

REPUTATION SCORING FAKE NEWS USING TEXT MINING

By

Ahmad Firdaus
2-1551-015

MASTER'S DEGREE
In

INFORMATION TECHNOLOGY
FACULTY OF ENGINEERING & INFORMATION TECHNOLOGY



SWISS GERMAN UNIVERSITY
Prominence Office Tower
Tangerang 15143
Indonesia

Revision after Thesis Deffense 27th July 2017

STATEMENT BY THE AUTHOR

I hereby declare that this submission is my own work and to the best of my knowledge, it contains no material previously published or written by another person, nor material which to a substantial extent has been accepted for the award of any other degree or diploma at any educational institution, except where due acknowledgement is made in the thesis.

Ahmad Firdaus_____
Student_____
Date

Approved by:

Dr. Maulahikmah Galinium, M.Sc

Thesis Advisor_____
Date

Alva Erwin, M.Sc, MTI, ST

Thesis Co-Advisor_____
Date

Dr. Ir. Gembong Baskoro, M.Sc

Dean_____
Date

ABSTRACT**REPUTATION SCORING FAKE NEWS USING TEXT MINING**

By

Ahmad Firdaus

Dr. Maulahikmah Galinium, M.Sc, S.Kom Advisor

Alva Erwin, M.Sc, MTI, ST Co-Advisor

SWISS GERMAN UNIVERSITY

The classification of hoax news or news with incorrect information is one of the text categorization applications. Like text-based categorization of machine applications in general, this system consists of pre-processing and execution of classification models. In this study, experiments were conducted to select the best technique in each sub-process by using 1200 articles hoax and 600 article no hoax collected manually. This research Tried experimenting to determine the best preprocessing stages between stop removal and stemming and showing the results of the deception Tree algorithma achieving an accuracy of 100% concluded above naive bayes more stable level of accuracy in the number of datasets used in all candidates . Information gain, TFIDF and GGA based on using Naive Bayes algorithm, supporting Vector Machine and Decision Tree no significant percentage change occurred on all candidates. But after using GGA (Optimize Generation) feature selection there is an increase of accuracy level The results of a comparison of classification algorithms between Naive Bayes, decision trees and Support Vector machines combined with the GGA feature selection method for classifying the best result is generated by the selection of GGA + Decission Tree feature on candidate 2 (Paslon2) 100% and in the selection of the Information Gain + Decission Tree Feature selection with the lowest accuracy Candidate 3 at 36.67%, but overall improvement of accuracy Occurred on all algorithma after using feature selection and Naive bayes are faster in processing time and Decision Tree is the longest for processing time and Naive bayes more stable level of accuracy in the number of datasets used in all candidates.

Keyword Classification, Pre-processing, Feature Selection, Accuracy

COPYRIGHT



© Copyright 2017
By Ahmad Firdaus Student
All rights reserved

DEDICATION

I dedicated this thesis to my lovely family who have prayed for me every day, my parents for all the support and encouragement, colleagues in MIT Swiss German University batch 17.



ACKNOWLEDGEMENTS

Gratitude and thanks to Allah SWT matchless so I could finish this thesis on time. In addition also to all the colleagues and interviewee and expert panel who provide support and time either in the interview or discussion are:

1. Ika Suryani and Arsyia – My lovely Wife and Son
2. All friend Batch 17 Swiss German University

Thanks are expressed Dr. Maulahikmah Galinium, M.Sc as my Advisor Alva Erwin, M.Sc as Co-Advisor.



Contents

| | |
|--|----|
| ABSTRACT | 3 |
| DEDICATION | 5 |
| ACKNOWLEDGEMENTS | 6 |
| List of Figures | 8 |
| List of Tables | 9 |
| Chapter 1 – Introduction | 10 |
| 1.1 Background..... | 10 |
| 1.2 Research Problem | 11 |
| 1.3 Research Objective | 11 |
| 1.4 Research Question | 11 |
| 1.5 Hypothesis | 12 |
| 1.6 Significance of Study..... | 12 |
| 1.7 Scope and Limitation..... | 12 |
| 1.8 Thesis Structure | 12 |
| Chapter 2 – Literature Review | 13 |
| 2.1. Hoax..... | 13 |
| 2.2. Automated Deception Detection | 13 |
| 2.3. Deception Detection for News Verification | 14 |
| 2.4.1. Data Representation..... | 15 |
| 2.4.2. Deep Syntax..... | 15 |
| 2.4.3. Machine Learning Techniques | 15 |
| 2.5. Text Categorisation..... | 16 |
| 2.5.1. Linked Data | 16 |
| 2.5.2. Social Network Behavior | 16 |
| Chapter 3 Methodology | 27 |
| Chapter 4 – Result and Discussion | 37 |
| 4.1. Research Design | 37 |
| 4.2. Result..... | 37 |
| 4.2.2. Experiment for selection of feature selection techniques, type selection feature Gain Information, TFIDF and GGA..... | 46 |
| 5.1. Conclusion..... | 58 |
| 5.2. Recommendation | 58 |

| | |
|----------------------------------|----|
| References..... | 59 |
| APPENDIX E Curriculum Vitae..... | 63 |

List of Figures

| | |
|---|----|
| Figure 1 Separation of two data classes with maximum margins..... | 18 |
| Figure 2 Illustration 10-Fold Cross Validation..... | 19 |
| Figure 3 Confussion Matrix..... | 20 |
| Figure 4 Three Types of Fake News form Three Sub-Tasks in Fake News Detection | 24 |
| Figure 5 Methodology..... | 28 |
| Figure 6 Flowchart Case Folding..... | 29 |
| Figure 7 Flowchart Tokenization..... | 30 |
| Figure 8 Flowchart Stopword Removal..... | 31 |
| Figure 9 Example Unigrams and Bigrams..... | 31 |
| Figure 10 Flowchart NGrams | 32 |
| Figure 11 Flowchart Feature Selection..... | 33 |
| Figure 12 Flowchart Model Development..... | 34 |
| Figure 13 Flowchart Model Assessment..... | 34 |
| Figure 14 Flowchart 10-fold Validation | 36 |
| Figure 15 Flowchart Classification..... | 36 |
| Figure 16 Process Document From Files..... | 39 |
| Figure 17 Text Document Selection | 39 |
| Figure 18 List Text Directory | 40 |
| Figure 19 Stopword Removal | 40 |
| Figure 20 Validation Operator..... | 41 |
| Figure 21 Selecting Machine Learning..... | 42 |
| Figure 22 Composition of Classification Processes in Rapidminer..... | 42 |
| Figure 23 Confussion Matrix Stopword Removal Preprocessing..... | 42 |
| Figure 24 Graph Results of Stopword removal and Machine Learning all Candidate | 44 |
| Figure 25 Stemming Process | 44 |
| Figure 26 Confussion Matrix Stemming Preprocessing | 45 |
| Figure 27 Graph Results of Stemming and Machine Learning all candidate | 46 |
| Figure 28 Graph Results of Time Processing Stemming and Stopword Removal in all candidate | 47 |
| Figure 29 Sequence of classification process Information Gain | 48 |
| Figure 30 Confussion Matrix GI+NB Preprocessing | 48 |
| Figure 31 Graph Results Information Gain + Naive Bayes all candidate | 49 |
| Figure 32 Sequence of Classification process TFIDF | 50 |
| Figure 33 Confussion Matrix TFIDF+NB Preprocessing..... | 50 |
| Figure 34 Graph Results of Feature Selection TFIDF + Machine Learning all candidate | 51 |
| Figure 35 Sequence of Classification process GGA..... | 52 |