DEEP LEARNING FOR ARABIC TEXT SENTIMENT ANALYSIS

By

Rana Mahmoud Kamel AbdelMoneim Kamel

A Thesis Submitted to the
Faculty of Engineeringat Cairo University
in Partial Fulfillment of the
Requirements for the Degree of
MASTER OF SCIENCE
in
Electronics and Communication Engineering

FACULTY OF ENGINEERING, CAIRO UNIVERSITY
GIZA, EGYPT
2019

DEEP LEARNING FOR ARABIC TEXT SENTIMENT ANALYSIS

By

Rana Mahmoud Kamel AbdelMoneim Kamel

A Thesis Submitted to the
Faculty of Engineeringat Cairo University
in Partial Fulfillment of the
Requirements for the Degree of
MASTER OF SCIENCE
in
Electronics and Communication Engineering

Under the Supervision of

Prof. Dr. Mohamed Nafie
Professor
Electronics and Communication Engineering Department
Faculty of Engineering, Cairo University

Dr. Karim Seddik
Associate Professor
Communications and Computer department
Faculty of Engineering, American University in Cairo

FACULTY OF ENGINEERING, CAIRO UNIVERSITY
GIZA, EGYPT
2019

DEEP LEARNING FOR ARABIC TEXT SENTIMENT ANALYSIS

By

Rana Mahmoud Kamel AbdelMoneim Kamel

A Thesis Submitted to the
Faculty of Engineeringat Cairo University
in Partial Fulfillment of the
Requirements for the Degree of
MASTER OF SCIENCE
in
Electronics and Communication Engineering

Approved by the Examining Committee:

Prof. Dr. Mohamed Nafie, Thesis Main Advisor

Dr. Karim Seddik, Advisor
Associate Professor at faculty of engineering, American University in Cairo

---

Prof. Dr. Mohamed Khairy, Internal Examiner

Prof. Dr. Samhaa El-Beltagy, External Examiner
Professor at faculty of Computer Science, Cairo University

FACULTY OF ENGINEERING, CAIRO UNIVERSITY
GIZA, EGYPT
2019

Engineer's Name: Rana Mahmoud Kamel AbdelMoneim Kamel
Date of Birth: 01/03/1991
Nationality: Egyptian
E-mail: rana.mahmoudkamel@hotmail.com
Phone: +20 100 387 835 3
Address: 108, St. 104, Maadi, Cairo

Registration Date: 01/03/2015
Awarding Date: 2019
Degree: Master of Science
Department: Electronics and Communication Engineering

Supervisors:

Prof. Dr. Mohamed Nafie
Dr. Karim Seddik

Examiners:

Prof. Dr. Mohamed Nafie	(Thesis main advisor)
Dr. Karim Seddik Associate Professor at faculty of Engineering, American University in Cairo	(Advisor)
Prof. Dr. Mohamed Khairy	(Internal examiner)
Prof. Dr. Samhaa El-Beltagy Professor at faculty of Computer Science, Cairo University	(External examiner)

Title of Thesis:

Deep Learning for Arabic Text Sentiment Analysis

Key Words:

Deep Learning; Word Embeddings; Arabic Natural Language processing; Sentiment Analysis

Summary:
Nowadays, people express their opinions, reviews on products, movies, hotels, restaurants…etc publicly on the internet on social media platforms, blogs or forums. The number of Arabic speaking users on the internet has been increased in the last decade and the research for analyzing the Arabic text has gained a lot of attention.

In my work, deep learning techniques are used to classify Arabic tweets and reviews into two classes (positive/ negative) and three classes (positive/ negative/ neutral). Also, this work investigates whether deep learning can overcome ordinary machine learning algorithms and replace the effort of feature engineering in previous work. Finally, deep learning proved to have better results than machine learning techniques for most of the used datasets by using a data augmentation architecture.

Acknowledgements

This thesis would not exist without the support of many people. First of all, I would like to thank my supervisors Prof. Dr. Mohamed Nafie and Prof. Karim Seddik for the support provided during my work and reviewing the thesis.

I would like to dedicate a special thanks to Prof. Dr. Samhaa EL-beltagy for the continuous support and help I received during my thesis development as well as the encouragement and helpful feedback.

Finally, I would like to thank my lovely parents and husband for the continuous support and encouragement during my studies and specially in the hard times I passed by. I couldn't have done it without you.

Disclaimer

I hereby declare that this thesis is my own original work and that no part of it has been submitted for a degree qualification at any other university or institute.

I further declare that I have appropriately acknowledged all sources used and have cited them in the references section.

Name: Rana Mahmoud Kamel AbdelMoneim Kamel Date: 10/6/2019

Signature:

Table of Contents

List of Tables

List of Figures

Abstract

Nowadays, people express their opinions, reviews on products, movies, hotels, restaurants…etc publicly on the internet on social media platforms, blogs or forums. The number of Arabic speaking users on the internet has been increased in the last decade and the research for analyzing the Arabic text has gained a lot of attention.

In my work, deep learning techniques are used to classify Arabic tweets and reviews into two classes (positive and negative) and three classes (positive, negative and neutral). Also, this work investigates whether deep learning can overcome ordinary machine learning algorithms and replace the effort of feature engineering in previous work.

Finally, deep learning proved to have better results than machine learning techniques for most of the used datasets by using a data augmentation architecture.

Chapter 1: Introduction

In the last decade, the Internet has become one of the life essentials. People use the internet to do almost everything in life: paying bills, online shopping, hotel or restaurant reservation, jobs' application…etc. With the growth of the internet use, the development of smart phones and the availability and diversity of social media platforms has increased like Facebook, Twitter, Instagram…etc.

Nowadays, people become addicted to social networks where they can express their feelings, opinions, feedback and thoughts. For example, in Twitter, 6000 short messages (tweets) are posted by users every second.[1] [114]

As a result of the users' behavior towards social networks, researchers and companies began to think how to benefit from this freely available data by checking the feedback on their products, marketing some new items, classifying the audience based on opinions. At this point, sentiment analysis got a lot of attention in the field of data mining and natural language processing.

Research has focused on the field of natural language processing and how to analyze human language either verbal or written to extract useful information and to be used in other applications like sentiment analysis, speech recognition, machine translation and a lot more. Sentiment analysis is the task of analyzing and processing the text and detecting whether it expresses a positive or a negative opinion and in other cases detecting the polarity towards a specific topic. Sentiment analysis can be used by companies to gather users' reviews and opinions regarding certain product or service and take decision like continue to produce it or not.[117] [53]

Sentiment analysis is performed by the computers using machine learning and deep learning techniques as will be shown in the next section.

1.1 Problem Formulation

Machine learning is widely used for sentiment classification and recently, deep learning techniques showed a very good performance in various natural language processing (NLP) tasks. Deep learning is considered a part of machine learning methods. It is based on artificial neural networks consisting of an input layer and an output layer with many hidden layers between them where the name deep is derived. The role of these hidden layers is to extract features from input data automatically without manual feature engineering used in ordinary machine learning algorithms. Deep Learning has different techniques such as Deep Neural Networks (DNN), Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN) and Long Short-Term Memory (LSTM).

Deep Learning has gained a huge success in the field of image recognition and it began to be widely used in natural language processing and specifically sentiment analysis. Many research papers have been published in the field of sentiment analysis either using machine learning techniques like SVM, Naïve bayes, KNN…etc or deep learning techniques like CNN and RNN. Most of these researches were analyzing the English language which is the most used language across the world.

[1] `http://www.internetlivestats.com/twitter-statistics/`

The goal of this thesis is to implement different deep learning techniques with different architecture to evaluate the impact of variations of hyper-parameters and the network architecture on the overall performance. We used tweets and reviews from different websites for our datasets and we performed classification into positive and negative sentiments in addition to including the neutral class.

1.2 Motivation

Sentiment analysis is commonly used for English language while work on Arabic language is still in the early stages. Therefore, there is not too many Arabic resources (datasets and lexicons) when compared to the English language. However, analyzing the Arabic language has been the focus of many researchers in the past few years due to the increase of Arabic speaking users on the internet and specifically on social media forums.

Sentiment Analysis in Arabic is considered a challenging task because of two reasons. First, the Arabic language has a very complex morphology and structure. Second, the Arabic language has many variations such as the Classical Arabic (CA) which is mainly used in the Holy Quran, the Modern Standard Arabic (MSA) which is a modernized and simplified version of CA and Dialectal also known as colloquial Arabic (DA) where each region has its own dialect. MSA is understood by all Arabic countries due to its use in education, media and formal communication.

As stated in [46] and as per November 2015 by Internetworldstats[2], the number of people speaking Arabic as first language exceed 267 million and the number of people speaking Arabic as a second language exceeds 250 million. For the Arabic Internet users, there are around 168.1 million users with a user growth rate of 6,592.5. These statistics makes research of sentiment analysis on the Arabic language important. [46]

1.3 Thesis Contribution

In this work, machine learning and deep learning techniques with various architectures are used to classify Arabic tweets and reviews into two classes (positive/ negative) and three classes (positive/ negative/ neutral).

Convolutional Neural Network and Long Short Term memory networks have been implemented with different input like a small dataset and augmented dataset also called in this thesis Merged datasets. We also studied the impact of different architectures like changing number of layers and using multi-channel CNN. In addition, we used two different pre-trained word embeddings to build our input for the deep learning networks.

For the machine learning algorithms, we used the SVM and Naïve Bayes algorithms as proved to be the best classification techniques in the field of sentiment analysis.

Finally, this work investigates whether deep learning can overcome ordinary machine learning algorithms and replace the effort of feature engineering and compare both techniques in terms of performance or accuracy obtained.

[2] http://www.internetworldstats.com/stats7.html

1.4 Thesis Organization

The rest of the paper is structured as follows. Chapter 2 gives a theoretical background on sentiment analysis techniques, machine learning algorithms for supervised and unsupervised learning and the metrics used for evaluation; this chapter also includes the explanation of neural networks and backpropagation and finally discussion on deep learning. Chapter 3 presents the literature review and the work done by previous research in the field of machine learning and deep learning specifically for sentiment analysis in both English and Arabic languages. Chapter 4 presents the main contribution in the thesis by explaining the experiments we performed and the results obtained. Finally, we discuss the conclusion and future work in Chapter 5.