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