

I. IDENTIFICATION DATA

Thesis name:	Detecting Propaganda Articles by its Internet Distribution Pattern
Author's name:	Ondřej Bouček
Type of thesis :	[REDACTED]
Faculty/Institute:	[REDACTED]
Department:	Department of Computer Science
Thesis reviewer:	Carlos A. Catania
Reviewer's department:	Department of Computer Science, National University of Cuyo, Mendoza, Argentina

II. EVALUATION OF INDIVIDUAL CRITERIA

Assignment	[REDACTED]
<i>Evaluation of thesis difficulty of assignment.</i>	
This work attempts to provide a solution to the complex and significant problem of computational propaganda detection. The process for achieving the solution has required the integration of different computer science topics such as complex networks analysis, machine learning and software engineering.	

Satisfaction of assignment	[REDACTED]
<i>Assess that handed thesis meets assignment. Present points of assignment that fell short or were extended. Try to assess importance, impact or cause of each shortcoming.</i>	
The work in the thesis has met the assignment. The central aspects of a graph classification method have been developed. Several models and algorithms were discussed in the thesis. In particular, I would like to highlight the considerable effort in the development of the dataset for evaluating all the considered approaches.	

Method of conception	[REDACTED]
<i>Assess that student has chosen correct approach or solution methods.</i>	
The student has followed the correct methodology used in Machine Learning. He has analyzed state of the art and proposed different approaches for dealing with the graph classification problem. Besides the well-known SNA features, the student has proposed a set of specially designed features for nodes and edges of the graph. Then, a set of models were carefully designed for evaluating its hypothesis following machine learning standard procedure.	

Technical level	[REDACTED]
<i>Assess level of thesis specialty, use of knowledge gained by study and by expert literature, use of sources and data gained by experience.</i>	
The student has demonstrated valuable skills for dealing with a new and difficult problem and provided a valid solution using a diverse set of tools. He has shown expertise in several areas such as software development, machine learning, and the field of complex networks.	

Formal and language level, scope of thesis	[REDACTED]
<i>Assess correctness of usage of formal notation. Assess typographical and language arrangement of thesis.</i>	
In general, the thesis is informative and clear. The student expressed in a correct language the different aspects involved in the process of building a graph classification method for propaganda detection using formal notation when required. I have found just a couple of typos:	

1. pg 19: "This chapter describes how our dataset was **crated.**"
2. pg 23 "Every node corresponds to an Article, **it store publishing date,**"

Selection of sources, citation correctness

Present your opinion to student's activity when obtaining and using study materials for thesis creation. Characterize selection of sources. Assess that student used all relevant sources. Verify that all used elements are correctly distinguished from own results and thoughts. Assess that citation ethics has not been breached and that all bibliographic citations are complete and in accordance with citation convention and standards.

The student has always made reference to third party articles and software applications used for meeting the thesis assignment. All references used in the work followed the proper quality standards.

Additional commentary and evaluation

Present your opinion to achieved primary goals of thesis, e.g. level of theoretical results, level and functionality of technical or software conception, publication performance, experimental dexterity etc.

I would like to emphasize the importance of attacking challenging research topics such as this one. Despite being a very difficult research topic, the detection of computational propaganda has a great impact on the normal operation of our society. Any effort, no matter their size worth it.

III. OVERALL EVALUATION, QUESTIONS FOR DEFENSE, CLASSIFICATION SUGGESTION

Summarize thesis aspects that swayed your final evaluation. Please present apt questions which student should answer during defense.

In this thesis, the student has proposed a new method for dealing with computational propaganda detection analyzed as a graph classification problem. A distribution graph for an article was created with information on how it was referenced or duplicated. The resulting graph includes a set of specially designed features for nodes and edges.

The student was involved in all the different stages of building a machine learning solution for graph classification. The student has placed particular emphasis on the construction of a dataset. However, it is not clear if the dataset will be available to the research community. Three different algorithms were analyzed and carefully evaluated following the standard machine learning methodology. Finally, and perhaps the most valuable contribution of the present thesis is the demonstration that the Internet distribution pattern of an article is sufficient to achieve up to 80% accuracy on unseen articles.

Apt questions:

1) In section 4.1 the student mentioned that EUvsDisinfo has 1200 propaganda articles. However, the student has selected less than half. Is there any reason for not using 1200? Also, Is not clearly explained why the final dataset ends with only 117 propaganda articles?

2) ADG is a directed graph, and the Local Cluster Coefficient (LCC) is for an undirected graph. Have you considered using some of the LCC approaches for directed graphs?

3) Random Forest provides a feature importance score, have you analyzed the results? It would have been great to see the impact of each one of the features. In particular, the difference between edge, vertex, and NSA features.

4) What were the criteria for selecting the different feature sets for GNN (table 7.4)?

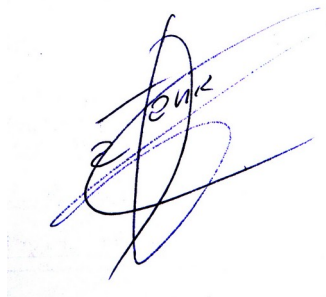
5) Have you considered GNN using the ADG as input, instead of the pre-computed features? By using ADG as the input you can let the GNN learn the best feature presentation for classifying ADG.

6) Cross-validation results were missing for SVM and RF. It would have been great to see them for comparison with GNN.

I evaluate handed thesis with classification grade [REDACTED]

Date: **06/08/2022**

Signature:



Dr. Carlos Catania (PhD)