

SENTIMENT ANALYSIS OF SINHALA NEWS COMMENTS

Isuru Udara Liyanage

(168243P)

Thesis/Dissertation submitted in partial fulfillment of the requirements for the degree
Master of Science

Department of Computer Science and Engineering

University of Moratuwa

Sri Lanka

May 2018

DECLARATION

I declare that this is my own work and this dissertation does not incorporate without acknowledgement any material previously submitted for a Degree or Diploma in any other University or institute of higher learning and to the best of my knowledge and belief it does not contain any material previously published or written by another person except where the acknowledgement is made in the text.

Also, I hereby grant to University of Moratuwa the non-exclusive right to reproduce and distribute my thesis/dissertation, in whole or in part in print, electronic or other medium. I retain the right to use this content in whole or part in future works (such as articles or books).

Isuru Udara Liyanage

Date

The above candidate has carried out research for the Masters thesis/ Dissertation under my supervision.

Dr. Surangika Ranathunga

Date

ABSTRACT

Mining sentiment values from unstructured text uncovers interesting patterns that can be effectively used for many applications. One interesting yet poorly explored area is online news comment analysis, in particular for Sinhala language. Despite the uptrend in online Sinhala news articles and related comments, no efficient method exists for analyzing and identifying the public sentiment associated with them. In this research our effort is to classify online Sinhala news comments according to its sentiment orientation.

Most of the sentiment analysis research is done for English language. As for Sinhala, only one research can be found for classification of Sinhala news comments according to its sentiment values. Since it is an initial attempt it lacks the use of advanced text analysis methods and localization, and hence can be improved in many ways.

In this research we build a complete Sinhala sentiment analysis system, from data collection to sentiment classification. First we gather a dataset by crawling through a popular online news site. Compiled dataset contains news items and related comments. Sufficient amount of comments are annotated according to its sentiment values. Finally sentiment analysis is carried out to identify sentiment values associated with each comment.

This research provides many valuable outputs to the research community, sentiment analysis for Sinhala text. Dataset, the labeled data set in particular, can be used for future Sinhala text analysis research. Finally direction and a baseline will be set for future research on sentiment analysis for Sinhala text.

ACKNOWLEDGEMENT

I would like to express my deepest appreciation to my thesis advisor Dr. Surangika Ranathunga for her continuous guidance throughout the research work. She steered me in the right direction from the beginning by providing valuable insights, resources and supervision. Without her guidance and motivation this wouldn't have been a success.

I am grateful to Dr. Charith Chitraranjan, Dr. Shehan Perera, Dr. Malaka Walpola and Dr. Indika Perera for their motivation and guidance for making this a success.

I would also like to thank my colleges at DirectFN for their support and the understanding during the MSc duration.

Finally I express my heartfelt gratitude for my parents for their love and support throughout my life.

TABLE OF CONTENT

DECLARATION.....	ii
ABSTRACT.....	iii
ACKNOWLEDGEMENT.....	iv
TABLE OF CONTENT.....	v
LIST OF FIGURES.....	vii
LIST OF TABLES.....	vii
LIST OF EQUATIONS.....	ix
LIST OF ABBREVIATIONS.....	ix
CHAPTER 1: INTRODUCTION.....	1
1.1 News and Comments.....	1
1.2 Sentiment Analysis.....	3
1.3 Problem and Motivation.....	3
1.4 Objective.....	4
CHAPTER 2: LITERATURE REVIEW.....	5
2.1 Sentiment Analysis.....	5
2.1.1 Formal Definition.....	8
2.2 Sentiment Classification.....	9
2.2.1 Unsupervised.....	9
2.2.2 Supervised.....	12
2.2.1 Other Languages.....	13
2.2.3 Sentiment Classification of News Comments.....	13
2.2.3.1 Sinhala Language.....	14
2.2.4 Features.....	16
2.2.3 Domain Adaptation (transfer learning).....	17
2.2.4 Cross-lingual Sentiment Classification.....	18
2.2.5 Sentiment Shifters.....	18
2.3 Lexicon.....	19
2.4 Focus Detection.....	21
2.5 Evaluation.....	22
2.6 Summary.....	23
CHAPTER 3: RESEARCH METHODOLOGY.....	25

3.1 Architecture and Implementation.....	26
3.1.1 Data Collection.....	26
3.1.1.1 Inter-rater Agreement.....	30
3.1.2 Preprocessing.....	32
3.1.2.1 Effect of Punctuations.....	34
3.1.3 Feature Selection.....	37
3.1.3.1 Word Embedding.....	38
3.1.4 Classification Algorithms.....	41
3.1.4.1 Naive Bayes.....	41
3.1.4.2 Logistic Regression.....	42
3.1.4.3 Decision Tree.....	42
3.1.4.4 Random Forest.....	43
3.1.4.5 Support Vector Machines.....	43
3.1.4.6 Convolution Neural Networks.....	44
3.2.3.7 Recurrent Neural Networks.....	45
3.2.3.8 CNN + SVM Hybrid Technique.....	47
CHAPTER 4: SYSTEM EVALUATION.....	48
4.1 Baseline Experiment.....	48
4.2 Preprocessing - Effect of Punctuations.....	49
4.3 Features.....	52
4.3.1 Word Embedding: Word2Vec.....	54
4.4 Classification Algorithms.....	56
4.4.1 Recurrent Neural Networks.....	58
4.4.2 Hybrid CNN + SVM.....	59
4.5 Error Analysis.....	60
4.6 Summary.....	61
CHAPTER 5: CONCLUSION.....	62
5.1 Future Work.....	63
REFERENCES.....	64

LIST OF FIGURES

Figure 1.1: Online news comments	2
Figure 1.2: News Comments.....	2
Figure 1.3: Interactive widgets.....	3
Figure 2.1: Components and Features [8]	7
Figure 3.1: Proposed system architecture	25
Figure 3.2: XML formatted news article	27
Figure 3.3: Web application to simplify annotation task	29
Figure 3.4: Annotated article	29
Figure 3.5: Exploring data	30
Figure 3.6: CSV formatted extracted comments	34
Figure 3.7: CNN input matrix	45
Figure 3.8: CNN model	45
Figure 3.9: RNN model	46

LIST OF TABLES

Table 2.1 Sentiment Breakdown	5
Table 2.2 Sinhala news comment breakdown	6
Table 2.3 Patterns of tags for extracting two-word phrases from reviews.	10
Table 3.1: Inter-rater agreement	30
Table 3.2: Word2Vec model properties	40
Table 4.1: Experiment results Medagoda et al [19]	48
Table 4.2: Initial experiment results	49
Table 4.3: Effect of punctuations	50

Table 4.4: Features with highest sentiment orientation (Logistic Regression)	51
Table 4.5: Logistic Regression classifier	53
Table 4.6: SVM classifier	53
Table 4.7: Random Forest classifier	53
Table 4.8: Naive Bayes classifier	54
Table 4.9: Decision Tree classifier	54
Table 4.10: Word2Vec model - Similar words	55
Table 4.11: W2V Continuous Bag of Word model (CBOW)	55
Table 4.12: W2V Skip gram model	55
Table 4.13: W2V vector dimension and features	56
Table 4.14: Logistic Regression parameter tuning (c value)	57
Table 4.15: Random Forest parameter tuning (# features, criterion, max features)	57
Table 4.16: SVM parameter tuning (kernel, c value)	57
Table 4.17: TF-IDF features	58
Table 4.18: W2V skip gram with word count features	58
Table 4.19: RNN-LSTM with W2V skip gram word count features	58
Table 4.20: Confusion matrix (SVM with W2V skip gram word count)	59
Table 4.21: Confusion matrix (RNN-LSTM with W2V skip gram word count) ...	59
Table 4.22: Hybrid CNN + SVM with Word2Vec	59
Table 4.23: Misclassified instances	60
Table 4.24: Best results of each of the algorithm	61

LIST OF EQUATIONS

Equation 2.1: Pointwise Mutual Information	10
Equation 2.2: Sentiment Orientation	11
Equation 2.3: Precision	22
Equation 2.4: Recall	23
Equation 2.5: Accuracy	23
Equation 2.6: F1 Score	23
Equation 3.1: Cohen's kappa	30

LIST OF ABBREVIATIONS

NLP	Natural Language Processing
POS	Part Of Speech
SVM	Support Vector Machines
SO	Sentiment Orientation
PMI	Pointwise Mutual Information
WSD	Word Sense Disambiguation
FNE	Focused Named Entities