PARTICLE SWARM OPTIMIZATION (PSO) APPROACH FOR FEATURE SELECTION IN SENTIMENT ANALYSIS FOR PROPAGANDA ISSUES

KAPT MUHAMMAD ZAKWAN BIN MUHAMAD RODZI

MASTER OF SCIENCE
(COMPUTER SCIENCE)

UNIVERSITI PERTAHANAN NASIONAL MALAYSIA

2022

PARTICLE SWARM OPTIMIZATION (PSO) APPROACH FOR FEATURE SELECTION IN SENTIMENT ANALYSIS FOR PROPAGANDA ISSUES

KAPT MUHAMMAD ZAKWAN BIN MUHAMAD RODZI

Thesis submitted to the Centre for Graduate Studies, Universiti Pertahanan Nasional Malaysia, in fulfilment of the requirements for the Degree of Master of Science (Computer Science)

2022

ABSTRACT

Feature selection, one of the main components of feature engineering, is the process of selecting the most important features to be input in machine learning algorithms. Feature selection techniques are employed to diminish the number of input variables by eliminating redundant or irrelevant features and narrowing down the set of features without deducting the predictive efficiency. The polynomial of feature selection in sentiment analysis is the main problem because it can decelerate sentiment classification accuracy and misfortunes to extort the optimum of features' subset. To overcome the problem, this research provides a text feature selection approach that requires an optimization method. Particle swarm optimization (PSO), one of the optimization approaches has been applied as feature selection in this research. PSO is the bio-inspired algorithm, and it is a simple classifier to search for an optimal solution in the solution space. It is different from other optimization algorithms in such a way that only the objective function is needed, and it is not dependent on the gradient, domain, or any differential form of the objective. It also has very few hyperparameters. This research presents the implementation of the PSO (feature selection) performance in the propaganda domain for sentiment analysis. The effectiveness of the PSO algorithm is tested using the datasets from Kaggle concerning Donald Trump and Hillary Clinton's tweets during US Election Presidential in 2016. Three experimental results have shown that PSO is better and more substantial in constructing optimum and quality feature subsets compared to other sentiment analysis tools, machine learning algorithms and swarm algorithms. Suggestion of the PSO algorithm implementation in others domain as tourism, medicine, product reviews etc.

ABSTRAK

Pemilihan fitur merupakan salah satu komponen utama kejuruteraan fitur di mana proses memilih fitur-fitur untuk dijadikan input dalam algoritma *machine learning*. Teknik pemilihan fitur digunakan bagi mengurangkan bilangan pembolehubah input dengan menyingkirkan fitur yang tidak berkaitan dan mengecilkan set fitur tanpa mengurangkan keberkesanan di dalam ramalan keadaan. Polinomial pemilihan fitur yang tinggi dalam analisis sentimen merupakan permasalahan yang utama kerana ia boleh mengurangkan ketepatan klasifikasi sentimen dan subset fitur secara optimum. Bagi mengatasi masalah tersebut, kajian ini menyediakan pendekatan pemilihan fitur teks yang memerlukan kaedah pengoptimuman. *Particle Swarm Optimization* (PSO), merupakan salah satu pendekatan pengoptimuman telah digunakan sebagai pemilihan fitur dalam penyelidikan ini. PSO adalah algoritma yang diilhamkan dari algoritma bio-inspired, dan ia merupakan pengelas mudah untuk mencari penyelesaian secara optimum. Ia berbeza daripada algoritma pengoptimuman yang lain kerana PSO hanya memerlukan fungsi objektif sahaja, dan tidak bergantung pada kecerunan, domain atau sebarang bentuk objektif yang berlainan bagi mencari penyelesaian pemilihan fitur. Ia juga mempunyai sangat sedikit hyperparameter. Penyelidikan ini adalah untuk menguji prestasi PSO di dalam pemilihan fitur terhadap domain propaganda bagi penganalisaan sentimen. PSO menggunakan set data daripada Kaggle yang melibatkan data ciapan Donald Trump dan Hillary Clinton dalam pilihan raya presiden Amerika Syarikat pada 2016. Tiga eksperimen telah dijalankan menunjukkan bahawa PSO lebih efisien dalam menghasilkan subset fitur yang optimum yang berkualiti dan lebih baik berbanding kaedah dari alatan analisis sentiment dan pembelajaran mesin. Kajian ini

mencadangkan penggunaan algoritma PSO pada masa hadapan dalam domain yang lain seperti pelancongan, perubatan, penggunaan produk dan sebagainya.

ACKNOWLEDGEMENTS

In the name of Allah, the Most Gracious and Most Merciful, peace and blessings to our beloved Prophet Muhammad (PBUH). Thank you to the Almighty with His help and blessings, my thesis for Master of Science (Computer Science) has managed to be completed in time.

This project would not have been possible without the support of many people. Many thanks to my main supervisor, Dr Siti Rohaidah who read my numerous revisions and helped make some sense of the confusion. Also, thanks to my co-supervisor, Dr Arniyati who offered guidance and support.

Thanks to the Bahagian Staf Perisikan Pertahanan (BSPP) for providing me with the financial means to complete this project. And finally, thanks to my beloved wife (Siti Nadiah), parents, sister (Lina Hazwani), and numerous friends especially to Kapt Wan Nur Hidayat who endured this long process with me, always offering support and love.

APPROVAL

The Examination Committee has met on **16 June 2022** to conduct the final examination of **Kapt Muhammad Zakwan bin Muhamad Rodzi** on his degree thesis entitled **Particle Swarm Optimization (PSO) Approach for Feature Selection in Sentiment Analysis for Propaganda Issues.**

The committee recommends that the student be awarded the of Name of Degree (Computer Science).

Members of the Examination Committee were as follows.

Prof Madya Dr. Zuraini binti Zainol
Faculty of Defence Science & Technology
Universiti Pertahanan Nasional Malaysia
(Chairman)

Dr. Hassan bin Mohamed
Faculty of Defence Science & Technology
Universiti Pertahanan Nasional Malaysia
(Internal Examiner)

Prof. Madya. Dr. Suhaila binti Zainudin
Faculty of Science & Technology
Universiti Kebangsaan Malaysia
(External Examiner)

APPROVAL

This thesis was submitted to the Senate of Universiti Pertahanan Nasional Malaysia and has been accepted as fulfilment of the requirements for the degree of **Master of Science (Computer Science)**. The members of the Supervisory Committee were as follows.

Dr Siti Rohaidah binti Ahmad
Faculty of Defence Science & Technology
Universiti Pertahanan Nasional Malaysia
(Main Supervisor)

Dr Arniyati binti Ahmad
Faculty of Defence Science & Technology
Universiti Pertahanan Nasional Malaysia
(Co-Supervisor)

UNIVERSITI PERTAHANAN NASIONAL MALAYSIA

DECLARATION OF THESIS

Student's full name : Kapt Muhammad Zakwan bin Muhamad Rodzi

Date of birth : 2nd October 1991

Title : Particle Swarm Optimization (PSO) Approach for Feature Selection in Sentiment Analysis for Propaganda Issues

Academic session : 1/2019

I hereby declare that the work in this thesis is my own except for quotations and summaries which have been duly acknowledged.

I further declare that this thesis is classified as:

☐ **CONFIDENTIAL** (Contains confidential information under the official Secret Act 1972)*

☐ **RESTRICTED** (Contains restricted information as specified by the organisation where research was done)*

☐ **OPEN ACCESS** I agree that my thesis to be published as online open access (full text)

I acknowledge that Universiti Pertahanan Nasional Malaysia reserves the right as follows.

1. The thesis is the property of Universiti Pertahanan Nasional Malaysia.
2. The library of Universiti Pertahanan Nasional Malaysia has the right to make copies for the purpose of research only.
3. The library has the right to make copies of the thesis for academic exchange.

Signature	**Signature of Supervisor/Dean of CGS/ Chief Librarian
Click here to enter text.	Click here to enter text.
IC/Passport No.	**Name of Supervisor/Dean of CGS/ Chief Librarian
Date:	Date:

*If the thesis is CONFIDENTAL OR RESTRICTED, please attach the letter from the organisation with period and reasons for confidentiality and restriction.

** Witness

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

PSO	-	Particle Swarm Optimization
SA	-	Sentiment Analysis
FS	-	Feature Selection
TDR	-	Typed Dependency Relations
POST	-	Part-of-Speech Tagging
NLP	-	Natural Language Processing
VADER	-	Valence Aware Dictionary for Sentiment Reasoning
SVM	-	Support Vector Machine
TF-IDF	-	Term Frequency - Inverse Document Frequency
ABC	-	Artificial Bee Colony
ACO	-	Ant Colony Optimization
GA	-	Genetic Algorithm
RDA	-	Red Deer Algorithm
KNN	-	K-Nearest Neighbour
NB	-	Naïve Bayes

LIST OF SYMBOLS

c	-	Learning Rates
P_i	-	Swarm's Best Previous Location
r	-	Random Numbers
v	-	Velocity
w	-	Inertia Weight
T	-	Features
T_r	-	Training Dataset
m	-	Number of Features
n	-	Documents

CHAPTER 1

INTRODUCTION

1.1 Background

Social media like Facebook, Twitter and Instagram is an internet-based form of a communication platform that expedite users to have conversations, share thoughts and information, create web content, and have discussions on politics or current issues. Billions of people around the world use these platforms to share a lot of information and make virtual connections between them. Social media also can be divided into three levels which are personal, professional and company level where it helps people use social media within the boundary of every level. This technology has facilitated the extent of the information rapidly and precisely to users or targets. Sometimes, it will be used as a method to disseminate the propaganda. Propaganda is expected to consist of true or false information used to inspire people to blow away countless agendas in order to achieve clear goals.

According to (Jowet, 2009), three prerequisites for diverse types of social communication for propaganda are the character of the target society, communication experience, and the propagandist. In (Fellows, 2017), the definition of propaganda indicates that each communication utilized is to assist and support specifically or by

implication, the government’s, or the opposition's points in inducing and affecting an individual's or any group's feeling, conduct, and actions for the good thing about the advocate. Various modes of communication have been used as a propaganda method to have an impact on the thoughts of millions of individuals all around the world (Harold D, 2012). To aim at gaining or turning the vote for the opposing party, propaganda is frequently interrelated to the distribution of negative messages (Jowet, 2009). Despite in (Abbasi et al., 2008), propaganda also can be used to disseminate positive messages for example encouraging people to vote, championing women’s rights, community health, racial equality and so on. In the nutshell, propaganda is a very dynamic and flexible domain as it can convey the sentiment either positive or negative.

Sentiment Analysis (SA) is a type of contextual mining in which subjective information is identified and extracted from the source material. The results of SA can help various organizations in multiple fields, such as politics, business, productions, and marketing to make decisions. As an example, a businessman can understand the social sentiment of their product, brand and service while overseeing online conversations in social media. Besides, certain politicians can also use SA to identify and determine the message included in voter messages, whether it is positive or otherwise in order to win the general election. Overall, the purpose of this study is to thoroughly examine the effectiveness of SA for issues of propaganda utilizing the Feature Selection (FS) method, particularly in political focus during elections.

1.2 Problem Statement

Recently, SA became one of the most trending analysis techniques to make accurate decisions and predictions in all aspects such as the result of an election, weather forecast and business marketing. FS is the main method used in the process of SA. FS is a method for extracting the best characteristics from large datasets (Agrawal et al., 2021). Research on the reasons for and consequences of this shift has focused on objective measures of FS in propaganda domain. As tweets and other short links on the Internet are just used to refer individuals to the actual message, which may be sometimes a longer article rather than a video or image, the content is always needed for propaganda. It turns out that propaganda is effective not only in disseminating the message, but also able to customize the content so that it will be more appealing to readers and tend to have strong ideology respectively. This research aims to better understand what the features are will be used in propaganda domain.

Xie et al., (2021) was stated that FS is used to remove any unwanted, irrelevant, or redundant characteristics from the dataset, which will not help improve the model's accuracy but may rather reduce it. One of the best approaches to resolve this problem by using optimization algorithm, PSO as FS. PSO is needed to detect relevant and non-relevant features in order to perform the classification process and increase the reading accuracy in the domain's context. For example, in (Kumar et al., 2016), three domains have been tested by PSO which are Product Reviews, Governmental Decisions and Restaurants Review while in (Sivaganesan Professor & Aggarwal, 2022) was focused on Amazon product reviews. Furthermore, PSO also has been used in hotel and laptop datasets to classify positive and negative customer opinions (Shang

et al., 2016). Most of the research only focuses on the product reviews domain. As a result of the current literature review analysis, there is no propaganda domain related to PSO as FS.

Several research used PSO as FS acquired good results on accuracy, precision, recall and F-score evaluation in difference domain respectively. As an example, Nurcahyawati, (2021) conducted a series of experiments on the tourist review dataset from an online data source to improve the performance of existing sentiment analysis models and evaluate the model's usefulness. Nurcahyawati, (2021) showed the best result in precision, recall and F-score values using PSO as FS on tourist reviews domain. One of the goals of this study is to see how the PSO as FS affects the rule of propaganda results in terms of accuracy, precision, recall, F-score evaluation since PSO never been choose as FS in propaganda domain.

1.3 Research Question

(a) Features selection always focusing on the highlighted issues in various domain. Hence, what are the features will be used in propaganda domain for sentiment analysis?

(b) PSO mostly used in customer reviews domain for feature selection. Can the PSO algorithm be used for feature selection in propaganda domain?

(c) Several swarm intelligence algorithms have been used as FS in propaganda domain and acquire the outstanding performance on accuracy, precision, recall and F-score evaluation. Regarding the literature review, there is no propaganda domain related to PSO. Hence, how about the evaluation result of swarm intelligence algorithm using PSO towards propaganda domain?

1.4 Objective

Specifically, the research objectives are proposed as follows:

(a) To investigate the FS technique in identifying features in the propaganda agenda.

(b) To apply a PSO algorithm for FS in sentiment classification in propaganda domain.

(c) To evaluate the performance of the PSO as FS on propaganda issues based on accuracy, precision, recall and F-score.

1.5 Scope

This research scope is designed regarding the use of attribute's method for sentiment analysis which is one of the data mining techniques in order to extract the precise features from the vast data in the dataset using Particle Swarm Optimization (PSO) algorithm. This research used over 6000 labelled tweets from Donald Trump

(DT) and Hillary Clinton (HC) in the Kaggle dataset which are categorized into two groups, 'Quote Status' and 'Retweet' for classification model test in propaganda domain. Only 741 data were chosen for the classification process in this study once the text pre-processing step was completed and suitable for propaganda domain. 3 experiments of the accuracy comparison had been done in this research that involved sentiment analysis tools, machine learning and swarm intelligence algorithms. All tests were run on a workstation equipped with an Intel Core i5 quad-core processor, 8 Gb of RAM and employed python programming to implement PSO algorithm.

1.6 Organization of The Thesis

The various chapters in the thesis are organized as follows:

Chapter 1 describes the background of research, problem statements, objective of research and scope of the research.

Chapter 2 presents literature reviews of a general overview of propaganda domain, feature selection, sentiment analysis and particle swarm algorithm. Besides, it also highlights the sentiment classification techniques from previous studies.

Chapter 3 introduces used in this research contains five phases. There is text pre-processing, FS, the association between features and sentiment polarity, sentiment classification and testing, evaluation, and analysis.

Chapter 4 discusses the previous experiments done by this research: The comparison of PSO in accuracy, precision, recall, and f-score and between sentiment analysis tools, machine learning and swarm intelligence algorithms. This chapter

describes how experiment is done with discussion on the choice of applied algorithm model then evaluate the result. Lastly, the chapter describes the result's analysis of these experiments for further study.

Chapter 5 presents a review of the thesis, the summary, contributions, and several possible areas that have potential for further work.

1.7 Conclusion

In this concluding chapter, this research is focus on feature selection in propaganda domain for sentiment analysis. The main purpose of this chapter is to give the briefly understanding of problem statement, research questions and research objectives for the whole of thesis writing. 3 research objectives had been proposed in order to answer the research questions of the thesis.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

The relationship between the propaganda domain, SA, and FS tactics is briefly discussed in this chapter. This is the most fundamental hypothesis that will be used in this investigation. In (Jowet, 2009), propaganda is a type of communication that is used to persuade communities, or individuals in general, to support a particular agenda. Propaganda is distributed in a variety of ways nowadays, including social media posts, graphics, cartoons, animations, articles, and TV and radio shows. The topic of this paper is electoral propaganda. Through the analysis of associated documents, social media, publications, and forums, SA is then utilised to determine the positive and bad parts of the propaganda itself. This study discusses the many methodologies employed by past researchers in the field of propaganda utilising SA, such as feature selection to eliminate unnecessary characteristics and sentiment methods to detect sentiment in texts or other sources. Because textual information has a high measurement classification, feature selection is a dominating side in sentiment analysis. This can compromise SA classification interpretation.

2.2 Propaganda

The anticipating of propaganda's result either positive, negative, or neutral by using SA always becomes one of the main challenges to train a classification model and perform SA on tweet sentences for a vital circumstance such as an election (Bagheri & Islam, 2017). The act of spreading rumours, tales, stories, and gossip, whether genuine or false, in order to persuade and influence the general public is known as propaganda. Falsified facts are frequently used in propaganda to deceive people and affect their thinking. Propaganda is commonly employed in literature to manipulate and influence the opinions of society. In the past, propaganda was employed to affect public judgement and perception to fulfil the influencer's political or economic goals (Fitzmaurice, 2018). The term "propaganda" began to transform in its objective and function, shifting from religious indoctrination to political aspects or opinions, reflecting the transfer in societal authority from Church to State during the First World War (Fellows, 2017). Propaganda became more political and partisan during this period, intending to force or persuade a large audience to accept a certain viewpoint or point of view.

Propaganda can only be generated by deliberate preparation and the use of existing propagandist feelings, according to the following definitions. Several doubts have been raised about the existence of propaganda. In (Jowet, 2009), the goal of propaganda may be to persuade others to adopt the propagandist's beliefs and attitudes, or to persuade people to engage in certain patterns of behaviour, such as monetary contributions, group membership, or spontaneous demonstrations for a cause. According to them, another propaganda's objective is to safeguard and sustain the

authority of the organisation or group represents in order to ensure the legitimacy of its campaigns and events. As a result, one may argue that propaganda is created solely to further the propagandist's doctrine. As a result, propaganda efforts play a crucial role in persuading individuals to adopt the propagandist's viewpoint.

2.3 Sentiment Analysis (SA)

The method of analysing texts that contain opinions and feelings is known as SA. Any text that expresses an opinion or emotion has a negative, positive, or neutral component. Opinion or information exchange, as well as propaganda production, are commonly employed on the internet (Abbasi et al., 2008). Even though there are other text classifications, SA was chosen for two primary reasons, first according to (Abbasi et al., 2008), web pages are dense with emotional material and opinions. Second, written information provides opinions or directions that may impact public perception in decision-making. SA is capable of recognising falsehoods, as well as rage or hatred conveyed through community channels. Furthermore, Abbasi et al., (2008) claimed that SA can assist in the creation of an effective analysis as well as a better knowledge of how extremist groups use the internet to spread information and propaganda.

D. Liu & Lei, (2018) used SA in politics to examine sentiments expressed in election candidate speeches, issues, and election result prediction. The military, the media, politicians, advertisers, governments, corporations, and others employ propaganda in elections, marketing, and other settings. The purpose of this study is to see how prior studies have employed SA to analyse the disinformation in campaign speeches.