Supervised Link Prediction on the WikiLinkGraphs Dataset

Helbert Paat

University of the Philippines
hapaat@up.edu.ph

January 12, 2021

Abstract

Link prediction is one of the challenging tasks in network analysis. Wikipedia is one of those networks where link prediction is very important. Like Wikipedia, the evolution of real world networks over time introduce the addition of new nodes and new links. in the context of Wikipedia, new articles are made and new hyperlinks are created connecting these articles, formulating the network of Wikipedia. In this study, we tackled the problem of link prediction using the WikiLinkGraphs dataset. We showed that the performance of Random Forest Classifier using a combination of heuristics (Common Neighbors and Distance (CND), Sorensen Index (SI), Resource Allocation (RA), Hub Promoted Index (HPI), and Common Neighbor and Parameterized Algorithm (CCPA)) as features outperformed the set of four heuristic features (Common Neighbors (CN), Adamic/Adar (AA) Coefficient, Jaccard Index (JI), and Preferential Attachment (PA)), basic degree features, triad features, and distance features in terms of accuracy, F-Measure, and AUC score.

I. INTRODUCTION

The advent of the internet has made it easy for people to learn valuable information and gain new knowledge. Wikipedia is one of the popular sites that made information sharing accessible. The network of articles in Wikipedia can be visualized as graph between multiple nodes that represent entities such as topics. Topics are linked together if they are related to each other. Due to the expansion of contents and topics in Wikipedia, analysis of network has become an area of research in the recent years.

Like Wikipedia graph, the evolution of real-world networks over time introduce the addition of new nodes and links. One of the many tasks in the analysis of networks is link prediction. It is a method used to predict connections between nodes in the future using the current network patterns [1]. Understanding current patterns and historical structure of the graph is helpful to predict the appearance of new links in the network. Link prediction now plays a vital role in many fields such as recommender systems.

In Wikipedia, link prediction is important to determine what hyperlinks must be added in each topic page. The prediction can be based on the topological structure of the graph or content of each article. In e-commerce and online shopping, link prediction is used to recommend products to the user using the data on the user's or similar users' previous purchases, search items, and online activities. In co-authorship graphs, link prediction is used to understand how authors are associated.

In this paper, link prediction in WikiLinkGraphs dataset is investigated. For this purpose, we propose a solution to link prediction in WikiLinkGraphs dataset that is based on a new combination of five well-known and studied heuristics: Common Neighbors and Distance (CND), Sorensen Index (SI), Resource Allocation (RA), Hub Promoted Index (HPI), and Common Neighbor and Parameterized Algorithm (CCPA). The performance of using these set of features was compared to other studied set of features: basic degree features, triad features, distance features, and set of four heuristics (Common Neighbors (CN), Adamic/Adar (AA) Coefficient, Jaccard Index (JI), and Preferential Attachment (PA)). Given a network in two consecutive years, three different learning algorithms are trained to determine the possibility of link formation in the newer network. The classification probability determined by the machine learning model was used as the link prediction heuristic for a pair of nodes. The test results on 7 WikiLinkGraphs datasets were reported. Their performances in three metrics were compared: accuracy, F-Measure, and AUC (Area under ROC Curve) score. The goal is to understand how this new combination of five features performs for link prediction on the WikiLinkGraph.

The paper is organized as follows. Section II details the previous works on link prediction. Section III presents brief description of the supervised machine learning methods, the dataset used, and the performance measures. Section IV presents the experimental

study and the analysis of results. Lastly, section V concludes the paper and summarizes the results.

II. RELATED WORKS

The efficiency and performance of traditional machine learning algorithm has made them suitable for the task of analyzing the behavior of entities in any network and predict link formation. There have been different studies that use traditional machine learning algorithms for link prediction in many areas. In this section, the previous works used supervised machine learning methods for link prediction using different features and methodology.

Approaches for link prediction include node-based and topology-based methods. Node similarity is used as an important factor to determine the possibility of two nodes to have a link. Common Neighbors, Jaccards Coefficient, and Sum of Neighbors are some of the features extracted in the topological method. These features are used to find similarity scores of pairs of nodes in the network.

By extracting graph topological and domain-specific proximity features on the data on the BIOBASE and DBPL co-authorship networks, applied supervised learning was applied to solve the link prediction problem [12]. In a work by Zhou et al., six real-world networks were studied and the the performance of local similarity heuristics was evaluated [13]. They also describe a new local similarity measure which was based on resource distribution through common neighbors [13]. Moreover, Liben-Nowell and Kleinberg evaluated the performance of link prediction heuristics to understand network growth over time by using topological heuristic scores as link predictors for arXiv co-authorship networks [11]. In a work by Leskovec et al., the problem of edge sign detection in real-world social networks was tackled by applying supervised learning methods and by applying partial triads as similarity feature of two nodes [14].

A study that is directly significant to this paper is the study of Julian et al. [15] where they studied supervised learning for link prediction by proposing triad features as predictors together with basic degree features and a set of four heuristic features (CN, AA, JI, and PA). In the same study, it was shown that including the triad features to the basic features as model predictors improved classification accuracy and in some cases exceeded the accuracy of AA.

In [3], link prediction is applied to determine future possible co-authorship by considering the network as a projected graph: a projection of a bipartite graph linking authors to publications. The problem is then converted into linking or non-linking class discrimination problem which was solved using classical machine learning techniques. Results show that it performs well in the computer science bibliography dataset.

III. SUPERVISED LINK PREDICTION

This paper evaluates the effectiveness of a proposed set of heuristic features for link prediction in the WikiLinkGraph dataset. Three well-known supervised machine learning methods were used, which include Decision Trees, Random Forests, and Adaboost Classifier. Their performance was measured on the WikiLinkGraph dataset. In this section, we briefly review the supervised machine learning techniques that we use in our experiments.

A. Supervised Machine Learning Methods

The supervised machine learning techniques used in the experiments are the following.

Decision Tree (DT): One of the most effective machine learning algorithms for classification is decision tree [7]. Decision Tree is a non-parametric supervised learning algorithm where the goal is to create a model that predicts the value of a response variable by understanding simple decision rules in the data features.

Random Forest (RF) Classifier: Random Forest is an ensemble machine learning algorithm that is widely popular because of its excellent performance in many classification and regression tasks. In this method, a huge number of decision trees are determined from boostrap samples from the training dataset. In classification, Random Forest works in such a way that a vote is cast by each tree in the forest which specifies thee classification of the new sample. The predicted probability vector is determined by computing the proportion of votes in each class across the ensemble.

Adaboost Classifier: Adaboost Classifier is an iterative ensemble method where multiple classifiers are combined in order to improve prediction accuracy. It works by combining poorly performing classifiers to get a classifier with strong predictive performance. For every iteration, weights of incorrectly classified instances are adjusted. The subsequent classifiers then concentrate on fitting to difficult cases.

B. Feature Selection

For the WikiLinkGraphs dataset graph G = (V,E), we generated the training and testing examples via the following process:

1. Dataset graphs G_n and G_{n+1} for two consecutive years are considered.

2. New subgraphs $G_n^* = (V_n^*, E_n^*)$ and $G_{n+1}^* = (V_n^*, E_{n+1}^*)$ are induced by removing the nodes of the graphs with degree 3 and by including only nodes that are present in both graphs.

3. Topological features of pairs of nodes that do not form an edge in G_n^* but form an edge in G_{n+1}^* were generated. These are the positive examples. Moreover, we also generated the features of pairs of nodes that do not form an edge in G_n^* but have a common neighbor. A portion of these pairs of nodes are the negative examples. Note that the dataset is highly imbalanced. Hence, we use Synthetic Minority Over-sampling Technique (SMOTE) to fix the imbalanced dataset problem.

4. A 70-30 training-testing split to the all these examples was applied.

In this paper, we propose a solution to the link prediction that is based on a set of five studied heuristics. The following set of heuristics are used as input features of the model.

1. Common Neighbor and Distance (CND): Two key structural properties of a complex network which are common neighbor and distance are used in this feature. To reflect the chance that a link will be formed between the links x and y, the score for any nodes x and y without link is given by the following formula:

$$S_{xy} = \begin{cases} \frac{CN_{xy}+1}{2} & \text{if } \Gamma(x) \cap \Gamma(y) \neq \emptyset \\ \frac{1}{d_{xy}} & \text{otherwise} \end{cases} \quad (1)$$

Note that CN_{xy} refers to the number of common nodes between node x and y, $\Gamma(x)$ is the set of neighbors of node x, and d_{xy} is the shortest path distance between x and y.

2. Sorensen Index (SI): In this measure, twice of the number of common neighbors of nodes x and y is divided by the sum of their degrees as shown in Eq. 2.

$$S_{xy} = \frac{2|\Gamma(x) \cap \Gamma(y)|}{k_x + k_y} \quad (2)$$

3. Resource Allocation (RA): The amount of resource node y gives to node x through indirect links defines the similarity index. A unit of resource is contributed by each intermediate link. The formula for the score is shown in Eq. 3.

$$S_{xy} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{k_z} \quad (3)$$

4. Hub Promoted Index (HPI): Hub Promoted Index (HPI) is the ratio of common neighbors of nodes x and y to the minimum of degrees of the nodes. Formula is shown in Eq. 4.

$$S_{xy} = \frac{\Gamma(x) \cap \Gamma(y)}{min\{k_x, k_y\}} \quad (4)$$

5. Common Neighbor and Centrality Based Parameterized Algorithm (CCPA) [17]: CCPA is based on two node properties which are the number of common neighbors and their centrality. The score is shown in Eq. 5. Note that N is the number of nodes in the graph and the parameter α controls the importance of the node properties common neighbor and centrality.In this study, α is chosen to be 0.50. CITE THIS.

$$S_{xy} = \alpha \cdot (|\Gamma(x) \cap \Gamma(y)|) + (1 - \alpha) \cdot \frac{N}{d_{xy}} \quad (5)$$

We compared the results with the performance of the model when we use eight degree features (basic features), eight triad features, and four popular heuristic features which is discussed in a paper by Julian et al [15]. The eight degree features are $d_{in}(x)$, $d_{out}(x)$, $d_{in}(x)/d_{out}(x)$, $d_{out}(x)/d_{in}(x)$, $d_{in}(y)$, $d_{out}(y)$, $d_{in}(y)/d_{out}(y)$, and $d_{out}(y)/d_{in}(y)$ where x and y are the pairs of nodes. The triad features are defined by the number of partial triads formed by x and y with a common neighbor w. The eight triad features are t_1, t_2, t_3, t_4, $t_1/C(x,y)$, $t_2/C(x,y)$, $t_3/C(x,y)$, and $t_4/C(x,y)$ where $C(x,y)$ is the total number of common neighbors. See Table **??** for a visualization of the four triad types.

$x \to w \to y$	$x \to w \leftarrow y$	$x \leftarrow w \to y$	$x \leftarrow w \leftarrow y$
t_1	t_2	t_3	t_4

Table 1: *Four triad types*

These are the heuristic features used for the purpose of comparison.

1. Common Neighbors (CN): We compute the link prediction score by finding the number of common neighbors of nodes x and y. Formula is shown in Eq. 6.

$$S_{xy} = |\Gamma(x) \cap \Gamma(y)| \quad (6)$$

2. Adamic/Adar coefficient (AA): The formula for the score is given in Eq. 7.

$$S_{xy} = \sum_{w \in \Gamma(x) \cap \Gamma(y)} \frac{1}{log|\Gamma(w)|} \quad (7)$$

3. Jaccard Index (JI): Common neighbors between the nodes are considered in this similarity measure as

shown in Eq. 8 cite this.

$$S_{xy} = \frac{|\Gamma(x) \cap \Gamma(y)|}{|\Gamma(x) \cup \Gamma(y)|} \tag{8}$$

4. Preferential Attachment (PA): The score for this similarity measure is shown in Eq. 9. Note that in this measure, the degrees of the pair of nodes x and y are the basis for the score. Note that the degree of a node x is represented by k_x.

$$S_{xy} = k_x \cdot k_y \tag{9}$$

Moreover, distance features are also used. The three distance features used are the following: shortest path distance, the number of paths of length 2 from the source node to the destination node, and the number of paths of length 2 from the destination node to the source node.

C. The Dataset

The WikiLinkGraphs dataset is used in our experiments to evaluate the performance of our models. WikiLinkGraphs is a complete dataset of the network of internal Wikipedia links. The dataset spans 17 years and contains graph data for the 9 largest language editions: German (de), English (en), Spanish (es), French (fr), Italian (it), Dutch (nl), Polish (pl), Russian (ru), and Swedish (sv) [10]. The dataset contains the following fields:

- *page_id_from*: an integer, the source article page identifier (used by MediaWiki)
- *page_title_from*: a string, source Wikipedia article title
- *page_id_to*: an integer, the target page identifier (used by MediaWiki)
- *page_title_to*: a string, target Wikipedia article title

Figure 1 shows a sample extract of the file *enwiki.wikilink_graph*.2018 − 03 − 01.*csv.gz*.

page_id_from	page_title_from	page_id_to	page_title_to
10	AccessibleComputing	411964	Computer accessibility
12	Anarchism	5013592	6 February 1934 crisis
12	Anarchism	2181459	Abstentionism
12	Anarchism	839656	Adolf Brand
12	Anarchism	2731583	Adolf Hitler
12	Anarchism	192008	Adolphe Thiers
12	Anarchism	729048	Affinity group
12	Anarchism	30758	Age of Enlightenment
12	Anarchism	627	Agriculture
12	Anarchism	710931	AK Press

Figure 1: *Sample extract of the WikiLinkGraphs dataset*

This dataset has been produced by the following: Cristian Consonni from DISI, University of Trento, Trento, Italy; David Laniado from Eurecat, Centre Tecnologic de Catalunya, Barcelona, Spain, and Alberto Montresor from DISI, University of Trento, Trento, Italy.

D. Evaluation Criteria

The following performance measures were used for the evaluation of our models.

Accuracy: Accuracy is commonly used in evaluating the performance of machine learning models. This measure helps to determine which method achieve better results. It is defined as the ratio of instances correctly classified to the total number of instances in a given data set. Accuracy is defined to be the following:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

where TP (True Positive) is the number of instances correctly classified as positive, TN (True Negative) is the number of instances correctly classified as negative. FP (False Positive) indicates the number of instances that have been incorrectly classified as positive, and FN (False Negative) shows the number of instances that have been incorrectly classified as negative.

Area Under ROC Curve (AUC): The performance of a certain classification model at all different possible thresholds using TPR (True Positive Rate) and FPR (False Positive Rate) is displayed by the receiver operating characteristic curve (ROC curve). However, in calculating the full two-dimensional area beneath the ROC curve, AUC is a better criteria [9]. If the AUC is more than 0.5, a model performs better than chance. Generally, the performance of a model is better if the AUC is higher.

F-Measure: F-Measure is considered to be the harmonic mean of recall and precision. Precision is the ratio of instances truly classified as positive to the number of positive instances.

$$Precision = \frac{TP}{TP + FP}$$

On the other hand, Recall measures the proportion of True Positives that were correctly classified.

$$Recall = \frac{TP}{TP + FN}$$

By taking the harmonic mean of the precision and recall, F-Measure is define to be the following:

$$F1 = \frac{2}{\frac{1}{recall} + \frac{1}{precision}}$$

IV. DATA PREPARATION AND EXPERIMENTAL RESULTS

This section explains data preparation and feature extraction conducted in this study. The details of the experiments and analysis of results were also presented.

A. Data preparation

In this study, the WikiLinkGraphs dataset was treated as a graph $G =< V, E >$ where V and E are the set of nodes and edges, respectively. In this dataset, the set of nodes V contains the different titles pertaining to different articles in Wikipedia. If there is a hyperlink to another article in that page, then there is a directed edge from that article to other article in the set of edges E. In our implementation, the WikiLinkGraphs dataset is represented as a graph, using adjacency lists. As indicated in the dataset, an article title is assigned a unique page identifier which is an integer. If there is a hyperlink for article j in article i, then there is a directed edge (i,j) connecting them.

The WikiLinkGraphs dataset, most especially the data in the most recent years, has a very huge storage size. Since it will take a very long time to process these files, due to computational limitations, we only selected a pair of consecutive years for for various languages.

The details of the data set used in our experiments were shown in Table 2.

In the next stage, the data sets for the classification task were constructed. The experiments were implemented in Python. After extracting the relevant features, a class is assigned to each instance in such a way that if two nodes are connected, class 1 is assigned. otherwise, class 0 is assigned. Thus, our problem is a binary classification.

B. Experimental Results

In this section, experimental details are discussed. Our goal is to evaluate the effectiveness of using the extracted features for link prediction in WikiLinkGraphs. In other words, we want to determine the performance of machine learning models, where a model is trained on various features.

In the data preprocessing step, the data was first shuffled. Then, the data is split into two: 70% for training data and 30% for test data.

After data preparation and feature extraction, three different supervised machine learning algorithms for link prediction are used. These methods are Decision Tree, Random Forest, and Adaboost Classifier.The Random Forest Classifier is parameterized with 100 trees in the forest, no maximum depth, minimum of 2 samples to split internal node, minimum of 1 sample to be at the leaf node, and the Gini impurity was used in measuring the quality of the split.For the Decision Tree, the Gini impurity was also used to measure split quality. Moreover, there is also no maximum depth, a minimum of 2 samples to split an internal node and a minimum of 1 sample required to be at the leaf node. For Adaboost Classifier, we use the Decision Tree as the base estimator with maximum depth of 1 and learning rate of 1. Boosting is also terminated if it reached the maximum number of estimators which is 50. All of these machine learning models are implemented in Scikit-learn Python module. [16].

B.1 Performance of Machine Learning Models

Table 3 shows the accuracy scores, F-Measure, and AUC scores of the three different classifiers on the WikiLinkGraphs French dataset for the years 2003-2004 using all the features mentioned in this paper.

The Random Forest Classifier yielded the best result with an accuracy of 95.55%, F-Measure value of 76.27%, and AUC score of 96.41%. For the Decision Tree, it achieved an accuray of 93.49%, F-measure of 68.89% and AUC score of 84.09%. The Decision Tree achieved accuracy and F-Measure scores that are higher than the Adaboost Classifier. For Adaboost Classifier, the accuracy, F-Measure, and AUC score are 91.12%, 54.08%, and 93.70%, respectively. For the AUC score, the Adaboost Classifier achieved better performance than Decision Tree. Note that in this paper, hyperparameter tuning was not conducted.

From the same dataset, the following set of features were also derived and used separately to determine the features that yield the best results: basic features, distance features, set of four heuristic features (CN, AA, JI, PA), set of five heuristic features (CND, SI, RA, HPI, CCPA), and a set combining these 9 heuristic features.

Using the Random Forest Classifier and the set of all heuristic features yielded the best result with accuracy of 92.90%, F-Measure of 66.48%, and AUC value of 93.31%. The set of five heuristic features is second in terms of performance with accuracy of 92.24%, F-Measure of 65.37%, and AUC value of 93.38%. In all these metrics, the set of five heuristic features achieved better performance than the remaining sets of features.

Using the Adaboost Classifier, the set of all heuristic features performed best in terms of accuracy (accuracy = 86.75%). While the set of five heuristic features achieved the best F-Measure and AUC value (F-Measure = 54.31% and AUC = 91.77%). In all these metrics, the set of five heuristic features achieved higher scores than the basic, distance, triad, and set of four heuristic features.

Using Decision Tree, the set of all heuristic features achieved the highest accuracy (91.89%) which was followed by the distance feature and set of five heuristic

Dataset	Years	Number of Nodes	Training Examples	Test Examples	Positive Examples	Negative Examples
French (fr)	2003-2004	4 972	274 345	117 575	39 192	352 728
German (de)	2003-2004	11 158	663 902	284 528	94 843	853 587
English (en)	2002-2003	18 715	884 983	379 277	126 426	11 37 834
Spanish (es)	2004-2005	12 297	378 890	162 380	54 127	487 143
Italian (it)	2004-2005	5 202	328 308	140 702	46 901	422 109
Dutch (nl)	2003-2004	3 670	180 237	77 243	25 748	231 732
Polish (pl)	2003-2004	6 658	223 056	95 594	31 865	286 785

Table 2: *Our data sets extracted from the WikiLinkGraphs data set*

	Accuracy	F-Measure	AUC
Decision Tree	93.49	68.89	84.09
Random Forest	**95.55**	**76.27**	**96.41**
Adaboost Classifier	91.12	54.08	93.70

Table 3: *Accuracy, F-Measure and AUC of the models on the French WikiLinkGraphs dataset (2003-2004) using all the features*

features. Moreover, the set of all heuristic features also achieved the highest F-Measure (63.22%) which was followed by the set of five heuristic features and the set of four heuristic features. In terms of AUC, the set of triad features achieved the best result (AUC = 88.92%) which was followed by the distance features and the set of five heuristic features.

Tables 4, 5, and 6 show the performance scores for the basic features, distance features, triad features, four heuristic features, five heuristic features and a combination of all heuristic features using Random Forest, Adaboost Classifier, and Decision Tree, respectively.

	Accuracy	F-Measure	AUC
Basic	90.09	52.14	88.19
Distance	91.72	60.11	84.32
Triad	88.42	56.00	90.08
Heuristic (*4 Features*)	91.97	63.45	92.65
Heuristic (*5 Features*)	92.24	65.37	93.28
Heuristic (*All Features*)	**92.90**	**66.48**	**93.31**

Table 4: *Accuracy, F-Measure, and ROC scores of different features using the Random Forest Classifier on the French dataset (2003-2004)*

	Accuracy	F-Measure	AUC
Basic	73.11	36.88	83.71
Distance	86.59	50.27	84.40
Triad	75.248	40.04	83.05
Heuristic (*4 Features*)	86.56	53.71	90.96
Heuristic (*5 Features*)	86.89	**54.31**	**91.77**
Heuristic (*All Features*)	**86.75**	53.82	91.61

Table 5: *Accuracy, F-Measure, and AUC values of different features using the Adaboost Classifier on the French dataset (2003-2004)*

	Accuracy	F-Measure	AUC
Basic	88.94	45.77	71.77
Distance	91.72	60.084	84.251
Triad	88.45	55.75	**88.92**
Heuristic (*4 Features*)	91.20	60.89	81.83
Heuristic (*5 Features*)	91.39	62.26	83.10
Heuristic (*All Features*)	**91.89**	**63.22**	82.80

Table 6: *Accuracy, F-Measure, and AUC scores of different features using the Decision Tree Classifier on the French dataset (2003-2004)*

B.2 Performance based on Heuristic Features

From the previous results, the set of five heuristic features performed better than the set of four heuristic features in most cases. To understand the performance of individual heuristic features on the same dataset, an experiment was conducted. Using these individual heuristic functions, the performance of Random Forest classifier is evaluated.

In terms of accuracy, CN achieved the best accuracy (94.64%) while PA achieve the lowest (74.11%). In terms of F-Measure, CCPA achieved the best result (F-Measure = 65.22%) with PA achieving the lowest score again (F-Measure = 27.87%). In terms of AUC score, RA obtained the best value (AUC = 88.34) while PA achieved the poorest result (67.52%).

Figure 2 shows the performance scores of the individual heuristic features.

To determine the performance of the five heuristic features over the four heuristic features and other set of features across the different WikiLinkGraphs dataset, we ran an experiment on the seven datasets specified in Table 2. In terms of accuracy, the set of five heuristic features achieved the best result for five out of the seven datasets. In terms of F-Measure, the set of five heuristic features obtained the best score for six out of seven datasets. In terms of AUC, the set of five heuristic features achieved the highest score for six out of seven datasets. Figures 3, 4, and 5 show the performance of Random Forest Classifier on different languages of the WikiLinkGraphs dataset.

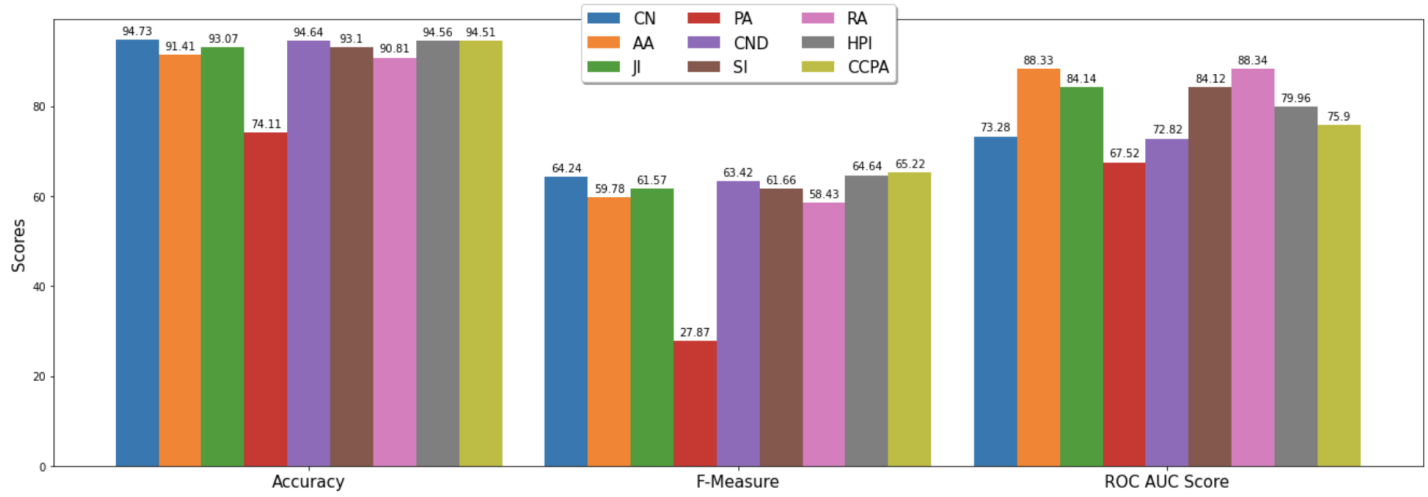

Figure 2: *Performance of the individual heuristic features using the Random Forest Classifier on the French WikiLinkGraphs dataset (2003-2004)*

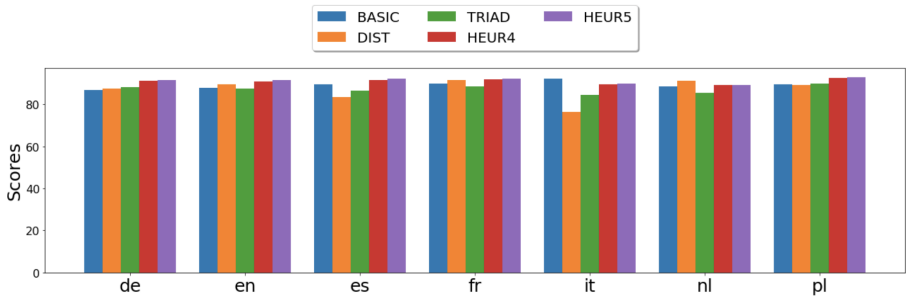

Figure 3: *Accuracy of Random Forest Classifier on the seven datasets using different sets of features*

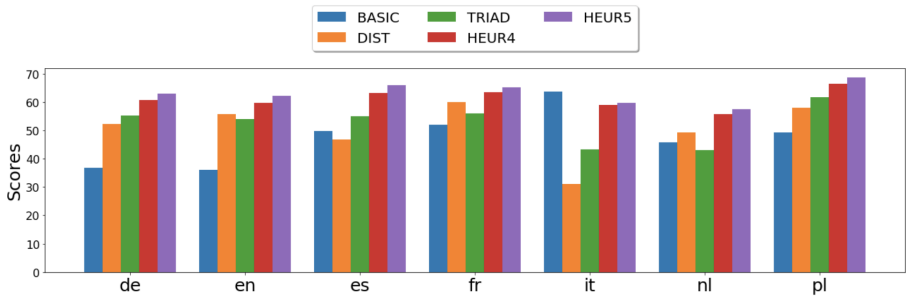

Figure 4: *F-Measure of Random Forest Classifier on the seven datasets using different sets of features*

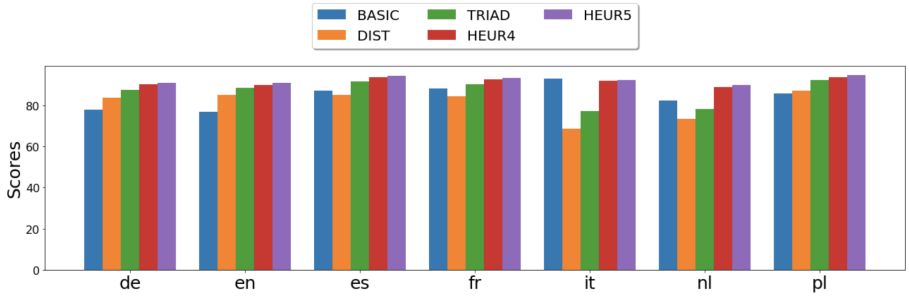

Figure 5: *AUC Scores of Random Forest Classifier on the seven datasets using different sets of features*

C. Analysis of the results

From Tables 3, 4, 5, and 6, it can be concluded that Random Forests is the best classifier to use as it yielded the highest accuracy, F-measure, and AUC scores among the three classifiers. The good performance of Random Forests, most especially when using heuristic features, verifies the fact that the decision trees are integrated easily because heuristics function similarly to decision boundaries [15] Moreover, using the three different classifiers, the five heuristics features achieved better accuracy, F-measure, and AUC score than the four heuristic features. For most cases, combining all the heuristic features and using them as input to the model achieves better result.

The individual performance scores of the nine heuristic features were evaluated using Random Forests on the French dataset (2003-2004) and is shown in Figure 2. CN, CND, HPI, and CCPA achieved the highest accuracy with almost similar accuracy scores. CCPA achieved the highest F-measure which is followed by HPI, CN, and CND. AA and RA, with almost similar scores, achieved the highest AUC score. On the other hand, PA has the lowest accuracy, F-measure, and AUC score. The poor performance of PA may be attributed to the fact that in the WikiLinkGraphs, huge number of current connections of a node does not guarantee more link formation in the future.

Using the Random Forest Classifier, it can also be observed that the individual performance of CN, JI, CND, SI, HPI, and CCPA in terms of accuracy exceeded the accuracy of the model (accuracy = 92.90%) when all heuristic features are used although the difference is small. This is for the benefit of achieving F-Measure and AUC which are higher than all the individual performance in terms of the said metrics. Since a huge portion of the the test set are negative examples, an increase in F-Measure and AUC is more advantageous than an increase in accuracy because high accuracy can be attributed to a large number of True Negatives.

Random Forests performed consistently in WikiLinkGraphs dataset across different languages, most especially when using heuristic features. This means that there is strong correlation among Wikipedia graphs and that Random Forests can generalize to different graphs of similar network structures.

V. Conclusion

In this paper, the performance of three machine learning algorithms (Decision Tree, Random Forest, and Adaboost Classifier) on different sets of features for link prediction was determined using the French WikiLinkGraphs dataset (2003-2004). Moreover, the individual performance of the heuristic features are also determined using the same dataset and the Random Forest Classifier. Finally, the performance of Random Forest Classifier using different sets of features is also evaluated using seven WikiLinkGraphs dataset of different languages.

Extracting all the features mentioned and using the machine learning methods, we have determined that Random Forests Classifier performs the best in terms of accuracy, F-measure, and AUC score for the French dataset of the WikiLinkGraphs for the years 2003-2004. Based from our experiment, in most cases, the set of five heuristic features (Common Neighbors and Distance (CND), Sorensen Index (SI), Resource Allocation (RA), Hub Promoted Index (HPI), and Common Neighbor and Centrality Based Parameterized Algorithm (CCPA)) used as features of Random Forest Classifier for link prediction provide results better than the set of four heuristic features (Common Neighbors (CN), Adamic/Adar Coefficient (AA), Jaccard Index (JI), and Preferential Attachment (PA)), basic features, distance features, and triad features in terms of accuracy, F-measure, and AUC score.

For future studies, we recommend using more heuristics as predictors to the machine learning model for link prediction. Moreover, the performance must be evaluated on different datasets of different networks.

References

[1] H. R. de Sa and R. B. C. Prudencio, Supervised link prediction in weighted networks, in The 2011 International Joint Conference on Neural Networks, 2011.

[2] C. Ahmed, A. ElKorany, and R. Bahgat, A supervised learning approach to link prediction in Twitter, Social Network Analysis and Mining, vol. 6, no. 1, May 2016.

[3] N. Benchettara, R. Kanawati, and C. Rouveirol, A supervised machine learning link prediction approach for academic collaboration recommen- dation, in Proceedings of the fourth ACM conference on Recommender systems - RecSys 10, 2010.

[4] D. Liben-Nowell. An Algorithmic Approach to Social networks. PhD thesis, M.I.T., june 2005.

[5] P.Wang,B.Xu,Y.Wu,andX.Zhou,Link prediction in social networks: the state-of-the-art, Science China Information Sciences, vol. 58, no. 1, pp. 138, Dec. 2014.

[6] Y. Yang, R. N. Lichtenwalter, and N. V. Chawla, Evaluating link prediction methods, Knowledge and Information Systems, vol. 45, no. 3, pp. 751782, Oct. 2014.

[7] Patil, T.R., Sherekar, S.S., Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification. Int. J. Comput. Sci. Appl. 6. 256-261, 2013.

[8] A. Popescul and L. H. Ungar, Structural Logistic Regression for Link Analysis, in 2nd Workshop on Multi-Relational Data Mining (MRDM 2003), S. Dzeroski, L. D. Raedt, and S. Wrobel, Eds., August 2003.

[9] Powers, D.M.W., Evaluation: from Precision, Recall and F-measure to ROC, Informedness, Markedness and Correlation. Journal of Machine Learning Technologies, 2(1), 37-63, 2011.

[10] Cristian Consonni, David Laniado, & Alberto Montresor. (2019). WikiLinkGraphs: A complete, longitudinal and multilanguage dataset of the Wikipedia link networks (Version 1.0) [Data set]. Presented at the The 13th International AAAI Conference on Web and Social Media (ICWSM-2019), Munich, Germany: Zenodo. http://doi.org/10.5281/zenodo.2539424

[11] Liben-Nowell, David, and Jon Kleinberg. "The link-prediction problem for social networks." *Journal of the American society for information science and technology* 58.7 (2007): 1019-1031.

[12] Al Hasan, Mohammad, et al. "Link prediction using supervised learning." *SDM06: workshop on link analysis, counter-terrorism and security.* 2006.

[13] Zhou, Tao, Linyuan L, and Yi-Cheng Zhang. "Predicting missing links via local information." The European Physical Journal B 71.4 (2009): 623-630.

[14] Leskovec, Jure, Daniel Huttenlocher, and Jon Kleinberg. "Predicting positive and negative links in online social networks." *Proceedings of the 19th international conference on World wide web*. ACM, 2010.

[15] Kyle D. Julian and Wayne Lu. "Application of Machine Learning to Link Prediction." 2016

[16] Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.

[17] Ahmad,I. ,Akhtar, M.U.,Noor,S. & Shahnaz,A.Missing link prediction using common neighbor and centrality based parameterized algorithm. Sci. Rep. 10, 1-9 (2020).