



AIN SHAMS UNIVERSITY
FACULTY OF ENGINEERING
Computer Engineering and Software Systems

Semi-supervised Language-independent Sentiment Analysis

A Thesis submitted in partial fulfillment of the requirements of
Master of Science in Electrical Engineering
(Computer Engineering and Systems)

by

Mohammad Hassan Hanafy

Bachelor of Science in Electrical Engineering
(Electronics Engineering and Electrical Communications)
Faculty of Engineering, Cairo University, 2012

Supervised By

Prof. Hazem Mahmoud Abbas

Prof. Mahmoud Ibrahim Khalil

Cairo, 2019



AIN SHAMS UNIVERSITY
FACULTY OF ENGINEERING
Computer Engineering and Systems

Semi-supervised Language-independent Sentiment Analysis

by

Mohammad Hassan Hanafy

Bachelor of Science in Electrical Engineering
Electronics Engineering and Electrical Communications
Faculty of Engineering, Cairo University, 2012

Examiners' Committee

Name and affiliation

Signature

Prof. Hassen Taher Dorrah

Electrical Power and Machines Engineering
Faculty of Engineering, Cairo University.

.....

Prof. Hoda Korashy Mohamed

Computer Engineering and Systems
Faculty of Engineering, Ain Shams University.

.....

Prof. Hazem Mahmoud Abbas

Computer Engineering and Systems
Faculty of Engineering, Ain Shams University.

.....

Dr. Mahmoud Ibrahim Khalil

Computer Engineering and Systems
Faculty of Engineering, Ain Shams University.

.....

Date: 18 July 2019

Statement

This thesis is submitted as a partial fulfillment of Master of Science in Electrical Engineering, Faculty of Engineering, Ain shams University. The author carried out the work included in this thesis, and no part of it has been submitted for a degree or a qualification at any other scientific entity.

Mohammad Hassan Hanafy

Signature

.....

Date: 18 July 2019

Researcher Data

Name: Mohammad Hassan Hanafy Mahmoud

Date of Birth: 10/12/1990

Place of Birth: Cairo, Egypt

Last academic degree: Bachelor of Science

Field of specialization: Electrical Engineering

University issued the degree : Cairo University

Date of issued degree : 2012

Current job : Senior Software Engineer

Thesis Summary

Sentiment analysis plays an important role in research and industry as extracting the opinions of people could be beneficial in several domains. Millions of active users express their opinions and sentiments daily in blogs, social networks and different other platforms. Twitter allows users from all the globe to express their feelings and opinions freely in a unit of text called tweet. With millions of tweets get published daily, twitter has attracted many researchers and organizations to exploit its data.

Early works on sentiment analysis have used rule based approaches then machine learning classifiers were introduced as it surpassed the former one, but most of these works have been built for a certain language or certain domain. Being a global platform that is used in almost all the countries creates new challenges to be faced. Users express their sentiments with different languages, tend not to use the formal language, do not stick to grammar rules, use slang words and new expressions are continuously added. that kept the door open for further innovations and systems to solve these problems.

In this thesis, we build a semi-supervised language-independent technique that does not depend on any feature of a certain language. It uses emoticons that is used heavily in twitter as heuristic labels to build the training set from raw tweets. Statistical and unsupervised approaches i.e bag of words and word2vec are used as feature representation for the classifiers.

Two main models are proposed in this work, both combine typical and deep learning classifiers i.e SVM, Max.Ent. ,CNN and LSTM. The first model used more core classifiers than the second one, and focused on tuning the combination of them to overcome their limitations. The second model used fewer classifiers but focused more on the feature representation, specially word2vec and how to make use of its models i.e skip-gram and continuous bag of words. The proposed models are very efficient regards memory and time as it used only 10% of training dataset compared to other approaches on the same test dataset. The results also show that both approaches are performant as they achieve the state-of-the-art accuracy of 86.37%.

Key words: Sentiment Analysis, NLP, Deep Learning, Machine Learning, Twitter

Acknowledgment

Mohammad Hassan Hanafy
Computer Engineering and System
Faculty of Engineering
Ain Shams University
Cairo, Egypt
July 2019

Abstract

Sentiment analysis plays an important role in research and industry as extracting the opinions of people could be beneficial in several domains. Millions of active users express their opinions and sentiments daily in blogs, social networks and different other platforms. Twitter allows users from all the globe to express their feelings and opinions freely in a unit of text called tweet. With millions of tweets get published daily, twitter has attracted many researchers and organizations to exploit its data.

Early works on sentiment analysis have used rule based approaches then machine learning classifiers were introduced as it surpassed the former one, but most of these works have been built for a certain language or certain domain. Being a global platform that is used in almost all the countries creates new challenges to be faced. Users express their sentiments with different languages, tend not to use the formal language, do not stick to grammar rules, use slang words and new expressions are continuously added. that kept the door open for further innovations and systems to solve these problems.

In this thesis, we build a semi-supervised language-independent technique that does not depend on any feature of a certain language. It uses emoticons that is used heavily in twitter as heuristic labels to build the training set from raw tweets. Statistical and unsupervised approaches i.e bag of words and word2vec are used as feature representation for the classifiers.

Two main models are proposed in this work, both combine typical and deep learning classifiers i.e SVM, Max.Ent. ,CNN and LSTM. The first model used more core classifiers than the second one, and focused on tuning the combination of them to overcome their limitations. The second model used fewer classifiers but focused more on the feature representation, specially word2vec and how to make use of its models i.e skip-gram and continuous bag of words. The proposed models are very efficient regards memory and time as it used only 10% of training dataset compared to other approaches on the same test dataset. The results also show that both approaches are performant as they achieve the state-of-the-art accuracy of 86.37%.

Key words: Sentiment Analysis, NLP, Deep Learning, Machine Learning, Twitter

Contents

Contents	xii
List of Figures	xvi
List of Tables	xvii
Abbreviations	xviii
Symbols	xix
1 Introduction	1
1.1 Definitions	2
1.1.1 Sentiment	2
1.1.2 Sentiment Analysis	2
1.2 Twitter Sentiment Analysis	3
1.2.1 Challenges	3
1.3 Thesis Contribution	4
1.4 Thesis Organization	4
Chapter 2: Literature Review:	4
Chapter 3: Theory and Algorithms:	5
Chapter 4: Proposed Models:	5
Chapter 5: Conclusions and Future work:	5
2 Literature Review	6
2.1 General Sentiment Analysis	7
2.2 Twitter Sentiment Analysis	8
2.3 Deep Learning Sentiment Analysis	9
3 Theory and Algorithms	12
3.1 Methodology	14
3.1.1 Data processing	14
3.1.2 Auto-Labeling	16
3.1.3 Feature Extraction and dimension reduction	17
3.1.3.1 Bag of Words	17
3.1.3.2 Term Frequency-Inverse Document Frequency	18
3.1.3.3 Word2Vec	18
3.1.4 Classifiers	21
3.1.4.1 Support Vector Machine	21

3.1.4.2	Maximum Entropy	21
3.1.4.3	Long-Short Term Memory	22
3.1.4.4	Convolutional Neural Networks	25
3.1.4.5	Voting Ensembles	26
3.2	Experimentation Tools	28
	Data processing	29
	Auto-labeling	30
3.2.1	Feature Extraction	30
	BOW with Tf-idf	30
	Word2Vec	30
3.2.2	Classifiers	30
	Support Vector Machine	30
	Maximum Entropy	30
	Long-Short Term Memory	30
	Convolutional Neural Networks	31
4	Models	32
4.1	Datasets	36
4.1.1	STS dataset	36
4.1.2	CIKM dataset	36
4.2	The Experimental Protocol	37
4.3	Core Models	38
4.3.1	Model 1: BOW with Tf-idf	38
	4.3.1.1 Model	38
	4.3.1.2 Motivation	38
	4.3.1.3 Results	39
4.3.2	Model 2: Aggregation of word2vec	40
	4.3.2.1 Model	40
	4.3.2.2 Motivation	41
	4.3.2.3 Results	41
4.3.3	Model 3: LSTM	41
	4.3.3.1 Model	41
	4.3.3.2 Motivation	42
	4.3.3.3 Results	43
4.3.4	Model 4: CNN	43
	4.3.4.1 Model	43
	4.3.4.2 Motivation	44
	4.3.4.3 Results	44
4.4	Proposed Models	45
4.4.1	Model 1: Weighted Voting ensemble	45
	4.4.1.1 Motivation	45
	4.4.1.2 Model	46
	4.4.1.3 Results	46
4.4.2	Model2: Rule based voting	50
	4.4.2.1 Motivation	50
	4.4.2.2 Model	51
	4.4.2.3 Results	51

4.5	Previous Results	51
4.6	Analysis	54
	Word2vec Embedding size and Training data size	55
	BOW window size	55
	Text size	55
	Training data size	56
4.7	Proposed Models Contributions	57
4.7.1	Memory	57
4.7.2	Time	59
5	Conclusions and Future Work	60
5.1	Conclusions	60
5.1.1	Semi-supervised	61
5.1.2	Language Independence	61
5.1.3	Models	61
5.1.4	Result	62
5.2	Future Work	62
5.2.1	Data	62
5.2.2	Classifiers	63
5.2.3	Languages	63
5.2.4	Architecture	63
	 Bibliography	 65