Detection of Anti-Social Behaviour in Online Communication in Arabic

Author:
Azalden Alakrot

Supervisors:
Dr. Nikola Nikolov
Dr. Liam Murray

A thesis submitted in fulfillment of the requirements for the degree of Doctor of Philosophy in the
Department of Computer Science and Information Systems

May 7, 2019
Declaration of Authorship

I, Azalden Alakrot, declare that this thesis titled, “Detection of Anti-Social Behaviour in Online Communication in Arabic” and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at University of Limerick.

- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.

- Where I have consulted the published work of others, this is always clearly attributed.

- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.

- I have acknowledged all main sources of help.

- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed:

________________________________________

Date:

________________________________________
Abstract

Azalden Alakrot

Detection of Anti-Social Behaviour in Online Communication in Arabic

Anti-social behaviour on social media cannot be easily ignored as it affects a large and growing percentage of the world’s population. It often has a negative effect on people’s lives; incidents of online abuse that may seem insignificant can have a cumulative impact on mental health. An increasing number of incidents of suicide and violence have been reportedly provoked by anti-social behaviour on social media. Most of the existing machine-learning approaches for detection of offensive language are specifically tailored for online communication in English. Solutions targeting Arabic language are rare, while, as we also demonstrate in this thesis, offensive language is widely spread in Arabic social media as well. Our hypothesis has been that Arabic may require a specific approach different from the solutions for English due to the specific linguistic characteristics of Arabic text and the unique to Arabic mixture of dialects frequently observed within the same conversation on social media.

The objective of this thesis is to contribute to the work on the automatic prevention of anti-social behaviour in online written communication in Arabic by introducing a large dataset of YouTube comments and proposing a text-mining pipe-line for training a binary classifier. The main challenge to automatic detection of offensive language is the absence of appropriate training datasets. Thus, as part of this work we undertook to collect data
from Arabic social media (Arabic YouTube channels) and construct a labelled dataset. Then we utilised this dataset to experiment with a variety of text pre-processing techniques, feature-selection methods, and classification machine-learning algorithms in order to recommend a process for automatic detection of offensive language in online written communication in Arabic. Our results are encouraging; they suggest Support Vector Machines classifier can be effectively deployed for the detection of offensive language in online written communication in Arabic. We believe that the proposed text-mining process will open the door for further research in this direction and will eventually result in effective automatic prevention of incidents of verbal abuse on Arabic social media.
In memory of my father

To my mother
With love and eternal appreciation

Azalden Alakrot
Acknowledgements

I would like to express my gratitude to everyone who supported me throughout the course of this PhD research. I am thankful for their aspiring guidance, invaluable constructive criticism and friendly advice during the work. First and foremost, I would like to thank The Libyan Ministry of Higher Education and Scientific Research for offering a scholarship to me for completing this work. I also would thank my host, the Department of Computer Science and Information Systems (CSIS) at the University of Limerick for the respectful atmosphere and the excellent working place, where I spent most of my time in the last few years, on my disk behind my computer working on this research project, and where I also met many kind people. I am indebted to my co-supervisors Dr Nikola Nikolov, who has guided me through this academic journey with unlimited support, Dr Liam Murray, who also guided me through this journey and gave me big encouragement to complete this work. I would also like to think Prof. Tiziana Margaria, Head of CSIS, for her support, especially with my conference fees.

I would like to thank Dr Arash Joorabch, for the invaluable discussions and advice that helped me throughout my work. I would also like to thank Dr Farshad Toosi, Eimhear O’Brien, Dr Haiyang Zhang, Dr Abobaker Mohmed, Dr Muftah Fraifer, Yahya Albalawi and Debbie Thompson, Head of University of Limerick Language Centre, for her support when I arrived in Ireland.

Last but not least I am indebted to my beloved Mother who has supported and encouraged me through this entire journey and my whole life. In the memory of my dear father, I pray to God for his pure spirit of mercy and forgiveness. I thank my beloved wife Ebtasm for her support at all times and
my lovely little children.
# Contents

**Declaration of Authorship**

**Acknowledgements**

1. **Introduction to the Thesis**
   1.1 Introduction ........................................... 1
   1.2 Motivation ............................................. 4
   1.3 Research Objectives .................................... 6
   1.4 Overview of the Research Methodology .................. 8
   1.5 Thesis Structure ....................................... 9
   1.6 List of Publications ................................. 10
      1.6.1 Summary of the Listed Publications .......... 11

2. **Social Networking Sites and Offensive Language**
   2.1 Introduction ........................................... 13
   2.2 Demographics of the Web ............................. 15
   2.3 Demographic of Arabs and Arabic Language on the Web ... 16
   2.4 Communities Position on Offensive Language .......... 17
   2.5 Virtual Communities and Cybercrime .................. 20
   2.6 Presence of Offensive Language in Social Networking Sites ... 22
   2.7 Legislation against Offensive Language ............... 24
   2.8 Studies on Offensive Language Detection ............. 24
   2.9 Summary ............................................... 26

3. **Existing Methods for Offensive Language Detection**
   3.1 Introduction .......................................... 27
   3.2 Text Classification .................................. 27
3.2.1 Methods Employed by Existing Abusive Language Detection Tools ............................................. 28
  3.2.1.1 Content-Based Feature Extraction ................................................................. 30
  3.2.1.2 Syntactic Features ......................................................................................... 31
  3.2.1.3 N-gram Features ......................................................................................... 32
  3.2.1.4 User-level Extraction Feature for Offensiveness Detection ................................. 32
  3.2.1.5 Other Features ............................................................................................. 32

3.3 Elements of Insults on the Internet ......................................................................................... 34

3.4 Abusive Language Definition ............................................................................................ 34

3.5 Research Motivation ........................................................................................................ 36

3.6 Study Direction .................................................................................................................. 37

3.7 Proposed Solutions ............................................................................................................ 38

3.8 Summary .......................................................................................................................... 39

4 Arabic Language and Text Mining ......................................................................................... 40

4.1 Introduction ........................................................................................................................ 40

4.2 Arabic Language ................................................................................................................. 41

4.3 Arabic Dialects ................................................................................................................... 43

4.4 Overview of Previous Research in Mining Arabic Dialect Text ........................................ 45

4.5 Arabic Language Features Affecting the Text Mining Process ........................................... 46
  4.5.1 Orthography of the Arabic Language ...................................................................... 46
  4.5.2 Morphology of the Arabic Language ...................................................................... 48
  4.5.3 Arabizi ................................................................................................................... 48

4.6 Text Pre-Processing ............................................................................................................ 49
  4.6.1 Feature Engineering ............................................................................................... 49
  4.6.2 Tokenization .......................................................................................................... 50
  4.6.3 Filtering .................................................................................................................. 50
  4.6.4 Normalisation ......................................................................................................... 52
  4.6.5 Stemming or Lemmatisation .................................................................................. 53
  4.6.6 Document Modelling and Representation .............................................................. 55
    4.6.6.1 The N-gram Model ....................................................................................... 56
4.6.6.2 Vector Space Model ........................................ 58
4.6.6.3 Part of Speech Tags ........................................ 59
4.7 Machine Learning Algorithms (ML) ............................. 60
  4.7.1 Feature Selection ........................................... 60
  4.7.1.1 Singular-Value Decomposition (SVD) ............... 61
  4.7.1.2 Extra-Trees Algorithm .................................. 62
  4.7.2 Naive Bayes ............................................... 62
  4.7.3 Support Vector Machines .................................. 63
  4.7.4 Logistic Regression ....................................... 64
  4.7.5 Regularisation ............................................. 65
  4.7.6 Decision Tree Algorithms ................................. 66
  4.7.7 Random Forests (RF) ..................................... 67
4.8 Evaluation of Text Classification ................................. 68
  4.8.1 K-Fold Cross-Validation .................................. 69
  4.8.2 ROC Curves ............................................... 69
  4.8.3 Boxplots ................................................... 70
4.9 Summary ......................................................... 71

5 Dataset Construction .............................................. 73
  5.1 Introduction .................................................. 73
  5.2 Previous Work in Datasets Utilisation for Offensive Language Detection ........................................ 75
  5.3 Dataset Collection ........................................... 79
  5.4 Sampling ....................................................... 81
  5.5 Descriptive Analysis of the Dataset .......................... 82
    5.5.1 Word Frequency ......................................... 82
    5.5.2 Further Discovery in the Dataset ....................... 83
  5.6 Annotation ...................................................... 83
  5.7 Limitations ..................................................... 88
  5.8 Summary ....................................................... 89

6 Machine Learning Approach to Detection of Offensive Language in Online Communication in Arabic 91
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.1</td>
<td>Introduction</td>
<td>91</td>
</tr>
<tr>
<td>6.2</td>
<td>Pilot ML Experiments</td>
<td>92</td>
</tr>
<tr>
<td>6.2.1</td>
<td>Text Pre-Processing</td>
<td>92</td>
</tr>
<tr>
<td>6.2.1.1</td>
<td>Tokenization</td>
<td>93</td>
</tr>
<tr>
<td>6.2.1.2</td>
<td>Filtering</td>
<td>93</td>
</tr>
<tr>
<td>6.2.1.3</td>
<td>Normalisation</td>
<td>94</td>
</tr>
<tr>
<td>6.2.1.4</td>
<td>Extra Normalisation</td>
<td>95</td>
</tr>
<tr>
<td>6.2.1.5</td>
<td>Experimental Setup</td>
<td>95</td>
</tr>
<tr>
<td>6.2.2</td>
<td>Results and Discussions</td>
<td>96</td>
</tr>
<tr>
<td>6.3</td>
<td>ML Experiments with Additional Syntactic and Linguistic Features</td>
<td>98</td>
</tr>
<tr>
<td>6.3.1</td>
<td>Feature Space</td>
<td>100</td>
</tr>
<tr>
<td>6.3.2</td>
<td>Experimental Setup</td>
<td>101</td>
</tr>
<tr>
<td>6.3.2.1</td>
<td>Initial Experiment</td>
<td>102</td>
</tr>
<tr>
<td>6.3.2.2</td>
<td>Experiment I</td>
<td>104</td>
</tr>
<tr>
<td>6.3.2.3</td>
<td>Experiment II</td>
<td>105</td>
</tr>
<tr>
<td>6.3.2.4</td>
<td>Experiment III</td>
<td>107</td>
</tr>
<tr>
<td>6.4</td>
<td>Discussion</td>
<td>109</td>
</tr>
<tr>
<td>6.5</td>
<td>Summary</td>
<td>112</td>
</tr>
<tr>
<td>7</td>
<td>Conclusion and Recommendation</td>
<td>113</td>
</tr>
<tr>
<td>7.1</td>
<td>Summary</td>
<td>113</td>
</tr>
<tr>
<td>7.2</td>
<td>Contributions</td>
<td>114</td>
</tr>
<tr>
<td>7.3</td>
<td>Future Work</td>
<td>117</td>
</tr>
<tr>
<td>A</td>
<td>List of Available Stemmers for Arabic Language</td>
<td>119</td>
</tr>
<tr>
<td>B</td>
<td>R Code to Compute Inter-Annontator Agreement</td>
<td>121</td>
</tr>
<tr>
<td>C</td>
<td>List of Features Selected by the Employed ML Methods</td>
<td>123</td>
</tr>
<tr>
<td>C.1</td>
<td>LR-L1</td>
<td>123</td>
</tr>
<tr>
<td>C.2</td>
<td>RFE</td>
<td>124</td>
</tr>
<tr>
<td>C.3</td>
<td>Union of LR1 and RFE</td>
<td>125</td>
</tr>
</tbody>
</table>
List of Figures

1.1 Statistics of online harassment in 2014 ........................................ 5
1.2 The proposed process to achieve the outlined goals. .............. 8

2.1 Internet users by region and country, 2016 from ITU Website . 17
2.2 Type of speech considered to be offensive language ............ 18

4.1 Example of MSA from BBC Arabic news. .......................... 42
4.2 Illustration of the result of SVM. ................................. 64
4.3 A binary decision tree ............................................... 66
4.4 Examples of ROC curves ............................................. 70
4.5 Structure of a boxplot. ................................................. 71

5.1 The first phase in the components of the proposed system pro-
cess. .......................................................... 76
5.2 YouTube comment scraper interface. .............................. 80
5.3 Frequency of nationalities mentioned in the first 30,000 words. 83
5.4 Frequency of countries’ names mentioned in the first 30,000 words. ........................................................................................................... 84

6.1 ROC curves for the two classes when an SVM classifier is applied 98
6.2 The final stage in the components of the proposed system, the
text mining process .................................................... 99
6.3 Results of 5-fold cross validation of five models trained with
features selected by LR-L1 ............................................. 104
6.4 Results of 5-fold cross validation of five models trained with
features selected by RFE ............................................ 106
6.5 Results of 5-fold cross validation of five models trained with the combined features by LR-L1 and RFE .......................... 108

C.1 Penn Treebank part-of-speech tags. ................................. 127
# List of Tables

2.1 The number of monthly active users for the top social media platforms by August 2018. .................................................. 15

4.1 Distribution of the number of Arabic speakers of different dialects (Ridout, 2018). ......................................................... 44

4.2 Examples of words separated by commas only. ..................... 51

4.3 Example of stop words in English and Arabic languages. ...... 52

4.4 Some words in English and Arabic languages with their stems. 55

5.1 Examples of instances of offensive language in YouTube comments in Arabic. ................................................................. 80

5.2 Examples of comments made by people from different Arab regions. ................................................................................. 85

5.3 Examples of non-Arabic comments and Arabic comments written in non-Arabic alphabet. ................................................. 86

5.4 Number of agreements and disagreements between annotators in the labelled dataset. ..................................................... 87

5.5 Number of positives in the two scenarios. ................................. 87

5.6 Inter-annotator agreement using kappa statistics. ................. 87

6.1 Examples of words misspelled by interchangeably using phonetically similar letters. ............................................................ 95

6.2 Comparative performance of trained SVM classifiers. ........... 97

6.3 Excluded proper nouns and a country names that frequently occur in the dataset. .......................................................... 100

6.4 Results from training an SVM classifier with a variety of alternative feature-selection methods. ........................................ 103
6.5  Accuracy of the LinearSVC classifier trained with different number of features selected by RFE.  

6.6  Distribution of the features selected by LR-L1 and RFE.  

6.7  Accuracy of the LinearSVC model trained with the combined features selected by LR-L1 and RFE.
List of Abbreviations

(NLP) Natural Language Processing
(ROC curve) Receiver Operating Characteristic curve
(MSA) Modern Standard Arabic
(ITU) International Telecommunication Union
(SVM) Support Vector Machine
(NB) Naive Bayes
(BoW) Bag of Words
(tf-idf) term frequency-inverse document frequency
(PoS) Part of Speech
(VSM) Vector Space Model
(OCR) Optical Character Recognition
(ML) Machine Learning Algorithms
(SVD) Singular Value Decomposition
(RF) Random Forests
(OOB) Out Of Bag
(p) precision
(r) recall
(tp) true positives
(fp) false positives
(fn) false negatives
(F1) F-measure or F1-score
(CV) Cross Validation
(AUC) Area Under the Curve
(PJ) Perverted Justice Foundation
(PAN) Plagiarism Analysis, Authorship IdeNtification
(NLTK)  Natural Language ToolKit
(ARLSTem)  ARabic Light STemmer
(LR)  Logistic Regression
(RFE)  Recursive Feature Elimination
(LR-L1)  Logistic Regression with L1 regularisation
(LinearSVC)  Linear Support Vector Classification
(MultinomialNB)  Multinomial Naive Bayes
(FBM)  Fundacion Barcelona Media
Chapter 1

Introduction to the Thesis

1.1 Introduction

The last decade has witnessed the evolution of the internet in a stunning and rapid manner, allowing the world to become an open book for the human being. There are rich sources of information available on the internet, to the point of flooding. The computer and its applications have become an essential part of the daily lives of those who inhabit modern societies. Computer-based information technology has invaded every facility of life. This technology has altered the various aspects of existence in record time. The internet was born out of the womb of this technology, bringing about an information revolution. Individuals can now access the world of knowledge rapidly; a speed almost as fast as the rate at which new technological devices are invented.

In the present time, friendships and relationships are often formed through a broad range of electronic devices. Most daily greetings, friendly gatherings and even household conversations happen behind a screen today. This dissertation focuses upon and describes the emergence of an electronic community in the virtual world. It discusses the manifold electronic platforms that facilitate the initiation and maintenance of relationships and communication between individuals and society, as with each passing day these interactions become more widely accepted as the level of new standards of
online interaction emerging within this new virtual environment. Despite all of the transformations that point to the emergence of a virtual community, the complexity of human nature has remained as it is; and, as in any real society, good and offence often coalesce. Most of the time, individuals communicate with each other for support and friendship; nonetheless, enmity and hatred have often been components of human behaviour, and they have had a decisive effect on societal history. The virtual community is no exception: offensive incidents, human beings misconducts and examples of malfeasance are present within humans in these virtual societies also. The differences are small and principally linked to the fact that in the virtual situation the perpetrator is enabled with typical features of the virtual world, such as anonymity with regard to the misconduct that affects people in their own homes.

One of the forms of online misconduct that has a major impact on society, with many undesirable consequences, is known as online offensive/abusive language. Traditional abuse is aimed at harming others, humiliating them and degrading their dignity by directing inappropriate and stray words to harm, defame and undermine a persons social standing. Abuse, historically, could only take place in a specific area and at a certain time; however, unlike traditional abuse, online abuse is not confined within any boundaries for both location and time. With the development of online technology, abuse has no bounds and it can be started and spread over the entire world in an instant, having the possibility of remaining online forever. Online abuse can simply be defined as a deliberate action carried out through electronic technology to insult others.

There is a wide range of research topics which focus upon offensive/abusive language on the internet, such as topics related to cyberbullying, hate speech, flaming, defaming and harassment. These studies have mostly been conducted in the English language. However, studies on the same topics from the technical domain in Arabic are rare. Moreover, for almost all of the scarce technical works conducted on offensive language in Arabic, some common
gaps apparent in these studies can be highlighted (Abozinadah, Mbaziira, and Jones, 2015; Mubarak, Darwish, and Magdy, 2017). Firstly, the datasets utilised in these studies are relatively small compared to similar studies in English (Yin et al., 2009; Dadvar, 2014), in conjunction to the fact that they often employ a small number of obscene words as keywords to collect their data. This method of data collection means that researchers can only find posts containing these keywords, with the result that many other sources of content are ignored; resulting in the loss of data which could be useful when it comes to enriching the dataset. Secondly, results presented by these prior studies could be improved in terms of accuracy, precision and recall. For example, precision and recall in (Abozinadah, Mbaziira, and Jones, 2015) is relatively low, and in (Mubarak, Darwish, and Magdy, 2017) the recall is extremely low.

In this study, the overall goal is to contribute to the detection of offensive Arabic language on social media platforms. In order to be able to suggest solutions that can contribute to reducing the risk and impact of offensive language on the internet in the Arabic domain, this research project has investigated many features in Arabic that appear particularly often in online conversation. This study also investigates society’s attitude towards offensive language on the internet and its growth over time, as well as the role of technology in the emergence of this type of virtual behaviour and the possibility of reducing the extent of social anxieties that it raises.

In addition to the aforementioned points, this research also strives to access and examine the applicability of approaches in the scope of information technology, particularly in the area of natural language processing (NLP) in the design of measures and solutions for the automatic detection of instances of online offensive language. In relation to the detection of online offensive incidents, the analysis of textual content published on social networking sites is a challenge. We began this work with an assessment of the applicability of varied choices of NLP approaches harnessed for other tasks, including spam filtering, document classification and clustering tasks. Then we used
methods that rely upon machine learning algorithms and text mining and applied these to content collected from social networking sites to detect offensive cases.

This study contributes to increasing the potential of natural language processing and data-driven methods for successful dissemination in the fight against anti-social behaviours in the virtual environment, especially against offensive language over the internet. This dissertation can also be understood as evidence of how text mining techniques can be used to improve the detection of abuse-related information in user-generated content, with the use of both machine-learning algorithms and data-driven methods.

1.2 Motivation

The emergence of new internet-based technologies has had a significant impact on modern life. One of these important developments is social networking platforms that play a key role in connecting people together around the world. There is a variety of social networking sites/apps, such as Facebook, YouTube, Instagram, Qzone and Twitter, that people use for entertainment and many other possible activities. The arrival of these new technologies often comes with excitement and optimism about the benefits they can bring to human life, in addition to ideas concerning how the standard of living can be improved for the better. However, after a while, some disturbing consequences, expected or unexpected, often become evident. The development of social networks has greatly influenced relationships in the global community. These changes, however, are not uniformly agreeable to entire communities.

One of the problems that make parts of societies uncomfortable with the new changes brought about by these innovations is the increasing online anti-social behaviour. Anti-social conduct has moved from societies to cyberspace and its negative effect has become more painful. Some of the behaviours that have a negative impact on internet users are offensive and abusive language
in online communities, and these anti-social acts are growing and worsening in social networks. A study carried out by the Internet & Technology sector of the Pew Research Centre revealed that about three-quarters of the internet users who participated in their study (73%) had seen or experienced online harassment. Offensive name-calling and embarrassment were the most popular form of harassment. About 92% of the participants agreed that cyberspace allowed individuals to be harsher toward each other than when they are face-to-face (Duggan, 2014). Figure 1.1 shows statistics related to online harassment in 2014, taken from the Pew Research Centre Website. A study by Wang et al. (2014) discovered that swearing words appear at a rate of 0.80% on Twitter and 7.73% of all tweets in their dataset contain swear words. They also found that the top seven swear words accounted for over 90% of all offensive tweets.

![Pie chart showing 60% not experienced harassment and 40% experienced harassment.](image)

**Figure 1.1**: Four-in-ten Internet users are victims of online harassment, varying degrees of severity. Source, American Trends Panel (wave 4). Survey conducted May 30 - June 30, 20014. Number of participants=2,839

Pew Research Centre (Duggan, 2014).

Combined, these facts, figures and reports have raised the query of what appropriate solutions can be provided for this issue, in addition to highlighting questions surrounding what is missing from current strategies to deal with
cyber incidents such as offensive and abusive language. The proposed solution is to develop a detection system that, when integrated, would enable social networks to discover offensive language with a pointed accuracy and then choose a mechanism to deal with it. Such mechanisms could operate by detecting and may prevent certain comments or posts from appearing, or block users who are sending such harmful content; also, it could warn the social network administrators about these incidents.

1.3 Research Objectives

The importance of the problem of Arabic offensive language over the internet and the scarcity of the amount of research in this sphere have given encouragement to the author of this research to seek a means to fill the gap associated with this phenomenon and to contribute to reducing the gap in the knowledge. This can be achieved by introducing new methods and techniques, possibly specific to the Arabic language, or adapting existing ones from related studies in English language communication that can be deployed in detecting and preventing offensive language occurrences in online communication in Arabic.

We present a view on the spread of offensive language over the internet that affects the individual and society. The goal is to clarify the dynamics of communication and relationships introduced with the arrival of the internet in daily life. This study aims to illustrate that virtual environments represent and act as a society in which members exhibit behaviour similar to what can be seen in an actual-life community, arguing that as a result, the interventions and precautions against anti-social behaviour, such as offensive language, have to be similar to the ones known to be effective in physical-life communities.

- **Objective 1**: To build a comprehensive dataset for studies of offensive language on the internet devoted to the Arabic language.
One of the major challenges faced throughout this study is the scarcity of appropriate and available datasets for studying offensive language in Arabic communication, in addition to the lack of electronic tools that could contribute to its inhibition. The required dataset should contain a balanced number of offensive comments and inoffensive comments from a variety of participants on social networking platforms. Moreover, the training dataset needs to be relatively large so that algorithms can learn from it and produce efficient predictive classifiers. The other challenge is the labour effort required to categorise this dataset.

In the context of this goal, the next research question was devised and investigated:

- **Research Question 1:** Is there evidence of the existence of offensive language in Arabic social media platforms?

- **Research Question 2:** What kind of offensive language is present in Arabic social media? Is it Modern Standard Arabic (MSA) or Arab dialects, and what other languages may be found?

- **Objective 2:** To enhance the accuracy of classifiers for the discovery of offensive Arabic comments on social media platforms. In the context of this goal, the next three research questions were devised:

  - **Research Question 3:** What is the impact of removing noisy data and, in general, data pre-processing on the accuracy of detecting offensive incidents on social media platforms?

  - **Research Question 4:** What features selection methods are effective in improving the accuracy of the detection of offensive incidents on social media platforms?

  - **Research Question 5:** Which machine learning algorithms achieve top performance in detecting abusive language in user-generated Arabic text?
Figure 1.2 illustrates an overview of the proposed process to achieve the outlined goals and answering the research questions.

**Figure 1.2:** The proposed process to achieve the outlined goals.

### 1.4 Overview of the Research Methodology

The study methodology utilised in this work is described and summarised in the following points.

**Data gathering:** The first stage is to gather the data, i.e. to build a corpus of online communication in Arabic. This step is achieved through the compilation of offensive comments from various Arab YouTube channels that have subscribers from different Arab countries (see Chapter 5).

**Data Analysis Techniques:** Four essential characteristics have been considered: *availability, representativeness, heterogeneity* and *balance* of the dataset to
be utilised in this study. The researcher examines these four characteristics during the data collection process in order to ascertain the appropriateness of the dataset for offensive language detection in Arabic. As well as this, a statistical analysis of the dataset is conducted regarding its size, the frequency of appearance of some words and the results of the annotation process. This statistical analysis is also used to assist in identifying the computational linguistic techniques needed for the feature selection task.

**Modelling:** The main objective of this stage of the study is to produce a predictive model suitable for the detection and maybe prevention of the offensive language, which include, preparing a profane words dictionary, generating and selecting appropriate features, building the classifier model and using visualisation techniques. Therefore, the components of the proposed model are specified.

**Experiments:** The proposed model is trained and tested on the collected dataset by performing real experiments with a view to showing how it is able to detect offensive/abusive comments.

**Evaluation:** Precision, recall and F-measure are employed to evaluate the efficiency of the proposed model. Additionally, comparisons are conducted with various baselines and previous work. Also, the evaluation of the classifier performance is conducted utilising 10-fold cross validation and a ROC curve.

### 1.5 Thesis Structure

This thesis is organised in seven chapters as follows.

**Chapter 2:** Provides an introduction to the offensive language phenomenon in real-life communities and social networking sites. This includes demographics of internet users, also the position of various communities towards anti-social behaviour and offensive language.
Chapter 3: Provides a review on previous research that includes several types of anti-social behaviour detection, such as cyberbullying, hate speech and offensive language detection. We also highlight the approaches adopted in prior works for feature selection and supervised learning.

Chapter 4: Provides an introduction to the Arabic language and highlights the differences between Modern Standard Arabic (MSA) and Arabic Dialects. Moreover, it contains descriptions of the well-established text mining techniques (feature selection and classification) that are employed in this study.

Chapter 5: Presents details about the data collection and labelling process with a statistical analysis of the dataset. Furthermore, it includes explanation of the reasons behind choosing the data sources, as well as the selection of annotators.

Chapter 6: Presents the implementation of the proposed predictive model, employing the newly gathered data for training it and testing its effectiveness in terms of its ability to detect offensive content. Provides the performance evaluation results, as well as the limitations of the model.

Chapter 7: Summarises this research and offers a number of solid conclusions as well as suggesting future avenues of research in this emerging field.

1.6 List of Publications

This work presented in this thesis is linked to the publications listed below with another one ready for publishing.


1.6.1 Summary of the Listed Publications

Our research has shown that in recent years, many studies target anti-social behaviour such as offensive language and cyberbullying in online communication. Typically, these studies collect data from various reachable sources, the majority of the datasets being in English. However, to the best of our knowledge, there is no dataset collected from the YouTube platform targeting Arabic text and overall there are only a few datasets of Arabic text, collected from other social platforms for the purpose of offensive language detection. Therefore, in the first paper listed above we contribute to this field by presenting a dataset of YouTube comments in Arabic, specifically designed to be used for the detection of offensive language in a machine learning scenario. Our dataset contains a range of offensive language and flaming in the form of YouTube comments. We document the labelling process we have conducted, taking into account the difference in the Arab dialects and the diversity of perception of offensive language throughout the Arab world. Furthermore, statistical analysis of the dataset is presented, in order to make it ready for use as a training dataset for predictive modelling.
In the second paper listed above, we present the results of predictive modelling for the detection of anti-social behaviour in online communication in Arabic, such as comments which contain obscene or offensive words and phrases. We used the dataset introduced in the first paper listed above to train a Support Vector Machine classifier and experimented with combinations of word-level features, N-gram features and a variety of pre-processing techniques. We summarise the pre-processing steps and features that allow training a classifier which is more precise, with 90.05

The two posters and a technical report listed above present the initial results from the study presented in the first two papers.
Chapter 2

Social Networking Sites and Offensive Language

2.1 Introduction

There has since been the goal of working to improve the computer interface for online social interaction. Many social networking sites have been created to allow people to form online communities. These sites focus on encouraging and facilitating interaction between various individuals by bringing them together in discussion boards and chat rooms. In the 1990s, most internet communication involved e-mail and chat rooms (Edosomwan et al., 2011) and the Instant Messaging IM in the late of 1990s (Boneva et al., 2006). Since then, the online social interaction methods have evolved. A large portion of the communication with family members and friends, which in the past occurred face-to-face, has moved into the online world. Social media received a significant boost with the arrival of many social networking sites. MySpace (2003), Facebook (2004), Flickr (2004), YouTube (2005), Twitter (2006) and a broad range of ensuing platforms commenced providing web tools that sparked new online communication interactions (Van Dijck, 2013). There has been an extraordinary increase in social media platforms, as they race to dominate as much as possible in this digital space. Hence, social networking sites have become the largest online communities with billions of users. The popularity of these platforms has quickly risen as they offer various benefits like
announcing activities, sharing news, and having conversations with friends, family members, and even strangers.

The Internet that would come to sustain online sociality and creativity was still mostly undiscovered space, as the frontier between various mediation activities had yet to be defined. It was a new environment, where the rules and laws of the existing world were no longer practical and new laws were not yet in place. The advancement in communication technology has increased awareness of the adverse effects of these technologies (Wellman and Haythornthwaite, 2008). The internet has changed almost all aspects of life including communication, education, economy, politics, and social life. The misuse of social networking sites including the publishing of insults, defamatory statements, slander, theft, and the spreading of racist and other harmful ideas to achieve illegal objectives has become a significant issue. As a result, social networking sites have become a weapon for some. While there are many positive aspects of social media, some of the harmful uses of these tools have become a significant concern. It is critical to determine the way new technology impacts society and how some of the negative consequences of social media can be overcome.

One of the negative consequences of these new technologies includes anti-social behaviour and how it impacts the online community. Anti-social behaviour, including offensive language when engaging in online communication, has become an issue which may be harmful to everyone with internet access. This chapter presents demographics of Arabic language and users on the web, also the communities attitude towards offensive language. It also shows examples of anti-social behaviour on the internet, and finally, a number of studies targeting offensive language online. In particular, this chapter addresses some aspects of the anti-social behaviour online in Arabic.
2.2 Demographics of the Web

The internet is an integral part of life for the majority of people throughout the world and is a crucial source of information including news, communication, entertainment, and social networking. As of 2016, there were approximately 3,385 million internet users (ITU-statistics, 2018), and there has been a substantial growth in the number of users on social networking since. Table 2.1 presents the number of users accessed several social media platforms by August 2018, who use the site at least once a month (Kallas, 2018).

<table>
<thead>
<tr>
<th>The social media platform</th>
<th>The number of users accessed the platform</th>
</tr>
</thead>
<tbody>
<tr>
<td>Facebook</td>
<td>2.23 billion</td>
</tr>
<tr>
<td>YouTube</td>
<td>1.9 billion</td>
</tr>
<tr>
<td>Instagram</td>
<td>1 billion</td>
</tr>
<tr>
<td>Qzone</td>
<td>563 million</td>
</tr>
<tr>
<td>Twitter</td>
<td>366 million</td>
</tr>
<tr>
<td>Reddit</td>
<td>330 million</td>
</tr>
<tr>
<td>Pinterest</td>
<td>200 million</td>
</tr>
<tr>
<td>Ask.fm</td>
<td>160 million</td>
</tr>
<tr>
<td>Tumblr</td>
<td>115 million</td>
</tr>
<tr>
<td>Flickr</td>
<td>112 million</td>
</tr>
<tr>
<td>Google+</td>
<td>111 million</td>
</tr>
</tbody>
</table>

The amount of time consumed using the internet varies based on country and community. Additionally, other factors such as user background, age, education, and economic status influence the amount of time on and the type of use of the internet (Nie and Erbring, 2000).
2.3 Demographic of Arabs and Arabic Language on the Web

In a study by Lazarinis (2009) argues that more than 60% of the online population is comprised of non English speakers. Overall, the number of non-English speakers is growing faster than the number of English speakers. The International Telecommunication Union (ITU) reports that over 3,385 million people use the internet around the world; of this number, 162.1 million are from Arab states (ITU-statistics, 2018). Arabic users are the seventh most abundant group of internet users, and Arabic is the fourth of the top languages used in the web, as well as the fastest growing language on the internet (Miniwatts-Marketing-Group, 2018). Research by Darwish and Magdy (2014) states that in 2000, there were 2.5 million Arab internet users. This number increased to 128.2 million in 2010 and 162.1 million in 2016 (ITU-statistics, 2018). In Arab countries, approximately 54% of the population is online. As of June 30, 2017, there were 141,290,000 Facebook users in Arabic-speaking countries (Miniwatts-Marketing-Group, 2018). These statistics show that internet use in the Arab world is steadily increasing. There is currently no verified information about the amount of Arabic content on the internet, but it is believed that Arabic content accounts for 1.5% of all global content online (Darwish and Magdy, 2014). That is, the relative amount of Arabic content online is most likely disproportionately small compared to English online content. On the other hand, while the quality of English online content is generally deemed to be low, Arabic content is deemed to be of high quality (Darwish and Magdy, 2014). Figure 2.1 shows the distribution of Arab internet users by countries.

Based on these facts, there is a definite need to learn more about the communication habits of Arabic individuals as they transition from face-to-face communication and letter writing to electronic communication including sending emails and communicating on social media.

In the following sections, definitions of offensive language, taboos, and social
Chapter 2. Social Networking Sites and Offensive Language

2.4 Communities Position on Offensive Language

Social interaction involves people using words to convey information. Some words can be rude or awkward if spoken directly and could make others unhappy or uncomfortable. Additionally, certain euphemisms and taboo topics are issues that some people refuse to talk about or are very reluctant to discuss. Some people may use words that are not part of the norm specifically because those words can cause strong negative emotions or be construed as abusive. Various social groups differ from one another in the way they behave and what topics they find offensive or inappropriate. Respect for the use of language is the key to creating true mutual communication (Gao, 2013). Taboo is defined as "a social or religious custom prohibiting or restricting a particular practice or forbidding association with a particular person,"
Fairman (2009) states that all cultures have behaviours and words that are considered taboo. Andersson and Trudgill (1990) write that taboo differs based on culture. For example, in northern Australia, certain tribes are not allowed to say a dead man’s name as that action is considered a taboo (Andersson and Trudgill, 1990). In our study, only the taboo types shown in Figure 2.2 are considered. Offensive language includes cursing, swearing, insulting, profanity, obscenity, rudeness, impoliteness, or other forms of bad language.

![Offensive Language Diagram](image)

**Figure 2.2:** Type of speech considered to be offensive language.

A study carried out by the Pew Research Centre Internet & Technology revealed that approximately 92% of all online users believe that the internet allows individuals to use harsh, offensive language more freely and easily than the real, offline world. The research also found that social media platforms are the most common areas where abusive language is found. Additionally, 66% of all online users report they have been harassed on a social
media platform, 22% have been harassed in the comment section of a website, and online gaming, email, online discussion or dating sites each account for 16% of online harassment experience/app (Duggan, 2014).

The need to address the issue with offensive language in online communication has become an issue addressed governmental authorities. For example, Senator Jeff Flake addressed offensive language in social media and the media in general when he said “We have to model behaviour that we would be proud that our kids are watching” (Rishab Nithyanand and Gill, 2017). For most, society does not accept obscene words and alienates this type of behaviour since they can cause social and psychological harm. Obscenity also gives a negative image to any community (Soliman, 2017; Fatima Zaid Al Zayed, 2016), we define the community in Section 2.5. While this type of reasoning is logical, even some individuals resort to obscenity seemingly out of habit when they are angry, frustrated, or engaged in conflict. When others see this type of behaviour, they may change the way that they view the individual using obscenity. Some believe that just not responding to certain attacks may be the best response, but it is very difficult not to respond when others are attacking. It is often difficult for people to retain their typical personality. Regardless of the reason for using obscene language, society does not accept the use of these words (Soliman, 2017).

General vision of society: Communities have a general view of what is and is not acceptable. A society that allows obscenity in its daily interactions will often be considered immoral and perverse. A society that uses respectable words is generally thought of as more civilised, elegant, and advanced. When there are respectful interactions, there will be fewer problems and misunderstandings, and society would benefit as a whole (Soliman, 2017).

Use of obscene words and rudeness: People often do not sympathise with rude individuals and are often afraid of these people. When a person continually uses abusive words, that individual is considered rude and as someone who does not care about anyone. This type of conduct is not suitable for the
majority of society. While an increasing number of societies are generally considered free and open, there are still some limits in place, and respect is something that should be shown for most individuals. The ability to control what is being said is a skill that allows people to manage their interactions with others responsibly (Soliman, 2017).

**The frivolous person:** A frivolous person is someone who often uses abusive words when angry. Many people see obscene words as words that are indicative of a frivolous person. The majority of successful people are those who are able to communicate effectively without the use of obscenity. The use of obscene words is something that can be harmful to the reputation of the speaker and could lead to the individual being alienated (Soliman, 2017; Fatima Zaid Al Zayed, 2016).

In summary, obscene words do not help and actually harm. For this reason, society does not accept this type of speaking and would often prefer that people not speak instead of speak using obscenity (Soliman, 2017). Offensive language is unproductive and unacceptable behaviour in a healthy community. In online communities, the practice of using obscenity must be curtailed.

The following sections illustrate how social networks have evolved and created a virtual society that engages in online communication and, eventually, a virtual community. In addition, we review the extent to which offensive language exists in the virtual world and the way that the government legislates offensive language in Section 2.7.

### 2.5 Virtual Communities and Cybercrime

Virtual communication occurs in an area often referred to as cyberspace (Porter, 2013). In cyberspace, digital interactions occur in a way that is not confined by traditional borders like time and place. Instead, digital interactions form the basis of the “virtual society” (Porter, 2013). Rheingold (1993) defined
the virtual community as social aggregations that emerge from the Net when enough people carry on discussions long enough, with sufficient human feeling, to form webs of personal relationships in cyberspace (Rheingold, 1993).

Virtual communities involve interactions based on shared interests, problems, and concerns. Personal behaviours are often altered to suit the particular circumstances present in a virtual community. There are also some indicators of respect, love, and good treatment while there are other indicators that illustrate harassment, violence, and hostility are occurring. These indicators are expressed through written words that help to elicit certain feelings and reactions.

Like in the real world, in a virtual community, certain crimes, threats, and bad behaviour can occur. As a result, cybersecurity is needed to protect against cybercrime (Von Solms and Van Niekerk, 2013). The International Telecommunication Union (ITU-statistics, 2018) defines cybersecurity in ITU-T X.1205 as the collection of tools, policies, security concepts, security safeguards, guidelines, risk management approaches, actions, training, best practices, assurance and technologies that can be used to protect the cyber environment and organisation and users assets. Many countries consider offensive language used in public to be a crime as an example for that electronic Irish Statute Book (electronic Irish Statute Book). Like conventional crime, cybercrime involves a wide range of scenarios.

The development of technology has led to changes in the type and scope of cybercrime. The definition of cybercrime varies based on the people involved (victim, offender and eyewitness), as well as with the development of an electronic environment where these crimes take place (Gordon and Ford, 2006). There are various types of crimes in the electronic environment, and each type needs to be tackled differently from the others. Our study considers the type of electronic crime related to anti-social behaviour and offensive language on social media platforms. Timothy Jay states that cursing is a rich emotional, psychological and sociocultural phenomenon (Jay, 2009b). His
statement has encouraged researchers in linguistics, psychology, and sociology to further study this phenomenon (McEnery, 2004; Jay, 2009a; Jay and Janschewitz, 2008; Christie, 2013). Our study approaches the issue of offensive language in online communication from a data analytics perspective.

2.6 Presence of Offensive Language in Social Networking Sites

Since online communication using email and chat rooms was introduced, social networking and communication began to change. Local social interaction moved to the broader, virtual world through internet-based communication.

There have been discussions about the effects, benefits, and implications of the internet. Some studies defined the term flame as hostile intentions characterised by words of profanity, obscenity, curse and insults resulting from reckless behaviour and which hurt a person or a group of people (Alonzo and Aiken, 2004). Some studies suggest that flaming is a social or cultural tendency (Kayany, 1998). Other studies suggest that flaming depends on the topic being debated, the confidence that individuals have in their anonymity, and the proximity and familiarity of individuals with group members (George et al., 1990). Unlike in-person communication, electronic communication does not require eye contact or any physical presence which would dictate a certain social etiquette. Those who are geographically separated can attack one another using the virtual world without any fear of physical harm or confrontation (Chapman, 1995).

With the improvement of the communication technologies and the greater popularity of social media, the general belief about the negative impact of these platforms remains unchanged. Many recent studies discuss the ubiquity of offensive language on social networking sites. Omernick and Sood (2013) find that the use of written obscenity increases with higher levels of
anonymity, while higher levels of identification lead to less swearing in the comments. In their study on the use of profanity on Yahoo! Buzz platforms, Sood, Antin, and Churchill (2012a) report that different communities such as political or sports forums use profanity at the same frequency, but in different ways. Wang et al. (2014) study the issue of cursing on Twitter by investigating the frequency of swearing and the preferred obscene words being used. The authors find that obscenity occurs at a rate of 0.80% on Twitter and that 7.73% of all tweets contain obscene words. They also find that seven particular curse words account for over 90% of all offensive tweets. On the other hand, swearing is also used to express feelings of anger, joy, sadness, love, and thankfulness. This study shows that swearing on Twitter is closely linked to both sorrow and anger.

Researchers report that swearing occurs far more in the digital world, especially on Twitter than in the physical world. Social networking sites allow people not to see one another and make people feel comfortable enough to say things that they would not otherwise say in the physical world.

Social media communication is characterised by certain distinctive features that distinguish it from direct, real-world interaction. Messages on social networks are often public and can spread quickly and widely due to the interconnected system. In contrast, offline discourse often remains between the individuals involved in the conversation directly (Wang et al., 2014). An example of the rapid, significant impact of social media is the role that it played in the Arab uprisings (Adi, 2014) that began in a small city and spread rapidly to other Arab countries through social media, in particular, via Facebook and Twitter.

In the Arabic context, some experts, clerics, and lawyers found that social networking sites have become a platform for defamation and insult. These individuals have emphasised that some of the actions are crimes that are punishable by law (Fatima Zaid Al Zayed, 2016). Lieutenant-Colonel Hayat Abdul Majeed, head of the cyber-crime department in Bahrain, said that through the first third of 2016, there were 81 cases of defamation, cursing, and insulting
that occurred on social networking sites (Fatima Zaid Al Zayed, 2016). These above cases help to illustrate the widespread prevalence of cursing and insulting that happens on social networking sites and how it has evolved into an unhealthy phenomenon that needs to be addressed.

2.7 Legislation against Offensive Language

Zeviar-Geese (1997) reports that California was the first state to pass a stalking law. Arab countries like Egypt, Saudi Arabia, Qatar, and the UAE have all enacted legislation against those who use offensive language. The legislation in the UAE, in particular, goes further than in the other countries and imposes stronger penalties for offensive language used on social media platforms. This particular law addresses information technology crimes and says that the punishment for swearing, cursing, and spreading offensive content is to be a fine of not less than AED 250,000 and not exceeding AED 500,000 (Hosani, 2018; Alaitihad, 2014). This legislation stresses the punishment for committing this type of crime on social networking sites since insulting or abusing others on the internet can quickly spread throughout the world in a matter of seconds. Other Arab countries like Kuwait and Bahrain have also considered making laws regarding internet crimes. Ultimately, laws that are specially made to address cybercrime do not matter, provided that laws that govern traditional abuse and insults exist and can be applied in the case of cybercrime.

2.8 Studies on Offensive Language Detection

With the increasing number of reports on the alarming consequences of the spread of the offensive language in cyberspace and especially the danger for young people, many text mining studies have been devoted to the detection
of offensive language in its variations, such as hate speech and cyberbullying. Studies targeted abusive and offensive language (Chen et al., 2012; Nobata et al., 2016; Xiang et al., 2012), harassment (Yin et al., 2009; Aggarwal and Zhai, 2012), and cyberbullying (Reynolds, Kontostathis, and Edwards, 2011; Dadvar et al., 2013; Al-garadi, Varathan, and Ravana, 2016). Only a few studies were found for offensive language in Arabic, one of which makes two datasets available, a dataset of 1,100 manually labelled tweets as well as a dataset of 32K user comments from a popular Arabic news site, both containing data entries deemed to be inappropriate language (Mubarak, Darwish, and Magdy, 2017). Another study applies manual labelling of 500 Twitter accounts, with half of these 500 accounts labelled as abusive (Abozinadah, Mbaziira, and Jones, 2015).

A study by Lieberman, Dinakar, and Jones (2011) suggests that most cyberbullying occurs during the discussion of topics of race, ethnicity, sexual identity, physical appearance, intelligence, social acceptance, and rejection. The authors identify specific approaches that can identify if a comment relates to these topics and if the comment is positive or negative in terms of cyberbullying. Another study by Dinakar, Reichart, and Lieberman (2011) focuses on detecting cyberbullying in YouTube comments. The messages were classified based on sensitive/negative topics and profanity. The conducted experiments show that the use of a binary classifier outperforms a multi-classifier for detecting sensitive messages (Dinakar, Reichart, and Lieberman, 2011).

Based on these studies, the problem of offensive language detection can be looked at it as a classification problem. The next chapter presents a range of studies employing text mining methods and document classification to tackle this issue.
2.9 Summary

The growth of the internet has allowed it to reach a broad and diverse audience as it allows people to access tools that make their life easier. For example, people are now able to communicate more easily and can also use the internet for entertainment, news, banking, shopping, and research, among other pursuits. The advantages of this are apparent, but they also come with some drawbacks. In the broader online community, many engage in poor behaviour and commit various types of crime. Cybercrime has become widespread and includes theft, fraud, hacking, virus dissemination, phishing, drug trading, paedophilia, cyberbullying, hate speech, and offensive language.

This chapter presents facts about the increasing number of users online and in the Arab world, mainly. The community and government positions on offensive language are also discussed along with applicable legislation. There is also an overview of some of the existing studies in this sphere and the methods used by prior studies.
Chapter 3

Existing Methods for Offensive Language Detection

3.1 Introduction

This chapter provides an introduction to the text mining process, which includes document classification, text processing, and features extraction; and how these techniques have been employed in various previous studies to detect anti-social behaviour. There has been much research carried out to determine ways to detect abusive language written in English. At the current time of writing this thesis, there are very few studies about the detection of offensive language written in Arabic. This study examines the use of offensive language posted in Arabic on social networking sites. It also presents abusive/offensive language definition and the direction of this study along with the proposed solutions. In particular, this chapter addresses some aspects of the anti-social behaviour online in Arabic.

3.2 Text Classification

Aggarwal and Zhai (2012) pointed out that the problem of text classification finds applications in a broad range of fields in text mining. Some examples of the text classification include news filtering and organisation, document
organisation and retrieval, opinion mining, email classification and spam filtering. The email classification and spam filtering are to classify emails to identify whether the email is a valid or junk in an automated manner. A broad range of techniques for text classification has been designed to treat these problems. The most commonly used methods for text classification including Support Vector Machine (SVM), Naive Bayes (NB), decision tree, random forests, logistic regression and nearest neighbour. Offensive language may be treated similarly to spam filtering; it is possible to use many methods used for spam filtering for offensive language detection.

Aggarwal and Zhai (2012) also stated that feature selection is an essential problem for document classification. In feature selection, the attempt is being made to identify the most relevant features for the classification. This is important due to some words being much more possible to be correlated to the interested class than others. Thus, a broad range of features has been utilised in the literature. We will review the most important ones and the study that uses it.

3.2.1 Methods Employed by Existing Abusive Language Detection Tools

This section reviews feature extraction methods as well as offensive language detection research. Several studies have found that the primary point of supervised learning is the selection of useful features that can strongly impact the machine learning process (Cavnar and Trenkle, 1994; Sanchez and Kumar, 2011; Chen, Mckeever, and Delany, 2017).

- **Filtering Methods for Offensive Language in Social Media:** Popular online social networking sites use specific techniques to prevent the appearance of offensive language. *Facebook* allows users to create a blacklist and comments that contain the words in the blacklist are treated as spam (Vairagade and Fadnavis, 2016). *YouTube* uses a safe mode that hides comments which contain offensive language. All comments
that feature the offensive language prior to turning on the safe mode will continue to appear. Twitter provides a policy that allows users to block accounts that post offensive content (Waseem and Hovy, 2016; Mubarak, Darwish, and Magdy, 2017). YouTube also uses safe mode to hide videos that may contain inappropriate content as identified by user flags. Most popular social media sites use a dictionary-based filter that finds abusive language (Waseem and Hovy, 2016). The dictionary is either predefined such as the one used by YouTube or it could be user-generated such as in the filter used by Facebook. Most sites rely on user reports to find the abusive language and react accordingly. The use of an automatic filter to stop profane words and sentences can make the accuracy of the filter low and could result in many false positive alerts for content that is not abusive (Waseem and Hovy, 2016).

- **Text Mining Methods for Detecting Offensive Language in Social Media:** Abusive language identification in social media is a difficult task because it is working to detect language in an unstructured format that uses informal speech patterns that could have grammar and spelling errors. While the methods currently adopted by social media sites are generally considered inadequate, many studies have proposed methods to detect abusive comments using text mining and natural language processing (NLP) methods (Waseem and Hovy, 2016). A study by Chen et al. (2012) points out that text mining methods are implemented to analyse data requires for data acquisition and pre-process, feature extraction and classification.

The majority of these studies successfully use content-based features, sometimes accompanied by other features. Chen et al. (2012) propose a combination of lexical and syntactic features to detect offensive language in YouTube comments to protect users from receiving potentiality offensive comments. To determine if a post is offensive, the authors used features including style, structure, and content-specific features. These features helped to compile specific comment statistics including the length of the comment in tokens,
average word length, number of insults, and blacklisted words. The authors also considered the writing style of commenters through an analysis of certain words and what part of speech these words were (nouns, adverbs, verbs, or adjectives). A recent study conducted by Nobata et al. (2016) applied these methods except they also tracked spelling and grammar factors. These markers were considered a sign that could help to detect abuse.

Nobata et al. (2016) argued that offensive language often features proper grammar. The authors felt that there are many examples of offensive language being noisy (i.e. great deal of unnecessary data that can hardly make sense). The study found that most research focuses on finding abusive language in English, but that many other languages have not been studied and no conclusive evidence suggests that tools used to determine the abusive language in English can or cannot be used to detect similar abusive language in different languages.

The following sections provide important details about the range of features used in prior studies that analyse offensive language in online interactions.

### 3.2.1.1 Content-Based Feature Extraction

Many NLP studies focus on content-level features. Lexical features have been used by many research studies in abusive language identification. Lexical features such as particular words or phrases are often employed in textual classification tasks (Calvo, Lee, and Li, 2004). This type of analysis features the extraction of each word as an independent feature, a process that is sometimes called word level feature. The technique involves words being treated with the number of their appearance, regardless of location in the sentence in a bag of words (BoW) fashion (Chen et al., 2012; Vandersmissen, 2012; Abozinadah, Mbaziira, and Jones, 2015; Chen, Mckeever, and Delany, 2017). In their harassment detection study Yin et al. (2009) shows variety of attributes including local features, sentiment features, and context features. The authors report that they had obtained significant improvement in the accuracy
of their classifier by using term frequency-inverse document frequency feature (tf-idf) (Baeza-Yates and Ribeiro-Neto, 1999). On the other hand, a study by Waseem and Hovy (2016) reports that the use of the Bag of Words (BoW) method alone produces low accuracy and high false positives rate in abusive language detection.

Many tools to detect abusive language use a lexicon, dictionary, and blacklist. A study by Chen et al. (2012) used multiple features, including lexical features, to construct a dictionary of abusive words. The words are selected based on their “strength” and the way that they are used Xu and Zhu (2010). The same method is utilised in the studies conducted by Sood, Antin, and Churchill (2012b), Xiang et al. (2012), and Nobata et al. (2016). Yin et al. (2009) employ a method that tracks the occurrence of certain profane words along with second- and third-person pronouns.

A recent study by Nobata et al. (2016) uses linguistic features and other features presented in the following sections. The linguistic features include the number of words in the comments, average word length, punctuation in the comment, number of one-letter tokens, number of capital letters, number of URLs, number of kind words, the number of insults modal, and profane words.

### 3.2.1.2 Syntactic Features

The study by Nobata et al. (2016) uses the linguistic features described in the previous section as well as syntactic features, distributional semantic features, and N-gram features. The syntactic features are based on the part of speech (PoS) tags and parsing dependency, a term which identifies dependency-based words derived from analyses or dependency trees. This technique is used in various research studies (Liu, 2012).
Chapter 3. Existing Methods for Offensive Language Detection

3.2.1.3 N-gram Features

The N-gram feature approach is described in Section 4.6.6.1 has been used by many studies. Nobata et al. (2016) use N-gram features along with other linguistic and syntactic features. The authors examined the efficacy of character N-grams features to detect offensive language in user online comments. They report that the two N-gram features were powerful in English; they suggest applying them in other languages with enough training data that would perhaps give good results too.

A study by Warner and Hirschberg (2012) focuses on hate speech and did not put as much emphasis on finding the offensive language. They use a support vector machine classifier, trained with range of features including unigram and the part-of-speech trigram. Yin et al. (2009) utilise a combination of features including word N-grams (N=1, 2, and 3) for profane words with what they called contextual features to improve results.

3.2.1.4 User-level Extraction Feature for Offensiveness Detection

In the study of cyberbullying detection in YouTube comments Dadvar (2014) conducts experiments with a range of features such as content, gender, and age features. The results show that there was a slight improvement in performance that could be achieved through the inclusion of features that represent previous user behaviour. Also, the author found out that the incorporation of gender information can improve the detection of cyberbullying incidents. On the other hand, Dadvar (2014) points out that the age information is often incorrect; therefore, the study suggests the use of age prediction algorithms.

3.2.1.5 Other Features

Contextual and distributional semantics features are also used to detect offensive language. In the study of detection of harassment on web Yin et al. (2009) found that most online comments are not made to harass others and
that most comments made with the intent of harassment look different than other comments. Based on these findings, the authors came up with what they called contextual features. The authors point out that some people when they discuss topics that they have a strong opinion about it tend to use words that make it resemble harassment conversation. Likewise, some casual amiable conversations between some group of people it seems like harassing. The authors called these kinds of conversation as harassment-like. The contextual features distinguish between harassment comment and harassment-like comment. For harassment-like comment, they suggest that the appearance of personal pronouns and the abusive language together, not always a sign of harassment (Yin et al., 2009).

Nobata et al. (2016) utilise distributional semantics features for abusive language detection in cyberspace. Distributional semantic feature derives estimates of semantic similarities between words from large text corpora. The idea behind this comes from the hypothesis that semantically similar words tend to show in similar contexts (Bruni, Tran, and Baroni, 2011). This technique has been used successfully in NLP applications (Mikolov et al., 2013; Le and Mikolov, 2014; Djuric et al., 2015b).

Another work by Djuric et al. (2015a) uses the paragraph2vec approach, proposed by Le and Mikolov (2014), to classify user comments as abusive or not abusive. This approach reportedly achieves better accuracy than the BoW approach. Nobata et al. (2016) examine multiple features, including syntactic and embedding features. The authors found that these features were effective when combined with NLP features and also determined that the N-gram character tool alone works well in a noisy corpus.

Mehdad and Tetreault (2016) employ light-weight features, instead of deeper linguistic features such as PoS tags. Light-weight features are computationally much less expensive than syntactic or discourse features. The authors found it very challenging to pre-process noisy text to extract more in-depth linguistic features. Online user-generated text can be considered noisy text.
3.3 Elements of Insults on the Internet

Online insults consist of several elements that impact how the insults take place and how they are formed. The elements being studied must be clarified and selected to ensure that the differences are considered and that the proposed approaches match the nature of each element.

The essential element is the people engaged in the incident, which can be classified as people who use profanity, obscenity, and insults, people who are the recipients of the insults, and the people who are the witnesses to the incidents of insult. The witnesses are people who are not involved in the incident and are only viewing it (Dadvar, 2014).

The platform used to send the insults is another key element in the process. Most online communication occurs on social media platforms that allow users to see other users profiles, which can be a mean that provides an opportunity for the predators to attack people. Platforms typically provide users with the ability to interact with each other using email or some form of messaging. Additionally, a platform often allows photograph and video sharing, some form of reaction to posts, such as rating or likes, as well as the ability to follow posts made by others (Dadvar, 2014).

Another element involves the content and the form of the insult. There are several ways for users to interact with one another and the offensive content can be posted using various methods including a video, image, or text (Dadvar, 2014). The goal of this study is to target text that uses profanity. The following section defines abusive and offensive language.

3.4 Abusive Language Definition

For the purpose of developing a machine learning model that can assist in detecting abusive/offensive language, there is a need for a definition that can help annotators to put their labels in the labelling process. This process
is needed to develop a labelled dataset that can be used for machine learning task. Yin et al. (2009) define online harassment as a kind of action in which a user intentionally annoys one or more other users in a web community. Berry and Kogan (2010) state that the task of detecting misbehaviour involves detecting inappropriate activity that some users in a virtual community find to be offensive or rude. Warner and Hirschberg (2012) point out that a definition of hate speech is needed to annotate their corpus properly. Warner and Hirschberg (2012) define hate speech as speech that humiliates, assaults, attacks or diminishes a person or group of people with mutual characteristics such as race, sex, religion or disability. There are also some other types of online harassment. For example, Yin et al. (2009) state that user attempts to establish an online link to users that do not want to have a connection with the user can be considered harassment.

The process of online offensive language is done by sending offensive, inappropriate, and disrespectful content across social networking sites to one person or group of people or spreading obscene language and profanity in social networks.

Offensive language has been defined by several previous studies such as (Yin et al., 2009; Berry and Kogan, 2010; Razavi et al., 2010). The definition of offensive language adopted in this study is inspired by prior research includes the disrespectful phrases include language that makes a person feel insulted, intimidated, or embarrassed. Often, these types of phrases contain references to sexual issues, mental or physical disabilities. This type of language is used for the degradation of others and could include calling them by other names, including the names of animals. Discrimination and racism are directed at individuals from a particular ethnic group, nationalities, religions, gender, disability, class, ideologies, and job titles. Terms and phrases, referring to these topics, will often provoke anger.
3.5 Research Motivation

Several factors motivate the need for designing a dataset to detect offensive language in Arabic. Since the amount of online communication in Arabic is significant (see Section 2.3), the problem cannot be solved with manual moderation. Thus there is a need for automatic detection.

Second, not much attention has been paid in addressing offensive language detection for the case of Arabic language, as the sets of labelled training data utilised in the research on this topic are not large (see Section 2.8). Therefore, a classifier can hardly be built with confidence. Thus, we have been motivated to build a corpus and sizeable labelled dataset as part of this study (details Section 5.3). There are many incidents of offensive language occurring in private environments on the internet where access is restricted, such as Facebook for example, which otherwise would be a good source for such data. Therefore, we took the opportunity to collect a lot of such comments, from a source with no restrictions, and build a dataset suitable for training predictive models.

Arabic is still not thoroughly studied regarding text mining, and many aspects have not been examined yet (Darwish and Magdy, 2014). Previous studies, which deal with Arabic, are predominantly concentrated on the Arabic language called Modern Standard Arabic (MSA), and little attention has been paid to Arabic dialects so far (for details see Chapter 4). However, the Arabic text generated in social networking sites is mostly composed of Arabic dialects. In this regard, we are interested in detecting online user-generated offensive content in Arabic. To achieve this, two apparent hurdles can be seen. The first one is the co-occurrence of many different dialects in the content; the second is the casual writing style with no strict rules, which is the predominant writing style in social media. In this work, we explore what can help to improve a predictive classifier for detecting text that contains offensive content in casual-style Arabic text, possibly a mixture of multiple Arabic dialects.
3.6 Study Direction

Despite the impressive number of studies that deal with online abuse in its many variations, and although these studies are continuing to develop multiple methods to deal with a variety of demands, the solutions are generally for English. With the increasing number of Arab users of the internet and social networking sites (see details in Section 2.8), there is a paucity of studies on addressing the problems mentioned in the Arabic context. This scarcity of research makes dealing with the subject of offensive language detection in Arabic an urgent necessity. For the purpose of finding solutions that work for text in Arabic, there are a number of points that deserve attention. We discuss them in the remainder of this section.

Arabic text has many characteristics different from English; these characteristics need methods explicitly designated to deal with them. In Chapter 4, we review the Arabic text characteristics and the existing methods to deal with them. Another important point is that Arabic text on social networking sites is typically written in an informal manner with the majority of users employing their local dialects. In section 5.5.2 Table 5.3 provides examples of aspects of the different types of complexities that appear in the comments of Arab commentators on social networking platforms.

A study by Alruily (2012) examines the use of text mining to identify crime patterns in an Arab crime report corpus. The study is conducted on Arabic text obtained from newspapers to determine patterns of crime. This text is entirely different from the texts used in user comments on social networking platforms in terms of quality of writing, expressions used, and the type of language used. The text used in the newspapers is grammatically correct, there are mostly no spelling mistakes, and the writing follows the linguistic parsing rules. The opposite is true in a text written on social media platforms where few linguistic rules are followed correctly. Another study by Elyezjy and Elhaless (2015) investigates crimes using text mining and network analysis of the Arabic text. The study utilises 777 real investigation documents
about theft crimes from a police department as a source of their dataset. Both studies target the identification of crime patterns which is different from the goal of our study.

Recent studies by Abozinadah, Mbaziira, and Jones (2015) and Abozinadah and Jones Jr (2017) approach abusive Arabic language detection on Twitter using text mining and statistical learning methods, respectively. They collected a dataset in which the positives are tweets containing five Arabic swear words. Another study by Mubarak, Darwish, and Magdy (2017) conducts its experiments on a dataset collected from Twitter using some beginnings of abusive phrases that are usually used in offensive languages, such as: ٍٍٍٍٍ، ٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍٍ‌
our proposed approach is the integration of human cognition and sensitivity to individuals behaviours in society with algorithms and technical solutions. We mainly focus on two tasks. The first is the automatic detection of offensive language in social networks using text mining methods along with machine learning algorithms (details Chapter 4). The second task is to utilise human cognition for determining whether content generated by users online is obscene (details Chapter 5). The approach proposed in this thesis provides a new element in the scope of preventive measures in the fight against anti-social behaviour; in particular, the offensive language on the internet.

3.8 Summary

Many studies have been conducted to tackle the speared of offensive language in cyberspace, but limited amount of research has been carried out for the field of detecting offensive language online in Arabic.

Based on previous studies, the most apparent issues in this regard are as follows:

- The detection of offensive language in the Arabic domain has not been studied adequately.
- There are only small datasets in Arabic used to detect offensive language. Machine learning systems must use a large annotated corpus to achieve proper information extraction.

In particular, we review the variety of features employed in previous text mining studies highlighting the most appropriate features that could be used in our study. We also give a definition of abusive/offensive language. Finally, the objectives and direction of this study are introduced along with an outline of the proposed solutions.
Chapter 4

Arabic Language and Text Mining

4.1 Introduction

At present, the quantity of both structured and unstructured digital data increases quickly with the unstructured data, such as websites, news reports, emails and online forums, being estimated as 85-90% of it (Weiser, Biros, and Mosier, 2006; Alruily, 2012). Accordingly, the development of effective text mining methods for extracting useful information from unstructured data has become an increasingly important area of research. The vast majority of text mining methods developed in recent years are for the English language, due to the fact that most of the digital text data is in English. Nonetheless, the current growth of non-English digital textual data makes it more urgent than ever to adapt existing text mining methods and develop new tools for other languages, such as Arabic.

The main steps in the text mining process are text pre-processing, text mining operations, evaluation and information discovery (Zhang, Chen, and Liu, 2015). The pre-processing step typically involves data collection, data cleaning, and generally is the transformation of text data into forms ready for text mining operations, such as document classification, document clustering and discovery of association rules. These text mining operations are typically executed

1https://w3techs.com/technologies/overview/content_language/all
in the form of machine learning, the output of which can be viewed as a structural description of the text data, generally referred to as a model. Models can be seen as short summaries of the data which can be used for both better understanding of the data by humans and automatic predictions (Zhang, Chen, and Liu, 2015). Finally, text mining operations are followed by evaluation and a final interpretation of the model which may also include information visualization (Zhang, Chen, and Liu, 2015).

This chapter presents an introduction to the Arabic language and text mining methods that can be applied to Arabic text. In this regard, the main research focus is on text mining of the official Arabic language which is known as Modern Standard Arabic (MSA). The research results on mining Arabic dialects are still very slim, despite the fact that these dialects are the main-tongues across the Arab homelands and are widespread in casual online writing (Zaidan and Callison-Burch, 2014). In general, effort on various aspects of mining Arabic text is still lacking and is very little compared to other languages. We review the existing methods for processing Arabic text (MSA), and the ability to apply these methods to text containing a mixture of colloquial Arabic dialects. These include pre-processing methods, feature selection, machine learning algorithms and evaluation methods. The difference between MSA and colloquial Arabic dialects is also explained in this chapter. We review a few text-mining studies that do consider Arabic dialects in their work.

### 4.2 Arabic Language

Arabic is spoken by over 250 million people and is the official language in more than 20 countries in the Arab world spread from North Africa to the Arabian Peninsula (Zaidan and Callison-Burch, 2014). Arabic is the fourth most used language on the web². The Arab world is characterised by a high degree of linguistic and cultural continuity. Nevertheless, the contemporary

Arabic language has a degree of diversity; in this section, the focus is on highlighting the two most important types of spoken Arabic.

The first type, Modern Standard Arabic (MSA), is the language of all formal printed texts such as books, formal media, including newspapers and magazines. It is also the language used in formal spoken communication, e.g. public speaking and broadcasting on radio and television. That is, in the Arab world one needs to be able to understand both written and spoken forms of MSA (Ryding, 2005). Figure 4.1 illustrates a script written in MSA, which is a report from BBC Arabic News.

The other type of Arabic is the vernacular one used in informal spoken and written conversations, and for it there is no agreed standard of utterance. Native Arabic language speakers are fluent in at least one colloquial form of Arabic (mother tongue), and they understand a broad range of others (Ryding, 2005). There are several major Arabic dialects such as Gulf, Egyptian,

---

3http://www.bbc.com/arabic
Levant, Iraqi and Maghrebi, and each of these regional dialects or slangs co-exist with MSA (Ryding, 2005). A dialect is more flexible and changeable than the formal language; it easily accepts foreign expressions as well as expressions for the latest cultural concepts and trends. Arabs use colloquial language to develop forms of spontaneous linguistic arts, such as popular songs, jokes and folk tales. Moreover, colloquial Arabic is generally used in blogs, forums as well as on various social networking platforms (Zaidan and Callison-Burch, 2014). Arabic dialects may differ greatly from one another owing to the geographic distance between them. The colloquial dialects of adjacent areas such as Egypt and Sudan are easily understandable by people who inhabit these lands; however, regional dialects of separated regions, such as Moroccan and Kuwaiti, show large differences, which may require people from different regions to adjust their daily language to a more widespread level when communicating with one another (Ryding, 2005).

The pioneer of Arabic Language studies Anis Ibrahim (1992) was an expert and member of the Arabic Language Academy Cairo states that there are two main factors attributed to the formation of dialects in the world. The first one is environmental isolation between people who speak the same language, and the second one is language conflict as a result of invasion and migration. Both factors have played role in the formation of Arabic dialects. In the Arab world, areas with dense population are separated by sea and desert, and both displacement and movement of population have occurred throughout history. Moreover, languages spoken by neighbours of the Arab world may have influenced the formation of dialects as well (Ibrahim, 1992).

### 4.3 Arabic Dialects

With the expansion of online social networking sites in the Arab world, various Arabic dialects began to spread out via the internet. There are six predominant dialects, each with a number variations, and tens of less-spoken
Multi dialects factor has led to the complexity of the interaction between people from different parts of the Arab world. Since the proliferation of social networking platforms, large groups of people have started to express their opinions by writing on social media platforms. In spite of MSA being the traditional way of writing in Arabic, people tend to employ Arabic dialects in their informal communication on social media sites. Arabic dialects may differ in terms of lexical uses to express concepts; despite that, in many cases the lexicon used has proper Arabic roots (Darwish and Magdy, 2014). Moreover, the pronunciation of the same letters is sometimes different from one dialect to another. The letter qa (ٛ (/q/)) is the one that mostly varies in pronunciations (Habash, 2010). It is pronounced as (Uh) in Egyptian and Levantine Arabic, while in Gulf Arabic it is pronounced as /g/. Another example is the letter jeem (چ(j)), which is pronounced as a soft g as in gavel in Egyptian and some Yemini dialects, and as j like in John in most other dialects (Darwish and Magdy, 2014).

Thus, there is need to analyse whether the dataset introduced in Chapter 5 represents a large segment of the Arab nationalities, and also what sort of language is used on social media, whether MSA or dialect vernacular. This analysis along with several samples of Arabic dialects from many different Arabic states are provided in Section 5.5.1. Some notable observations in this dataset are:

### Table 4.1: Distribution of the number of Arabic speakers of different dialects (Ridout, 2018).

<table>
<thead>
<tr>
<th>Dialect</th>
<th>Areas Spoken</th>
<th>Number of Speakers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Egyptian</td>
<td>Egypt</td>
<td>55,000,000</td>
</tr>
<tr>
<td>Gulf</td>
<td>Bahrain, Kuwait, Oman, Qatar, Saudi Arabia, UAE</td>
<td>36,056,000</td>
</tr>
<tr>
<td>Hassaniya</td>
<td>Mauritania, southern Morocco, south western Algeria, Western Sahara</td>
<td>3,000,000</td>
</tr>
<tr>
<td>Levantine</td>
<td>Lebanon, Jordan, Palestine, Syria</td>
<td>21,000,000</td>
</tr>
<tr>
<td>Maghrebi</td>
<td>Algeria, Libya, Morocco, Tunisia</td>
<td>70,000,000</td>
</tr>
<tr>
<td>Mesopotamian/Iraqi</td>
<td>Iraq, eastern Syria</td>
<td>35,000,000</td>
</tr>
<tr>
<td>Sudanese</td>
<td>Sudan, Southern Egypt</td>
<td>40,000,000</td>
</tr>
<tr>
<td>Yemeni</td>
<td>Yemen, Somalia, Djibouti, southern Saudi Arabia</td>
<td>15,000,000</td>
</tr>
</tbody>
</table>
• Arabic words in Latin alphabet transcription, using phonetically similar characters (see Table 5.3). This style of writing is called Arabizi (see Section 4.5.3).

• Arabic mixed with English, French, Spanish and many other languages in Latin alphabet transcription (see examples in Table 5.3).

• English mixed with Arabic, both in Arabic alphabet transcription (see examples presented in Table 5.3).

• Words spelt in accordance with their pronunciation in dialects (see examples in Table 5.3).

• Creative spelling, misspellings, word elongations and Arabic and non-Arabic abbreviations, such as the abbreviation LOL used as لول (lwl)

All these factors could possibly affect the results of text mining, notably within the context of communication on social media sites.

4.4 Overview of Previous Research in Mining Arabic Dialect Text

Most of the text mining research targeting Arabic dialects is relatively recent. The main problems being targeted are the detection of Arabic dialects and sentiment analysis (Zaidan and Callison-Burch, 2014; Soliman et al., 2014; Aldayel and Azmi, 2016).

In their study, Zaidan and Callison-Burch train and evaluate a simple probabilistic classifier, based on smoothed N-grams, for automatic dialect identification (Zaidan and Callison-Burch, 2014). They use a relatively large training dataset of 100,000 sentences for readers comments on online versions of three newspapers (Zaidan and Callison-Burch, 2011).
More studies exist for sentiment analysis of Arabic. Aldayel and Azmi recommend a hybrid approach to sentiment analysis of tweets written in an Arabic dialect (Aldayel and Azmi, 2016). In summary, they apply unsupervised learning to automatically label their tweets based on lexical-based classifier and then train an SVM classifier with the labelled data for polarity detection in tweets. Another two studies propose sentiment analysis models for Facebook comments made by young adults (Hedar and Doss, 2013; Soliman et al., 2014). Both of them also suggest using an SVM classifier, the second, proposed by Soliman et al. (2014), being a Gaussian-kernel SVM classifier. The biggest obstacle for sentiment analysis of Arabic text is often recognised as the absence of a preconceived, open-source, labelled dataset (Al-Kabi et al., 2013).

4.5 Arabic Language Features Affecting the Text Mining Process

The Arabic language poses several specific challenges to the text mining process. When it comes to formal Arabic texts, such as news reports that are written in MSA, most of the challenges are orthographic and morphological. Content, generated by users of social media platforms, introduces additional challenges related to the use of dialects, text decorations, abbreviations, word elongations, misspelling, and inconsistency in writing style. The text mining process of Arabic content in forms of speech and written text is influenced by the orthographic and phonological characteristics of Arabic. In the following subsections we explore these challenges in depth.

4.5.1 Orthography of the Arabic Language

Arabic is a Semitic language with a script from right to left, and it has twenty-eight letters that form words; the shape of the letters change based upon their
place in words. Letters might or might not contain diacritics, depending on the author. There are eight basic diacritic marks employed in Arabic, alongside others that are rarely used (Darwish and Magdy, 2014). Specific to Arabic writing is the hamza sign that can reshape three letters; it can reshape the letter to either of the three forms, , , ; the letter to the form and the letter to the form. Another specificity of Arabic orthography is the kashida (or tatweel) character used for elongation of other characters in the form of a variable-length line, occasionally used inside words for justification. Some Arabic letters are interchangeably used because of similarity of their shapes and phonetics leading to common orthography mistakes and morphological transformations (Darwish and Magdy, 2014).

Moreover, there are two types of numerals utilised across Arab countries, Arabic numerals and Arabic-Indic numerals. The Arabic numerals are 0 1 2 3 4 5 6 7 8 9, and the Arabic-Indic numerals are 9 8 7 6 5 4 3 2 1. The two types of numerals differ in their Unicode representation, which might also affect the text mining process (Darwish and Magdy, 2014). The Arabic letters are also used in other languages, such as Persian and Urdu (Atallah and Omar, 2008; Darwish and Magdy, 2014). Because of these similarities, some letters from these languages are sometimes used in Arabic online conversations too.

Although the use of the described orthographic elements can be of great importance in terms of letters formation, morphology and grammar (in the case of diacritics and hamza), the orthographic variety they introduce complicates the text mining methods and may reduce their effectiveness (Darwish and Magdy, 2014). When processing electronic text, most of the aforementioned orthographic features are smoothly treated by employing letter normalisation and diacritic and kashida removal. However, such normalisation sometimes increases ambiguity (Darwish and Magdy, 2014).
4.5.2 Morphology of the Arabic Language

Arabic is mainly based on three types of words; nouns, verbs and particles (Abdul-Al-Aal and Ashamil, 1987; Darwish, 2002). Words are created from roots (generally referred to as stems in text mining) by adding prefixes and suffixes, inserting infixes or doubling consonants. The general form of an Arabic word is: Prefix(es) + Stem + Suffix(es) (Benajiba, Rosso, and Benedíruiz, 2007). Furthermore, stems might accept multiple prefixes/suffixes. Prefixes can be coordinating conjunctions, determiners, and prepositions, while suffixes are attached pronouns and gender and number markers (Darwish and Magdy, 2014). For example, the stem word كتب (kutib, wrote) can appear in various forms, with a prefix as كتب (aktub, write), with a suffix as كتبت (katabt, she wrote) and with an infix as كتاب (كتاب, kitab, book).

4.5.3 Arabizi

Sometimes, Arabic users use a combination of the Latin alphabet letters and numerals to spell Arabic words in online environments. Numerals are utilised to represent Arabic letters which may not have phonetic equivalent in a Latin alphabet language, thus they are substituted by numerals with similar shapes to the corresponding Arabic letters. For instance, 2 and 3 represent the letters ﯽ and ﯾ, respectively. This form of writing is known as Arabizi and is commonly used in Arabic social media. Arabizi first appeared in response to the need to type Arabic in computer systems that supported only the Latin alphabet in the past before the switch to Unicode. Despite the growing support for Arabic in many systems nowadays, Arabizi is still popular because some users are accustomed to mastering the use of the Latin alphabet keyboard layouts as opposed to an Arabic keyboard layout (Darwish and Magdy, 2014).
4.6 Text Pre-Processing

4.6.1 Feature Engineering

Text pre-processing can be viewed as a process of extracting features from raw, unstructured text data with the goal of representing the data in a form suitable for machine learning or other type of analysis. This form is typically a table, each row of which represents a document in a corpus, and each column is either a numeric or a categorical attribute of a document. These attributes are commonly referred to as features. Thus, text pre-processing can be seen as feature engineering, a process of employing human knowledge about the text data and the machine-learning algorithm to be applied to it (Chollet, 2017).

Pre-processing of text typically involves (Neto et al., 2000; Iiritano and Rufolo, 2001; Mathiak and Eckstein, 2004):

- tokenisation,
- filtering,
- normalisation,
- document modelling and representation.

Generally, the pre-processing techniques utilised for mining text in English can be applied to Arabic as well. However, there are also many points of dissimilarity between Arabic and English text that need to be observed in the context of text mining. In the following sections, we present these points as well as the methods that are required to handle the particular characteristics of Arabic text. Moreover, there are differences between Modern Standard Arabic (MSA), and Arabic dialects which also need to be observed. Typically, the pre-processing steps include: treating a number of the orthographic features; conducting morphological analysis or stemming; identifying stop words; and handling lexical and writing system variations.
4.6.2 Tokenization

Words can be formed from a stream of letters and separated by a set of delimiters (Witten et al., 2016). The first step in text pre-processing is to separate the alphabetic sequences into tokens. While tokenisation may seem easy, it can be a quite complicated process when the tokens are words (Weiss et al., 2010; Witten et al., 2016). A complicating factor can be the nonalphabetic characters, for which a decision needs to be made whether to be discarded or retained. For example, numbers are typically retained as tokens together with any non-alphabetic characters which are part of the number, such as the sign of the number and the decimal point. Similarly, alphanumeric sequences may be regarded as tokens. Commonly, the separators between tokens are punctuation characters as well as white space (typically, sequences of either space, tab, or new-line characters). Other punctuation characters such as ( ) ¡ ¿ ! ? , are more likely to be delimiters and may also be tokens; moreover, the characters . , - : ’ are delimiters only depending on the application (Weiss et al., 2010; Witten et al., 2016). Arabic text has the same characteristics as English in this respect. Arabic, similar to English, is a segmented language, i.e. words in Arabic text are separated by white space. However, in casual online writing sometimes words are separated by comma only (without white space), and in this work, we also consider the comma character as a delimiter to handle this style of informal writing. Table 4.2 contains examples of comma-separated words in our dataset. Occasionally, in informal online writing, there is little attention paid to the correct use of delimiters between words.

4.6.3 Filtering

Filtering is the further removal of punctuation marks as well as the removal of diacritics (in Arabic) and selected words from the documents. Standard filtering also excludes the most frequent words from a document as they carry little useful information; these are called stop words. Sometimes it is also
beneficial to discard low-frequency (Kilgarriff and Grefenstette, 2003; Blanchard, 2007; Witten et al., 2016). This process helps in minimising the size/dimensionality of the dataset, which would otherwise represent impediments to the mining (Saad and Ashour, 2010). The following is a list of items that are typically filtered out:

**Diacritics:** Diacritics assist in clarifying the meaning of words. The most widely employed method is the removal of all diacritics despite the risk of increased ambiguity (Darwish and Magdy, 2014).

**Kashidas:** Kashidas are elongated characters for justifying a block of text. They are usually eliminated (Darwish and Magdy, 2014).

**Word Elongation:** In casual text on social networking websites, authors commonly elongate words by iterating some of the characters in the word to convey emotions or significance. Many examples appear in text generated in social media websites, such as wwwwooooo and loooool in English posts. In informal Arabic text, for example, the pair of letters م (LA) meaning “no”, are often iterated multiple times. Sequences of the same letter, whether grammatically correct or not, are typically reduced to a single letter (Darwish, Magdy, and Mourad, 2012). This is similar to, for instance, the word “happy” in English, pre-processed to hapy.

**Stop words:** Stop words (also known as function words) are a predetermined
Table 4.3: Example of stop words in English and Arabic languages.

<table>
<thead>
<tr>
<th>Language</th>
<th>Stopwords</th>
</tr>
</thead>
<tbody>
<tr>
<td>English language</td>
<td>a, the, and, to, in, on, at, also, above, about, almost, usually, often, any</td>
</tr>
<tr>
<td>Arabic language</td>
<td>في كل، لم، لن، له، من، هو، هي، كما، لها، منذ، وقد، لا، نفسه، لقاء، مقابل، هناك، قال، كان، نهاية، وقالت، وكانت، فيه، لكن، منها، يمكن، أكد</td>
</tr>
</tbody>
</table>

list of commonly used words, such as articles, personal pronouns and prepositions that may be discarded (see Table 4.3). Appendix A contains references to websites that list common stop words for the purpose of text mining. The removal of stop words can increase the predictive accuracy in some cases and decreases it in others. The difficulty in dealing with Arabic words is that they accept both prefixes and suffixes (Darwish and Magdy, 2014). Pronouns like (ني (-ni), me, كأ (-ka), you) as well as coordinating conjunctions like (fa) and (oa)) can be attached to a preposition. Thus, identifying stop words may require stemming (see Section 4.6.5) in advance. Stop words removal has been shown to be effective in retrieving relatively long news documents (Chen and Gey, 2002; Xu, Fraser, and Weischedel, 2002); though, it led to low performance in retrieving Arabic nanoblog posts (Darwish, Magdy, and Mourad, 2012). The low performance could be due to the short length of nanoblog posts which become even shorter after stemming and stop word removing.

4.6.4 Normalisation

The goal of normalisation is to avoid or lessen the problem of sparseness in the dataset. As mentioned in Section 4.5.1, the variation of the usage of certain letters in Arabic can result in the writing of certain words in several different ways mistakenly or not, which increases the number of features in
the representation space. For example, a common typo is to write Ireland as آيرلندا instead as أيرلندا with a hamza. Due to the lack of rules for exact spelling of foreign names, such names may allow alternative correct spellings, such as Belgrade which can be written as بلغراد or maybe as بلجارد. At the normalisation step, alternative spellings of the same word are replaced by a single one.

Due to the diversity of the dialects of Arab users in social media, the majority of the users spell words as they are pronounced in their local dialects (Darwish and Magdy, 2014). For example, the word (صدق) صدق - meaning truth or honestly) is frequently seen as (صح) صح as pronounced in Gulf Arabic. According to Darwish et al., this issue is not yet been address in the existing text mining research (Darwish and Magdy, 2014). Another problem related to the variety of dialects is that there are different words for some concepts in different dialects. An example of such concept is a lot, the word a lot is شديد (shaydid) in Soudanese Arabic, كثير (kythir) in Levantine Arabic, مرة (mara) in Saudi Arabic, واحد (Wyde) in Gulf Arabic, ياسير (yaysir) in Mauritanian Arabic, هليا (halba) in Libyian Arabic, برااف (bizzaf) in Algerian and Moroccan Arabic, برشا (bersha) in Tunisian Arabic, and أوي (awiue) in Egyptian Arabic. Also, letters with similar pronunciation are incorrectly used interchangably, such as using the letters ص and ص interchangeably. Examples of such use are provided in Table 6.1.

4.6.5 Stemming or Lemmatisation

While, typically, Arabic words have three-letter roots, Arabic has complicated grammatical rules and it is very rich in its derivational system, which makes it particularly challenging for morphological analysis and computational processing (Darwish and Magdy, 2014). In Arabic (as in English), a word can have more than one form. For example, book and books are two forms of the same word in English. For text mining, it is often useful to reduce words to a stem, which can be either the root or the lemma of the word,
however, the benefit of this operation can be application-dependent (Weiss et al., 2010). Stemming decreases the number of tokens, for example, these two words 

\[
\text{تاريخ} \quad \text{و تاريخ}
\]

they are two tokens if treated without stemming, but they are one token 

\[
\text{تاريخ}
\]

if the stemming were applied. See some examples of words and their roots in Table 4.4.

The process of transforming a word to a stem is called *stemming* (Giridhar et al., 2011). In Arabic, stemming can be done by stripping prefixes, suffixes and infixes of words. Since stemming leads to decreasing the total number of different words, it also reduces the dimensionality of the feature space for text mining (Al-Kabi, Al-Shawakfa, and Alsmadi, 2013).

Early stemmer algorithms used for English text are the Porter stemmer (Porter, 1980) and the Lovins stemmer (Lovins, 1968). Some stemmer algorithms have also been developed for Arabic, such as the Khoja stemmer (Khoja and Garside, 1999) and the Buckwalter stemmer (Buckwalter, 2002) and these are root-based stemme. Root-based stemmer algorithms find the three-letter roots of Arabic words, with the different words derived from the same root being grouped together (Khoja and Garside, 1999; Al-Shalabi et al., 2007; Momani and Faraj, 2007). The main shortcoming of root-based stemmers is that they affect the meaning of the words (Froud, Lachkar, and Ouatik, 2012). To alleviate this effect, lemma-based stemmers was proposed as well and know as light stemming for Arabic (Aljlayl et al., 2001; Darwish and Oard, 2003; Larkey, Ballesteros, and Connell, 2002). In the light stemming approach, prefixes, suffixes and infixes are removed only if they match entries in a lists of common prefixes and suffixes, respectively. The advantage of this method is that it requires no morphological processing. Two frequently employed light stemmers are:

- Al-Stem (Darwish and Oard, 2003), which is aggressive to some extent. It lightly weeds out the following prefixes in order from right to left:

  

\[
\begin{align*}
\text{فألا، وا، فا، لا، ل، ب، ب، م، م، ل، نت، نت،ام، م، كم، كم، لم، لم، ل، نت).
\end{align*}
\]

In English, the Arabic words are translated as follows:

\[
\begin{align*}
\text{فألا، وا، فا، لا، ل، ب، ب، م، م، ل، نت، نت،ام، م، كم، كم، لم، لم، ل، نت).
\end{align*}
\]
TABLE 4.4: Some words in English and Arabic languages with their stems.

<table>
<thead>
<tr>
<th>Language</th>
<th>Word</th>
<th>Stem</th>
</tr>
</thead>
<tbody>
<tr>
<td>English language</td>
<td>provider, providers, providing, provided, provides</td>
<td>provid</td>
</tr>
<tr>
<td></td>
<td>leader, leaders, leading, leaded, leads</td>
<td>lead</td>
</tr>
<tr>
<td>Arabic language</td>
<td>زرود</td>
<td>سيطر، يسيطر، يسيطرون، يسيطران، أسيطر، تسيطر، تُسيطرن</td>
</tr>
</tbody>
</table>

plus the following suffixes starting from right to left, too:

أ، و، ون، وه، أن، تي، نه، ن، هم، هن، ها، هية، تكل، نا، بين، يه، ة، ه، ي، ا.

- Umass light10 stemmer (Larkey, Ballesteros, and Connell, 2002), which weeds out the prefixes الي، وال، قال، قال، قال، قال، quoi، et from the beginning of words plus the suffixes ها، أن، أت، ون، بين، يه، هية، تكل، نا، بين، يه، ة، ه، ي، ا from the end.

- ARLSTem (Abainia, Ouamour, and Sayoud, 2017), which weeds out prefixes, suffixes and infixes from a word. This is the stemmer we utilise in our work, as it reportedly compares well to other available stemmers.

A list of links to various available stemmers is provided in Appendix A.

4.6.6 Document Modelling and Representation

Since the best known methods for supervised and unsupervised learning are typically not applicable directly to unstructured text data, a document representation is needed in order to utilise these methods for descriptive and predictive text analysis (Buckwalter, 2002). Accordingly, the text needs to be
transformed into a proper structure, a table in which the data rows are examples and the columns are the features of the examples. The widely used text document representations for this purpose are the N-gram model, and the vector space model (VSM) (Buckwalter, 2002).

4.6.6.1 The N-gram Model

An N-gram in computational linguistics is a sequence of n items from a given text. Most frequently, the items are either words or characters (Nahm and Mooney, 2000). According to Cavnar and Trenkle, the employment of N-gram features is a straightforward, inexpensive and highly effective method used for modeling documents in a document classification scenario (Cavnar and Trenkle, 1994). There is a wide range of applications using this model, such as spelling check, language identification, document classification, optical character recognition (OCR) and information retrieval (Cavnar and Trenkle, 1994; Nahm and Mooney, 2000; Abou-Assaleh et al., 2004; Xun et al., 2017). N-grams have been used in the English language linguistic and information retrieval research at least since the 1950s, starting with an early study on predicting the next letter based on prior letters (Shannon, 1951) as well as later in the popular work of Katz on speech recognisers (Katz, 1987).

The N in N-grams is a whole number greater than 0, constrained by the available memory and the computational complexity. There are other names for the model based on the value of N, for instance, unigram model (N = 1), bigram model (N = 2), trigram model (N = 3), quadrigram (N = 4). When the items in an N-gram are words, the models are also denoted as $w_N$ (unigram), $w_{N-1}w_N$ (bigram), $w_{N-2}w_{N-1}w_N$ (trigram), $w_{N-3}w_{N-2}w_{N-1}w_N$ (quadrigram), where w stands for a word.

The unigram model when the items are words is also known as the bag of words (BoW) model, the word bag pointing to the fact that any syntactic structure in the document is discarded, while keeping only the words, i.e. the vocabulary (Machine Learning Mastery).
The employment of character trigrams and quadrigrams has been shown to be highly efficient in Arabic information retrieval (Darwish and Oard, 2002; Mayfield et al., 2002). Darwish and Oard (2002) report that the usage of character N-gram models far outweighs the benefits of light stemming. They suggest three reasons for this:

- Character N-grams frequently match stems of words.
- Character N-grams modelling does not require a predefined dictionary of stems.
- Character N-grams that contain prefixes and suffixes of words occur more frequently than N-grams that contain stems. The weight of these then can be significantly reduced by tf-idf modelling.

Darwish and Magdy (2014) point out as well that the employment of character N-grams needs to be completed in conjunction with kashida and diacritics removal and letters normalisation. They also state that the main drawback of utilising character N-gram models is the expanded processing of text and the size of the required storage space. For instance, a six-letter word is represented by four tokens when character trigrams are used.

The N-gram model is used in probabilistic language modelling with the objective of computing the probability of a sentence or a sequence of words (Jurafsky and Martin, 2017). Consider a sequence of words \( W = w_1w_2w_3...w_{n-1}w_n \). The probability \( P(W) \) of \( W \) occurring in a text document can be expressed according to the so-called chain rule as shown in Equation (4.1).

\[
P(w_1w_2w_3...w_{n-1}w_n) = \prod_i P(w_i|w_1w_2w_3...w_{i-1})
\]  

(4.1)

For example, if \( W = \text{its colour is so bright} \), then

\[
P(W) = P(\text{its}) \ P(\text{colour} — \text{its}) \ P(\text{is} — \text{its colour}) \ P(\text{so} — \text{its colour is}) \ P(\text{bright} — \text{its colour is so})
\]
The posterior probabilities in the above example can be estimate by Equation (4.2).

\[ P(w_i|w_1w_2w_3...w_{i-1}) = \frac{\text{count}(w_1w_2w_3...w_{i-1}w_i)}{\text{count}(w_1w_2w_3...w_{i-1})} \] (4.2)

Consider an N-gram model and let \( k = N - 1 \). Then according to the Markov assumption, the posterior probabilities can be approximated as shown in Equations (4.1) and (4.2).

\[ P(w_i|w_1w_2w_3...w_{i-1}) \approx P(w_i|w_{i-k}...w_{i-1}), k \geq 1 \]

\[ P(w_i|w_1w_2w_3...w_{i-1}) \approx P(w_i), k \geq 1 \]

It can be noted that the sparsity in a dataset increases with increasing \( N \), as typically few documents in a corpus would contain N-grams with a high value of \( N \) (Xun et al., 2017).

### 4.6.6.2 Vector Space Model

The vector space model (VSM) is the next step in transforming a document corpus into a form that can be used as the input to machine learning algorithms (Chowdhury, 2010). Assuming each document in the corpus is represented as a set of N-grams, also called terms, VSM further transforms the document into a vector of the frequency of occurrence of each term in the document. In their simplest form, these frequencies can indicators (either 0 or 1) of the presence of the term in a document. The size of each document vector is equal to the number of terms in the entire documents corpus. VSM is both simple in terms of data structure and an efficient document representation method for analysing a big collection of documents.

**tf-idf:** Consider a document corpus \( C \) of \( |C| \) documents, containing \( m \) terms in total, a typical practice is to build a document-term matrix \( DT = (dt_{ij}) \in \mathbb{R}_{|C| \times m} \), where \( dt_{ij} \) is a numeric representation of term \( i \) in document \( j \). According to VSM, \( dt_{ij} \) is the frequency of occurrence \( f_{ij} \) of term \( i \) in document \( j \), i.e. the number of occurrences of term \( i \) in document \( j \) divided by the total
number of terms in document $j$. In some cases the frequencies are logarithmically scaled, i.e. $dt_{ij} = \log(1 + f_{ij})$.

When the documents in corpus $C$ are considered independent from each other, the frequency of a term can be a misleading measure of its importance. If the frequency of term $i$ is relatively high in most of the documents in corpus $C$, then its high frequency in a particular document $j$ does not carry significant information about document $j$. In this case, it is beneficial to correct $f_{ij}$ by multiplying it by a factor inversely proportional to the number of documents $C_i \subseteq C$ containing term $i$. This corrected measure is known as term frequency-inverse document frequency (tf-idf) and it is typically calculated as

$$f_{ij} \times \log \frac{|C|}{|C_i|}$$

(4.3)

### 4.6.6.3 Part of Speech Tags

The representation of a document can be further enriched by parts-of-speech (PoS) tags (also known as word classes, or syntactic categories). Tagging a word as a noun or a verb, for example, provides a lot of information about potential neighbouring words and the syntactic structure surrounding that word. PoS tags are particularly beneficial for the discovery of named entities, such as persons or organisations in text documents (Jurafsky and Martin, 2017). On top of a corpus annotated with PoS tags, a tree structure of each sentence can be created as well. The collection of such trees for a corpus is known as a *treebank*. A large number of treebanks for benchmark corpora, including some Arabic corpora, are currently available in the annotated corpus collection of the Linguistic Data Consortium.\footnote{https://catalog.ldc.upenn.edu/byproject}

The treebanks in this collection were previously known as the Teebank Penn Project (Jurafsky and Martin, 2017).
4.7 Machine Learning Algorithms (ML)

Offensive comments filtering is an application of text classification with two classes, one of which is the class of offensive comments to be filtered out. A range of ML algorithms, including Naive Bayes (NB), Support Vector Machines (SVM) and Logistic Regression (Davidson et al., 2017), have been successfully used for offensive language detection in English. Thus, it is sensible to assume that the same ML algorithms can perform well for detection of offensive comments in Arabic. All of the listed ML algorithms require labelled training and testing datasets. In our case, these are datasets of comments which have been labelled as either offensive or inoffensive in advance. Before proceeding with an overview of the main ML algorithms used for document classification, we discuss the process of feature selection which plays a major role in building accurate classifiers.

4.7.1 Feature Selection

Feature selection is the process of selecting a subset of features from the training dataset and discarding others with the goal of maximising the accuracy of the predictive model. Since there were some features when they used to represent the training set, they increase the classification error on new data, these are called noise features. By excluding noise features this helps to avoid overfitting. Some ML algorithms show poor performance when learning from a lot of features; therefore, it is vital to select the ones that represent the dataset most. With a fixed number of training examples, the accuracy of a classifier first increases as the number of features grows, but after some point, it starts to decrease dramatically. This phenomenon is known as the Hughes effect (Pal and Foody, 2010). Thus, selecting the best features is essential for building an accurate document classifier.

It can also be useful to create compound features that are combinations of the raw features in the training dataset. For some classifiers, such as logistic
regression and Naive Bayes, these compound features need to be designed manually in advance before the training process. Some other machine learning classifiers can model the combinations between features automatically. This is also known as embedded feature selection, as opposed to the filter and wrapper methods. Such classifiers include variations of random forest and Support Vector Machines (SVMs), as well as deep learning algorithms (Jurafsky and Martin, 2017).

4.7.1.1 Singular-Value Decomposition (SVD)

Singular Value Decomposition (SVD) is a matrix decomposition method that can be applied for dimensionality reduction. It consists in representing a typically sparse matrix by its constituent parts with a view of simplifying certain subsequent matrix calculations.

Assume $A \in \mathbb{R}^{m \times n}$ is a matrix (e.g., representing $m$ documents and $n$ terms). SVD consists of representing $A$ as the product of three matrices $U$, $\Sigma$ and $V$, i.e. $A = U \cdot \Sigma \cdot V^T$.

- $U, \Sigma, V$: unique
- $U, V$: column orthonormal.
  - $U^T U = I$, $V^T V = I$ ($I$: identity matrix)
  - Columns are orthogonal unit vectors.
- $\Sigma$: diagonal
  - Entries (singular values) are positive, and sorted in decreasing order.

These three matrices can be found as following:

- $U$ is an $m \times k$ matrix.
• $\Sigma$ is an $k \times k$ matrix, all elements of which, except the main diagonal, are zeros. The nonzero elements on the main diagonal are positive real numbers called singular values.

• $V$ is an $k \times n$ matrix.

### 4.7.1.2 Extra-Trees Algorithm

The Extra-Trees algorithm is other tree-based ensemble techniques has been proposed by Geurts, Ernst, and Wehenkel (2006). This technique is useful for feature selection in addition to being effective classifiers. It builds out an ensemble of decision or regression trees with two main differences comparing to other tree-based techniques. These two differences are that splitting nodes by picking up the cut-points randomly and employing the entire learning sample to develop the trees, rather than a bootstrap sample only. The Extra-Trees splitting procedure has two parameters, the number of features selected randomly at each node and the minimum sample size for splitting a node. Ensemble model is generated from the whole training sample (Geurts, Ernst, and Wehenkel, 2006).

This technique is used for dimensionality reduction by generating a large constructed set of trees against a target feature and then use each features usage statistics to discover the most informative subset of features.

### 4.7.2 Naive Bayes

Naive Bayes (NB) is a widely used probabilistic ML algorithm based on the Bayesian decision theory (Platt, 1999). NB builds a probability classifier and can be implemented quite efficiently with a linear time complexity (Berry and Kogan, 2010), which makes it suitable for training with large datasets. It employs a simplistic or naive assumption all features (typically the columns in a dataset of examples) are statistically independent from each other. Even though this simplistic assumption is rarely true, especially in the context of
text mining (Zhang, 2004), NB is one of the most broadly utilised ML algorithms in document classification, known to far outweigh even highly sophisticated classification algorithms (Zhang, 2004).

The scikit-learn Python module includes three types of Naive Bayes classifiers. These are:

- Gaussian NB: The features are presumed to numeric with normal distribution.
- Multinomial NB: The features are presumed to be discrete.
- BernoulliNB: Implements NB for data with binary features. This model can be applied for a binary document-term matrix (1 a term appears in a document, 0 a term does not occur in a document).

4.7.3 Support Vector Machines

Support Vector Machines (SVM) (Cristianini and Shawe-Taylor, 2000) is a supervised machine learning algorithm that has become the norm for text classification (Berry and Kogan, 2010). SVM is based on the notion of decision hyperplanes that determine the frontiers of the classes. A decision hyperplane splits a set of data items (e.g., documents), so that ideally items belonging to the same class are all placed in the same half-space. It works by selecting two parallel hyperplanes to split up between the two classes and then maximises the margin between the two hyperplanes, for this reason, it is also known as the maximal margin classifier (Berry and Kogan, 2010).

In summary, SVM works as follows. Consider a given particular hyperplane that separates two classes of data items. The distances between the hyperplane and the closest data points from the two classes, called support vectors can be computed. Once the support vectors become known, the hyperplane is replaced by two parallel hyperplanes that separate the classes, such that the width of the strip between the two hyperplanes, called margin, is maximised. Within the margin, there are no data points. See the illustration in
Figure 4.2. The hyperplane in the middle between the two parallel hyperplanes is called the frontier separates between the two classes. SVM finds the optimal frontier, such that the margin between the two classes is as wide as possible (Weston et al., 2001). Although SVM is limited to the case when the classes are linearly separable, it is one of the most successfully applied algorithms for document classification (Davidson et al., 2017).

![Figure 4.2: Illustration of the result of SVM.](image)

### 4.7.4 Logistic Regression

Logistic regression is another ML algorithm that can be used for binary classification. Similar to Naive Bayes, it is a probabilistic classifier, and similar SVM it finds a hyperplane that separate two classes of data items. Logistic regression is based on the linear regression algorithm for numeric prediction. Assuming the class attribute $y$ is binary, logistic regression aims at expressing it as a linear combination of the predictor attributes (i.e. features) as $\sum_{i=1}^{m} w_i f_i$, where $f_i \in \mathbb{R}$ are the features and $w_i \in \mathbb{R}$ are their weights. Since the values of the linear function are in the interval $(-\infty, \infty)$, a logit transformation (see Equation (4.4)) is applied to obtain values in the interval $[0, 1]$ and these values are then interpreted as the probabilities of the examples to belong to
the class labelled as 1 (Friedman, Hastie, and Tibshirani, 2001; Jurafsky and Martin, 2017).

\[ p(y = 1|x) = \frac{1}{1 + e^{\sum w_i f_i}} \]  

(4.4)

The logistic regression can be prone to overfitting. Regularisation helps in controlling and reducing it.

### 4.7.5 Regularisation

Regularisation is a feature selection technique utilises shrinkage estimators to get rid of the redundant feature from data. It is a popular method of controlling and reducing overfitting. Two regularisation norms L1 and L2 are commonly used. L1 norm is also known as least absolute shrinkage and selection operator (lasso). It minimises the sum of the absolute differences \( S \) between the target values \( y_i \) and the predicted values \( f(x_i) \):

\[ S = \sum_{i=1}^{n} | y_i - f(x_i) | \]  

(4.5)

The lasso method penalises high values of the less important features to be zero for irrelevant ones. Therefore, it decreases the number of features for training the model, which can make the model less overfitted. The L2 norm is also known as ridge regression. It minimises the sum of the squared differences \( S \) between the target values \( y_i \) and the predicted values \( f(x_i) \):

\[ S = \sum_{i=1}^{n} (y_i - f(x_i))^2 \]  

(4.6)

Ridge regression brings down the values of less important features but not to zero. That is, it does not eliminate irrelevant features but rather minimises their impact on the trained model (Friedman, Hastie, and Tibshirani, 2001; Jurafsky and Martin, 2017).
4.7.6 Decision Tree Algorithms

A decision tree is in the form of a tree structural predictive model, the input of the algorithm is a set of data and output is a classifier rule. Each internal node (including the root) of the tree correspond to a condition on the value of a single feature, depending on which condition one of the available branches is followed down the tree. In their simplest form conditions can be hyperplanes that divide the input space. The leaf nodes correspond to classification decisions. Figure 4.3 shows an example of a binary decision tree, where the root node is the decision of the tree, $B_i (i = 1, 2)$ are branches (decision point) of the tree and $C_j (j = 1, 2, 3, 4)$ are the leaves (terminal nodes) of the tree.

A decision tree is typically built from the top down. A recursive procedure is applied for picking the feature at each internal node, starting from the root. The feature picked at each internal node is the one that would split the training dataset into most homogeneous groups of examples. Decision trees are known as handling well missing values and noisy data. Regarding the overfitting problem a number of tree pruning algorithms have been proposed in order to avoid it (Breiman, 2017). A typical pruned classification tree has three to twelve terminal nodes, which makes it also suitable for getting an insight into the structure of the training data (Cutler et al., 2007).
Random Forests (RF)

Random forests (RF) is a very commonly used and effective ML algorithm for both classification and regression problems. It has been proposed by (Ho, 1995) as an extension of previous work by (Breiman, 2001). The extension integrates Breiman’s “bagging” notion with random selection of features. The fundamental core of RF is the merging of \( Y > 1 \) binary decision trees built using various bootstrap samples generated from the training dataset which are selected randomly for each single tree using a subset of features. Bootstrap aggregation, also called bagging, is an approach for lessening the variance of an estimated prediction function. Some entries may appear multiple times in the bootstrapped dataset and others may not be included even once. Each decision tree is created out of a different random subset of features. Typically, the number of features for classification is the square root of the total number of features (Friedman, Hastie, and Tibshirani, 2001). For classifying a new data entry, each tree votes and RF chooses the class with the majority votes. In their early work, (L. Breiman, 1984) proposed that each single tree is granted an equal vote, however later versions of RF permit weighted voting.

RF Algorithm for Classification (Friedman, Hastie, and Tibshirani, 2001):

- For \( y = 1 \) to \( Y \):
  - Generate a bootstrap dataset \( BD \) of size \( N \) from the training dataset.
  - Train a random-forest tree \( RFT_y \) with \( BD \). Start from the root of the tree and build it downwards by iterating recursively steps (i) – (iii), until the minimum node size \( n_{min} \) is reached.
    - Select \( k \) features randomly from the features’ set \( f \).
    - Pick a feature to split the node on.
    - Split the node into two nodes.
  - The result is the ensemble of trees \( \{RFT_i : i = 1, \ldots, Y\} \).
To classify a new data entry $x$:

Classification: Let $C_y(x)$ be the class prediction of the $y_{th}$ random-forest tree. Then $C_{rf}^y(x) = \text{majority vote} \{C_y(x) : y = 1, \ldots, Y\}$.

When building a single decision tree, RF collates all data entries that are not included in the bootstrapped sample in an out-of-bag (OOB) set. They are used to obtain an unbiased estimate of the classification error when adding trees to the forest, known as OOB error. They are also beneficial for evaluating the accuracy of the classifier (an alternative to an evaluation with a test dataset or cross-validation). The OOB error of RF reflects the strength of the single trees in the forest and the correlation between these trees. Whenever the number of features used for splitting nodes is reduced, the correlation between any two trees lessens and the strength of a tree decreases (Klassen and Paturi, 2010).

### 4.8 Evaluation of Text Classification

The performance of a textual classifier can be evaluated in terms of its precision ($p$) and recall ($r$) measures. For a classifier and with regard to a particular class $c$, let the number of true positives be $tp$, false positives $fp$, and false negatives $fn$. Then the precision and recall are defined as (Berry and Kogan, 2010):

$$p = \frac{tp}{tp + fp}$$

$$r = \frac{tp}{tp + fn}$$

Precision is the proportion of positive identifications was correct. Recall is the proportion of actual positives was identified correctly. There is a trade-off between precision and recall, i.e. the higher the precision, the lower the recall and vice-versa. Thus, typically their weighted harmonic mean, known as F-measure (also known as F1-score) is also used. It is defined as (Berry and Kogan, 2010):

$$F1 = \frac{2pr}{p + r}$$
4.8.1 K-Fold Cross-Validation

Cross-validation (CV) is a holdout procedure employed to evaluate a machine learning classifier on a limited data sample. It assesses the ability of a predictive model to classify new data instances accurately (Witten et al., 2011; Machine Learning Mastery). CV has one parameter $k$ that refers to the number of partitions that a data is to be divided into. For that reason, the approach is often named k-fold CV. When a certain value is assigned to $k$, it may be used instead of $k$ in reference to the procedure, for instance when $k = 10$ (a frequently used value of $k$) the method is known as tenfold cross-validation. CV is a common approach because it generally produces a less biased or less optimistic estimate of a models accuracy (Machine Learning Mastery).

The steps of CV are (Machine Learning Mastery):

- Shuffle the dataset randomly.
- Divide the dataset into $k$ subsets.
- For each subset in turn:
  - Hold out the subset as a testing set.
  - Use all other subsets as a training set.
  - Test the predictive classifier with the testing dataset.
  - Save the evaluation score $E_i$ and repeat with the next subset.
- Calculate the average evaluation score $E = \frac{\sum_i}{k} E_i$ and report it as an evaluation score of a model trained with the whole dataset.

4.8.2 ROC Curves

Receiver Operating Characteristic (ROC) curve is popularly employed as a visualisation method that illustrates the performance of a binary probabilistic classifier and helps in choosing the best out of a few alternative classifiers (Lasko et al., 2005). For either of the two classes, the curve is a plot of
the false positive rate ($fpr$) against the true positive rate ($tpr$), assuming a variable decision threshold. In other words, a ROC curve shows the relative trade-off between benefit ($tpr$) and cost ($fpr$) (Fawcett, 2006). The ideal point in ROC plot is the top-left corner, which corresponds to the zero false positive rate, and 100% true positive rate. The closest to the top-left corner point in a ROC curve, thus, represents the ideal decision threshold for which the ratio $tpr/fpr$ is maximised.

ROC curves located around or below the diagonal (the *no power* line in Figure 4.4) correspond to random classifiers. A reasonably good classifier will give a ROC curve that is consistently better than random across all decision threshold choices. The most accurate classifiers would have ROC curves that closely approach the top-left corner (see Figure 4.4). The area under the curve (AUC) is also used as an indicator for a classifiers performance. If choosing between two classifiers, the one with a larger AUC is considered better (Fawcett, 2006).

![Figure 4.4: Example of ROC curves.](image)

### 4.8.3 Boxplots

Boxplots are a common standard technique for summarising a range of numeric values (Tukey, 1977; Potter et al., 2006), which depicts a five-number
summary of a data distribution; these numbers are the median, the lower and upper quartiles, and the upper and lower extreme values. In addition, it exhibits both the remarkably high and remarkably low values as outliers.

The generic structure of a boxplot is depicted in 4.5. The quartiles in a boxplot are either equivalent to or near the 25th and 75th percentiles. To create a boxplot, a scale is sketched based on the set of data, and a rectangle is laid on the scale with one end of the rectangle on the scale point of the 25th percentile and the other end on the 75th percentile. In addition, a line in the rectangle points to the median on the scale; the median is equivalent to the 50th percentile of the distribution. The rectangle that spans quartiles Q1 to Q3 (see 4.5) is called interquartile range (IQR). Two lines, called whiskers, are drawn along the scale from quartile Q1 and quartile Q3, respectively. They extend to the upper and lower extreme values, respectively, but only up to 1.5 quartile. Data points beyond 1.5 quartile outside the rectangle are generally plotted one by one as outliers (Behrens and Yu, 2003).

**Figure 4.5: Structure of a boxplot.**

### 4.9 Summary

The fast emergence of online communication has resulted in considerable and constantly increasing amounts of digital text data. For the purpose of the extraction of information from massive collections of text data, studies
have been conducted on several text mining approaches. When the text content is not structured according to a formal grammatical convention (which is typically the case with online communication), information extraction becomes a challenging task. The less structured the data is, the harder the task of descriptive or predictive modelling becomes. Natural language processing is a hard task in general, and even harder for a highly morphologically complex language such as Arabic; moreover, dealing with multiple dialects within the same document complicates the problem further.

In this chapter, an overview of the Arabic language has been presented along with a discussion of the challenges faced by text mining techniques when dealing with Arabic text. We discussed the presence of two kinds of Arabic language living side by side. These are MSA and the multitude of Arabic dialects. Also, detailed information is provided on the text mining preprocessing methods that might need to be applied to the Arabic text before utilising any machine learning techniques. Moreover, we discuss the applicability of these techniques to MSA and the difficulties that emerge when applying them to Arabic dialects. Finally, an overview of the machine learning algorithms, that are widely used for document classification, is presented as well together with a review of the evaluation techniques used for assessing classifiers performance.

Appendix A presents a list of links to Arabic resources that can be useful within a text mining workflow, such as stemmers and stop words lists.
Chapter 5

Dataset Construction

Warning: this chapter contains a range of words which may cause offence.

5.1 Introduction

Although the number of offensive language detection studies has increased in recent years, there are not many datasets specifically labelled for tackling this problem. Currently, and to the best of our knowledge, there are not many datasets publicly available to allow targeting the same issue in Arabic text. We found a few recent studies, one of which makes two datasets available, a dataset of 1,100 manually labelled tweets as well as a dataset of 32K user comments from a popular Arabic news site, both containing data entries deemed to be inappropriate language (Mubarak, Darwish, and Magdy, 2017). Another study applies manual labelling of 500 Twitter accounts, with half of these 500 accounts labelled as abusive (Abozinadah, Mbaziira, and Jones, 2015). In general, the labelled datasets in these studies are relatively small. In addition, these studies predominantly use data collected from Twitter (the maximum length of Twitter posts is 140 characters), while the length of the comments on other social media platforms, such as YouTube, can be irregular in terms of number of words (e.g., on YouTube the number of words per post can exceed thousands). Therefore, an initial goal of work has been to construct a suitable corpus, different and richer than the few ones available
in the research literature, which can potentially improve further the research results in detection of offensive language in online communication in Arabic.

In the design of corpora, there are essential characteristics that need to be considered such as *availability, representativeness, heterogeneity and balance* (Nguyen et al., 2012) with availability and representativeness being two crucial factors in studying offensive language detection. There are many incidents of offensive language occurring in private environments on the internet where access is restricted, such as *Facebook*, which otherwise would be a good source for such data. However, there are also other sources publicly available, such as *YouTube*, and incidents of abuse and offensive language happen regularly on these platforms as well. Furthermore, on public platforms, victims are humiliated in front of a larger segment of people, and more people take part in the abuse compared to platforms with a higher level of privacy. Thus, such platforms are a rich source of data which is both publicly available and representative.

*YouTube* is a popular platform for sharing videos, which provides many activities for its users. It allows users to comment on shared videos, and these comments occasionally contain offensive language and insults. *YouTube* has been of special interest in research on flaming and antagonism (Pihlaja, 2014), with flaming defined as posting negative comments online (Lange, 2007). A study by Moor, Heuvelman, and Verleur (2010) states that hostility by insulting, swearing or using otherwise offensive language appears to be extremely common on *YouTube*. The work of Lange presents a potential interpretation of the widespread of flaming on *YouTube*. It suggests that plenty of people assume that *haters* are users who do not publish videos themselves. That is, there is a category of *YouTube* users who tend to post comments, typically having little to do with the video they are commenting on, whilst having never to risk receiving any unpleasant criticism themselves. This opinion suggests the presence of a crowd of the *YouTube* users who simply enjoy offending others (Lange, 2007).
The minimum number of positives (i.e., profane comments) required depends on the employed data mining methodology (Indurkhya and Damerau, 2010). Ideally, a dataset of this kind should represent the diversity of text present in cyberspace and also generated by a variety of people. By text diversity we mean the variety of writing styles, where the style speaks of the personal intentions of the author. It is sensible to assume that the larger the number of diverse profane comments is present in a training dataset, the more accurate offensive language and harassment detection can be made by employing the dataset for predictive modelling. Furthermore, we aimed at a balanced number of positive and negative labelled comments in the dataset, which can help for minimising the false negatives in a predictive analytics scenario.

Another challenge in constructing a dataset of this kind is the data labelling process owing to the manual labour required for it. The labelling process can vary depending on the purpose of the study. Therefore, the definition and the specific instructions for labelling should be adhered to at all times during the process.

In this chapter, we present the dataset that we have collected for our experiments. We also discuss the methods of collecting and labelling the dataset, its structure as well as its suitability for offensive language and harassment detection in Arabic text. Figure 5.1 highlighted the parts to be covered in this chapter.

5.2 Previous Work in Datasets Utilisation for Offensive Language Detection

There is plenty of research conducted in offensive language detection and the related problems of cyberbullying and predator identification. In this section, we present an overview on the sources of the data utilised in previous work
with the goal to give a better grasp of what needs to be done in terms of preparing a dataset suitable for our study.

Previous studies use datasets collected from various social media platforms. Some studies choose to use the same benchmark dataset so they can compare the accuracy of their methods to others (Mehdad and Tetreault, 2016), while others studies choose to use newly collected datasets from different social networking sites, which would enable them to discover new characteristics. According to Berry and Kogan (2010) it is extremely difficult to find datasets containing cyberbullying and sexual predator activities owing to law enforcement practices. Here we list some of the datasets available in this regard.

Kontostathis (2009) reports that there are few reliable labelled datasets in predator communications. Many of the works that appeared in this regard are based on chat log transcripts from the website (Perverted-Justice-Foundation,
Chapter 5. Dataset Construction

2002) (PJ). PJ is a non-profit foundation to help in identifying cyber predators and paedophiles. This organisation recruits volunteers who pose as young people in online chat rooms and respond when an adult approaches them looking for a sexual relationship with minors. Police authorities use the information they acquire from tracking suspects for arresting violators. If the trial convicts the suspect based on these activities, then the chat logs are posted on PJ. There were 325 transcripts on the site as of July 2009 (Berry and Kogan, 2010) which went up to 623 chat transcripts as of July 2017, after more incidents had been detected. Examples of studies which employ data from this source are those by Potha and Maragoudakis (2014) and Parapar, Losada, and Barreiro (2012).

A second dataset was developed by Gauch, who collected chat logs for the task of detecting chat room topics (Bengel et al., 2004). Gauch’s project involved the development of a crawler that downloaded chat logs (ChatTrack). To the best of our knowledge the software ChatTrack is not available anymore. This dataset was used in some of the studies including an analysis of predator communications and is now considered outdated (Kontostathis, 2009).

A third dataset is publicly available on the website of the workshop Content Analysis for the Web 2.0 (FBM, 2009). This dataset was originally developed for the workshop with three challenges: text normalisation, opinion and sentiment analysis, and misbehaviour detection. The dataset is designed to be representative of what might be found in cyberspace. The data is collected from five different platforms, including Kongregate, Slashdot, Ciao, MySpace and Twitter. Since its release, this data has been used in many studies. Yin et al. (2009) utilise three of the CAW 2.0 workshop datasets: Kongregate, Slashdot and MySpace. Nandhini and Sheeba (2015) use two of the CAW 2.0 workshop datasets: Formspring.me and MySpace; while Huang, Singh, and Atrey (2014) and Singh, Huang, and Atrey (2016) employ the Twitter corpus from the CAW 2.0 dataset. But lately, I couldn’t reach the dataset on the provided link.
A fourth popular source of data is the community-driven project PAN (Bog-
danova and Rosso, 2012), where PAN stands for Plagiarism Analysis, Au-
thorship Identification, and Near-Duplicate Detection. PAN holds a series
of competitions on the application of automated text analysis for forensic re-
search with provided labelled training datasets. As an example, the PAN
competition involved sexual predator identification. For this particular task,
the datasets originate from diverse online sources:

- Conversations between convicted sexual predators and volunteers pos-
ing as minors from PJ (Perverted-Justice-Foundation, 2002).

- Logs of IRC discussions on various topics from IRC channels\textsuperscript{12}.

Examples of studies as an outcome of the PAN competition are those by (Villatoro-
Tello \textit{et al.}, 2012) and (Parapar, Losada, and Barreiro, 2012) which use the
dataset provided for sexual identification in the 2012 competition.

Other studies choose sources not mentioned above. For instance, (Reynolds,
Kontostathis, and Edwards, 2011) employ a dataset obtained directly from
Formspring.me. A number of studies choose Twitter as a source (Bellmore
\textit{et al.}, 2015; Huang, Singh, and Atrey, 2014; Sanchez and Kumar, 2011; Zhao,
Zhou, and Mao, 2016); for example, Sanchez and Kumar (2011) utilise Twit-
ter messages containing commonly used terms of abuse. Some works choose
more than one source such as Twitter and MySpace groups (Zhao and Mao,
2016; Soundar and Ponesakki, 2016; Zhang \textit{et al.}, 2016). Parime and Suri
(2014) also use a dataset obtained from MySpace. A few works employ data
collected from YouTube (Dinakar, Reichart, and Lieberman, 2011; Dadvar,
2014). For example, Dinakar, Reichart, and Lieberman (2011) employ a col-
lection of 4500 YouTube comments in their experiments. All the studies listed
in this section are for English.

\textsuperscript{1}http://www.irclog.org/
\textsuperscript{2}http://krijnhoetmer.nl/irc-logs/
5.3 Dataset Collection

According to Nalini and Sheela (2014), data collection for the study of cybercrime needs to focus mainly on selecting appropriate platforms to avoid both legal and technical issues. The YouTube platform has nearly two billion users (YouTube press statistics 2005) and the internet Live Stats website\(^3\) reports that there are 70,122 YouTube videos viewed in one second. YouTube does not prevent users from publishing offensive content, and in the case when such contents is published, it takes time to have it removed. Therefore, comments very likely contain a variety of speeches ranging from compliments to pejoratives, such comments are often available, which makes them a good source of data pertaining to cyber insulting.

Among various social media platforms, YouTube is the second-biggest social media platform with 1.8 billion internet users, after Facebook with 2.2 billion as of March 2018 (Kallas, 2018). YouTube is localised in 88 countries and can be accessed in 76 languages (YouTube press statistics 2005). It has a broad scope of users, from different age and gender groups (YouTube press statistics 2005). This diversity in types of users makes the material published on YouTube represent a wide range of societal attitudes and thus is appropriate as a source for investigating the interaction between people. Similar to other social media platforms, YouTube is a place for communication between people without limits, owing to the anonymity that is allowed, which opens the door to users to speak without restrictions and misbehave in their interaction with others. It is common to curse and offend others, and such incidents are increasing (Protalinski, 2011; Kawate and Patil, 2017). Many users, videos, and comments create a suitable environment for people to disturb and insult others through posting offensive comments in cyberspace. Table 5.1 contains some instances of offensive language in YouTube comments.

The comments in Table 5.1 suggest the presence of harassing. Thus, they are a proper source for the dataset required for our research.

\(^3\)http://www.internetlivestats.com/statistics/ [Online; accessed Jun 2018]
Table 5.1: Examples of instances of offensive language in YouTube comments in Arabic.

<table>
<thead>
<tr>
<th>Translation</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>A whore in every sense of the word</td>
<td>عاهرة بكل معنى الكلمة</td>
</tr>
<tr>
<td>He is mentally sick</td>
<td>هذا مريض نفسي</td>
</tr>
<tr>
<td>God’s curses on the one whose face looks like a monkey’s face</td>
<td>آلام خرا لعنة الله على وجه اللى مثل وجه القرد</td>
</tr>
<tr>
<td>She’s a fallen woman!</td>
<td>هي ساقطة اشتهتوضع منه</td>
</tr>
<tr>
<td>She’s a failure and artistically dead</td>
<td>فاشلة ومية فنيا من زمن</td>
</tr>
</tbody>
</table>

We utilised the open source tool for collecting the YouTube comments by Klostermann (2015), Figure 5.2 depicts the interface of this scraper.

Figure 5.2: The interface of the YouTube comment scraper used to collect the corpus.
5.4 Sampling

For the purpose of our study, we choose to select videos based on the YouTube channel they are posted in. It can be noticed that some of the channels that are keen to increase subscriber number are posting controversial videos about celebrities. This kind of videos attracts people who like to comment on rumours and they occasionally use insults in their comments. We picked videos with the highest number of comments from the selected channels, expecting that longer discussions would contain a high number of comments with offensive language, thus helping us in increasing the number of positives.

As the main target of our study is to detect comments containing offensive language, it is important to increase the number of positives (i.e. abusive comments) in the dataset and achieve balance between positives and negatives. The dataset would be imbalanced if the class of interest contains a small number of training instances (also named minority or positive class), while the rest of the most instances is the second class (also named majority or negative class) (Ali, Shamsuddin, and Ralescu, 2015).

The comments collected are written by casual spoken language. Usually, in this type of online communication different people have different writing style; moreover, there are frequent changes in the manner of how people communicate, in terms of their writing. These changes and the total lack of any structural rule make the processing of this kind of content a great challenge.

The comments in our dataset were collected in July 2017, and the upload dates for these videos range from 2015 to 2017. The dataset contains the following 14 attributes: CommentID, Username, Date, Timestamp, CommentText, Likes, HasReplies, NumberOfReplies, Replies.id, Replies.user, Replies.date, Replies.timestamp, Replies.likes, Replies.commentText. The next section presents an initial analysis of the dataset.
5.5 Descriptive Analysis of the Dataset

Our corpus consists of 167,549 YouTube comments posted by 84,354 users along with 87,388 replies posted by 24,039 users from 150 YouTube videos. These videos present controversial media footage of celebrities. This kind of footages provoke viewers, leading some of them to use offensive language in their comments. As it has been emphasised in the introduction, representativeness is an important factor. Thus, for learning more about what our dataset represents we conducted an analysis to identify whether or not the comments are written or read by people from a wide range of Arab nations. We consider the presence of people from different Arab countries in a conversation as a sign that these insults are understandable by the majority of them.

5.5.1 Word Frequency

As a first step, we computed the word frequency for all terms appearing in our dataset. A list of a total number of 250,382 unique words were obtained from the calculation of words frequency in the whole dataset. Then we sorted the list based on word frequency from the largest to smallest. We manually searched the first 30,000 words in the list for names of countries and nationalities and recorded their frequency. The choice of the manual search is due to misspellings that can be a reason to miss many words. The two histograms in Figures 5.3 and 5.4 illustrate the nationalities and countries, respectively. Figure 5.3 shows the frequency of nationalities mentioned in the first 30,000 words and Figure 5.4 shows the frequency of countries’ names mentioned in the first 30,000 words. To give an example, Table 5.2 presents some comments taken randomly from the dataset and referring to the nationality of some people. These examples of comments reflect the diversity of nationalities included in our dataset and suggests that people from the majority of the Arab region understand the insults in these comments.
Chapter 5. Dataset Construction

Figure 5.3: Frequency of nationalities mentioned in the first 30,000 words.

5.5.2 Further Discovery in the Dataset

We also discovered the use of profane words from languages other than Arabic in the collected comments. These occur in the form of a single word, a phrase or whole sentences in another language. Foreign words transcribed with the Arabic alphabet and Arabic words transcribed with a non-Arabic alphabet are also present; moreover, some sentences mix languages. We discovered 475 such comments, examples of which are presented in Table 5.3.

5.6 Annotation

Previous related studies employ a variety of strategies for labelling datasets. For example, Warner and Hirschberg (2012) manually label user comments and a corpus of websites. Huang, Singh, and Atrey (2014) choose to label about 13,000 messages to be positive or negative for cyberbullying detection. They asked three students to perform the job, and comments with disagreement in labelling were rejected. Dadvar (2014) proposes to ask three students
Figure 5.4: Frequency of countries’ names mentioned in the first 30,000 words.

to label posts to be either yes or no in terms of bullying. Posts on which at least two students agree are marked as positive, i.e. bullying. Reynolds, Kontostathis, and Edwards (2011) employ a dataset which includes 2,696 posts labelled with the use of Amazons Mechanical Turk service\(^4\). Kontostathis et al. (2013) hired three workers also on Amazons Mechanical Turk to label their dataset. Al-garadi, Varathan, and Ravana (2016) report that the dataset labelling is done with the assistance of three people as well. We have followed the same labelling process by employing three annotators. Details about our annotators and labelling process are stated next.

Out of the whole dataset, we picked nine videos with offensive comments which also have a relatively high total number of comments, assuming that the longer the conversation is the higher amount of offensive content it contains. These nine videos contain nearly 16,000 comments with the number of words in each comment ranging from 1 to 2,338 with average of 75 words. The labelling was performed on this sample out of the whole data collection. We assigned the labelling task to three annotators from three different

\(^4\)https://www.mturk.com/
### Table 5.2: Examples of comments made by people from different Arab regions.

<table>
<thead>
<tr>
<th>Comments and Translations</th>
<th>Arabic Translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frankly, Sama, you are a very kind person, you have a good heart, I wish God guides you to what he loves, I am Sudanese.</td>
<td>بصراحة يا سما أنتي شخصيتك زريزة جدا وباين عليك أنك طيب وقلبك أبيض فأنتي لك أن يهديك الله لما يحب ويرضى أنا سودانيه</td>
</tr>
<tr>
<td>I swore, you are the best and respected singer I have ever seen in my life. Greetings and my love from Libya.</td>
<td>والله العظيم أحسن فنانة واحترم فنانة شفتها في حياتي محبتي ليك حبيبي من ليبيا</td>
</tr>
<tr>
<td>Algerian, Ahlam has a good heart and I swear by God that the problem is that she doesn’t take into account what she says or how she behaves, which means that her impulsiveness could get her into a big mess.</td>
<td>جزائي إحلام دى قلبه طيب والله العظيم لكن لا يحسب لكلامها أو لنطقها أي ان عفويتها قد تؤدي بها إلى متهاات</td>
</tr>
<tr>
<td>Who told you he’s Iraqi! I heard him, his dialect is either Lebanese or maybe Syrian.</td>
<td>من كل لكم هذ عراقي أصلا سمعت لهجته!! لهجته اللبنانية مدري سورية</td>
</tr>
<tr>
<td>You’re Syrian, aren’t you?</td>
<td>آتي سوري مو</td>
</tr>
<tr>
<td>He’s Kuwaiti, isn’t she, why she doesn’t speak Kuwaiti... Mix of Egyptian, Lebanese and Syrian dialects.</td>
<td>هي مو كويتيه لبش منتحجي كويتي...مصري ع لبناني ع سوري</td>
</tr>
</tbody>
</table>

nationalities; one is Iraqi and the second is Egyptian, i.e. from the two nationalities most highly represented in the dataset as shown in Figures 5.3 and 5.4; moreover, they are from high-density urban areas. The third person is from Libya, a country and nationality with low representation in the dataset, and also from low-density urban area. The ages of three annotators are 44, 34 and 32, respectively. Two of them finished their third level education, one in information technologies, the other one in accounting and the third one quit university in his second year.

We asked the annotators to label offensive comments as positive and inoffensive comments as negative and leave unlabelled any comment they are
**Table 5.3:** Examples of non-Arabic comments and Arabic comments written in non-Arabic alphabet.

<table>
<thead>
<tr>
<th>Comment</th>
<th>Language</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fu** off</td>
<td>English</td>
</tr>
<tr>
<td>who the phu** is she ????? pffff</td>
<td>English (misspelled)</td>
</tr>
<tr>
<td>vous êtes trs trs mauvaise je sais pas pourquoi comporter comme a avec ahlam moi je l’aime bien c’est une femme vivante c’est trs mchant tous les gens qui sont jaloux d’elle il faut la laisser tranquille ahlam on t’aime les Marocains</td>
<td>French</td>
</tr>
<tr>
<td>poquito enferma</td>
<td>Spanish</td>
</tr>
<tr>
<td>inti asln min ma7loki mn li3rab chali fomk bsm ma thadri 3la lmgharba fhanti ya nakira</td>
<td>Arabic in Latin alphabet transcription</td>
</tr>
<tr>
<td>Tfu mnin awdi nti mlmrib tfuu</td>
<td>Arabic in Latin alphabet transcription</td>
</tr>
<tr>
<td>لا يا بنت تعني أنت مزجدي دين فنانة بنت بورنو دين تعمل أغاني سكس مش فديوكلوب</td>
<td>English mixed with Arabic, both in Arabic alphabet transcription</td>
</tr>
</tbody>
</table>

not sure about. We also explained to them what can be considered offensive language by providing them with a definition in Section 3.4, and we asked them to adhere to it all the time. The three annotators agreed on 10,715 comments, the inter-annotator agreement is 71% of the whole sample. The number of comments with at least one disagreement is 4,335, and the number of unlabelled comments by at least one person is 848. We excluded these 848 comments. A summary of these numbers is presented in Table 5.4. In addition to the attributes mentioned in the last paragraph at section 5.4, four more attributes have been added, three of which represent the opinion of the three annotators, and the fourth attribute is the final decision about the comment, whether it is offensive or not, based on the agreement between the three annotators.

For accomplishing the construction of the dataset we adopt the two following scenarios:

**Scenario 1:** Label as offensive the comments on which all annotators agree,
TABLE 5.4: Number of agreements and disagreements between annotators in the labelled dataset.

<table>
<thead>
<tr>
<th>Comments on which all annotators agree</th>
<th>Inter-annotator agreement</th>
<th>At least one annotator disagree</th>
<th>Comments unlabelled by at least one</th>
</tr>
</thead>
<tbody>
<tr>
<td>10715 comments</td>
<td>71%</td>
<td>4335 comments</td>
<td>848 comments</td>
</tr>
</tbody>
</table>

and label as inoffensive the rest.

**Scenario 2**: Label as offensive the comments on which at least two annotators agree, and label as inoffensive the rest.

In the first scenario the number of comments labelled as positives and negatives are 3,532 and 11,518, respectively, i.e. 23% positives. In the second scenario the number of comments labelled as positives and negatives are 5,817 and 9,233, respectively, i.e. 39% positives, as shown in Table 5.5. Table 5.5 summarises the number of positives in both scenarios for the whole sample. Moreover, the Inter-annotator agreements are calculated between each pair of annotators using kappa statistics and presented in Table 5.6. The R code used to calculate Inter-annotator agreements is presented in Appendix B.

<table>
<thead>
<tr>
<th>Scenario 1</th>
<th>Scenario 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>labelled offensive by three annotators</td>
<td>labelled offensive by two annotators</td>
</tr>
<tr>
<td>Percentage of positives</td>
<td>Percentage of positives</td>
</tr>
<tr>
<td>3532 comments</td>
<td>5817 comments</td>
</tr>
<tr>
<td>23%</td>
<td>39%</td>
</tr>
</tbody>
</table>

**TABLE 5.6**: Inter-annotator agreement using kappa statistics.

<table>
<thead>
<tr>
<th>Inter-annotator agreement between</th>
<th>Kappa statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Egyptian and Libyan</td>
<td>0.698</td>
</tr>
<tr>
<td>Iraqi and Libyan</td>
<td>0.579</td>
</tr>
<tr>
<td>Egyptian and Iraqi</td>
<td>0.512</td>
</tr>
</tbody>
</table>

**Interpretation of Cohen’s kappa coefficient** (McHugh, 2012; Ziai, 2017).

The Kappa statistic varies from 0 to 1, where:

- 0 = agreement equivalent to chance.
- 0.1 to 0.20 = slight agreement.
- 0.21 to 0.40 = fair agreement.
- 0.41 to 0.60 = moderate agreement.
- 0.61 to 0.80 = substantial agreement.
- 0.81 to 0.99 = near perfect agreement.
- 1 = perfect agreement.

5.7 Limitations

There are two notable limitations associated with this dataset. The first one is the use of a nine-video sample. Because of this, the results cannot be generalised to the entire Arab population. However, this sample introduces a large segment of Arab social media users from the Arab East, which has the highest percentage of internet users in the Arab world. Therefore, we believe that there is a strong confidence that the randomisation process along with the choice of highly popular videos related to celebrities known throughout the entire Arab world as well as the relatively high number of comments collected minimise the potential effect of limiting the data collection to nine videos.

Another limitation arises from the choice of annotators to label the dataset. Two of them are from nations highly represented among the authors of comments in this sample, and the third annotator is from a nation with a relatively lower representation in the dataset. We made this choice in order to ensure that comments labelled as offensive would be considered such throughout the entire Arab region. The inter-annotator agreement between the three annotators is 71%. This percentage is very reasonable, especially when we take into consideration another factor which is that different people have different perspectives on the same comment in terms of its offensiveness. As

it has been pointed out, the responses of participants towards potentially offensive language varies depending on the context (Ofcom, 2015). The annotators’ views on each comments were also influenced by their age, gender and personal experiences. At the same time, we want to point out that the results of this study shows moderate to high inter-annotator agreement with the kappa statistic between each pair of annotators being at an acceptable level (see Table 5.6), thus we believe we minimised the effect of the limitation associated with the choice of annotators.

5.8 Summary

In this chapter we introduce a dataset of YouTube comments in Arabic together with a statistical analysis of it. We collected the data according to the principles of availability, representativeness, heterogeneity and balance, thus ensuring that it can be applied for training predictive analytics models for detection of abusive language in online communication in Arabic. Along with conversations scripted in Arabic, this dataset also includes foreign words transcribed with the Arabic alphabet as well as Arabic words transcribed with a non-Arabic alphabet. To the best of our knowledge, this is the first dataset of YouTube comments in Arabic of this kind.

In terms of representativeness and heterogeneity, the dataset portrays a real case of offensive language comments in Arabic by a wide variety of Arab YouTube users from various nationalities. In terms of balance, our dataset contains 39% positives based on the agreement of two out of three annotators from three different Arab nationalities. It is interesting to note that our annotators agree on their evaluation of the language as either offensive or inoffensive in 71% of the cases. This is something we anticipated, and we find that it confirms the representativeness and heterogeneity that we were seeking.
Chapter 5. Dataset Construction

It can be noted that our dataset is a collection of casual online writing in Arabic dialects, which typically do not have formally defined grammars. Therefore, we believe that Nobata et al. (2016) argument could not be applied to our case.

We conclude that this dataset is appropriate for employment as a training dataset in the context of machine learning. The next chapter presents experiments of a variety of pre-processing techniques and machine learning algorithms for building accurate models for detecting offensive content in online communication in Arabic.
Chapter 6

Machine Learning Approach to Detection of Offensive Language in Online Communication in Arabic

6.1 Introduction

This chapter presents the setup and results of several ML experiments, conducted with the dataset introduced in Chapter 5. The experiments aim at studying the impact of various text pre-processing, feature-extraction and feature-selection techniques (described in Chapter 4) on the accuracy of a document classifier for detection of offensive language in online communication in Arabic. Regarding data pre-processing, our experiments focus on filtering out noisy characters and normalising inconsistencies present in casual online writing in Arabic. The combined effect of these data pre-processing techniques and a few feature-selection methods is then evaluated by training document classifiers. The ultimate goal of this study is to recommend an optimal workflow for training such document classifiers, and thus provide an answer to research question 3, introduced in Chapter 1.
6.2 Pilot ML Experiments

As a first objective in our text mining experiments, we set to examine the effect of text pre-processing to a classifiers accuracy when only word-level and N-gram features are employed. Previous research suggests that Support Vector Machines (SVM) is most successful and frequently used machine learning (ML) algorithm for document classification (Joachims, 1998; Burges, 1998; Cortes and Vapnik, 1995; Vapnik, 1998; Cristianini and Shawe-Taylor, 2000). Accordingly, we selected to build an SVM classifier using the Python ML library scikit-learn (Pedregosa et al., 2011) with both word-level and N-gram features (Buitinck et al., 2013).

6.2.1 Text Pre-Processing

Typically, text pre-processing is performed once, before applying a ML algorithm with a document corpus as a training dataset. It involves operations, such as tokenization, filtering and normalisation (Mathiak and Eckstein, 2004; Iiritano and Ruffolo, 2001; Neto et al., 2000). Arabic is a Semitic language with a script from right to left, and it has twenty-eight letters that form words; these letters are also used in other languages, such as Persian and Urdu (Atallah and Omar, 2008). Thus, letters from these languages may appear in casual Arabic conversations over the internet. The three predominant types of words in Arabic are nouns, verbs and particles (Darwish, 2002). Therefore, results in tokenization, filtering and normalisation from previous research with other languages (e.g., English) are generally applicable to Arabic too. However, there are some notable differences. For example, since some Arabic letters are similar phonetically, users on social media typically misspell words by using the wrong but phonetically similar letters (details presented in Section 6.2.1.4).
6.2.1.1 Tokenization

As introduced in Section 4.6.2, tokenization is the process of partitioning alphabetic sequences into tokens. In the tokenization phase, choosing characters as delimiters depends on the application as some characters may or may not be delimiters in different scenarios (Weiss et al., 2010; Witten et al., 2016). Characters such as space, tab and newline are most of the time regarded as delimiters and are not counted as tokens; these are often called white space. Other characters can be used as delimiters too, such as ( ) ¿ ! ? and ,. Arabic text has the same characteristics as English in this respect. Words in Arabic are separated by white space. However, in casual writing online, words are sometimes separated by a comma only (without white space). Some examples of words separated by comma/s only in the dataset introduced in Chapter 5 are "ék@Qå", "Pñ ®K", "éÊ t" and "brawahk". more of these examples are presented in Table 4.2. Occasionally, in informal writing online, there is little attention paid to the formally correct use of delimiters between words. In this work, we also consider the case when the comma character is the only delimiter between words.

6.2.1.2 Filtering

As it has stated in Section 4.6.3 filtering is the removal of punctuation marks, commas, diacritics (in Arabic) and selected words from the documents. Standard filtering excludes the most frequent words in the documents, such as articles, conjunctions and prepositions; these are known as stop words. Stop words occur excessively and typically do not contribute significant information for the purpose of text mining. Also, words that occur very rarely are likely to have no statistical significance and can be removed as well as outliers (Kilgarriff and Grefenstette, 2003; Blanchard, 2007; Witten et al., 2016). The process of filtering helps to minimise the size of the number of features in the dataset, which would otherwise represent impediments to text mining (Saad and Ashour, 2010).
The pilot experiment is conducted by removing the list of stop words provided by the Natural Language Toolkit (NLTK) (Bird, Klein, and Loper, 2009), which list contains 248 words. Additionally, we only keep alphabetic characters, both Arabic and Latin, while other characters, such as punctuation characters and all other special characters, including numbers, are removed. We also remove the kashidas, as kashidas are mere word elongation characters (Darwish and Magdy, 2014).

In addition, we utilized the stemmer ARLSTem (see Section 4.6.5) to remove prefixes, suffixes and infixes from words. This stemmer is available in NLTK library in Python. Its documentation\(^1\) states that “this stemmer was evaluated and compared to several other stemmers using Paices parameters (under-stemming index, over-stemming index and stemming weight), and the results showed that ARLSTem is promising and producing high performances”. It has also been mentioned that “this stemmer is not based on any dictionary, therefore it can be utilised online efficiently”.

### 6.2.1.3 Normalisation

Section 4.6.4 provides an introduction to the normalisation phase. In the normalisation, letters are replaced by other letters for further improvement of the performance of text mining operations. This includes removing diacritics, replacing ّ,ّ, ّ and ّ by ّ (Darwish and Magdy, 2014; Sallam, Mousa, and Hussein, 2016), as well as replacing the Arabic-Indic numerals with Arabic numerals.

Furthermore, we replaced the originally Persian and Urdu letters that appear in the text by equivalent Arabic letters (Darwish and Magdy, 2014).

\(^1\)https://www.nltk.org/api/nltk.stem.html
6.2.1.4 Extra Normalisation

As it has been stated in Section 4.3 there are some letters in Arabic which are confusing for many users, and they use them interchangeably by mistake, because of the phonetic similarity between them in some dialects, or perhaps as a result of poor spelling. Table 6.1 contains examples of misspelled words in our dataset. We experimented with additional normalisation for addressing this problem with an attempt to enhance the results of text mining. We refer to it as extra normalisation. It consists in replacing ص by ص by ص and ض by ض by ض.

Table 6.1 contains examples of misspelled words in our dataset.

<table>
<thead>
<tr>
<th>Translation of the word</th>
<th>Correction of the word</th>
<th>Misspelled word</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scandal</td>
<td>ضيحة</td>
<td>ضيحة</td>
</tr>
<tr>
<td>She laughs</td>
<td>ضحك</td>
<td>ضحك</td>
</tr>
<tr>
<td>Favourite</td>
<td>مفضل</td>
<td>مفضل</td>
</tr>
<tr>
<td>She learnt by heart</td>
<td>حافظة</td>
<td>حافظة</td>
</tr>
<tr>
<td>Great</td>
<td>عظم</td>
<td>عظم</td>
</tr>
<tr>
<td>Person name Kazem</td>
<td>كاظم</td>
<td>كاظم</td>
</tr>
<tr>
<td>Dirty</td>
<td>سماء</td>
<td>سماء</td>
</tr>
<tr>
<td>Slut</td>
<td>الفاسقة</td>
<td>الفاسقة</td>
</tr>
</tbody>
</table>

6.2.1.5 Experimental Setup

We trained an SVM classifier Cristianini and Shawe-Taylor (2000) using word-level features (see Section 3.2.1.1 for details). We first trained our model on a dataset without applying any data pre-processing. The outcome of this run is used as a baseline. Next, we trained an SVM classifier with a pre-processed version of our dataset (see Section 6.2.1). Also, we experimented training SVM with and without the extra normalisation described in Section 6.2.1.4. Finally, we experimented with N-gram level features, as well. For the purpose of training an SVM classifier, it is required to transform the corpus of
text documents (comments in our case) to a document-term matrix, an entry in which is the number of occurrences of a particular word in a particular document.

For evaluating a classifier’s accuracy, we split the dataset into a training set and a test set and followed the widely accepted approach to apply 5-fold cross validation (for details see Section 4.8.1) and observe precision, recall and the F1-score (Berry and Kogan, 2010). In addition, we plotted the receiver operating characteristics (ROC) curves. ROC curves are two-dimensional graphs in which the false positives rate $fp$ is plotted on the x-axis and the true positives rate $tp$ is plotted on the y-axis (for more details see Section 4.8.2). They show the relative trade-off between benefit $tp$ and cost $fp$ and are a further indicator for the accuracy of a predictive model (Fawcett, 2006).

6.2.2 Results and Discussions

In all our experiments we used the labelled dataset of 15,050 YouTube comments introduced in Chapter 5 as a training dataset. Our baseline experiment is conducted with word-level features and without applying any data pre-processing. Its precision and recall for the class of offensive comments are 0.83 and 0.65, respectively, while its overall accuracy is 0.85. Compared to the results of (Mubarak, Darwish, and Magdy, 2017), the recall is 20% better, while the precision is 15% worse. The difference in the results can be due to the different training datasets, thus it is hard to draw definite conclusions based on it.

A summary of our experimental results is presented in Table 6.2. We apply pre-processing including tokenization, filtering, normalisation and extra normalisation. When no stemming is applied, the precision is 2%, and recall 4% better than the baseline. Stemming further improves precision by 3%, and recall by 8%. We observed that the experiment with applying extra normalisation (see Section 6.2.1.4), however, it does not improve the results by
Chapter 6. Machine Learning Approach to Detection of Offensive Language in Online Communication in Arabic

Table 6.2: Comparative performance of trained SVM classifiers.

<table>
<thead>
<tr>
<th></th>
<th>Precision</th>
<th>Recall</th>
<th>F1-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>0.83</td>
<td>0.65</td>
<td>0.74</td>
</tr>
<tr>
<td>Pre-processing applied</td>
<td>0.85</td>
<td>0.69</td>
<td>0.76</td>
</tr>
<tr>
<td>Pre-processing applied with stemming</td>
<td>0.88</td>
<td>0.77</td>
<td>0.82</td>
</tr>
<tr>
<td>Pre-processing and stemming applied stop words not removed</td>
<td>0.88</td>
<td>0.77</td>
<td>0.82</td>
</tr>
<tr>
<td>N-grams (1-5)</td>
<td>0.83</td>
<td>0.80</td>
<td>0.81</td>
</tr>
<tr>
<td>N-grams (1-5) and stemming</td>
<td>0.81</td>
<td>0.78</td>
<td>0.80</td>
</tr>
</tbody>
</table>

more than 1% for precision and 2% for recall. Overall, the results give evidence of the usefulness of the data pre-processing step and noise eliminating in terms of stemming and extra normalisation. The classifier performance using 5-fold cross validation is 90.05%. ROC curves have also been used to illustrate the results. The ROC curves (for the two classes of offensive and inoffensive comments) in the case when the SVM classifier is built after data pre-processing and stemming are shown in Figure 6.1. The curves are closer to the left and top borders of the plot, which is another indicator that the accuracy of our classifier is high. With regard to the stop words removing that does not show any difference in our results.

Additionally, we conducted experiments to study the impact of N-gram co-occurrence as features on the classifier performance (see Section 4.6.6.1 for details). We attempted a range of values of N from 1 to 10 with the best results obtained for N ∈ (1..5). We observed a major improvement of the recall value, 15% improvement over the baseline; however, no improvement of precision. In comparison to the baseline, stemming combined with the use of N-gram features increases recall by 13%; though, precision decreases by 2%.

We report the impact of word-level features and popular pre-processing methods, including extra normalisation, on the performance of an SVM classifier trained to detect offensive comments. The results presented give an evidence that these methods could improve the classifier performance.
FIGURE 6.1: ROC curves for the two classes (offensive/class 0 and inoffensive/class 1 comments) when an SVM classifier is applied after pre-processing and stemming.

This experiment addresses Objective 3, Research Question 3: what is the impact of removing noisy data and, in general, data pre-processing on the accuracy of detecting offensive incidents on social media platforms? We have observed that data pre-processing with stemming can be leveraged to enhance the detection of offensive language in casual Arabic text used in social media platforms. In addition, the utilisation of N-gram features improves the classifier’s performance.

6.3 ML Experiments with Additional Syntactic and Linguistic Features

Typically, document corpora contain raw unstructured text. Thus, the majority of works in the sphere of document classification employ words as features, as the easiest and most straightforward features that can be extracted from raw text (Gabrilovich and Markovitch, 2006). Feature extraction involves one or more of the following methods:

- extraction of features from raw data (Bosch, 2017),
- transformation of existing features to new features (Motoda and Liu, 2002; Bosch, 2017), which may include combining of multiple existing features into a single one (Markovitch and Rosenstein, 2002; Bosch, 2017).

The remainder of this chapter addresses the components in the diagram depicted in Figure 1.2 in Chapter 1. These are highlighted in Figure 6.2. That is, we carry out experiments to examine the marked sections in Figure 6.2, which include feature extraction and selection with the goal of discovering the set of features that allow training an accurate predictive model. In this process, we study the impact of various feature sets on the models accuracy.

![Diagram](image)

**Figure 6.2:** The final stage in the components of the proposed system, the text mining process.
6.3.1 Feature Space

Similar to the pilot experiment in Section 6.2, we pre-process each comment and stem it using ARLSTem (Abainia, Ouamour, and Sayoud, 2017). Next, we employ Stanford Log-linear Part-Of-Speech Tagger for Arabic (Stanford, 2011; Diab, Hacioglu, and Jurafsky, 2004), and generate N-gram features for $N=1..4$, i.e. unigrams, bigrams, trigrams and quadrigrams features, each weighted according to the tf-idf model. Stop words are not removed.

In addition to that, proper nouns and a country name (Egypt) that occur extensively in the dataset are excluded during the feature-selection process. These nouns and a country names are listed in Table 6.3. As it can be observed in Table 6.3, parts of these words are chopped off. That is, their exclusion is done after pre-processing in order to exclude all their variations.

<table>
<thead>
<tr>
<th>The name in English</th>
<th>The name in Arabic</th>
<th>The name after pre-processing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ahlam</td>
<td>أحلام</td>
<td>حلم</td>
</tr>
<tr>
<td>Kazim</td>
<td>كاظم</td>
<td>كاظم</td>
</tr>
<tr>
<td>Saddam</td>
<td>صدام</td>
<td>صدام</td>
</tr>
<tr>
<td>Qaysar</td>
<td>قيسمر</td>
<td>قيسمر</td>
</tr>
<tr>
<td>Egypt</td>
<td>مصر</td>
<td>مصر</td>
</tr>
</tbody>
</table>

In addition, we extract other features (see Section 3.2.1.1), which include:

- number of characters in a comment,
- number of tokens in a comment,
- number of words (i.e. alphanumerical tokens in a comment after pre-processing, see Section 6.2.1),
- number of unique words in a comment (i.e. unique alphanumerical tokens in a comment after pre-processing, see Section 6.2.1),
- number of words with consecutive repeated characters in a comment,
- number of consecutive repeated characters in a comment,
- number of likes of a comment,
- number of obscene words in a comment. We used a list that includes the 415 such words used by (Mubarak, Darwish, and Magdy, 2017) plus 38 additional obscene words added by us. For each comment, we count the number of obscene words with and without pre-processing, and then we compare these two numbers, and pick the larger. With this we ensure that most of the obscene words in a comment are counted.

The total number of features generated is 9,435.

6.3.2 Experimental Setup

Due to the high dimensionality of our dataset, dimensionality reduction is essential for both reducing the training time and achieving high accuracy of the predictive model being trained. Thus, we conduct several experiments to find out the best method/s for dimensionality reduction. These include methods for feature transformation and feature selection. Several features selection and feature transformation methods, available in the scikit-learn Python module (Pedregosa et al., 2011), are examined. These include:

- training logistic regression (LR) and support vector machines (LinearSVC) models with L1 regularisation and selecting features based on their regularised weights;
- feature ranking with recursive feature elimination using logistic regression (RFE);
- a decision tree classifier (ExtraTreesClassifier), for details see Section 4.7.1.2;
- Singular-Value Decomposition (SVD) (see Section 4.7.1.1).

These methods have been chosen for a few different reasons. First of all, SVM’s reputation in the text classification research encourages us to examine its performance in feature selection. The next one, logistic regression with L1
regularisation (LR-L1) is used for feature selection in (Davidson et al., 2017) and reportedly results in high-accuracy models for English. We chose to test this method with Arabic text as well. In addition to experimenting with LR-L1 in the same way as (Davidson et al., 2017) (see Section 6.3.2.1 for details), we also applied as a method for feature ranking with recursive feature elimination (RFE). ExtraTreesClassifier is used for feature selection as one of the available tree-based ensemble methods (see Section 4.7.1.2). Finally, SVD is a very well-known technique for dimensionality reduction (see Section 4.7.1.1) that is also tested.

Out of the whole dataset, 15% is held out for testing, while the rest 85% are used as a training dataset.

6.3.2.1 Initial Experiment

With a set of features being selected, we build an SVM classifier to initially evaluate the importance of each of the five feature selection methods listed above. We choose SVM because it tends to be the ML method of choice in the majority of text mining studies (Joachims, 1998; Burges, 1998; Cortes and Vapnik, 1995; Vapnik, 1998; Cristianini and Shawe-Taylor, 2000). The results from this initial experiment are summarised in Table 6.4.

The first feature-selection method applied is logistic regression with L1 regularisation (LR-L1). This method has been applied in (Davidson et al., 2017) and reportedly leads to high-accuracy models. We apply it with the same settings in our exterminates. These include penalty=“l1” (for details see Section 4.7.5) and C=0.01 (a positive number which is an inverse of the regularisation strength, i.e. smaller values specify stronger regularisation 2). With these settings, the number of features being selected is 151, based on this we choose to select 150 the most important features for the rest of the feature selection methods. The second feature-selection method is feature ranking with recursive feature elimination using logistic regression (RFE). In this method,

Chapter 6. Machine Learning Approach to Detection of Offensive Language in Online Communication in Arabic

103

Table 6.4: Results from training an SVM classifier with a variety of alternative feature-selection methods.

<table>
<thead>
<tr>
<th>Feature selection/transformation method</th>
<th>Number of features</th>
<th>Accuracy of the initial SVM model</th>
<th>Precision</th>
<th>Recall</th>
<th>F1-score</th>
<th>Median Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>LinearSVC</td>
<td>150</td>
<td>0.88</td>
<td>0.70</td>
<td>0.78</td>
<td>0.83</td>
<td></td>
</tr>
<tr>
<td>LR-L1</td>
<td>151</td>
<td>0.88</td>
<td>0.74</td>
<td>0.80</td>
<td>0.84</td>
<td></td>
</tr>
<tr>
<td>RFE</td>
<td>150</td>
<td>0.95</td>
<td>0.54</td>
<td>0.69</td>
<td>0.79</td>
<td></td>
</tr>
<tr>
<td>ExtraTreesClassifier</td>
<td>150</td>
<td>0.88</td>
<td>0.73</td>
<td>0.78</td>
<td>0.83</td>
<td></td>
</tr>
<tr>
<td>SVD</td>
<td>150</td>
<td>0.75</td>
<td>0.68</td>
<td>0.71</td>
<td>0.73</td>
<td></td>
</tr>
</tbody>
</table>

the number of features to be selected needs to be set in advance. For the initial experiment we set this to be 150. The third dimensionality-reduction method is LinearSVC which takes the same parameter as LR-L1. The fourth dimensionality-reduction method is ExtraTreesClassifier. We keep the default settings and then we pick the 150 most important features captured by ExtraTreesClassifier in our experiments. The fifth dimensionality-reduction methods are SVD, we pick the 150 most important features as well. Table 6.4 presents results obtained from employing features harvested by these five methods.

From these five methods, we pick the two with the highest score (see Table 6.4). In the following subsections, we give further details about these two dimensionality reduction methods, as well as the scores that have been achieved by combining the features that have been selected by them.

Several ML algorithms, used in previous research, are also examined with the features selected by the two best methods. These are Logistic Regression, SVM, Naive Bayes, Decision Trees, and Random Forest (Xiang et al., 2012; Davidson et al., 2017). All our experiments are based on the ML algorithms available in the scikit-learn Python module (Pedregosa et al., 2011), namely LinearSVC (SVM), naive_bayes.MultinomialNB (Naive Bayes), DecisionTreeClassifier, RandomForestClassifier and LogisticRegression.

Each model is evaluated using standard 5-fold cross-validation. We record the accuracy of a model on each of the file folds and use these values to plot and compare the five models.
6.3.2.2 Experiment I

First, we trained LR-L1 to select the most relevant features out of 9,435 total features generated. It resulted in 151 features being selected, listed in Appendix C. Table 6.6 shows the distribution of these features. The 5 other features include number of words that have consecutive repeated characters, number of consecutive repeated characters, total number of characters in a comment, number of unique words and number of words in a comment, number of obscene words in a comment.

Next, we train the five ML algorithms listed in Section 6.3.2.1 with the 151 features selected by LR-L1. LinearSVC had the highest accuracy with precision, recall and F1-score for the class of offensive comments 0.88, 0.73 and 0.79, respectively. The performance of the five models illustrated by Boxplot, see Figure 6.3.

Discussion  The boxplot in Figure 6.3 shows that the most accurate model is LinearSVC with nearly 90% accuracy. The other from top to bottom are: MultinomialNB with above 85% accuracy, LogisticRegression with 85% accuracy,
DecisionTreeClassifier with below 80% accuracy, and finally RandomForest-Classifier with below 75% accuracy. Having said that, the medians in the box-plot indicate that the performance of LogisticRegression is better than MultinomialNB. Thus, if we take the median in consideration LogisticRegression jumps ahead of MultinomialNB. Interestingly, even though RandomForest-Classifier has the worst score, it has the lowest variance, while LinearSVC fluctuates the most.

### 6.3.2.3 Experiment II

In the second experiment, we used RFE as a feature-selection method. By default, RFE reduces the number of features by 50%, which is insufficient in our case. Alternatively, the number of RFE-selected features can be pre-defined. Given that 150 features were selected in Experiment I (see Section 6.3.2.2), we run RFE with a pre-defined number of selected features in the range 50-500, and evaluate each feature set by training a LinearSVC model. Table 6.5 summarises the characteristics of the trained LinearSVC models. It can be observed that the highest precision and recall are observed when RFE selects 200 features. Table 6.6 presents the distribution of these features. The feature listed as other features is number of words that have consecutive repeated characters.

The results of training the same five models as in Experiment I with the 200 selected features show that LinearSVC and MultinomialNB tend to perform better than the other three ML algorithms (see Figure 6.4). While the precision of LinearSVC is slightly higher than in Experiment I, the recall is significantly lower.

**Discussion**  The boxplots in Figure 6.4 demonstrate that the highest accuracy is achieved again by LinearSVC, just below 90%. MultinomialNB is very
Table 6.5: Accuracy of the LinearSVC classifier trained with different number of features selected by RFE.

<table>
<thead>
<tr>
<th>Number of features</th>
<th>Precision</th>
<th>Recall</th>
<th>f-measure</th>
<th>Median Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>0.95</td>
<td>0.32</td>
<td>0.48</td>
<td>0.72</td>
</tr>
<tr>
<td>100</td>
<td>0.95</td>
<td>0.32</td>
<td>0.48</td>
<td>0.72</td>
</tr>
<tr>
<td>150</td>
<td>0.95</td>
<td>0.54</td>
<td>0.69</td>
<td>0.79</td>
</tr>
<tr>
<td>200</td>
<td>0.95</td>
<td>0.55</td>
<td>0.70</td>
<td>0.81</td>
</tr>
<tr>
<td>250</td>
<td>0.94</td>
<td>0.62</td>
<td>0.75</td>
<td>0.83</td>
</tr>
<tr>
<td>500</td>
<td>0.91</td>
<td>0.70</td>
<td>0.80</td>
<td>0.82</td>
</tr>
</tbody>
</table>

Figure 6.4: Results of 5-fold cross validation of five models trained with the 200 features selected by RFE.
close to it and significantly more accurate than when run with features selected by LR-L1 in Experiment I 6.3.2.2. LogisticRegression has accuracy below 85%, which is worse than its accuracy in Experiment I. DecisionTreeClassifier has accuracy 80%, which is better than its accuracy in Experiment I 6.3.2.2. Finally, RandomForestClassifier has accuracy of below 70% which is worse than its accuracy in Experiment I. As in Experiment I, RandomForestClassifier has the lowest accuracy, but still the lowest variance, while LinearSVC and MultinomialNB fluctuate more than the other three.

### 6.3.2.4 Experiment III

In a third experiment, we combined the selected feature sets from the previous two experiments, resulting in 331 unique features, listed in Appendix C. Table 6.6 shows the distribution of these features; the 6 other features include number of words that have consecutive repeated characters, number of consecutive repeated characters, total number of characters in a comment, number of unique words and number of words in a comment.

We train the five models again with the same dataset and these 331 features. The results, depicted in Figure 6.5, show that LinearSVC tends to perform better than the other four again. Its precision, recall and F1-score are 0.89, 0.76 and 0.81, respectively (see Table 6.7).

---

**Table 6.6: Distribution of the features selected by LR-L1 and RFE.**

<table>
<thead>
<tr>
<th>Feature Type</th>
<th>LR-L1</th>
<th>RFE</th>
<th>LR-L1 ∪ RFE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word unigrams</td>
<td>140</td>
<td>73</td>
<td>194</td>
</tr>
<tr>
<td>Word bigrams</td>
<td>23</td>
<td>23</td>
<td></td>
</tr>
<tr>
<td>PoS tag unigrams</td>
<td>3</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>PoS tag bigrams</td>
<td>2</td>
<td>7</td>
<td>9</td>
</tr>
<tr>
<td>PoS tag trigrams</td>
<td>41</td>
<td>41</td>
<td></td>
</tr>
<tr>
<td>PoS tag quadrigrams</td>
<td>55</td>
<td>55</td>
<td></td>
</tr>
<tr>
<td>Other features</td>
<td>6</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>Total</td>
<td>151</td>
<td>200</td>
<td>331</td>
</tr>
</tbody>
</table>
Table 6.7: Accuracy of the LinearSVC model trained with the combined features selected by LR-L1 and RFE.

<table>
<thead>
<tr>
<th>The approach</th>
<th>Number of features</th>
<th>Precision</th>
<th>Recall</th>
<th>f-measure</th>
<th>Median Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>RFE ∪ LR-L1</td>
<td>200 ∪ 151 = 331</td>
<td>0.89</td>
<td>0.76</td>
<td>0.81</td>
<td>0.84</td>
</tr>
</tbody>
</table>

Figure 6.5: Results of 5-fold cross validation of five models trained with the combined 331 features by LR-L1 and RFE.

Discussion In this experiment, LinearSVC fluctuates less compared to Experiments I and II. The second best model, MultinomialNB, has accuracy between 85% and 90%. LogisticRegression has 85% accuracy once again; the same as in Experiment I 6.3.2.2. On the other hand, if we take the median in the consideration, MultinomialNB and LogisticRegression are at the same level. DecisionTreeClassifier has accuracy below 80%, which is the same as in Experiment I. RandomForestClassifier has the lowest variance once again, however at the cost of 70% accuracy, which is lower than its accuracy in both Experiment I and Experiment II.

The combination of features selected by RFE and LR-L1 improves the results of LinearSVC compared to the use of any of the two methods alone, but with only one drawback which is the negative effect on precision compared to using the features selected by RFE only.
6.4 Discussion

As already stated, the majority of the research in offensive language detection is in English. In our research we have been inspired by the work of Davidson et al. (2017) on automated hate speech detection and the problem of offensive language in English. They employ a range of features along with logistic regression with L1 regularisation for feature selection. Their best model has an overall precision of 0.91, recall of 0.90, and F1-score of 0.90.

Very little research has been done on this topic for Arabic. Mubarak, Darwish, and Magdy (2017) experiment with a dataset of total 1,100 instances, which consist of 100 tweets and 1000 replies to them. They experiment with two types of features: Log Odds Ratio (LOR) (Forman, 2008) generated for each word unigram and bigram that occurs at least 10 times, and a list of obscene words called SeedWords (SW). Their results when using both types of features (i.e., LOR for bigrams + SW) for precision, recall and F1-score are 89%, 45% and 60%, respectively. When only LOR for unigrams are used, the precision, recall and F1-score are 98%, 41% and 58%, respectively. A comparison of the results of Mubarak, Darwish, and Magdy (2017) to ours would not be indicative about the performance of our models due to the different scale of our study (i.e., the significantly larger training dataset utilized in our study and the larger size of YouTube comments compared to tweets).

Similar to Davidson et al. (2017), we experimented with logistic regression with L1 regularisation (LR-L1) and RFE for feature selection. The distribution of the 151 features, selected by LR-L1, is presented in Table 6.6. The majority (140) of the features in this group are word unigrams, most of which are offensive if used to describe a person either on their own or in combination with other words. The six features which we call other features are

- number of words with consecutive repeated characters in a comment,
- number of unique words in a comment,
- number of words,
Chapter 6. Machine Learning Approach to Detection of Offensive Language in Online Communication in Arabic

- number of consecutive repeated characters in a comment,
- number of characters in a comment and
- number of obscene words in a comment.

The remaining five selected features are PoS tags: three unigrams and two bigrams.

We discovered that when the LR-L1 selected features are used for training a classifier for detection of offensive language in Arabic, LinearSVC performed best out of the five classifier-training algorithms we experimented with, followed by Naive Bayes. The 5-fold cross validation results suggest, though, that the LinearSVC model has relatively high variance, as its highest accuracy (for one of the five folds) is about 90%, while its lowest accuracy is around 72%.

A second group of 200 features were selected using RFE. Their distribution is presented in Table 6.6. RFE selected 96 word-unigrams and -bigrams, most of which are offensive if used to describe a person either on their own or in combination with other words. In contrast with LR-L1, RFE did not pick any of the PoS tag unigrams, but it selected PoS bigrams, trigrams and quadrigrams with the total number of 103. Besides these, RFE has only picked one additional feature, which is number of words with consecutive repeated characters in a comment.

This second group of feature has resulted high precision but unfortunately very low recall, i.e. 95% and 55%, respectively. It is important to mention that recall is a pivotal factor in the problem of offensive language detection. Similar to the case of LR-L1 features, LinearSVC achieved highest accuracy compared to four other classifier-training algorithms, but again with a relatively high variance. The use of the features selected by RFE led to an improvement in the performance of Naive Bayes bringing it very close to LinearSVC.

The union of the features selected by RFE and LR-L1 contains 331 unique features, presented in Table 6.6. The intersection of RFE and LR-L1 selected
features is 20 features, 19 of which are word unigrams, and the 20th one is the number of words with consecutive repeated characters in a comment. Our experiments with training a classifier with the 331 unique features in the union show an improvement in the precision of LinearSVC compared to the use of LR-L1 alone, but it brings it down compared to the use of RFE only. In contrast, the combined features lead to higher recall compared to using either LR-L1 or RFE alone. This improvement is significant (by 21%) compared to using RFE features only, and not very significant (by 2%) compared to using LR-L1 features only. In addition to that the 5-fold cross validation shows that each model has lower variance compared to the use of either of the group of features solo. For example, the accuracy of LinearSVC ranges from 90% to above 77% with average of 86%.

Our experiments show evidence that a combination of multiple feature selection methods (LR-L1 and RFE in our experiments) may lead to better results when training a classifier for detecting offensive language in Arabic. The drop of precision compared to using RFE-selected features only is not too significant because it is outweighed by the improvement of the recall, the high values of which are more important than the high values of precision. When detecting offensive comments, it is sensible to recall as many offensive comments as possible by allowing lower precision, i.e. wrongly classifying inoffensive comments as offensive, as the harm of offence is arguably larger than the harm of misclassifying a comment as offensive.

The running time to build the proposed model is as follows:

- Generate part of speech (PoS) tags 156 min.
- Select features using RFE - 605 min.
- Select features using LR-L1 - 8 sec.
- Train LinearSVC model - 1.26 sec.

The computer specifications used are:
- Processor: Intel Core i7-2600 CPU @ 3.40Hz 3.40 GHz,
- Installed memory (RAM): 16.0GB,
- System type: 64-bit Operating System,
- Windows 7 Professional.

6.5 Summary

As the impact of anti-social behaviour in social networking platforms is growing with the increasing popularity of these platforms, it cannot be ignored. In this chapter, we conduct a range of experiments to examine the effectiveness of text mining methods for constructing a classifier for detection an offensive language in online communication. This study is undertaken to enrich the current results in finding a solution that would contribute to the reduction of this phenomenon specifically in Arabic.

In this work, we conduct ML experiments with a dataset of YouTube comments in Arabic. We report the impact of a range of feature-selection methods and popular pre-processing methods, including extra normalisation, on the performance of an SVM classifier trained to detect offensive comments. We have observed that data pre-processing with stemming can be leveraged to enhance the detection of offensive language in casual Arabic text used on social media platforms. In addition, the ML methods used for feature selection show that the employment of the part-of-speech tags is effective. These methods show that some other features, presented in Section 6.3.1, can be useful. In particular, it can be noted that the utilisation of N-gram features is effective. The results of the conducted experiments give evidence that we can construct classifier with reasonable overall accuracy.
Chapter 7

Conclusion and Recommendation

7.1 Summary

The harm of abusive language and harassment on social media has been well recognised in today's society. The work presented here is an attempt to tackle this problem, in particular, in the case of offensive language on Arabic social media. This thesis presents one of the few studies on machine learning approaches towards the automatic detection of offensive language in online communication in Arabic. Our work is a necessary stepping stone towards the development of effective tools for the prevention of such incidents. The contributions of this thesis include:

- the construction of a large (largest at the time of writing this thesis) training dataset for predictive modelling,
- design and experimental evaluation of a text mining process for building a classification model for the detection of offensive language in written communication in Arabic on a social media platform.

To the best of our knowledge, the dataset constructed as part of this work is the largest dataset of Arabic text, specifically collected for training predictive models for the detection of offensive language in online communication. The scarcity of appropriate training datasets is a shared problem in related studies on offensive language detections (mostly conducted for the English language) that not only prevents training effective classification models, but
it also makes it hard to compare the results of alternative approaches. In the absence of a suitable dataset in Arabic, we constructed a corpus of 15,050 labelled YouTube comments in Arabic (see Chapter 5). We collected data specifically from YouTube, as it is the second most popular social media platform after Facebook at the time of conducting this study. The central focus of this thesis is on the experimental work with a variety of techniques for data preprocessing, feature selection and supervised machine learning for training a model that classifies YouTube comments as either offensive or inoffensive. We made a particular effort to ensure that our classifier is robust enough to deal with the frequently occurring phenomenon of multiple Arabic dialects being used in a single conversation on social media between people from different parts of the Arab World (see Chapter 5). The presented experiments with five different machine learning algorithms suggest that, out of the five, SVM trains the most effective predictive models. We report the impact of word-level features and popular pre-processing methods, including extra normalisation, on the performance of an SVM classifier. In particular, we observed that data pre-processing with stemming and the utilisation of N-gram features can be leveraged to enhance the detection of offensive language in casual Arabic text.

7.2 Contributions

The aim of this study is the development of a method for the automatic detection of offensive language in informal written conversations in Arabic on social media. The main contributions of this thesis are as follows:

1. **Objective 1:** Build a comprehensive dataset for studies of offensive language on the internet devoted to the Arabic language.

   We constructed a dataset of 15,050 comments collected from YouTube, and we followed a commonly accepted methodology to label this dataset.
These comments are collected from discussions on social issues provoked by controversial videos about celebrities in the Arab World. These are videos that attract a relatively high number of comments with offensive language. The percentage of comments in our dataset, labelled as offensive by at least two out of three annotators, is 39%.

- **Research Question 1:** Is there evidence of the existence of offensive language in Arabic social media platforms?
  
  **Result:** The analysis of the dataset labelled by the selected annotators gives evidence of the presence of offensive language with 39% of 15050 YouTube comments labelled as offensive (Alakrot, Murray, and Nikolov, 2018a).

- **Research Question 2:** Furthermore and foremost, what kind of offensive language is used in Arabic social media? Is it Modern Standard Arabic (MSA) or Arab dialects, and what other languages may be found?
  
  **Result:** Our analysis suggests that the offensive comments are primarily written in Arabic dialect and foreign languages. Offensive foreign words and phrases are either in their original language orthography or sometimes in Arabic orthography (Alakrot, Murray, and Nikolov, 2018a).

There is a clear evidence of the presence of offensive language in Arabic social media (see Chapter 5).

2. **Objective 2:** Enhance the accuracy of classifiers for the discovery of offensive Arabic comments on social media platforms. In the context of this goal, the next two research questions were devised:

- **Research Question 3:** What is the impact of removing noisy data and, in general, data pre-processing on the accuracy of detecting offensive incidents on social media platforms?
Result: We examined the impact of removing noisy data and, in general, data pre-processing on the accuracy of detecting offensive incidents on social media platforms. We found out that the typical pre-processing methods used in Arabic text-mining, presented in Section 4.5.1, as well as extra normalisation (see Section 6.2.1.4) improve the accuracy of the proposed model (Alakrot, Murray, and Nikolov, 2018b).

- Research Question 4: What feature-selection methods are effective in improving the accuracy of the detection of offensive incidents on social media platforms?

Result: We conducted experiments with a range of feature-selection methods to test their impact on the accuracy of the detection of offensive incidents. We observed that when either logistic regression, or recursive feature elimination, or the combination of the two are used to select features, the trained LinearSVC classifier has the highest precision, recall and accuracy, i.e. 89%, 76% and 84%, respectively. Also we investigated the ability of other linguistic features to improve the classification accuracy and reduce false positives in detecting message-level offensiveness. In order to achieve that, a set of linguistic features was extracted from the dataset; some of these features were selected at the feature-selection phase, thus indicating that they directly affect the accuracy of the trained classifier. These features are as follows:

- number of characters in a comment,
- number of tokens in a comment,
- number of words in a comment,
- number of unique words in a comment,
- number of words with consecutive repeated characters in a comment,
• number of consecutive repeated characters in a comment,

• number of obscene words in a comment.

• **Research Question 5:** Which machine learning algorithms achieve top performance in detecting abusive language in user-generated Arabic text?

**Result:** We have experimented with five different machine learning algorithms and our experiments gave evidence that LinearSVC performs best compared to the four other machine learning algorithms we experimented with.

### 7.3 Future Work

With this study we open the doors for further and deeper studies on the factors that affect the accurate detection of offensive language in Arabic social media. Some of the ways in which our work can be extended are listed below:

• Number of comments in our dataset include either Arabic in Latin alphabet transcription, i.e. Arabizi, or foreigner words written in Arabic alphabet transcription. The number of these comments is not significant, thus features related to them did not get selected in our experiments. It might be useful if future studies consider including more of this type of comments which would enable a classifier to detect offensive language in them as well.

• While considering the variety of Arabic dialects, our study mostly covers offensive language used by people from eastern Arabic countries. This work can be extended to include offensive language used specifically by people from western Arabic countries to a greater extent.

• We chose to collect data from selected *YouTube* videos. Our dataset can be further enriched by adding labelled comments from a larger and
more diverse selection of YouTube videos, as well from other social media platforms.

- This study did not take into consideration emojis and potentially offensive abbreviations. These can be utilised for extracting additional features.

This work presents a comprehensive text-mining study on the detection of offensive language in Arabic social media. We present a dataset collected and pre-processed, specifically for the purpose of this study, and we present results of experiments with a variety of text-mining techniques for training an effective classifier for detection of offensive language in casual Arabic text that has a variety of Arabic dialects mixed with some of other foreigner languages. We believe that our work opens up a series of prospects for further important academic research in the detection and prevention of anti-social behaviour in social media platforms.
Appendix A

List of Available Stemmers for Arabic Language


2. Light10 stemmer:
   - implemented in Lemur: http://sourceforge.net/p/lemur/wiki/Parser%20Applications/
   - Solr: http://wiki.apache.org/solr/LanguageAnalysis


4. AMIRA 2.0: urlhttp://nlp.ldeo.columbia.edu/amira/

5. QCRI's Arabic processing library that includes a tokenizer, word segmenter, POS tagger, and NER http://alt.qcri.org/tools/

6. Al-Stem (Darwish and Oard, 2003)

7. Alexander Fraser (while at BBN): http://tides.umiacs.umd.edu/software/ stem_aggressive.tar

8. Buckwalter analyzer:
   - http://www.ldc.upenn.edu/Catalog/catalogEntry.jsp?catalogId=LDC2004L02
• with Java reimplementation: http://sourceforge.net/projects/aramorph/


10. MADA for dialects (Habash et al., 2013).

11. Sebawai (Darwish, 2002).

12. IBM word segmenter (Lee et al., 2003).

Appendix B

R Code to Compute Inter-Annnotator Agreement

Inter-annotator agreement is a measure of how well two (or more) annotators can make the same annotation decision for a certain category.

This R code has obtained from (TOM, 2014). This code is used to calculate the Inter-annotator agreement between decision made by every two annotators, as it has presented in Table 5.6.

```r
# install the library
install.packages("irr")

# load the required library
library(irr)

# read in the dataset
ds.full <- read.delim("Datasource , header=T , sep=\"\t\")

# combine the two columns of the annotators in a single data frame
ds.iaa <- data.frame(ds.full$attributive , ds.full$attributive.anno2)

# find observation that were annotated by both annotators
# here , we can only retain the annotations of annotator 2 ,
# because annotator 1 did all observations , whereas annotator 2
# only did a subset
```
ds.iaa.sharedobs <- droplevels(
  ds.iaa[ds.iaa$ds.full.attributive.anno2!="",]
)

# cross-tabulation
table(ds.iaa.sharedobs)
# Cohen's kappa
Appendix C

List of Features Selected by the Employed ML Methods

This Appendix presents the list of all features selected by our feature-selection methods. The abbreviations of the part-of-speech tags are listed in Section C.4.

C.1 LR-L1

Logistic Regression parameters are set as (class_weight='balanced', penalty="l1", C=0.01).

The number of features: 151

[...]
Appendix C. List of Features Selected by the Employed ML Methods

C.2 RFE

The number of features: 200

[num_w_repchars', 'num_unique_words', 'num_terms', 'num_rep_chars', 'Non-pro', 'num_chars', 'VBD', 'RP', 'NNP NNP', 'NN NN', 'IN']

Appendix C. List of Features Selected by the Employed ML Methods

C.3 Union of LR1 and RFE

The number of features: 331

Appendix C. List of Features Selected by the Employed ML Methods

<table>
<thead>
<tr>
<th>Feature</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>num_w_repchars</td>
<td>Number of repeated characters</td>
</tr>
<tr>
<td>num_unique_words</td>
<td>Number of unique words</td>
</tr>
<tr>
<td>num_terms</td>
<td>Number of terms</td>
</tr>
<tr>
<td>num_rep_chars</td>
<td>Number of repeated characters</td>
</tr>
<tr>
<td>No-pro</td>
<td>Pronoun absence</td>
</tr>
<tr>
<td>num_chars</td>
<td>Number of characters</td>
</tr>
<tr>
<td>WRB NN</td>
<td>Word-frequency bigram</td>
</tr>
<tr>
<td>WP VBP RP</td>
<td>Word-phrase bigram</td>
</tr>
<tr>
<td>WP VBP NN DTNN</td>
<td>Word-phrase bigram with noun class</td>
</tr>
<tr>
<td>WP PRP NN</td>
<td>Word-phrase bigram with pronoun</td>
</tr>
<tr>
<td>VBP VBP NN NNP</td>
<td>Verb-verb bigram with noun class</td>
</tr>
<tr>
<td>VBP PUNC</td>
<td>Verb-punctuation</td>
</tr>
<tr>
<td>VBP NNP NNP IN</td>
<td>Verb-noun-noun bigram with pronoun class</td>
</tr>
<tr>
<td>VBP NNP DTNNP</td>
<td>Verb-noun-noun bigram with noun class and pronoun class</td>
</tr>
<tr>
<td>VBP NNP NN DTNN</td>
<td>Verb-noun-noun bigram with noun class and noun class</td>
</tr>
<tr>
<td>VBP NN NN WP</td>
<td>Verb-noun-noun bigram with pronoun class</td>
</tr>
<tr>
<td>VBD NOUN DTNN</td>
<td>Verb bigram with noun class and noun class</td>
</tr>
<tr>
<td>VBD NNP JJ</td>
<td>Verb bigram with noun class and adjective</td>
</tr>
<tr>
<td>VBD NN WP VBP</td>
<td>Verb bigram with pronoun class and verb class</td>
</tr>
<tr>
<td>VBD NN VBP DTNN</td>
<td>Verb bigram with verb class and noun class</td>
</tr>
<tr>
<td>VBD NN NOUN</td>
<td>Verb bigram with noun class</td>
</tr>
<tr>
<td>VBD NN NNP DTNNP</td>
<td>Verb bigram with noun class and pronoun class</td>
</tr>
<tr>
<td>VBD NN DTNN</td>
<td>Verb bigram with noun class and noun class</td>
</tr>
<tr>
<td>VBD NN CC</td>
<td>Verb bigram with class</td>
</tr>
<tr>
<td>VBD</td>
<td>Verb bigram</td>
</tr>
<tr>
<td>VB IN NN NN</td>
<td>Verb bigram with noun class and noun class</td>
</tr>
<tr>
<td>RP VBP NN DTNN</td>
<td>Verb bigram with pronoun class and noun class</td>
</tr>
<tr>
<td>RP NNP RP</td>
<td>Verb bigram with pronoun class and pronoun class</td>
</tr>
<tr>
<td>RP JJ DTNN</td>
<td>Verb bigram with noun class and noun class</td>
</tr>
<tr>
<td>'RP JJ'</td>
<td>Verb bigram with noun class and noun class</td>
</tr>
<tr>
<td>RP</td>
<td>Verb bigram</td>
</tr>
<tr>
<td>PRP VN</td>
<td>Pronoun-verb-noun</td>
</tr>
<tr>
<td>PRP NN NNP DTNNP</td>
<td>Pronoun-noun-noun bigram with noun class</td>
</tr>
<tr>
<td>PRP</td>
<td>Pronoun bigram</td>
</tr>
<tr>
<td>ADJ</td>
<td>Adjective bigram</td>
</tr>
<tr>
<td>NOUN DTNN VBP</td>
<td>Noun bigram with verb class</td>
</tr>
<tr>
<td>NNS VBP IN</td>
<td>Noun noun bigram with pronoun class</td>
</tr>
<tr>
<td>NNS NN DTNNP</td>
<td>Noun noun bigram with noun class and noun class</td>
</tr>
<tr>
<td>NNS JJ NN DTNN</td>
<td>Noun noun bigram with noun class and noun class</td>
</tr>
<tr>
<td>'NNP RP NN PRP'</td>
<td>Noun noun bigram with pronoun class and pronoun class</td>
</tr>
<tr>
<td>'NNP NNP VBP JJ'</td>
<td>Noun noun bigram with verb class</td>
</tr>
<tr>
<td>'NNP NNP NNP RP'</td>
<td>Noun noun bigram with noun class and pronoun class</td>
</tr>
<tr>
<td>'NNP NNP DTJJ NN'</td>
<td>Noun noun bigram with noun class and noun class</td>
</tr>
<tr>
<td>'NNP NNP'</td>
<td>Noun noun bigram</td>
</tr>
<tr>
<td>'NNP NN RP VBP'</td>
<td>Noun noun bigram with verb class</td>
</tr>
<tr>
<td>'NNP NN PRP NNP'</td>
<td>Noun noun bigram with pronoun class and noun class</td>
</tr>
<tr>
<td>'NNP NN JJ RP'</td>
<td>Noun noun bigram with adjective class</td>
</tr>
<tr>
<td>'NNP NN IN NNS'</td>
<td>Noun noun bigram with noun class and noun class</td>
</tr>
<tr>
<td>'NNP JJ IN'</td>
<td>Noun noun bigram with noun class and noun class</td>
</tr>
<tr>
<td>'NNP IN JJ'</td>
<td>Noun noun bigram with noun class and noun class</td>
</tr>
<tr>
<td>'NNP DTNN DTJJ NN'</td>
<td>Noun noun bigram with noun class and noun class</td>
</tr>
<tr>
<td>'NN WP VBP VBP'</td>
<td>Noun bigram with verb class</td>
</tr>
<tr>
<td>'NN VBP WP VBD'</td>
<td>Noun noun bigram with verb class and verb class</td>
</tr>
<tr>
<td>'NN VBP VBP NNP'</td>
<td>Noun noun bigram with noun class</td>
</tr>
<tr>
<td>'NN VBD PRP'</td>
<td>Noun bigram with pronoun class</td>
</tr>
<tr>
<td>'NN VBD NOUN'</td>
<td>Noun bigram with noun class</td>
</tr>
<tr>
<td>'NN VBD NNS NN'</td>
<td>Noun bigram with noun class and noun class</td>
</tr>
<tr>
<td>'NN RP NNP NNP'</td>
<td>Noun noun bigram with noun class and noun class</td>
</tr>
<tr>
<td>'NN NNS RP'</td>
<td>Noun noun bigram with noun class</td>
</tr>
<tr>
<td>'NN NNS DT'</td>
<td>Noun noun bigram with noun class</td>
</tr>
<tr>
<td>'NN NNP WP VBP'</td>
<td>Noun noun bigram with verb class and verb class</td>
</tr>
<tr>
<td>'NN NNP VBP NNS'</td>
<td>Noun noun bigram with noun class and noun class</td>
</tr>
<tr>
<td>'NN NNP RP NN'</td>
<td>Noun noun bigram with noun class</td>
</tr>
<tr>
<td>'NN NNP NNP NOUN'</td>
<td>Noun noun bigram with noun class</td>
</tr>
<tr>
<td>'NN NN NNP NNS'</td>
<td>Noun noun bigram with noun class and noun class</td>
</tr>
<tr>
<td>'NN NN JJ NNP'</td>
<td>Noun noun bigram with noun class and noun class</td>
</tr>
<tr>
<td>'NN NN DTNN CC'</td>
<td>Noun noun bigram with noun class and noun class</td>
</tr>
<tr>
<td>'NN NN'</td>
<td>Noun noun bigram</td>
</tr>
<tr>
<td>'NN JJ RP JJ'</td>
<td>Noun noun bigram with adjective class</td>
</tr>
</tbody>
</table>
| 'NN

C.4 Penn Treebank Part-of-Speech Tags

<table>
<thead>
<tr>
<th>Tag</th>
<th>Description</th>
<th>Example</th>
<th>Tag</th>
<th>Description</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>CC</td>
<td>coordin. conjunction</td>
<td>and, but, or</td>
<td>SYM</td>
<td>symbol</td>
<td>+, %, &amp;</td>
</tr>
<tr>
<td>CD</td>
<td>cardinal number</td>
<td>one, two</td>
<td>TO</td>
<td>“to”</td>
<td>to</td>
</tr>
<tr>
<td>DT</td>
<td>determiner</td>
<td>a, the</td>
<td>UH</td>
<td>interjection</td>
<td>ah, oops</td>
</tr>
<tr>
<td>EX</td>
<td>existential ‘there’</td>
<td>there</td>
<td>VB</td>
<td>verb base form</td>
<td>eat</td>
</tr>
<tr>
<td>FW</td>
<td>foreign word</td>
<td>mea culpa</td>
<td>VBD</td>
<td>verb past tense</td>
<td>ate</td>
</tr>
<tr>
<td>IN</td>
<td>preposition/sub-conj</td>
<td>of, in, by</td>
<td>VBG</td>
<td>verb gerund</td>
<td>eating</td>
</tr>
<tr>
<td>JJ</td>
<td>adjective</td>
<td>yellow</td>
<td>VBN</td>
<td>verb past participle</td>
<td>eaten</td>
</tr>
<tr>
<td>JJR</td>
<td>adj., comparative</td>
<td>bigger</td>
<td>VBP</td>
<td>verb non-3sg pres</td>
<td>eat</td>
</tr>
<tr>
<td>JJS</td>
<td>adj., superlative</td>
<td>wildest</td>
<td>VBZ</td>
<td>verb 3sg pres</td>
<td>eats</td>
</tr>
<tr>
<td>LS</td>
<td>list item marker</td>
<td>1, 2, One</td>
<td>WDT</td>
<td>wh-determiner</td>
<td>which, that</td>
</tr>
<tr>
<td>MD</td>
<td>modal</td>
<td>can, should</td>
<td>WP</td>
<td>wh-pronoun</td>
<td>who, who</td>
</tr>
<tr>
<td>NN</td>
<td>noun, sing.</td>
<td>llama</td>
<td>WPS</td>
<td>possessive wh-</td>
<td>whose</td>
</tr>
<tr>
<td>NNS</td>
<td>noun, plural</td>
<td>llamas</td>
<td>WRB</td>
<td>wh-adverb</td>
<td>how, where</td>
</tr>
<tr>
<td>NNP</td>
<td>proper noun, sing.</td>
<td>IBM</td>
<td>$</td>
<td>dollar sign</td>
<td>$</td>
</tr>
<tr>
<td>NNPS</td>
<td>proper noun, plural</td>
<td>Carolinas</td>
<td>#</td>
<td>pound sign</td>
<td>#</td>
</tr>
<tr>
<td>PDT</td>
<td>preposition</td>
<td>all, both</td>
<td>“</td>
<td>left quot</td>
<td>“ or “</td>
</tr>
<tr>
<td>POS</td>
<td>possessive ending</td>
<td>’s</td>
<td>”</td>
<td>right quot</td>
<td>” or ”</td>
</tr>
<tr>
<td>PRP</td>
<td>personal pronoun</td>
<td>I, you, he</td>
<td>(</td>
<td>left parenthesis</td>
<td>[, {, &lt;</td>
</tr>
<tr>
<td>PRPS</td>
<td>possessive pronoun</td>
<td>your, one’s</td>
<td>)</td>
<td>right parenthesis</td>
<td>], }, &gt;</td>
</tr>
<tr>
<td>RB</td>
<td>adverb</td>
<td>quickly, never</td>
<td>,</td>
<td>comma</td>
<td>,</td>
</tr>
<tr>
<td>RBR</td>
<td>adverb, comparative</td>
<td>faster</td>
<td>.</td>
<td>sentence-final punct</td>
<td>. !</td>
</tr>
<tr>
<td>RBST</td>
<td>adverb, superlative</td>
<td>fastest</td>
<td>:</td>
<td>mid-sentence punct</td>
<td>: ... --</td>
</tr>
</tbody>
</table>

Figure C.1: Penn Treebank part-of-speech tags (including punctuation) (Jurafsky and Martin, 2017)


Aljlayl, Mohammed et al. (2001). “IIT at TREC-10.” In: TREC.


Alruily, Meshrif (2012). “Using text mining to identify crime patterns from arabic crime news report corpus”. In:


on geometrical models of natural language semantics. Association for Computational Linguistics, pp. 22–32.

Buckwalter, Tim (2002). “Buckwalter Arabic Morphological Analyzer Version 1.0”. In:


Dadvar, Maral (2014). “Experts and machines united against cyberbullying”. In:


Hedar, Abdel Rahman and M Doss (2013). “Mining social networks Arabic slang comments”. In: IEEE Symposium on Computational Intelligence and Data Mining (CIDM).


Lieberman, Henry, Karthik Dinakar, and Birago Jones (2011). “Let’s gang up on cyberbullying”. In: Computer 44.9, pp. 93–96.


Neto, Joel Larocca et al. (2000). “Document clustering and text summarization”. In:


Nobata, Chikashi et al. (2016). “Abusive language detection in online user content”. In: Proceedings of the 25th International Conference on World Wide


