بسم الله الرحمن الرحيم

MONA MAGHRABY

شبكة المعلومات الجامعية

التوثيق الالكتروني والميكروفيلم

MONA MAGHRABY

شبكة المعلومات الجامعية

التوثيق الإلكتروني والميكروفيلم

جامعة عين شمس

التوثيق الإلكتروني والميكروفيلم

قسم

نقسم بالله العظيم أن المادة التي تم توثيقها وتسجيلها
علي هذه الأقراص المدمجة قد أعدت دون أية تغيرات

يجب أن

تحفظ هذه الأقراص المدمجة بعيدا عن الغبار

MONA MAGHRABY

AIN SHAMS UNIVERSITY
FACULTY OF ENGINEERING
Computer and Systems Engineering Department

Using Transfer Learning Techniques for Analysing Text Documents

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

by

Amr Mohamed Hosny Anwar Keleg
Bachelor of Science in Electrical Engineering
(Computer and Systems Engineering Department)
Faculty of Engineering, Ain Shams University, 2017

Supervised By
Prof. Dr. Mahmoud Ibrahim Khalil
Computer and Systems Engineering Department
Faculty of Engineering, Ain Shams University.

Prof. Dr. Samhaa R. El-Beltagy
School of Information Technology
Newgiza University.

Cairo, 2021

AIN SHAMS UNIVERSITY
FACULTY OF ENGINEERING
Computer and Systems Engineering Department

Using Transfer Learning Techniques for Analysing Text Documents

by

Amr Mohamed Hosny Anwar Keleg
Bachelor of Science in Electrical Engineering
(Computer and Systems Engineering Department)
Faculty of Engineering, Ain Shams University, 2017

Examiners' Committee

Name and affiliation	**Signature**
Prof. Dr. Mohamed W. Fakhr Computer Engineering Department Arab Academy For Science Technology and Maritime Transport.	.....
Prof. Dr. Hani M. K. Mahdi Computer and Systems Engineering Department Faculty of Engineering, Ain Shams University.	.....
Prof. Dr. Mahmoud Ibrahim Khalil Computer and Systems Engineering Department Faculty of Engineering, Ain Shams University.	.....
Prof. Dr. Samhaa R. El-Beltagy School of Information Technology Newgiza University.	.....

Date: 18 August 2021

Statement

This thesis is submitted as a partial fulfilment 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.

Amr Mohamed Hosny Anwar Keleg

Signature

.....

Date: 18 August 2021

Researcher Data

Name: Amr Mohamed Hosny Anwar Keleg
Date of Birth: 22/02/1994
Place of Birth: Cairo, Egypt
Last academic degree: Bachelor of Science
Field of specialization: Computer and Systems Engineering
University issued the degree : Ain Shams University
Date of issued degree : 2017
Current job : Teaching Assistant

Thesis Summary

Summary

Transfer learning is becoming a widely used technique in the field of deep learning. In this thesis, it is used for detecting offensive text which is becoming a prevailing phenomenon on online social media. While the technique showed promising results in the task of offensive text classification, the results also showed how these models aren't robust to simple text substitution adversarial attacks. Moreover, hate speech is a specific type of offensive text that needs to be properly represented in the used data sets such that the model can correctly classify them as offensive text.

The thesis is divided into five chapters as listed below:

Chapter 1 is an introductory chapter demonstrating the motivation for using transfer learning for Arabic offensive text classification.

Chapter 2 gives an overview of the different transfer learning techniques that have emerged for building deep learning models especially in the field of natural language processing.

Chapter 3 first describes the offensive and hate speech data sets that are used in the various experiments done throughout the thesis. Then, the different transfer learning paradigms and the adversarial attacking algorithms are described.

Chapter 4 presents the results for the transfer learning experiments on OffensEval2020 data set. Additionally, it reports the results of attacking the best performing model using the new adversarial attacking method. At the end of the chapter, the zero-shot learning performance of the best performing model is reported using three different Arabic hate speech data sets.

Chapter 5 summarises the conclusions of the thesis and provides recommendations regarding future work that can be done to improve the fine-tuned models and to combat the new adversarial attacking methods.

Keywords: Natural Language Processing, Transfer Learning, Transformer models, Adversarial attacks, Arabic NLP

Abstract

Transfer learning is a rising technique that is being widely used to transfer linguistic knowledge from large pretrained models to new task specific models especially in the case of having small-sized annotated data sets related to this task. This technique is used to build machine learning models that can automatically detect abusive Arabic comments. Offensive text is a phenomenon that became ubiquitous on the different social media platforms. Automatic detection of these offensive text comments can be used to build a better understanding of how people are responding to and reacting with everyday events online. Moreover, building automated methods would allow social media sites to monitor the comments that are being shared on their online platforms.

Different methods of transfer learning were used to build the offensive text classification models. First, models based on using a pretrained Arabic dense word embedding as a way of transforming discrete tokens into dense vectors were used. Additionally, contextualised models based on attention mechanisms that are pretrained on large corpora were fine-tuned.

While contextualised models showed great ability in detecting offensive text, using simple adversarial attacking methods indicated how the unsupervised tokenisation techniques used in these models aren't robust to perturbations applied to bad words. Moreover, the quality and diversity of the task specific data set was proven to have a great impact on the generalisability of the fine-tuned model. Evaluating the model on three different Arabic hate speech data sets showed how the model isn't able to classify hate speech samples at the same level since the data set that is used for fine-tuning the model didn't contain a representative sample of these hate speech comments.