by dEFEND [Shu et al. 2019] and GCAL [Liao et al. 2021], which depend on the spread of the news on the social network. Second, when compared to a similar minimum requirements approach (SAFE) [Zhou et al. 2020], which uses only the news content (textual and visual), it presents superior results. However, the comparison with these works should be regarded as a reference since each of them uses a different number of news from Politifact, with different proportions of real and fake news.

Figure 3 shows the performance for the Fake News class. The best results are obtained when using the concatenation of the title and the text of the content, and the best algorithms are LR and RF, with $F1 = 0.89$ and $F1 = 0.871$, respectively. Since these differences are statistically significant, these results show the consistency of our approach, yielding similar good results for the most relevant class in this task (i.e., the Fake News class).

[9] https://scikit-learn.org/

[10] https://scikit-learn.org/stable/modules/grid_search.html

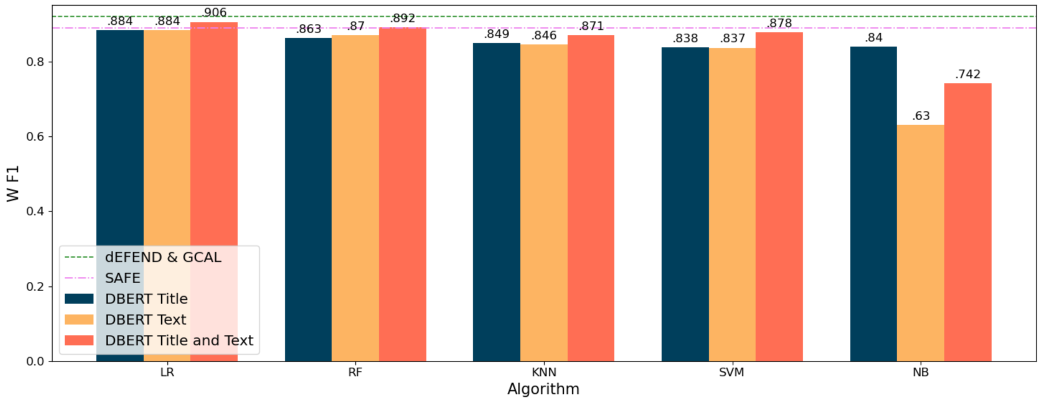

Fig. 2. Weighted Averaged F1 Score of news classification using DistilBERT representation of news textual contents

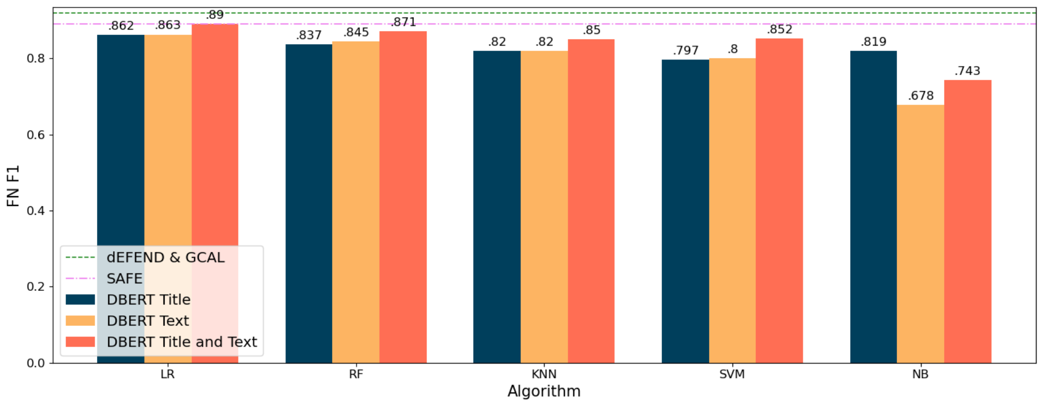

Fig. 3. F1 Score for the Fake News class using DistilBERT representations of news textual contents

We conclude that leveraging DilstilBERT to represent basic textual characteristics of the news allows us to reach a performance comparable to state-of-the-art approaches, but using considerably less data.

5.4 Experiments with topological metrics

This second set of experiments aims to analyze the discriminating power of topological metrics for fake news classification. For this purpose, we carried out several experiments, namely: a) assessment of the proper representation for topological metrics for classification purposes; b) an evaluation of the predictive power of the topological metrics; and c) a performance comparison between our previous work [Sáenz et al. 2020] and the models constructed with a refined set of topological metrics. These experiments consider the news for which we were able to collect the respective diffusion network, and thus the dataset includes only 659 news (Table I).

First, we experimented with two data representations: a) raw values, as extracted from the original diffusion graphs; and b) normalization considering z-score, which expresses the features in terms of units of standard deviations. The boxplots in Figure 4 show the results considering all executions using all the five algorithms and all metrics, which reveals that the best results were obtained with the z-score normalization. In general, we observe this behavior for all metrics with a few exceptions. For all algorithms, the differences in the results are statistically significant, with a single exception, RF, where the null hypothesis was not refuted (Table X). Thus, all results reported in the remaining of this section were obtained using the topological metrics represented as z-scores.

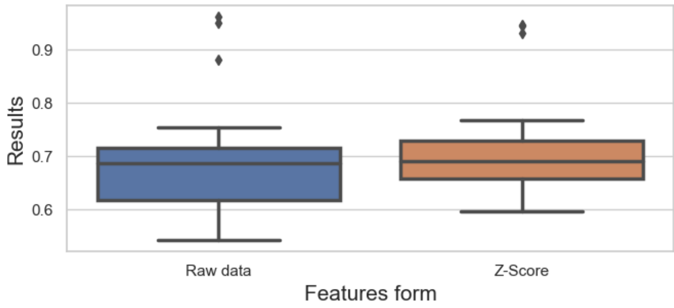

Fig. 4. Comparing results obtained from raw values and z-score normalized values

Table II. Topological metrics used as features per classifier

Topological Metric	**Acronym**	**Baseline**	**All_Metrics**	**Custom_Metrics**
Number of strongly connected components	scc	✓	✓	
Size of the largest strongly connected component	lscc	✓	✓	✓
Number of weakly connected components	wcc	✓	✓	
Size of the largest weakly connected component	lwcc	✓	✓	✓
Diameter of the largest weakly connected component	dwcc	✓	✓	✓
Clustering coefficient	cc	✓	✓	✓
K-Core number	kc	✓	✓	✓
Number of nodes	nodes		✓	✓
Number of edges	edges		✓	
Density of the network	density		✓	✓
Diameter of the network	diameter		✓	

In our previous work [Sáenz et al. 2020], we used the topological metrics originally proposed in [Pierri et al. 2020], which we consider our *baseline* for this set of experiments. In an attempt to improve these previous results, we considered four additional topological features. We shall refer to these classifiers as *all_metrics*. Table II details the metrics considered as features for each type of classifier.

The chart in Figure 5 compares the average results of the *baseline* and *all_metrics* classifiers, considering the weighted averaged precision, recall and F1 metrics (detailed results are summarized in Table XI). We observe improvements for all performance metrics and all algorithms. These improvements are statistically significant for all algorithms and metrics with two exceptions (Table XII). For the algorithms SVM and LR, the results for weighted precision are comparable. Improvements range from 0.2 pp (percentage points) to 5.6 pp in weighted precision; from 2.2 to 3.9 in weighted recall, and from 1.9 to 5.6 in weighted F1. In terms of algorithms, the best improvements were achieved on KNN classifiers, with an increment in the scores of near 3 pp for weighted precision, 4 pp for weighted recall and 5 pp for weighted F1. The most significant improvements were observed for the weighted F1 measure due to the improvements in both precision and recall. Thus, we conclude that these additional metrics improve FN classification.

Then we investigated if there were particularly influential topological metrics for FN classification. For this purpose, we used the leave-one-out technique, where we disregarded one metric each time and trained/tested the classifiers with all the other metrics. We did this for all topological metrics, and results are summarized in Table XIV. Figure 6 shows the differences for seven topological metrics for which we noticed any impact on the results when removed. The differences are calculated with regard to *all_metrics* classifiers. In general, the metrics with the greatest differences were *density* followed by the number of *nodes*, two of the newly proposed metrics. However, the behavior was very dependent on the classification algorithm used. For instance, for the RF algorithm, leaving out the *nodes* feature slightly improves the results (ranging between 1.4 and 1.6 pp). However, this

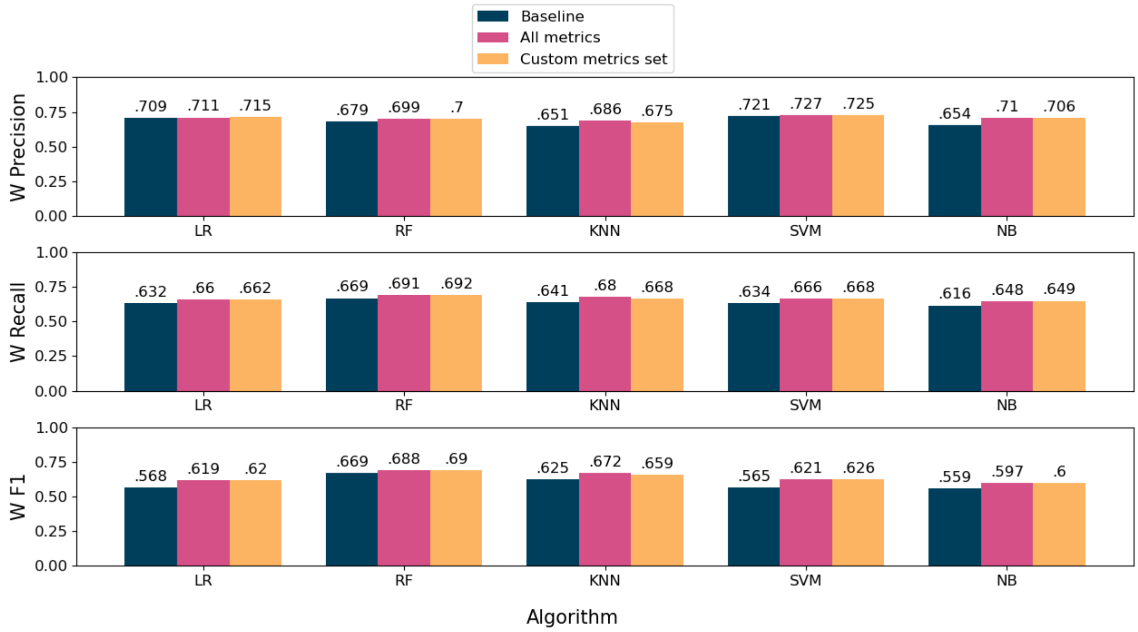

Fig. 5. Comparing weighted results from baseline, all_metrics and custom_metrics sets

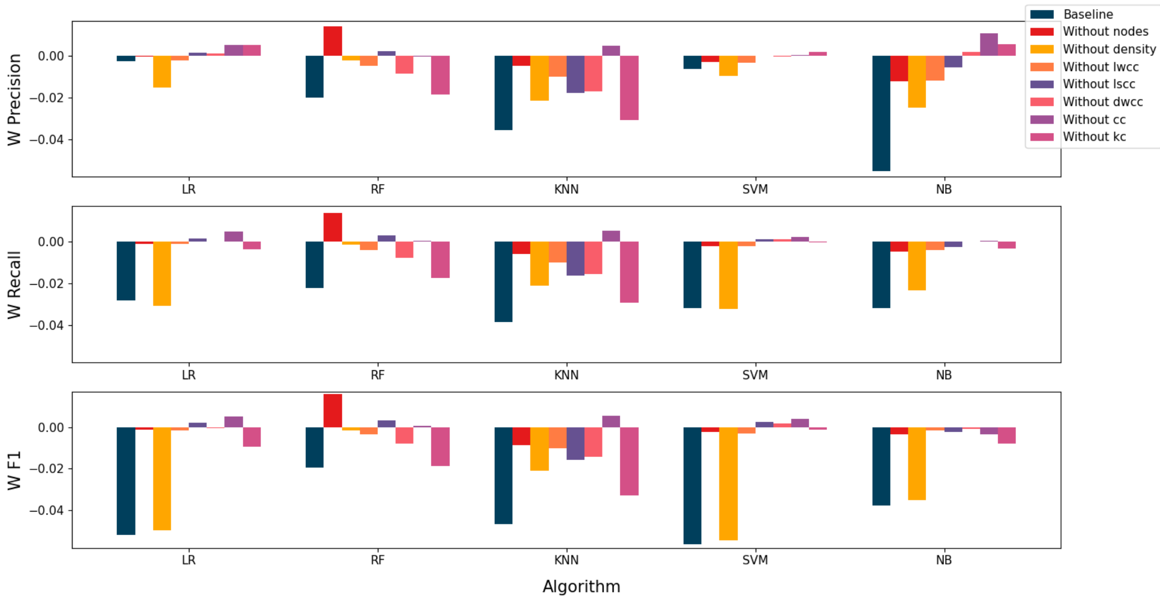

Fig. 6. Performance differences between all_metrics and baseline/Leave-one-out classifiers

behavior is not always consistent when compared to the other algorithms. Leave-one-out executions also yielded better results than *baseline* for most of the algorithms. Thus, we conclude that the value of the topological metrics can be verified when they are explored as a combination that represents the diffusion graph rather than individually.

Based on these results, we built classifiers using the topological metrics of which the leave-one-out exclusion experiments resulted in an impact, even if very small. We shall refer to classifiers based on this set of features as *custom_metrics*, where the metrics considered as features are listed in Table II. The chart in Figure 5 also compares the results of *custom_metrics* classifiers with the results of

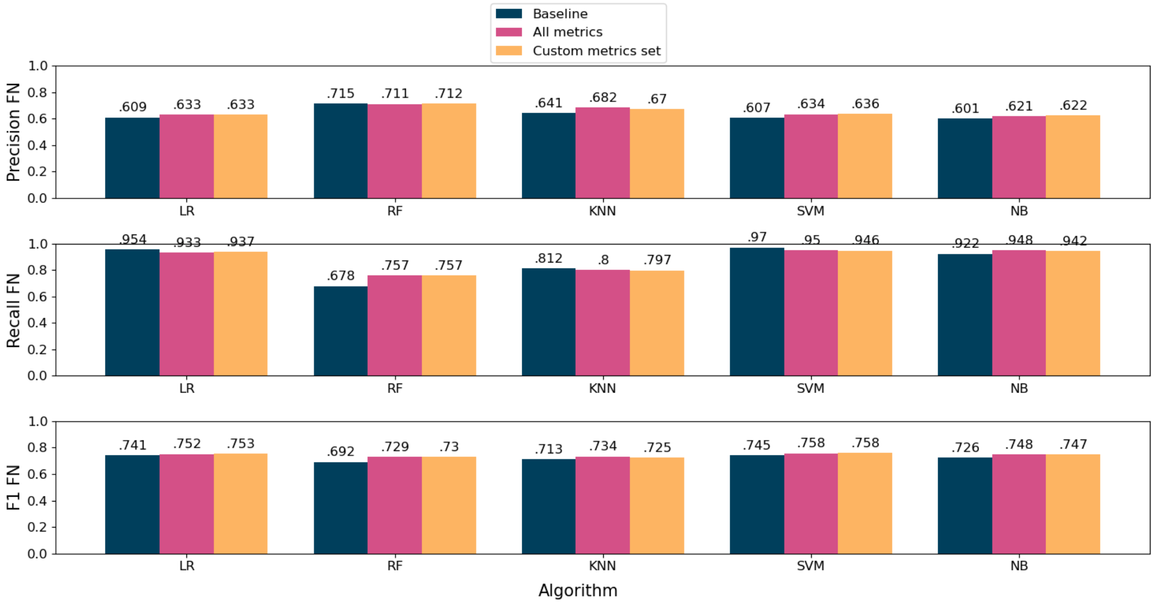

Fig. 7. Performance for the Fake News class (baseline, all_metrics and custom_metrics sets classifiers)

all_metrics and *baseline* classifiers. Like *all_metrics*, *custom_metrics* classifiers perform better than the *baseline* classifiers, where these improvements are statistically significant. However, the performance of *custom_metrics* and *all_metrics* is statistically comparable for all three metrics (Table XIII). Thus, we conclude that the excluded metrics were not relevant for the classification. The metrics included in *custom_metrics* highlight the size of the diffusion network, how separate/isolated the users are, and the existence and characteristics of potential news broadcasting groups.

Finally, we assessed the specific results for the Fake News class. Figure 7 details these results for precision, recall and F1. The statistical tests revealed that the F1 performance *all_metrics* and *custom_metrics* are comparable, but statistically superior to the baseline. Compared to the performance of the baseline classifiers, we observed that the use of more features (all_metrics or custom_metrics) increases the precision, sometimes at the expense of recall, with a few exceptions. In the specific case of the RF algorithm, for instance, the precision decreases, and the recall increases. By comparing the FN results with the averaged results, we can observe that the recall for FN detection is higher when compared to RN (real news) recall. For instance, while the recall for the LR classifier is close to 94%, the averaged recall is 30 pp lower for the same algorithm. However, the difference in terms of precision is much smaller, ranging from 1.5 pp to 8.9 pp. From these results, we conclude that the predictive models using topological metrics display a good performance for classifying fake news.

5.5 Experiments with combinations of features

This final set of experiments aim to assess the combination of DistilBERT representations of news and topological features. First, we assessed the differences of performance between classifiers using DistilBERT features only, and the concatenation of these features with topological ones. Notice the results are not the same reported in Section 5.3 since a different dataset is used. The second assessment involved the combination of classifiers in mixture-of-experts ensembles, using both voting and averaged probabilities to make a final prediction.

First, we compared the performance of the models constructed with and without the topological features. The later combine in a single vector the DistilBERT representations of title and/or text and the topological features (*Custom* metrics set in Table II). The averaged results are presented in Table

III. Among the models using topological features, the one combining the title and text yielded the best result, which is statistically significant (Table XV). When compared to their counterpart using textual features only, we observe that, in most cases, the concatenation of topological features with textual ones (title, news text, or both) improve the respective results achieved. The best absolute scores, considering the three performance metrics, were are achieved by using LR using the combination of the title, text and topological metrics (averaged $P = 88.9\%$, $R = 88.7\%$ and $F1 = 88.6\%$). However, improvements are statistically significant only for the SVM and Naïve Bayes algorithms, with a few exceptions. Nevertheless, this is a major improvement compared to our previous work [Sáenz et al. 2020], when the inclusion of topological metrics (*Baseline* metrics in Table II) affected negatively the results.

Regarding the algorithms, the best results were produced by LR and RF, (with no statistical differences between them), followed by KNN models (Table XVI). The performances achieved using these algorithms are consistent with those reported in [Reis et al. 2019] for fake news detection using distinct kinds of features.

Table III. Weighted averaged scores of news classification using title, text and topological features using ML algorithms

	LR	**RF**	**KNN**	**SVM**	**NB**
Weighted Precision					
Title	**.869 (± .006)**	.855 (± .006)	.826 (± .004)	.827 (± .003)	.828 (± .003)
Text	.870 (± .003)	**.872 (± .004)**	.826 (± .006)	.814 (± .006)	.719 (± .005)
Title and Text	**.888 (± .003)**	**.888 (± .006)**	.848 (± .005)	.843 (± .003)	.765 (± .005)
Title and Top.	**.872 (± .006)**	.856 (± .006)	.829 (± .003)	.832 (± .004)	.835 (± .003)
Text and Top.	.869 (± .005)	**.874 (± .005)**	.826 (± .005)	.820 (± .007)	.725 (± .007)
Title, Text and Top.	**.889 (± .004)**	.888 (± .003)	.845 (± .006)	.847 (± .002)	.773 (± .005)
Weighted Recall					
Title	**.864 (± .005)**	.849 (± .005)	.821 (± .005)	.809 (± .003)	.823 (± .004)
Text	.867 (± .003)	**.870 (± .004)**	.821 (± .005)	.798 (± .006)	.701 (± .004)
Title and Text	**.885 (± .004)**	**.885 (± .005)**	.844 (± .004)	.830 (± .003)	.756 (± .004)
Title and Top.	**.867 (± .006)**	.851 (± .006)	.825 (± .003)	.814 (± .004)	.831 (± .004)
Text and Top.	.867 (± .005)	**.871 (± .004)**	.819 (± .004)	.800 (± .006)	.709 (± .005)
Title, Text and Top.	**.887 (± .004)**	.885 (± .003)	.841 (± .006)	.834 (± .002)	.764 (± .004)
Weighted F1					
Title	**.864 (± .005)**	.848 (± .006)	.820 (± .005)	.803 (± .003)	.823 (± .004)
Text	.867 (± .003)	**.870 (± .004)**	.819 (± .005)	.793 (± .007)	.685 (± .003)
Title and Text	**.885 (± .004)**	.884 (± .005)	.843 (± .004)	.826 (± .004)	.751 (± .005)
Title and Top.	**.867 (± .006)**	.850 (± .007)	.825 (± .003)	.808 (± .005)	.831 (± .004)
Text and Top.	.867 (± .005)	**.871 (± .004)**	.817 (± .004)	.794 (± .006)	.695 (± .007)
Title, Text and Top.	**.886 (± .004)**	.884 (± .004)	.840 (± .006)	.830 (± .002)	.760 (± .005)

Finally, we explored mixture-of-expert ensembles to investigate if combinations of these classifiers would result in better performance. We constructed many ensembles as combinations of 2 to 5 base classifiers, all of them trained using the same dataset and set of features, but with distinct algorithms. To find the best ensemble, we tried the following variations: a) combinations of classifiers based on distinct algorithms, b) two combination rules (majority of votes and average probability), and c) variations on the features (i.e., only title, title and topological, only text, text and topological, title and text and title, text and topological).

Figure 8 presents the weighted F1 distributions, considering all ensembles using as learning function the average probability and majority voting. In average, the ensembles based on majority voting concentrate the best results, with the highest median ($WF1 = 85.4\%$), and first/third quartile values (83.8% and 86.6%, respectively). The probability-based ensembles presented a less consistent performance (median $WF1 = 84.9\%$, Q1 $WF1 = 83.1\%$ and Q3 $WF1 = 86.5\%$), with more outliers below Q1. A statistical t-test revealed these differences are significant ($p - value = 0.033$). However, a few average probability ensembles achieved the best absolute F1 scores. For both learning functions,

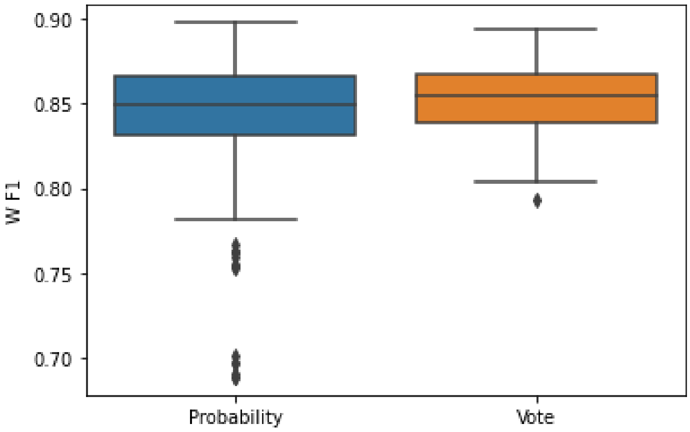

Fig. 8. Comparing classification committees grouped by majority vote or probability

Table IV. Performance comparison of the best algorithm with the best ensembles, according to the type of features

	LR	RF	KNN	Prob_LR_RF_KNN	Voting_LR_RF_KNN
Weighted Precision					
Title	.869 (± .006)	.856 (± .007)	.826 (± .004)	**.870** (± **.003**)	.865 (± .005)
Text	.870 (± .003)	.868 (± .007)	.826 (± .006)	**.872** (± **.005**)	**.872** (± **.005**)
Title and Text	.888 (± .003)	.888 (± .006)	.848 (± .005)	**.893** (± **.004**)	.892 (± .003)
Title and Top.	**.872** (± **.006**)	.857 (± .007)	.829 (± .003)	.870 (± .006)	.867 (± .007)
Text and Top.	.869 (± .005)	.874 (± .005)	.826 (± .005)	**.883** (± **.007**)	.879 (± .006)
Title, Text and Top.	.889 (± .004)	.887 (± .006)	.845 (± .006)	**.900** (± **.003**)	.896 (± .006)
Weighted Recall					
Title	.864 (± .005)	.851 (± .007)	.821 (± .005)	**.865** (± **.004**)	.860 (± .005)
Text	.867 (± .003)	.865 (± .008)	.821 (± .005)	.869 (± .005)	**.870** (± **.005**)
Title and Text	.885 (± .004)	.886 (± .005)	.844 (± .004)	**.891** (± **.004**)	.889 (± .003)
Title and Top.	**.867** (± **.006**)	.852 (± .007)	.825 (± .003)	.866 (± .005)	.863 (± .006)
Text and Top.	.867 (± .005)	.871 (± .006)	.819 (± .004)	**.880** (± **.007**)	.876 (± .006)
Title, Text and Top.	.887 (± .004)	.884 (± .006)	.841 (± .006)	**.897** (± **.003**)	.894 (± .006)
Weighted F1					
Title	.864 (± .005)	.850 (± .007)	.820 (± .005)	**.865** (± **.004**)	.859 (± .005)
Text	.867 (± .003)	.865 (± .008)	.819 (± .005)	.868 (± .005)	**.870** (± **.005**)
Title and Text	.885 (± .004)	.885 (± .005)	.843 (± .004)	**.891** (± **.004**)	.889 (± .003)
Title and Top.	**.867** (± **.006**)	.851 (± .006)	.825 (± .003)	.866 (± .005)	.862 (± .006)
Text and Top.	.867 (± .005)	.871 (± .006)	.817 (± .004)	**.880** (± **.007**)	.876 (± .006)
Title, Text and Top.	.886 (± .004)	.884 (± .006)	.840 (± .006)	**.897** (± **.003**)	.893 (± .006)

the best performances were achieved by ensembles composed of classifiers trained with LR, RF and KNN (89.3% for majority voting, and 89.7% for average probability). Models solely trained with LR, RF and KNN algorithms achieved the best scores in previous experiment (Table III), and their combination in an ensemble boosted the model performance.

Table IV compares the results in terms of weighted averaged precision, recall and F1 achieved by the best machine learning algorithm (LR), the best voting by majority ensemble and the best average probability ensemble. These results enable to compare the performance according to the set of features explored. It is possible to confirm that the best results are achieved by combining features representing the news title, text and topological features, an expected result given the individual performance of the base classifiers. These improvements are statistically significant, with a single exception (Table XVII). In terms of weighted F1, the improvements range from 0.2 to 0.7 pp points in the voting by majority ensemble, and 0.2 to 1.1 pp points in the average probability ensemble. These differences are consistent with increases in precision and recall.

Table IV reveals that the performance of the best voting and average probability ensembles (i.e., composed by title, text and topological) is improved by 0.4 and 0.6 pp, respectively, when compared

Table V. Positive class (FN) scores of news classification using title, text and topological features comparing best algorithm with best committees

	LR	RF	KNN	Prob_LR_RF_KNN	Voting_LR_RF_KNN
FN Precision					
Title	**.877 (± .007)**	.844 (± .008)	.825 (± .006)	.866 (± .004)	.857 (± .006)
Text	**.876 (± .005)**	.873 (± .008)	.813 (± .006)	.869 (± .007)	.870 (± .007)
Title and Text	**.896 (± .008)**	.887 (± .008)	.838 (± .004)	.895 (± .005)	.890 (± .005)
Title and Top.	**.884 (± .007)**	.845 (± .007)	.828 (± .005)	.868 (± .004)	.864 (± .009)
Text and Top.	**.876 (± .006)**	.873 (± .007)	.805 (± .003)	**.876 (± .007)**	.874 (± .005)
Title, Text and Top.	.897 (± .006)	.884 (± .005)	.834 (± .006)	**.902 (± .005)**	.895 (± .006)
FN Recall					
Title	.880 (± .007)	**.899 (± .005)**	.859 (± .006)	.897 (± .008)	.898 (± .006)
Text	.885 (± .003)	.885 (± .009)	.880 (± .006)	.899 (± .006)	**.900 (± .004)**
Title and Text	.898 (± .005)	.910 (± .007)	.891 (± .007)	.911 (± .005)	**.913 (± .007)**
Title and Top.	.877 (± .008)	**.898 (± .009)**	.864 (± .005)	.895 (± .009)	.894 (± .009)
Text and Top.	.883 (± .007)	.897 (± .006)	.890 (± .006)	**.912 (± .008)**	.906 (± .007)
Title, Text and Top.	.898 (± .004)	.910 (± .008)	.888 (± .010)	.914 (± .005)	**.915 (± .008)**
FN F1					
Title	.877 (± .005)	.869 (± .006)	.840 (± .003)	**.879 (± .004)**	.875 (± .006)
Text	.879 (± .002)	.878 (± .006)	.843 (± .004)	.883 (± .004)	**.884 (± .003)**
Title and Text	.895 (± .004)	.897 (± .005)	.862 (± .004)	**.902 (± .004)**	.901 (± .004)
Title and Top.	.879 (± .005)	.869 (± .007)	.844 (± .003)	**.880 (± .006)**	.877 (± .007)
Text and Top.	.879 (± .005)	.884 (± .005)	.844 (± .004)	**.893 (± .006)**	.889 (± .006)
Title, Text and Top.	.896 (± .004)	.896 (± .005)	.859 (± .006)	**.907 (± .003)**	.904 (± .006)

to their counterpart, trained using title or text only. However, this improvement is statistically significant only for the average probability ensemble. Thus, we conclude that the ensemble using average probability as meta-learning function, and base classifiers that concatenate textual and topological features yields the best result. We achieved superior results when compared to Safe, and very near results when compared to the state-of-the art dEFEND and GCAL. Recall that these works should be regarded as references, rather than baselines.

Finally, Table V details the results for the Fake News class only, also in terms of the best machine learning algorithm (LR), majority voting ensemble and average probability ensemble. These results enable to compare the performance according to the set of features explored. It endorses the benefits of the topological features combined with ensembles for the classification. In terms of F1, the improvements of the ensembles for textual features vary from 0.3 to 0.4 pp and the for the use of ensembles combined with topological features vary from 0.1 to 1.4 pp. These differences are consistent with precision and recall results. Again, the best result was achieved by the probability ensemble ($F1 = 90.7\%$) using title, text and topological metrics of news as features, which is statistically significant. Compared to the LR model, it improved 1.1 pp, highlighting the relevance of the use of ensembles and topological metrics.

Our experiments revealed that the use of ensembles combined with topological features improve the performance of textual DistilBERT models based on single classifiers and can reach near state-of-the-art results.

6. CONCLUSIONS

The present work proposed an FN classification process based on the compact representation of news content (title and text) using DistilBERT and the metrics representing their dissemination in the social network. We explored different combinations of these features, using five different classification algorithms and stacking ensembles.

By generating features using DistilBERT only on the textual attributes of the news, we achieved results comparable to the state of the art [Shu et al. 2019; Zhou et al. 2020], and superior to most works that used the Politifact data set (e.g., [Shu et al. 2020; Papanastasiou et al. 2019]. Among these works, there are both minimum requirement approaches (news content only) [Zhou et al. 2020], which are applicable in the context of early news detection, and approaches that extract information from the respective diffusion network [Shu et al. 2019; Shu et al. 2020; Papanastasiou et al. 2019]. Note that even applied only to the news title, the proposed approach achieves very good performance, denoting the discriminatory capacity of this feature.

We considered new metrics for representing the diffusion network, and assessed their contribution for fake news classification. Our results show that the inclusion of topological metrics as features improves the classification of fake news, mainly by improving the recall. These results were observed for individual machine learning algorithms and ensembles.

Future work includes, among other topics, the assessment of our proposal in other datasets; the execution of experiments using the original BERT model instead of the distilled version; the analysis of metrics representative of the news dissemination topology (e.g., centralities); the exploration of deep learning algorithms to combine multi-modal features; the investigation of more complex ensemble topologies and late fusion approaches for fake news classification; the study of features extracted from images; and the addition of interpretability mechanisms.

REFERENCES

Bauskar, S., Badole, V., Jain, P., and Chawla, M. Natural Language Processing based Hybrid Model for Detecting Fake News Using Content-Based Features and Social Features. *International Journal of Information Engineering and Electronic Business* 11 (4): 1–10, 2019.

Bondielli, A. and Marcelloni, F. A survey on fake news and rumour detection techniques. *Information Sciences* vol. 497, pp. 38–55, 2019.

Costa, L. d. F., Rodrigues, F. A., Travieso, G., and Villas Boas, P. R. Characterization of complex networks: A survey of measurements. *Advances in Physics* 56 (1): 167–242, Jan, 2007.

da Fonseca Vieira, V., da Silva Felix, L. G., Barbosa, C. M. G., and Xavier, C. R. Investigating the relation between companies with topological analysis of a network of stock exchange in brazil. *J. Inf. Data Manag.* 10 (3), 2019.

Demšar, J. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research* 7 (1): 1–30, 2006.

Devlin, J., Chang, M., Lee, K., and Toutanova, K. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proc. of the 2019 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, (NAACL-HLT)*, J. Burstein, C. Doran, and T. Solorio (Eds.). pp. 4171–4186, 2019.

Gadzicki, K., Khamsehashari, R., and Zetzsche, C. Early vs late fusion in multimodal convolutional neural networks. In *2020 IEEE 23rd International Conference on Information Fusion (FUSION)*. pp. 1–6, 2020.

Hansen, D. L., Shneiderman, B., Smith, M. A., and Himelboim, I. Chapter 3 - social network analysis: Measuring, mapping, and modeling collections of connections. In *Analyzing Social Media Networks with NodeXL (Second Edition)*, Second Edition ed., D. L. Hansen, B. Shneiderman, M. A. Smith, and I. Himelboim (Eds.). Morgan Kaufmann, pp. 31 – 51, 2020.

Leão, J. C., Laender, A. H. F., and de Melo, P. O. S. V. Overcoming bias in community detection evaluation. *J. Inf. Data Manag.* 11 (3), 2020.

Liao, H., Liu, Q., Shu, K., and xie, X. Fake news detection through graph comment advanced learning. *arXiv preprint arXiv:2011.01579*, 2021.

Murphy, K. P. *Machine Learning: A Probabilistic Perspective.* The MIT Press, 2012.

Papanastasiou, F., Katsimpras, G., and Paliouras, G. Tensor factorization with label information for fake news detection. *arXiv preprint arXiv:1908.03957*, 2019.

Pierri, F., Piccardi, C., and Ceri, S. Topology comparison of twitter diffusion networks effectively reveals misleading information. *Scientific Reports* 10 (1), Jan, 2020.

Polikar, R. Ensemble learning. *Scholarpedia* 4 (1): 2776, 2009.

Reis, J. C. S., Correia, A., Murai, F., Veloso, A., and Benevenuto, F. Supervised learning for fake news detection. *IEEE Intelligent Systems* 34 (2): 76–81, 2019.

Sanh, V., Debut, L., Chaumond, J., and Wolf, T. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

Shu, K., Cui, L., Wang, S., Lee, D., and Liu, H. Defend: Explainable fake news detection. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery Data Mining*. KDD '19. Association for Computing Machinery, New York, NY, USA, pp. 395–405, 2019.

Shu, K., Mahudeswaran, D., Wang, S., Lee, D., and Liu, H. Fakenewsnet: A data repository with news content, social context and dynamic information for studying fake news on social media. *arXiv preprint arXiv:1809.01286* vol. 8, 2018.

Shu, K., Mahudeswaran, D., Wang, S., and Liu, H. Hierarchical propagation networks for fake news detection: Investigation and exploitation. In *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 14. pp. 626–637, 2020.

Shu, K., Sliva, A., Wang, S., Tang, J., and Liu, H. Fake news detection on social media: A data mining perspective. *SIGKDD Explor. Newsl.* 19 (1): 22–36, Sept., 2017.

Shu, K., Zhou, X., Wang, S., Zafarani, R., and Liu, H. The role of user profiles for fake news detection. In *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. pp. 436–439, 2019.

Sáenz, C. A. C., Dias, M., and Becker, K. Combining compact news representations generated using distilbert and topological features to classify fake news. In *Anais do VIII Symposium on Knowledge Discovery, Mining and Learning*. SBC, Porto Alegre, RS, Brasil, pp. 209–216, 2020.

Takikawa, H. and Nagayoshi, K. Political polarization in social media: Analysis of the "twitter political field" in japan. In *2017 IEEE International Conference on Big Data (Big Data)*. pp. 3143–3150, 2017.

Wang, T., Lin, C., and Lin, H. Dga botnet detection utilizing social network analysis. In *2016 International Symposium on Computer, Consumer and Control (IS3C)*. pp. 333–336, 2016.

Wang, W. Y. "Liar, liar pants on fire
: A new benchmark dataset for fake news detection. *arXiv preprint arXiv:1705.00648*, 2017.

Zhang, C. and Ma, Y. *Ensemble Machine Learning: Methods and Applications*. Springer Publishing Company, Incorporated, 2012.

Zhang, L., Wang, S., and Liu, B. Deep learning for sentiment analysis: A survey. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* 8 (4), 2018.

Zhou, X., Wu, J., and Zafarani, R. Safe: Similarity-aware multi-modal fake news detection. *arXiv preprint arXiv:2003.04981*, 2020.

Zhou, X. and Zafarani, R. Network-based fake news detection: A pattern-driven approach. *SIGKDD Explor. Newsl.* 21 (2): 48–60, Nov., 2019.

Zhou, X. and Zafarani, R. A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Comput. Surv.* 53 (5), Sept., 2020.

APPENDIX A. DETAILED RESULTS FOR MODELS PERFORMANCES

APPENDIX A.1 Experiments with DistilBERT

Table VI presents the average performance of each model trained using compact representations of news textual features, with the respective standard deviation. Bold cells represent the best result per type of feature, and shaded cells the best results per performance metric. Tables VII and VIII presents the p-values of the t-tests performed to compare pair of models per type of feature and algorithms, respectively. Shaded cells indicate the cases where the null hypotheses was not refuted (i.e. there is no significant difference between the models). These conventions are adopted for all tables in this appendix.

Table VI. Average results for experiments using Title, Text and the combination of them

	LR	RF	KNN	SVM	NB
Weighted F1					
Title	.884 (± .004)	**.863 (± .005)**	.849 (± .005)	.838 (± .003)	.840 (± .004)
Text	.884 (± .005)	**.870 (± .006)**	.846 (± .008)	.837 (± .003)	.630 (± .002)
Title and Text	**.906 (± .005)**	.892 (± .004)	.871 (± .004)	.878 (± .002)	.742 (± .003)
FN F1					
Title	**.862 (± .005)**	.837 (± .007)	.820 (± .006)	.797 (± .004)	.819 (± .004)
Text	**.863 (± .005)**	.845 (± .008)	.820 (± .008)	.800 (± .004)	.678 (± .001)
Title and Text	**.890 (± .006)**	.871 (± .004)	.850 (± .005)	.852 (± .002)	.743 (± .002)

Table VII. p-values of pairwise t-test comparison of models using different types of textual features

	Title and Text					Title				
	LR	**RF**	**KNN**	**SVM**	**NB**	**LR**	**RF**	**KNN**	**SVM**	**NB**
Weighted F1										
Title	4.05E-09	5.12E-10	4.66E-09	5.03E-18	3.85E-22					
Text	5.51E-09	1.65E-07	4.99E-08	3.06E-18	5.47E-27	0.92	3.13E-02	0.36	0.43	8.30E-29
FN F1										
Title	1.66E-09	1.16E-09	7.67E-10	3.49E-19	1.16E-20					
Text	1.44E-09	9.10E-08	1.48E-08	3.46E-18	3.98E-23	0.70	0.08	0.86	0.06	3.36E-26

Table VIII. p-values of pairwise t-test comparison of models using different algorithms

	LR			RF			SVM		
	Title and Text	**Text**	**Title**	**Title and Text**	**Text**	**Title**	**Title and Text**	**Text**	**Title**
Weighted F1									
RF	1.39E-05	3.41E-04	2.55E-07						
SVM	2.22E-12	1.36E-15	8.91E-16	2.03E-08	1.55E-11	9.13E-11			
KNN	7.47E-13	1.70E-10	6.51E-12	7.25E-10	2.82E-07	2.60E-06	2.92E-05	4.18E-03	4.92E-05
NB	1.79E-25	1.13E-29	2.05E-14	3.57E-25	6.52E-27	1.46E-09	2.42E-28	6.08E-31	0.32
FN F1									
RF	6.39E-07	4.88E-05	8.37E-07						
SVM	1.62E-13	4.24E-17	3.62E-17	2.01E-10	2.79E-12	1.99E-12			
KNN	2.81E-12	4.67E-11	5.63E-12	5.70E-09	9.22E-07	2.74E-06	2.01E-10	2.79E-12	1.99E-12
NB	1.30E-23	3.72E-27	1.12E-13	4.69E-24	7.98E-23	2.41E-07	1.65E-26	3.09E-25	2.05E-10

APPENDIX A.2 Experiments with topological metrics

Table IX presents the average performance of each model trained using the raw or normalized values for the topological metrics, with the respective standard deviation, and Table X presents the p-values of the t-tests performed to compare pair of models according to these representations. Table XI displays the average performance of each model according to each set of topological metrics (baseline, all metrics and custom), and tables XII and XIII present the p-values of the t-tests performed to compare pair of models per topological metrics. Table XIV display the average results for the Leave-One-Out models, with the respective standard deviations.

Table IX. Average results for experiments using raw data and z-score normalized values for topological metrics

	LR	RF	KNN	SVM	NB
Weighted Precision					
Raw data	**.718 (± .013)**	.696 (± .012)	.681 (± .013)	.607 (± .007)	.709 (± .009)
Z-Score	.713 (± .006)	.696 (± .011)	.697 (± .010)	**.726 (± .009)**	.708 (± .007)
Weighted Recall					
Raw data	.633 (± .008)	**.689 (± .012)**	.674 (± .011)	.594 (± .006)	.630 (± .003)
Z-Score	.659 (± .005)	.688 (± .010)	**.691 (± .009)**	.663 (± .004)	.647 (± .003)
Weighted F1					
Raw data	.571 (± .009)	**.687 (± .012)**	.671 (± .011)	.543 (± .009)	.562 (± .004)
Z-Score	.618 (± .007)	.686 (± .010)	**.688 (± .010)**	.619 (± .006)	.596 (± .003)
FN Precision					
Raw data	.608 (± .004)	**.707 (± .009)**	.696 (± .010)	.590 (± .004)	.605 (± .002)
Z-Score	.631 (± .003)	**.707 (± .009)**	.703 (± .009)	.631 (± .003)	.620 (± .002)
FN Recall					
Raw data	.951 (± .012)	.755 (± .017)	.740 (± .012)	.881 (± .005)	**.962 (± .003)**
Z-Score	.932 (± .003)	.756 (± .015)	.768 (± .009)	**.947 (± .002)**	.944 (± .004)
FN F1					
Raw data	.739 (± .007)	.727 (± .012)	.713 (± .009)	.704 (± .003)	**.740 (± .002)**
Z-Score	.750 (± .004)	.726 (± .010)	.731 (± .008)	**.755 (± .002)**	.746 (± .002)

Table X. p-values of pairwise t-test comparison of models results using raw data and normalized data

		Z-Score				
		LR	**RF**	**KNN**	**SVM**	**NB**
Raw data	**Weighted Precision**	0.40	0.34	9.89E-03	8.22E-17	0.95
	Weighted Recall	1.60E-07	0.37	1.24E-03	1.48E-16	4.28E-11
	Weighted F1	4.12E-10	0.39	2.78E-03	1.88E-14	2.73E-14
	FN Precision	7.50E-11	0.32	0.12	2.74E-15	8.27E-12
	FN Recall	3.27E-04	0.56	2.68E-05	1.01E-18	9.07E-10
	FN F1	3.96E-04	0.45	3.57E-04	2.17E-18	1.83E-05

Table XI. Average results for experiments using all metrics, a custom metrics set and metrics from the baseline

	LR	RF	KNN	SVM	NB
Weighted Precision					
All metrics	.711 (± .006)	.699 (± .013)	.686 (± .010)	**.727 (± .009)**	.710 (± .007)
Baseline	.709 (± .017)	.679 (± .012)	.651 (± .010)	**.721 (± .005)**	.654 (± .012)
Custom metrics set	.715 (± .008)	.700 (± .007)	.675 (± .011)	**.725 (± .009)**	.706 (± .007)
Weighted Recall					
All metrics	.660 (± .005)	**.691 (± .013)**	.680 (± .009)	.666 (± .004)	.648 (± .003)
Baseline	.632 (± .006)	**.669 (± .011)**	.641 (± .010)	.634 (± .002)	.616 (± .003)
Custom metrics set	.662 (± .004)	**.692 (± .007)**	.668 (± .010)	.668 (± .005)	.649 (± .002)
Weighted F1					
All metrics	.619 (± .007)	**.688 (± .014)**	.672 (± .010)	.621 (± .006)	.597 (± .003)
Baseline	.568 (± .005)	**.669 (± .011)**	.625 (± .010)	.565 (± .004)	.559 (± .004)
Custom metrics set	.620 (± .005)	**.690 (± .007)**	.659 (± .011)	.626 (± .006)	.600 (± .002)
Precision					
All metrics	.633 (± .003)	**.711 (± .010)**	.682 (± .009)	.634 (± .003)	.621 (± .002)
Baseline	.609 (± .003)	**.715 (± .012)**	.641 (± .010)	.607 (± .002)	.601 (± .002)
Custom metrics set	.633 (± .002)	**.712 (± .008)**	.670 (± .011)	.636 (± .003)	.622 (± .002)
Recall					
All metrics	.933 (± .003)	.757 (± .017)	.800 (± .009)	**.950 (± .002)**	.948 (± .004)
Baseline	.954 (± .012)	.678 (± .011)	.812 (± .015)	**.970 (± .002)**	.922 (± .004)
Custom metrics set	.937 (± .005)	.757 (± .007)	.797 (± .012)	**.946 (± .003)**	.942 (± .003)
F1					
All metrics	.752 (± .004)	.729 (± .013)	.734 (± .008)	**.758 (± .002)**	.748 (± .002)
Baseline	.741 (± .005)	.692 (± .010)	.713 (± .011)	**.745 (± .002)**	.726 (± .003)
Custom metrics set	.753 (± .003)	.730 (± .006)	.725 (± .009)	**.758 (± .003)**	.747 (± .002)

Table XII. p-values of pairwise t-test comparison of models using All metrics and Baseline metrics

		All metrics				
		LR	**RF**	**KNN**	**SVM**	**NB**
Baseline	**Weighted Precision**	0.25	3.83E-02	4.05E-06	0.31	3.79E-10
	Weighted Recall	4.22E-10	1.38E-02	9.99E-07	1.25E-13	6.08E-15
	W F1	3.44E-13	4.47E-02	6.82E-06	4.78E-15	1.07E-14

Table XIII. p-values of pairwise t-test comparison of models using different sets of topological metrics

	Custom metrics				
	LR	**RF**	**KNN**	**SVM**	**NB**
Weighted Precision					
Baseline	0.05	3.09E-03	3.22E-04	0.35	6.60E-10
All metrics	0.11	0.64	0.09	0.94	0.48
Weighted Recall					
Baseline	4.61E-11	5.29E-04	2.14E-04	6.37E-13	3.06E-16
All metrics	0.47	0.62	4.18E-02	0.41	0.22
Weighted F1					
Baseline	5.18E-15	2.69E-03	1.57E-03	8.35E-15	3.89E-16
All metrics	0.84	0.58	4.41E-02	0.30	7.70E-03
Precision					
Baseline	1.47E-14	0.07	0.70	3.15E-14	9.46E-14
All metrics	0.83	0.60	3.30E-02	0.26	4.20E-02
Recall					
Baseline	6.18E-03	3.54E-13	6.81E-09	4.37E-13	7.24E-09
All metrics	3.53E-02	0.66	0.92	0.11	4.70E-03
F1					
Baseline	5.42E-07	5.34E-08	3.20E-06	4.53E-09	7.04E-13
All metrics	0.30	0.56	0.13	0.70	0.67

Table XIV. Average results for baseline and leave-one-out experiments

	LR	RF	KNN	SVM	NB
Weighted Precision					
Baseline	.709 (± .017)	.679 (± .012)	.651 (± .010)	**.721 (± .005)**	.654 (± .012)
without nodes	.711 (± .006)	.713 (± .007)	.682 (± .010)	**.724 (± .008)**	.698 (± .008)
without edges	.711 (± .006)	.699 (± .011)	.679 (± .010)	**.722 (± .009)**	.700 (± .009)
without diameter	.715 (± .008)	.701 (± .012)	.680 (± .013)	**.730 (± .009)**	.712 (± .007)
without density	.696 (± .007)	.697 (± .011)	.665 (± .007)	**.718 (± .011)**	.685 (± .010)
without scc	.711 (± .006)	.698 (± .009)	.686 (± .010)	**.726 (± .010)**	.710 (± .007)
without lscc	.713 (± .008)	.701 (± .012)	.669 (± .006)	**.727 (± .009)**	.704 (± .006)
without wcc	.711 (± .006)	.698 (± .008)	.686 (± .010)	**.726 (± .010)**	.710 (± .007)
without lwcc	.709 (± .006)	.694 (± .008)	.676 (± .010)	**.724 (± .008)**	.698 (± .009)
without dwcc	.712 (± .008)	.690 (± .012)	.669 (± .014)	**.727 (± .009)**	.712 (± .006)
without cc	.716 (± .008)	.698 (± .009)	.691 (± .012)	**.727 (± .008)**	.720 (± .008)
without kc	.716 (± .010)	.680 (± .012)	.656 (± .010)	**.729 (± .010)**	.715 (± .010)
Weighted Recall					
Baseline	.632 (± .006)	**.669 (± .011)**	.641 (± .010)	.634 (± .002)	.616 (± .003)
without nodes	.659 (± .006)	**.705 (± .007)**	.674 (± .009)	.664 (± .004)	.643 (± .004)
without edges	.660 (± .005)	**.691 (± .011)**	.673 (± .008)	.663 (± .004)	.646 (± .004)
without diameter	.662 (± .005)	**.693 (± .013)**	.674 (± .011)	.668 (± .004)	.648 (± .003)
without density	.629 (± .005)	**.690 (± .011)**	.659 (± .008)	.633 (± .005)	.625 (± .003)
without scc	.660 (± .005)	**.690 (± .009)**	.680 (± .009)	.666 (± .004)	.648 (± .003)
without lscc	.662 (± .006)	**.694 (± .011)**	.664 (± .005)	.667 (± .004)	.646 (± .002)
without wcc	.660 (± .005)	**.691 (± .007)**	.680 (± .009)	.666 (± .004)	.648 (± .003)
without lwcc	.659 (± .004)	**.687 (± .008)**	.670 (± .008)	.663 (± .004)	.644 (± .004)
without dwcc	.660 (± .005)	**.683 (± .011)**	.664 (± .011)	.667 (± .005)	.648 (± .002)
without cc	.665 (± .004)	**.691 (± .009)**	.685 (± .011)	.668 (± .004)	.649 (± .003)
without kc	.656 (± .005)	**.673 (± .012)**	.651 (± .008)	.665 (± .004)	.645 (± .003)
Weighted F1					
Baseline	.568 (± .005)	**.669 (± .011)**	.625 (± .010)	.565 (± .004)	.559 (± .004)
without nodes	.618 (± .008)	**.704 (± .007)**	.664 (± .010)	.619 (± .005)	.593 (± .006)
without edges	.619 (± .006)	**.689 (± .011)**	.664 (± .009)	.619 (± .005)	.598 (± .005)
without diameter	.620 (± .005)	**.690 (± .013)**	.666 (± .012)	.624 (± .006)	.596 (± .004)
without density	.570 (± .005)	**.687 (± .011)**	.651 (± .008)	.567 (± .005)	.562 (± .004)
without scc	.619 (± .007)	**.688 (± .009)**	.672 (± .010)	.623 (± .006)	.597 (± .003)
without lscc	.621 (± .008)	**.692 (± .011)**	.657 (± .006)	.624 (± .006)	.594 (± .002)
without wcc	.619 (± .007)	**.688 (± .007)**	.672 (± .010)	.623 (± .006)	.597 (± .003)
without lwcc	.618 (± .006)	**.685 (± .008)**	.662 (± .009)	.618 (± .006)	.595 (± .006)
without dwcc	.619 (± .006)	**.680 (± .011)**	.658 (± .012)	.623 (± .006)	.596 (± .003)
without cc	.624 (± .005)	**.689 (± .009)**	.678 (± .011)	.625 (± .006)	.593 (± .005)
without kc	.610 (± .005)	**.670 (± .013)**	.639 (± .008)	.620 (± .005)	.589 (± .003)

APPENDIX A.3 Experiments with combinations of features

Tables XV and XVI present the p-values of the t-tests performed to compare pair of models according to the features combined and algorithms. Table XVII present the p-values of the t-tests performed to compare them.

Table XV. p-values of pairwise t-test comparison of models with different combinations of features

	Title and Topological					Text and Topological					Title ,Text and Topological				
	LR	RF	KNN	SVM	NB	LR	RF	KNN	SVM	NB	LR	RF	KNN	SVM	NB
Weighted Precision															
Title	0.34	0.82	0.06	9.28E-03	1.12E-04										
Text						0.81	0.66	0.92	0.09	2.98E-02					
Title and Text											0.43	0.39	0.30	4.67E-03	2.37E-03
Title and Top.											9.40E-07	6.95E-11	1.07E-06	2.24E-08	1.55E-17
Text and Top.											2.07E-08	1.10E-07	1.12E-06	6.86E-10	1.32E-12
Weighted Recall															
Title	0.27	0.91	2.45E-02	5.25E-03	3.71E-04										
Text						0.82	0.88	0.62	0.42	1.63E-03					
Title and Text											0.52	0.27	0.20	8.00E-03	2.72E-04
Title and Top.											1.29E-07	9.86E-12	1.29E-06	1.97E-10	1.24E-18
Text and Top.											7.25E-08	3.17E-08	3.35E-08	3.32E-12	1.99E-15
Weighted F1															
Title	0.28	0.94	2.78E-02	1.14E-02	5.70E-04										
Text						0.86	0.93	0.41	0.65	1.55E-03					
Title and Text											0.53	0.29	0.21	1.19E-02	1.11E-03
Title and Top.											1.62E-07	1.98E-11	2.12E-06	2.61E-10	4.73E-18
Text and Top.											7.17E-08	4.13E-08	1.56E-08	4.62E-12	4.41E-15

Table XVI. p-values of pairwise t-test comparison of LR and RF models performances against the other algorithms

	LR						RF					
	Title	Text	Title and Text	Title and Top.	Text and Top.	Title, Text and Top.	Title	Text	Title and Text	Title and Top.	Text and Top.	Title, Text and Top.
Weighted Precision												
RF	5.41E-04	0.35	0.41	3.34E-05	0.22	0.44						
KNN	2.15E-12	3.02E-14	1.22E-13	2.42E-13	1.03E-12	3.75E-13	4.68E-10	1.25E-13	2.51E-11	9.39E-10	1.09E-13	4.12E-13
SVM	7.47E-13	1.41E-15	3.16E-16	5.71E-12	1.13E-12	6.33E-17	2.11E-10	5.55E-15	4.13E-13	3.42E-08	1.83E-13	2.92E-17
NB	9.01E-13	1.17E-24	1.41E-22	6.50E-12	1.06E-20	3.31E-22	2.87E-10	1.30E-23	2.80E-20	9.13E-08	3.20E-21	2.10E-22
Weighted Recall												
RF	3.56E-05	0.44	0.40	9.22E-06	0.39	0.64						
KNN	3.76E-13	8.66E-15	4.99E-14	1.66E-13	3.92E-14	2.02E-13	2.70E-10	3.34E-14	1.55E-12	2.46E-09	1.69E-15	1.10E-13
SVM	2.93E-16	3.49E-17	5.13E-17	8.76E-15	1.65E-15	1.30E-17	6.19E-14	1.25E-16	2.15E-15	1.55E-11	1.75E-16	1.68E-18
NB	3.73E-13	4.96E-27	3.42E-23	4.83E-12	2.46E-22	6.28E-23	4.45E-10	1.10E-25	8.43E-22	2.69E-07	3.18E-23	1.35E-23
Weighted F1												
RF	2.58E-05	0.46	0.35	7.67E-06	0.45	0.55						
KNN	6.78E-13	3.95E-15	5.76E-14	1.44E-13	1.43E-14	2.12E-13	8.14E-10	1.56E-14	1.62E-12	7.30E-09	7.77E-16	1.40E-13
SVM	1.25E-16	4.21E-17	2.96E-17	3.90E-15	8.64E-16	4.74E-18	2.68E-14	1.19E-16	1.01E-15	8.98E-12	1.37E-16	8.34E-19
NB	6.94E-13	7.51E-28	9.86E-23	5.02E-12	3.04E-22	3.26E-22	1.58E-09	1.66E-26	1.47E-21	9.40E-07	6.61E-23	1.20E-22

Table XVII. p-values from T-Test pairwise comparison of the inclusion of topological metrics

	Title and Topological metrics		Text and Topological metrics		Title, Text and Topological metrics	
	Prob_LR_RF_KNN	**Voting_LR_RF_KNN**	**Prob_LR_RF_KNN**	**Voting_LR_RF_KNN**	**Prob_LR_RF_KNN**	**Voting_LR_RF_KNN**
Weighted Precision						
Title	0.90	0.56				
Text			8.61E-04	1.82E-02		
Title and Text					1.08E-03	0.09
Weighted Recall						
Title	0.82	0.30				
Text			6.92E-04	1.48E-02		
Title and Text					1.73E-03	0.09
Weighted F1						
Title	0.80	0.26				
Text			7.15E-04	1.44E-02		
Title and Text					1.78E-03	0.08
FN Precision						
Title	0.37	0.08				
Text			0.06	0.15		
Title and Text					5.47E-03	0.10
FN Recall						
Title	0.77	0.23				
Text			9.89E-04	3.10E-02		
Title and Text					0.17	0.59
FN F1						
Title	0.84	0.58				
Text			7.34E-04	2.06E-02		
Title and Text					5.92E-03	0.17