

SENTIMENT ANALYSIS OF INDONESIAN LOW COST GREEN CARS WITH
TWITTER DATA

By

Avin Mohanza Kasim
12110007

A thesis submitted to the Faculty of
ENGINEERING AND INFORMATION TECHNOLOGY

in partial fulfillment of the requirements
for the
BACHELOR'S DEGREE
in

INFORMATION TECHNOLOGY



SWISS GERMAN UNIVERSITY
EduTown BSD City
Tangerang 15339
Indonesia

July 2014

Revision after Thesis Defense on July 16th 2014

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.

Avin Mohanza Kasim

Student

Date

Approved by:

James Purnama, M.Sc

Thesis Advisor

Date

Alva Erwin, M.Sc.

Thesis Co-Advisor

Date

Dr. Ir. Gembong Baskoro, M.Sc

Dean

Date

ABSTRACT**SENTIMENT ANALYSIS OF INDONESIAN LOW COST GREEN CARS WITH
TWITTER DATA**

By

Avin Mohanza Kasim
James Purnama, M.Sc, Advisor
Alva Erwin, M.Sc, Co-Advisor

SWISS GERMAN UNIVERISTY

Social Media has become a critical source for marketing's tool to obtain varies analysis research and results and sentiment analysis includes as an example. Sentiment Analysis is an analysis method for determining classifications of data sets into two classes; positive and negative. This research examines the sentiment analysis of Low Cost Green Car such by using tweets for measuring the satisfaction of people; focused on those who lived in Indonesian region, opinion and implicit facts of Low Cost Green Car. This research will conclude the study of sentiment analysis of tweets and methodology of retrieving the sentiment analysis regarding to Low Cost Green Car in Twitter. The insight from gathered tweets will be processed to retrieve the tendency of public's sentiment for each model such as Toyota Agya, Brio, Karimun, etc. The final result of this research may assist concerned industries and companies for later decision of their products, marketing strategy and business plan.

Keywords: Low Cost Green Cars, Sentiment Analysis, Rapidminer, Text Classification



SWISS GERMAN UNIVERSITY

DEDICATION

I dedicate this works for the future of the country I loved: Indonesia



ACKNOWLEDGEMENTS

First of all I wish to thank to Allah for giving me time to live and strength so I can finish my thesis and also my family, my mom who always supports me and cheers me up when I am down. Special thanks to my dad for guiding me a better way of life and to be a better person, I wish you were happy there.

I wish to thank the members of my committee for their support, patience and good humor. Their gentle but firm direction has been most appreciated. James Purnama, M.Sc was particularly helpful in guiding me in this thesis, Alva Erwin's patience toward my stubborn but still guides me until the end.

Thank to IT songo and three idiots for your presence, helps and the entertainment. Adrian Rhesa, Antonio Alfredo, Ari Aulia Hakim, Calvin Kwan, Harmando Taufik Gemilang, Kafin Bahfen, Muhammad Fadhil Mudjiono, Muhammad Ghaffar Adipridhana, Norman Jaya Subrata, Pandu Prakoso Tardan, Richard Kartiyanta and Sayid Ali Hadi. I cannot find these types of friend in another side of this world.

Thanks to Byondina Primasiwi for your support and patience during this research, thanks for reminding me the importance of nutrition in my life so that I get the power to finish this research. Also thanks to Alvita Subagyo, Benita Cecilia and Intan Rahayu Permatasarie

I would like give my gratitude to Kak Fatimah Wulandini and Chodijah Laras Putri for giving me the right way to do this research.

And last but not least, thanks to all people around me, I cannot mention all of you but your contribution makes me mature and makes my life colorful.

Table of Contents

STATEMENT BY THE AUTHOR	2
ABSTRACT.....	3
DEDICATION.....	5
ACKNOWLEDGEMENTS.....	6
LIST OF FIGURES.....	10
LIST OF TABLES.....	14
CHAPTER 1 – Introduction.....	15
1.1 General Statement of Problem Area.....	15
1.2 Research Purpose.....	15
1.3 Research Scope and Limitations	15
1.4 Research Problem.....	16
1.5 Significance of Study	16
1.6 Research Question and Hypothesis	17
1.6.1 Questions	17
1.6.2 Hypothesis.....	17
1.7 Methodology.....	18
1.7.1 Literature Review.....	18
1.7.2 Data Collection.....	18
1.7.3 Data Preprocessing.....	18
1.7.4 Creating Lexicon	19
1.7.5 Sentiment Analysis Process	19
1.7.6 Evaluation Analysis	19
1.7.7 Experiment.....	19
1.7.8 Publication.....	19
1.8 Design and Instrumentation	20
1.9 Data Analysis.....	20
CHAPTER 2 – LITERATURE REVIEW	22
2.1 Sentiment Analysis	22
2.2 Low Cost Green Car	23

2.3	Twitter.....	25
2.3.1	Characteristic of Twitter.....	26
2.3.2	Twitter as Corpus Sentiment Analysis.....	27
2.4	Emoticon Sentiment Analysis	27
2.5	Methodology Classification.....	28
2.5.1	Bag Words of Models	28
2.5.2	Naïve Bayesian Classifiers	28
2.5.3	Support Vector Machine	29
2.6	Support Vector Machine and Text Classification	29
2.7	Machine Learning Methods	30
2.7.1	Naïve Bayes.....	31
2.7.2	Maximum Entropy	31
2.7.3	Support Vector Machine	32
2.8	Related Work.....	33
CHAPTER 3 - METHODOLOGY		34
3.1	Data Collection	35
3.1.1	Web Crawling.....	35
3.1.2	Data Selection.....	40
3.1.3	Data Storing	43
3.2	Data Processing.....	46
3.2.1	Removing Promotion or Advertising tweets.....	47
3.2.2	Removing Foreign Language	49
3.2.3	Removing Link.....	49
3.2.4	Removing Instatweet.....	50
3.2.5	Removing Youtube Link.....	51
3.2.6	Removing Duplicate Tweet	51
3.3	Creating Sentiment Lexicon Dictionary	53
3.4	Sentiment Analysis Process.....	55
3.5	Evaluation Analysis	57
3.6	Experiment	57
CHAPTER 4 - Result.....		59
4.1	Result Overview.....	59
4.2	Sentiment Lexicon Dictionary Classification.....	59

4.2.1	Process Document from File.....	59
4.2.2	Validation.....	62
4.2.3	Store.....	63
4.2.4	Sentiment Lexicon Result.....	63
4.3	Processing Training Set	65
4.3.1	Retrieve	65
4.3.2	Process Document from File.....	65
4.3.3	Validation.....	65
4.3.4	Store.....	66
4.3.5	Training set result.....	66
4.4	Low Cost Green Cars Tweets Classification.....	68
4.4.1	Retrieve	70
4.4.2	Process Document from File.....	70
4.4.3	Low Cost Green Cars Tweets Result	70
4.4.4	Experiment.....	75
4.5	Data Analysis.....	86
4.5.1	LCGC analysis	86
4.5.2	Algorithm Analysis.....	89
CHAPTER 5 - CONCLUSION		92
5.1	Conclusion.....	92
5.2	Further Work.....	93
5.3	Contributions.....	93
GLOSSARY		96
APENDIX A		97
APENDIX B		99
CURICULUM VITAE.....		105

LIST OF FIGURES

Chapter 1

Figure 1. 1 Mthodology Overview 19

Chapter 2

Figure 2. 1 Honda Brio and Daigatsu Ayla..... 25

Figure 2. 2 Toyota Agya and Suzuki Karimun Wagon R 25

Figure 2. 3 Total Character in Twitter 26

Chapter 3

Figure 3. 1 Methodology Overview Diagram..... 34

Figure 3. 2 Data Collection Overview 35

Figure 3. 3 Twitter API Migration Warning 36

Figure 3. 4 Communication Between Two Web Services 37

Figure 3. 5 Twitter and Google Account Permission..... 38

Figure 3. 6 Search Keyword 38

Figure 3. 7 Configuring Spreadsheet Column..... 39

Figure 3. 8 Configuring Zap 39

Figure 3. 9 Search Daihatsu Ayla 40

Figure 3. 10 Search Honda Brio 40

Figure 3. 11 Search Suzuki Karimun Wagon R 40

Figure 3. 12 Search Toyota Agya 40

Figure 3. 13 Search Honda Brio Murah 41

Figure 3. 14 Search Ayla Mahal 42

Figure 3. 15 Keyword interior is connected to all low cost green cars 42

Figure 3. 16 Zapier Communication Pattern..... 43

Figure 3. 17 Google spread sheet file contain tweets for each lcgc..... 44

Figure 3. 18 Excel file exported from google spread sheet..... 44

Figure 3. 19 Honda Brio google spreadsheet contain..... 45

Figure 3. 20 Honda Brio nyaman google spread sheet contain	46
Figure 3. 21 Data preprocessing overview	47
Figure 3. 22 Promotion tweets in brio's google spread sheet	48
Figure 3. 23 Advertising tweets in agya's google spread sheet	48
Figure 3. 24 Foreign language tweets	49
Figure 3. 25 Tweet with link contains positive sentiment	50
Figure 3. 26 Tweet with instgram link	50
Figure 3. 27 Tweet with youtube link	51
Figure 3. 28 Duplicate tweet	52
Figure 3. 29 Creatng sentiment dictionary overview	53
Figure 3. 30 Sentiment lexicon dictionary	54
Figure 3. 31 Emoticon lexicon in dictionary	55
Figure 3. 32 Sentiment analysis process overview	56
Chapter 4	
Figure 4. 1 Dictionary training model in rapid miner	59
Figure 4. 2 Process document from file parameter	60
Figure 4. 3 Positive and negative directories	61
Figure 4. 4 Text processing	61
Figure 4. 5 Regular expression tokenization.....	62
Figure 4. 6 Validation using svm algorithm	62
Figure 4. 7 Validation using naïve bayes algorithm	63
Figure 4. 8 Dictionary accuracy result from svm algorithm	64
Figure 4. 9 Dictionary accuracy result from naïve bayes algorithm	64
Figure 4. 10 Training set model in rapidminer	65
Figure 4. 11 Validation using svm algorithm	66
Figure 4. 12 Validation using naïve bayes algorithm	66
Figure 4. 13 Accuracy training set result with svm algorithm	66

Figure 4. 14 Precision training set result with svm algorithm	67
Figure 4. 15 Recall training set result with svm algorithm.....	67
Figure 4. 16 Accuracy training set result with naïve bayes algorithm	67
Figure 4. 17 Precision training set result with naïve bayes algorithm	68
Figure 4. 18 Recall training set result with naïve bayes algorithm	68
Figure 4. 19 Toyota Agya 55 tweets	69
Figure 4. 20 Training model for dataset.....	70
Figure 4. 21 Honda Brio result using svm algorithm	71
Figure 4. 22 Toyota Agya result using svm algorithm.....	71
Figure 4. 23 Daihatsu Ayla result using svm algorithm.....	72
Figure 4. 24 Suzuki Karimun wagon R result using svm algorithm.....	72
Figure 4. 25 Honda Brio result using naïve bayes algorithm	73
Figure 4. 26 Toyota Agya result using naïve bayes algorithm	73
Figure 4. 27 Daihatsu Ayla result using naïve bayes algorithm	74
Figure 4. 28 Suzuki Karimun wagon R result using naïve bayes algorithm ...	74
Figure 4. 29 Success and error percentage classification Toyota Agya	76
Figure 4. 30 Success and error percentage classification Daihatsu Ayla	77
Figure 4. 31 Success and error percentage classification Honda Brio	78
Figure 4. 32 Success and error percentage classification Suzuki Karimun Wagon R	79
Figure 4. 33 Overall success and error percentage classification.....	80
Figure 4. 34 Success and error percentage Toyota Agya	81
Figure 4. 35 Success and Error percentage Daihatsu Ayla	82
Figure 4. 36 Success and error percentage Honda Brio.....	83
Figure 4. 37 Success and Error percentage Suzuki Karimun Wagon R	84
Figure 4. 38 Overall success and error percentage	85
Figure 4. 39 Overall statistical sentiment.....	86
Figure 4. 40 Toyota agya tagword positive(lef) and negative(right)	87

Figure 4. 41 Honda Brio tagword positive(left) and negative(right) 88

Figure 4. 42 Daihatsu Ayla tagword positive(left) and negative(right)..... 88

Figure 4. 43Suzuki karimun wagon R tagword positive(left) and negative(right)
..... 89



LIST OF TABLES

Chapter 3

Table 3. 1 Positive and Negative Words.....	41
Table 3. 2 Brio tweets before filter and after filtered	52
Table 3. 3 Agya tweets before filter and after filtered	52
Table 3. 4 Ayla tweets before filter and after filtered.....	52
Table 3. 5 Karimun tweets before filter and after filtered	53

Chapter 4

Table 4. 1 Agya tweets details.....	75
Table 4. 2 Ayla tweets detail.....	76
Table 4. 3 Brio tweets detail	77
Table 4. 4 Wagon R tweets detail.....	78
Table 4. 5 Overall tweets detail	79
Table 4. 6 Agya tweets detail.....	81
Table 4. 7 Ayla tweets detail.....	81
Table 4. 8 Brio tweets detail	82
Table 4. 9 Wagon R tweets detail.....	83
Table 4. 10 Overall tweets detail	84
Table 4. 11 Low cost green cars tweets statistic	86
Table 4. 12 Svm and nbc accuracy, precision and recall comparison	90
Table 4. 13 Svm and nbc accuracy comparison	91