King Saud University
College of Computer and Information Sciences
Computer Science Department

Arabic Text Dialect Recognition

Authors

Mohand Al-Rasheed	439101298
Khalid Albader	439101990
Abdulrahman Alshawi	439101980
Abdullah Alsuwailem	439101690
Musaad Alqubayl	439101884

Supervised by: Dr. Nasser Alsadhan

Research project for the degree of Bachelor in Computer Science
First/Second Semester 1443
Autumn/Spring 2021

Contents

Acknowledgements

We would like to express our great gratitude to Dr. Nasser Alsadhan for his valuable suggestions. and his aid throughout the writing of this report. His willingness to give his time so generously has been very much appreciated.

English Abstract

The Arabic language is one of the oldest languages widely used today, and as a result of that, many Arabic speaking regions have formed dialects exclusive to their own. For example, many countries surrounding the Arabic Gulf have formed a dialect different to countries in the Levantine region. We intend on identifying and systematically determining the dialect of a piece of text.

This research has many applications in Arabic text analysis, such as helping in identifying the regions customers most often come from by analyzing a product's reviews and comments and breaking them down by region, which provides useful intel for a business. It also helps in narrowing the nationality of an anonymous writer of a piece of text by predicting their region. One of the major challenges in dialect recognition is dividing data into classes of dialects. Saudi Arabia and the UAE have dialects that differ widely from each other when solely considered, though they feel very similar in comparison to a Levantine dialect. The researchers will determine a classification easy enough for a machine to detect, but sophisticated enough to be useful.

We intend to build a machine learning powered classifier that distinguishes between a set number of different Arabic dialects (e.g. Egyptian, Levantine, Gulf, etc.) when given a piece of text. We'll use state of the art technologies in the field of NLP (natural language processing) in order to train an effective classifier that understands the differences between dialects.

Arabic Abstract

اللغة العربية من أقدم اللغات المستخدمة بكثرة حاليا، ونتيجة لذلك، الكثير من المناطق المتحدثة للعربية أنشأت لهجات مخصصة بمناطقهم. فعلى سبيل المثال، الكثير من المناطق المجاورة للخليج العربي تتحدث لهجة مختلفة بشدة عن لهجات المناطق الشامية. يعتزم الباحثون على أتمتة عملية التعرف على اللهجات من خلال تحليل قطعة من النص.

البحث له العديد من التطبيقات، وأهمها هو في تحليل النصوص العربية،

فمثلا استخدامه في التعرف على مناطق عملاء جهة معينة عن طريق تحليل التقاييم والتعليقات المضافة على منتجاتهم، مما يمكن الجهة على التعرف على عملائهم بشكل أدق. كذلك يمكن استخدامه للتنبؤ بمنشأ مرسل رسالة مجهولة عن طريق التعرف على منطقة نشأته.

من أهم التحديات في تصنيف اللهجات هي تقسيم البيانات لأصناف من اللهجات. فعلى سبيل المثال، المملكة العربية السعودية والإمارات العربية المتحدة يتحدثون بلهجات مختلفة إذا حصرنا النظر عليهم، ولكن يشبهون بعض حين تتم مقارنتهم مع اللهجات الشامية. سيختار الباحثون مجموعة مناسبة من اللهجات حيث تكون سهلة للنظام في التعرف عليها، ولكن معقدة كفاية لكي تكون مفيدة.

في هذا المشروع ننوي بناء مصنف (classifier) مدعوم بتقنيات تعلم الآلة لكي يصنف ما بين مجموعة من اللهجات المحددة (مثل اللهجة المصرية، والشامية، والخليجية، وغيرها) إذا أعطي قطعة من النص. سيستخدم الباحثون أحدث التقنيات في مجال تحليل اللغات الطبيعية (NLP) لكي يدربوا مصنف فعال، يفرق بين اللهجات العربية.

1 Introduction

As languages develop across regions far apart from each other dialects begin to take shape, machine learning researchers became interested in classifying text in some language to it's proper dialect. This is because its connected to more insightful text analysis.

A dialect is the variation of a language in grammar, pronunciation and vocabulary. Every individual has their own way of talking that is affected by dialect, accent, background and many other factors[5]. The Arabic language has a variety of dialects throughout the Arabic world, dialects could differ not only across countries but also in the same country or even city. Arabic dialects differ from one another in pronunciation and vocabulary, different dialects have different words or different variations of a word that could refer to the same meaning, which sometimes make it a bit difficult to understand each other, and it can make it harder for non-Arabic speakers who are trying to learn Arabic.

Machine Learning is a field of study that is concerned with developing algorithms that utilize data with the intent of solving tasks traditional methods cannot solve, in a way similar to how humans approach complex problems[8]. It is a rapidly growing field, many countries are racing each other to adapt machine learning technologies and develop smart and automated systems, applications, and adapt them into our daily lives as well as numerous varieties of fields that could benefit from them. Natural language processing, abbreviated as *NLP* is a branch of machine learning that is primarily focused on analyzing text. Numerous companies are racing to develop programs that utilize NLP to analyze user behaviour. One of the difficulties facing companies developing using NLP for Arabic speakers is the numerous varieties of dialects in Arabic.

1.1 Problem statement

Dialects are formed mainly due to regional separation between the Arab world. This separation reduces interaction between different regions, and as a result of that, many Arabic speaking regions have formed dialects exclusive to their own. For example, many countries surrounding the Arabic Gulf have formed a dialect different to countries in the Levantine region. The research's main problem is how to identify and predict dialect types from text.

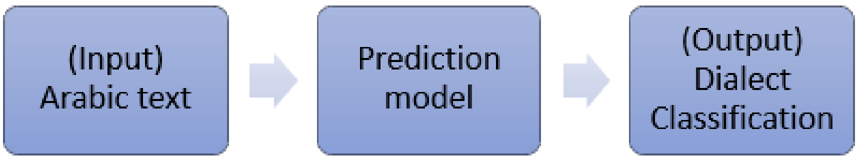

Figure 1: Illustration of the problem

1.2 Goals and objectives

The goal of this research is to analyze and understand Arabic text to classify the dialect of any piece of Arabic text. The objective is to implement the most appropriate state of the art NLP model that helps in achieving the best possible accuracy which correlates to correctly classifying what dialect the text is from.

1.3 Proposed solution

This research will contribute in solving Arabic dialect detection by using one of the latest advancements in the field of natural language processing.

1.4 Research scope

The scope of this research is mainly focused about analyzing, preprocessing and modeling a state of the art NLP model to classify Arabic text into a set of dialects.

2 Background

2.1 Natural language processing

"Natural Language Processing is a theoretically motivated range of computational techniques for analyzing and representing naturally occurring texts at one or more levels of linguistic analysis for the purpose of achieving human-like language processing for a range of tasks or applications." [16]. The field of NLP is an active area of research and development solely for the purpose of computerizing the process of analyzing written/spoken text in a human-like way.

In recent years NLP has become an essential part for many technologies that are relevant today. Companies are taking advantage of the abundance of data that is flooding the internet every day and are developing numerous NLP technologies and applications that we use everyday.

Human languages are surprisingly complex and ambiguous in nature, there are languages that are easier to process for computers than others due to various reasons. German for example relies heavily on morphology and compositional word-building that aids in generalizing to unseen words[1]. However, there has been continuous advancements done on computational techniques that will try to solve challenges around the ambiguity of languages.

2.1.1 Preprocessing

Preprocessing refers to the manipulation of raw data to format it in a way that is easier for computers to process and analyze. It is a technique that is crucial for any NLP task to perform well, it can directly impact the accuracy and performance of any kind of task performed on it. It is the first step taken for any NLP task. Some operations of preprocessing include, *normalization* of data, *segmentation* of data, *tokenization* of text, *stemming* of words and *noise removal*. When dealing with Arabic text usually the first step is filtering out non-Arabic content from text especially when you are getting the content from social media.

In this section we will discuss the most important steps in preprocessing, such as tokenization and converting text to embeddings.

2.1.1.1 Tokenization

Tokenization is the process of breaking down input text into smaller components called tokens so that its easily analyzable for computers. It is an important step in preprocessing text for any NLP task. There are several methods for performing tokenization, such as white tokenization, subword tokenization and others. White space tokenization breaks sentences into words that we call tokens, while this is useful for languages like English and French, it is needed to perform some additional steps for languages like Chinese and Japanese where words are not separated by spaces. While subword tokenization breaks down words into different tokens, so for example, ”Unfriendly” is broken down to ”Un”, ”friend” and ”ly” [21].

Tokenization also has limitations for the Arabic language, owing to the complexity of the language, words like ”عقد” and ”جد” depending on the context or pronunciation could lead to different meanings, So the word ”عَقَدَ”. means to tie, which is different from ”عَقَّدَ”. which means to over-complicate. Also there are huge differences in formal and informal Arabic (more on that in section 3.1), as well as different dialects having vastly different sentence structure. and not just in Arabic this is also true for most languages, and that is one of the challenges of tokenization.

2.1.1.2 Word embedding

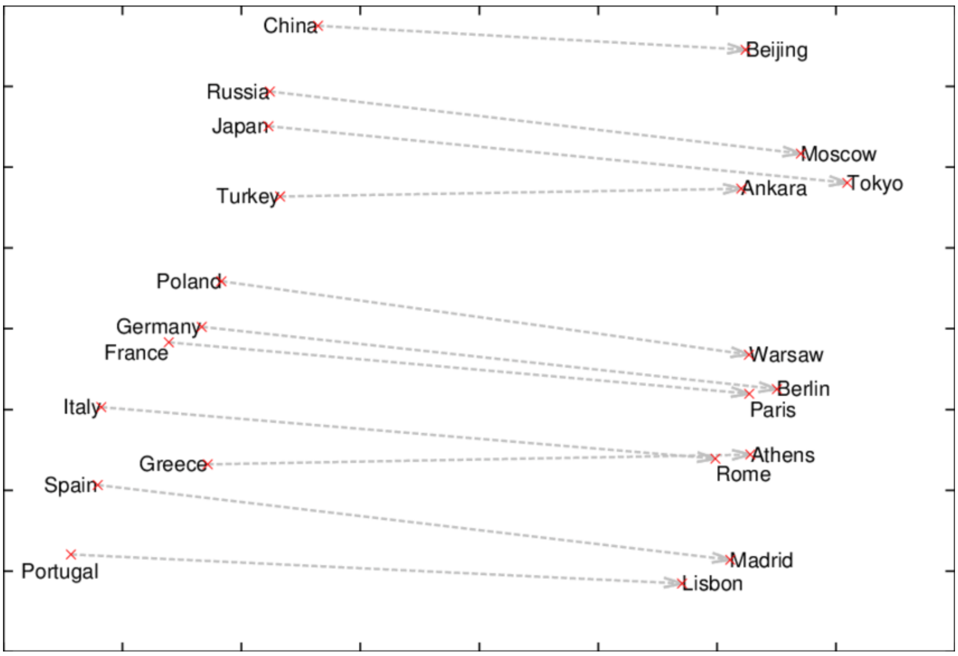

Figure 2: Country and capital vectors projected by PCA [18]

Computers can’t understand natural language, so in order to make computers understand it we have to create a representation for a language that a computer can process, and that is what word embedding do. Word embedding is a representation of words that encodes the semantic meaning of words in vectors, such that words that are similar in meaning are probably going to be close in vector space [13]. There are several word embedding models, and generally all models share the concept of context to determine how close are words to each other, "You shall know a word by the company it keeps!" (Firth, J. R. 1957:11). Figure 2 shows a model that learnt the relationships between countries and their capitals without information of what a capital city means.

2.2 Neural networks

Neural networks are a sub-field of machine learning, and also the parent field of deep learning. Neural networks are made up of layers of neurons and work like interconnected nodes inspired by the neurons inside the brain. By taking in data, they are able to recognize hidden patterns and correlations in unprocessed data and use said patterns to cluster, classify and predict the data, among other applications.

A typical neural network architecture contains the following:

1. Input layer: takes the initial data.
2. Hidden layer(s): a layer, or more, placed between input and output which captures the non-linearity of the data.
3. Output layer: produce the outcome of the prediction.

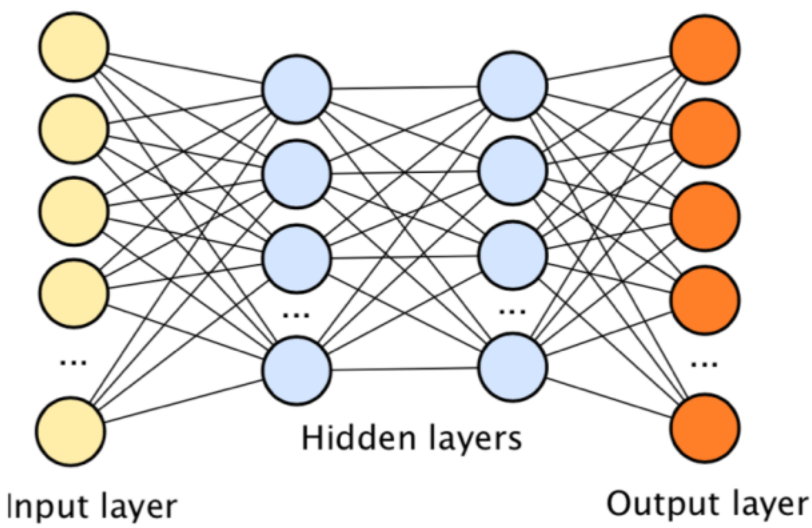

Figure 3: Neural networks architecture [17]

Neural networks also are ideally fitted to assist humans solve complicated issues in real-life situations. They can examine and model the relationships among inputs and outputs which can be nonlinear and complicated as well as make generalizations and inferences,

Neural networks are now prevalent in NLP such text classification (which is the objective of this research), machine translation, semantic parsing, which extracts useful bits of information in a large text, and many more.

2.2.1 Deep learning

Deep learning is a machine learning model that utilizes large neural networks. deep learning have dramatically developed the state-of-the-art in speech recognition, object detection and a number of different domains along with genomics and drug discovery[15].

While deep learning isn't accurately defined, what differentiates deep learning models from other neural networks models is primarily the number of layers and the time it takes to train. An example of a deep learning model is *Convolutional Neural Networks* (CNNs) which are commonly used with images. Extracting meaning from a 2D structure such as images can be quite hard for traditional machine learning algorithms because of the inherent complexity of the patterns in image data, this complexity can be tackled by deep learning models though they require a large number of layers as well as long training time.

2.2.2 Transformers

Since its introduction in 2017 by the google research team, its rapid growth dominated the NLP field and became a standard for any encoder/decoder model today[23]. the transformer takes advantage of parallelization unlike Recurrent Neural Networks (RNNs), which process data in a sequential order, which is computationally more expensive compared to the transformer model[22]. In a high level overview, its model architecture can be divided into two major components, an encoder and a decoder. An encoder maps the input sequence to a numeric representation that holds information about the input sequence, the decoder given the output of the encoder generates a sequence of symbols one element at a time, the model consumes the previously generated symbols as additional input when generating the next[22].

There are many models used today that are built on the transformer architecture especially in NLP, for example, *Bidirectional Encoder Representations from Transformers* (BERT) is a popular transformer-based model. Also, OpenAI's *Generative Pre-trained Transformer* (GPT) models are transformer models that garnered wide attention for being excellent in imitating human produced text.

2.3 Dialect prediction approaches

One can approach the problem of dialect prediction in a number of ways, We will define and discuss some different dialect recognition approaches that differ in how they work.

2.3.1 Rule-based approach

Rule-based approach relies in written curated instructions made by humans to identify selected parts of the text that match a certain logic or found in dictionaries, a popular example in text classifications is to count the number of each word that relate to a category and the highest word count for a category classifies the text in that category.

Another example would be the Lexical Functional Grammar (LFG). "The LFG system incorporates a richly annotated lexicon containing functional and semantic information."[19].

2.3.2 Automatic machine-learning approach

The automatic approach in dialect recognition is based on machine learning, where it tries to build a statistical model that learns by analysing the training data after choosing an appropriate algorithm and applying NLP techniques. The most prominent algorithms in text classification would be support vector machines (SVMs), Naïve Bayes and deep learning methods.

2.3.3 Hybrid approach

"Hybrid systems combine a machine learning-trained base classifier with a rule-based system, used to further improve the results. These hybrid systems can be easily fine-tuned by adding specific rules for those conflicting tags that haven't been correctly modeled by the base classifier."[14]

2.4 Performance metrics

In binary[1] classification problems, we can test the performance of our results by matching the output of our model, the predicted label, to the real label in our data. This measure is known as the *accuracy* of our model according to the data. However there are more sophisticated measures that one can observe. We'll talk about two of those measure, mainly *precision* and *recall*.

First, we must define 4 quantities, **True Positive**, abbreviated TP consists of *true* which refers to the data belonging to class 1, while

[1]We can use precision and recall in multiclass classification by considering one class, A, at a time and lumping all other classes as $notA$

positive refers to the model's prediction belonging to class 1. And **False Negative** is similar to TP but in the context of class 0. We can mix and match T, F, P and N to get 4 different quantities.

Here we define precision and recall in the following way:

$$Precision = TP/(TP + FP)$$

$$Recall = TP/(TP + FN)$$

We can tweak the model's threshold of classification in order to achieve a different Precision and Recall metrics

3 Literature review

The discussed problem in this research has been tackled by many researchers over the years with varying results. in this section we intend to highlight the most important results regarding the Arabic language, the Arabic corpora and dialect classification methods that have been concluded from past research.

3.1 The Arabic language

Arabic speakers often use Modern Standard Arabic (*MSA*) when they're in a formal setting such as reading the news, though they have a regional dialect that they talk with in informal settings. In this section we'll detail the work made to document and break down different dialects into regions they belong to.

3.1.1 Arabic dialects

Dividing Arabic into different dialects is not a standardized task as dialects shift and change depending on the time and how much precision we intend to administer in our breakdown. Researchers working on this problem have found various breakdowns that we'll discuss.

Habash has suggested the following breakdown, while adding "and should not be taken to mean that all members of any dialect group are completely homogenous linguistically" [12].

1. Egyptian Arabic (EGY) which spans Eygpt and Sudan
2. Gulf Arabic (GLF) which spans the Arabic peninsula, Habash adds "although there is a wide range of sub-dialects within it." And "Omani Arabic is included some times."
3. Levantine Arabic (LEV) which spans the Levantine region

4. "North African (Maghrebi) Arabic (Mag) covers the dialects of Morocco, Algeria, Tunisia and Mauritania. Libyan Arabic is sometimes included." [2]
5. "Iraqi Arabic (IRQ) has elements of both Levantine and Gulf"
6. "Yemenite Arabic (Yem) is often considered its own class"

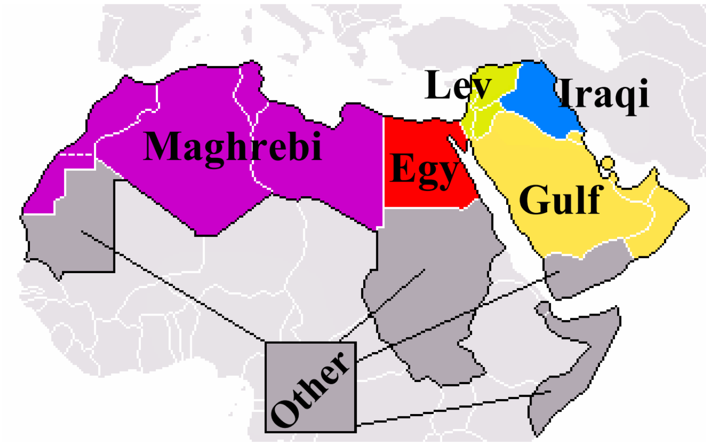

Figure 4: Zaidan and Callison-Burch (2011) gave a similar breakdown[24] to Habash's

Alshutayri also gave a similar breakdown, which is GLF (including Oman), EGY, LEV, NOR (which includes Morocco, Algeria, Tunisia and Libya) and IRQ. Although the breakdown is somewhat general and imprecise, its general enough to be useful in data collection and in classification[4].

3.2 Existing Arabic text corpora

The problem of dialect classification has been studied in the past with many studies building their own corpora, here we'll examine the most prominent of corpora.

In 2015 Shoufan and Alameri conducted a literature review, in which they summarised the advancements in NLP for dialectal Arabic in the following comprehensive table[20]. Bear in mind that the table includes more than text analysis and also includes speech analysis.

[2] Many other researchers abbreviate North African dialects as "NOR"

	Basic Language Analyses			Building Language Resources		Dialect Identification and Recognition		Semantic Analysis	
	Morph.	Syntax	Orthog.	Lexica	Corpora	From Text	From Speech	M. Translation	Others
Gulf	(Almeman & Lee, 2012), (Abuata & Al-Omari, 2015)		(Darwish, 2013), (Masmoudi et al., 2015)		(Zaidan & Callison-Burch, 2011), (Almeman et al., 2013), (Cotterell& Callison-Burch, 2014)	(Zaidan & Callison-Burch, 2011), (Sadat, Kazemi, & Farzindar, 2014), (Zaidan & Callison-Burch, 2014)	(Belgacem et al., 2010), (Zaidan&Callison-Burch, 2012), (Zhang et al., 2013), (Biadsy et al., 2009), (Akbacak et al.,2011)	(Jehl et al., 2012), (Salloum & Habash, 2012), (Sawaf, 2010)	(Mourad & Darwish, 2013)
Kuwaiti					(Mubarak & Darwish, 2014)	(Sadat, Kazemi, & Farzindar, 2014)	(Iskra et al., 2004)		
Saudis					(Mubarak & Darwish, 2014)	(Sadat, Kazemi, & Farzindar, 2014)	(Alghamdi et al., 2008), (Iskra et al., 2004)	(Sawaf, 2010)	
UAE					(Mubarak & Darwish, 2014)		(Lei & Hansen, 2009), (Iskra et al., 2004)	(Khamis, 2007)	
Qatari					(Mubarak & Darwish, 2014), (Zaghouani et al., 2014)	(Sadat, Kazemi, & Farzindar, 2014)	(Iskra et al., 2004)	(Al- Mannai et al., 2014)	
Bahraini						(Sadat, Kazemi, & Farzindar, 2014)	(Iskra et al., 2004)		
Omani						(Sadat, Kazemi, & Farzindar, 2014)	(Iskra et al., 2004)		
S. A. Peninsula								(Sawaf, 2010)	
Yemeni					(Belgacem et al., 2010)				
Sana´ani								(Al- Gaphari & Al Yadoumi, 2012)	
North Africa	(Almeman & Lee, 2012), (Habash et al., 2013)		(Masmoudi et al., 2015), (Darwish, 2013)		(Almeman & Lee, 2013)				
Egyptian	(Duh & Kirchhoff 2005), (Habash et al., 2012), (Almeman & Lee, 2012), (Al-Sabbagh & Girju, 2012a), (Salloum &Habash, 2014)		(Dasigi & Diab, 2011), (Habash, Diab, & Rambow, 2012), (Bies et al., 2014)	(Hedar & Doss, 2013)	(Habash et al.,2008), (Diab et al., 2010), (Benajiba & Diab, 2010), (Zaidan & Callison-Burch, 2011), (Al-Sabbagh & Girju, 2012), (Elfardy& Diab, 2012b), (Elfardy& Diab,2012c), (Almeman& Lee,2013), (Mubarak& Darwish, 2014), (Cotterell& Callison-Burch,2014), (Maamouri et al., 2014), (Hawwari et al., 2014), (Maamouri et al.,2014)	(Diab et al., 2010), (Zaidan & Callison-Burch, 2011), (Elfardy & Diab, 2012), (Elfardy & Diab, 2013), (Zaidan & Callison-Burch, 2012), (Habash et al., 2008b), (Zaidan & Callison-Burch, 2014), (Darwish et al., 2014)	(Belgacem et al., 2010), (Zhang et al., 2013), (Lei & Hansen, 2009), (Biadsy et al., 2009), (Akbacak et al., 2011), (Kirchhoff & Vergyri, 2005), (Iskra et al., 2004)	(Zbib et al.,2012), (Salloum & Habash, 2011), (Jehl et al., 2012), (Bakr et al.,2008), (Salloum & Habash, 2012), (Sawaf, 2010), (Mohamed et al., 2012), (Jeblee et al., 2014)	(Pasha et al., 2013), (Hedar & Doss, 2013), (El- Fishawy et al., 2014), (Ibrahim et al., 2015), (Mourad & Darwish, 2013), (Zirikly & Diab, 2014/2015), (El-Beltagy & Ali, 2013), (Darwish & Gao, 2014)
Cairene				(Al-Sabbagh & Girju, 2010)					
Morrocan				(Graff & Maamouri , 2012)	(Benajiba & Diab, 2010), (Diab et al., 2010), (Tratz et al., 2013) , (Mubarak & Darwish,	(Sadat, Kazemi,& Farzindar, 2014)	(Elfardy & Diab, 2012a), (Belgacem et al., 2010), (Iskra et al., 2004)	(Sawaf, 2010), (Tachicart & Bouzoubaa,	

Table 1: Dialectical Arabic NLP- Literature Overview[20]

	Basic Language Analyses			Building Language Resources		Dialect Identification and Recognition		Semantic Analysis	
	Morph.	Syntax	Orthog.	Lexica	Corpora	From Text	From Speech	M. Translation	Others
					2014)			2014)	
Tunisian	(Zribi, Khemakhem, & Belguith, 2013), (Boujelbane et al., 2014)		(Zribi et al., 2013), (Zribi et al., 2014)	(Boujelbane et al., 2013)	(Boujelbane et al., 2013), (Zribi, Graja, et al., 2013)	(Sadat, Kazemi, & Farzindar, 2014)	(Belgacem et al., 2010), (Boujelbane et al., 2013), (Iskra et al., 2004)	(Sawaf,2010), (Sadat, Mallek, et al., 2014)	
Libyan				(Graja et al., 2010)		(Sadat, Kazemi, & Farzindar, 2014)	(Iskra et al., 2004)	(Sawaf, 2010)	
Sudani	(Almeman & Lee, 2012)				(Mubarak & Darwish, 2014)	(Sadat, Kazemi, & Farzindar, 2014)		(Sawaf, 2010)	
Algerian					(Harrat et al., 2014)	(Harrat et al., 2015), (Sadat, Kazemi, & Farzindar, 2014)	(Iskra et al., 2004)		
Maghrebi*					(Cotterell & Callison-Burch, 2014)	Zaidan & Callison-Burch, 2012), (Zaidan & Callison-Burch, 2014)			
Levantine	(Habash &Rambow, 2006), (Habash &Rambow,2007) (Almeman & Lee, 2012),	(Chiang et al., 2006), (Maamouri et al., 2006)	(Habash &Rambow, 2007), (Dasigi & Diab, 2011), (Darwish, 2013), (Masmoudi et al., 2015)	(Duh & Kirchhoff 2006)	(Maamouri et al., 2006), (Diab et al., 2010), (Benajiba & Diab, 2010), (Soltau et al., 2011), (Zaidan & Callison-Burch, 2011), (Elfardy& Diab, 2012b), (Almeman& Lee,2013), (Almeman et al., 2013), (Cotterell & Callison-Burch, 2014)	(Habash et al., 2008), (Habash et al., 2008b), (Diab et al., 2010), (Zaidan & Callison-Burch, 2011), (Zaidan & Callison-Burch, 2012), (Elfardy & Diab, 2012c), (Zaidan & Callison- Burch, 2014)	(Elfardy & Diab, 2012a), (Zhang et al., 2013), (Biadsy et al., 2009), (Akbacak et al., 2011), (Iskra et al., 2004)	(Zbib et al., 2012), (Salloum & Habash, 2011), (Jehl et al., 2012), (Salloum & Habash, 2012), (Soltau et al., 2011)	(Mourad & Darwish, 2013)
Syrian				(Graff & Maamouri , 2012)		(Harrat et al., 2015), (Sadat, Kazemi, & Farzindar, 2014)	(Belgacem et al., 2010), (Lei & Hansen, 2009), (Iskra et al., 2004)		
North Syrian								(Sawaf, 2010)	
Damascus								(Sawaf, 2010)	
Lebanese						(Sadat, Kazemi, & Farzindar, 2014)	(Iskra et al., 2004)	(Sawaf, 2010)	
Jordanian	(Salloum &Habash, 2014)					(Sadat, Kazemi, & Farzindar, 2014)	(Iskra et al., 2004)	(Sawaf, 2010)	(Duwairi et al., 2014)
Palestinian					(Jarrar et al., 2014)	(Harrat et al., 2015), (Sadat, Kazemi, & Farzindar, 2014)	(Lei & Hansen, 2009), (Iskra et al., 2004)	(Sawaf, 2010)	
Iraqi	(Almeman & Lee, 2012)		(Masmoudi et al., 2015), (Darwish, 2013)	(Graff et al., 2006), (Rytting et al., 2011), (Graff & Maamouri 2012), (Cavalli-Sforza et al., 2013)	(Diab et al., 2010), (Habash et al., 2008a), (Benajiba & Diab, 2010), (Elfardy & Diab, 2012b), (Cotterell& Callison-Burch, 2014)	(Zaidan & Callison-Burch, 2012), (Zaidan & Callison-Burch, 2014), (Sadat, Kazemi, & Farzindar, 2014)	(Elfardy & Diab, 2012), (Belgacem et al., 2010), (Zhang et al., 2013), (Lei & Hansen, 2009), (Biadsy et al., 2009), (Akbacak et al., 2011)	(Condon et al., 2010), (Salloum & Habash, 2012)	
South Iraqi								(Sawaf, 2010)	
North Iraqi								(Sawaf, 2010)	
Baghdadi								(Sawaf, 2010)	

Table 2: Dialectical Arabic NLP- Literature Overview[20]

The most prominent corpora collected is the Arabic Online Commentary (AOC) dataset which gathered millions of comments from three newspapers[24].

Though the AOC dataset was big enough, it was not annotated fully, which might harm a predicting model's results. There has been work in creating an annotated dataset built from the AOC dataset alongside North African dialectical data collected from the Tunisian Arabic Corpus[3]. Then the researchers annoatated the collected data by using Amazon's *Mechanical Turk* (MTURK), which hires online annotators[9].

Another improvment of the AOC dataset came from Cotterell and Callison-Burch, in which they extended the AOC newpaper dataset to include about 550K words from 5 newspapers "**Al-Youm Al-Sabe'**, a Saudi-Arabian newspaper **Al-Riyadh**, a Jordanian newspaper **Al-Ghad**, an Algerian newspaper, **Ech Chorouk El Youmi** and an Iraqi newspaper **Al-Wefaq**.". As well as 660k words scraped from twitter tweets. After collecting the extended dataset, they manually annotated them using Amazon's Mechnical Turk[6].

There has been work in using social media as a valid source of dialectic data, creating the Social Media Arabic Dialect Corpus (SMADC) dataset, which scraped and annotated data from Twitter and Facebook[3].

Another dataset is the The Dialectal Arabic Tweets (DART), which manually annotated over 25k tweets in Maghrebi, Egyptian, Levantine, Iraqi, and Gulf[2].

3.3 Dialect classification results

There has been many attempts in solving the problem of this research, many of which use similar strategies. In this section we'll review the highlights of past literature's results.

Zaidan-Burch, the researchers behind the AOC dataset, mentioned in section 3.2 the results they found as well as their methodology. They used a "SRILM toolkit to build word trigram models, with modified Kneser-Ney as a smoothing method, and report the results of 10-fold cross validation"[4]. They have achieved an accuracy of 69.4% at classifying "MSA vs. LEV vs. GLF vs. EGY"[24].

Cotterell-Burch have extended the AOC data, also mentioned in section 3.2 and trained using two algorithms, SVM and Naive Bayes using unigram, bigram and trigram features[6]. The results are displayed in figure 5.

Alshutayri used the SMADC dataset to classify dialects to GLF, NOR, LEV, EGY and IRQ. They used Sequential Minimal Optimization (SMO) algorithm with multinomial Naive Bayes (MNB) with different

[3]http://www.tunisiya.org/

[4]The SRI Language Modeling Toolkit (SRILM) is a toolkit for building statistical language models

	Egy.	Lev.	Mag.	Gulf	Iraqi
NB Uni	**.89**	**.79**	**.92**	**.88**	**.87**
NB Bi	.88	.78	.89	.84	.66
NB Tri	.88	.77	.88	.84	.65
SVM Uni	.88	.78	.89	.85	85
SVM Bi	.87	.75	.87	.82	.79
SVM Tri	.87	.74	.87	.82	.79

Figure 5: "Experiments on newspaper commentary data (accuracy reported)."[6]

tokenizers, run via the data analysis tool WEKA to achieve an accuracy of 60.68%[4].

Alshutayri have also tried lexical methods. One of which is simple voting, which was described in section 2.3.2. Using the SMADC dataset, the researcher achieved an accuracy of 69.19% The researcher also experimented with weighted voting, filtering based on MSA words and more[4].

Words	NOR	EGY	IRQ	LEV	GLF
هههههههههه	0	1	1	1	0
خليتني	0	0	0	0	0
ليش	1	0	1	1	1
بهالطريقة	0	0	0	0	0
بس	1	1	1	1	1
للحين	0	0	0	0	1
بهالشكل	0	0	0	0	0
Total	**2**	**2**	**3**	**3**	**3**

Figure 6: Simple voting matrix representation of هههههههههه خليتني ليش بهالطريقة بس للحين بهالشكل[4]

3.3.1 Deep learning dialect classification results

Most of the research around dialect classification uses traditional text classification methods and deep learning methods are scarcely used, however, the surge of deep learning research has reinvigorated the interest in deep learning text classification.

Elaraby and Abdul-Mageed have used many different algorithms including deep learning algorithms such as CNN, CLSTM, LSTM, BiLSTM, BiGRU and Attention BiLSTM which they explain in their paper as well as traditional classifiers such as SVMs, Naive Bayes and others[10].

On the AOC dataset Elaraby and Abdul-Mageed used this dialect split "MSA vs. Egyptian vs. Gulf vs. Levantine" to obtain an accuracy of 82.45% using the Attention BiLSTM with Abdul-Mageed, et al. embeddings[10].

It's also noteable that the traditional classifiers won over deep learning classifiers only on the "EGY, GLF, and LEV" three way classification split[10].

4 Methodology

4.1 The SMADC dataset

In this research we'll be using the Social Media Arabic Dialect Corpus (*SMADC*) dataset that we talked about in section 3.2. We'll briefly note important details about its collection, filtration and annotation.

4.1.1 Collection and filteration

SMADC's corpus is collected from three different sources, Facebook, Twitter and online newspapers. We'll briefly go over the details of collection for each source.

For Twitter documents, the researchers collected 210,915 tweets, then proceeded to label tweets based on the existence of pre-defined seed words and the location of the tweet's sender as well as the Geolocation of the tweet. For Facebook documents, the researches scraped 2,888,788 comments from 422,070 Facebook posts. They annotated the comments based on the country of the account the post was from. For online newspapers, the researchers collected 10,096 comments from 25 newspapers and were automatically labeled based on the newspaper's origin country. The researchers filtered Facebook and Twitter documents automatically by removing hashtags, emojies, redundant characters and so on.[4]

Though this simple annotation was manually checked and corrected if it were to be false, we'll discuss more of that in section 4.1.2.

4.1.2 Annotation

After automatically annotating the documents in the way we described earlier, the researcher has used novel manual annotation techniques to annotate a part of the dataset. They had created an interactive online quiz where users would log in and manually annotate a number documents. Control documents was placed to check if the user is not randomly choosing options, and annotation conflicts were resolved by choosing majority voting. Resulting in 24,060 manually annotated documents. [4]

4.2 Preprocessing

The researchers used some preprocessing techniques that helps in increasing the models preformance and speed, like Tokenization, Segmentation and Embeddings which we will discuss each in the next sections.

4.2.1 Tokenization

Tokenization is a common task of NLP, tokenization is done by separating a text document into smaller units which is called tokens.

Tokens can be anything the implementer wants but there are some common tokens that are used like separating by word, character or sub-words, an example of arabic tokenization for اللغة العربية sentence would be, for word tokenization it will be like this اللغة #العربية where # works as a separator and for character tokenization it will be like this ا #ل #ل #غ #ة #ا #ل #ع #ر #ب #ي #ة and for sub-words tokenization it will be like this ال #لغة #ال #عربية.

Tokenization is an important part of the preprocessing becasue of how most deep-learing models work, for example the Transformers based models (which the researchers use) process raw text at token level.

The researchers didn't have to make a tokenization function because we made use of the tokenization function that is internally integrated in AraBERT (maybe cite here).

4.2.2 Segmentation

Word segmentation is one of the most important pre-processing steps for many NLP task, especially for rich languages like Arabic.

Arabic word segmentation work by separating the suffix and prefix that is attached to a stem word, for example the word العربية can be segmented to ال + عربي + ة, in this example we can see that the prefix in this word is ال and the suffix is ة and the stem word is عربي, segmentation has shown to have significant impact in many NLP application such as context understanding, because it gives more information to the model.

The researchers didn't have to make a segmentation function because we made use of the segmentation function that is internally integrated in AraBERT which uses Farasa segmenter[11].

4.3 Data representation

In this section the researchers will discuss how did they represent the text so that it will be understood by the computer.

4.3.1 Embeddings

Computers can't understand normal text because it can only see it as a bunch of binary numbers and because of that it can't understand the meanings of words, so the solution to that is embeddings, embeddings works by assigning a unique vector of numbers to a specific word, ideally an embedding captures some of the semantics of the text by placing semantically similar text close together in the embedding space, this works by finding what other words it often appears next to, so for example after finding the embeddings list or dictionary and you want to know what is the most similar words to العربية you would have to calculate it's vector and other vectors distance in the list to measures the similarity and relationship between them.

One problem of normal embeddings is that it doesn't capture the contextual semantic for the word, for example ذهبنا الى البر ثم اكلنا خبز البر here البر doesn't have the same meaning because of the context but in normal embeddings it will have and this is a problem, so the solution to that is what is known as contextual embeddings.

4.3.1.1 Contextual Embeddings

The diffrenace between normal embeddings and contextual embeddings is that the word embeddings capture word semantics in context such that it can represent differently under different context even though it is same word, one example of that is ELMo which BERT uses it internally.

ELMo gained its language understanding from being trained to predict the next word in a sequence of words, ELMo also goes a step further and tries to predict previous word so it can get a sense of the whole sentnce.

4.4 BERT

4.4.1 BERT overview

[7]

4.4.2 Using BERT in text classification

5 Experimental design

5.1 Datasets

5.2 Algorithms

References

1. Felipe Almeida and Geraldo Xexéo. Word embeddings: A survey, 01 2019.

2. Israa Alsarsour, Esraa Mohamed, Reem Suwaileh, and Tamer Elsayed. DART: A large dataset of dialectal Arabic tweets. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA).

3. Areej Alshutayri and Eric Atwell. Classifying arabic dialect text in the social media arabic dialect corpus (smadc). 01 2021.

4. Areej Odah O. Alshutayri. *Arabic Dialect Texts Classification*. PhD thesis, The University Of Leeds, 2018.

5. Fadi Biadsy. *Automatic Dialect and Accent Recognition and Its Application to Speech Recognition*. PhD thesis, Columbia University, USA, 2011.

6. Ryan Cotterell and Chris Callison-Burch. A multi-dialect, multigenre corpus of informal written Arabic. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 241–245, Reykjavik, Iceland, May 2014. European Language Resources Association (ELRA).

7. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018.

8. IBM Cloud Education. What is machine learning?, Jul 2020.

9. Mahmoud El-Haj, Paul Rayson, and Mariam Aboelezz. Arabic dialect identification in the context of bivalency and code-switching. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA).

10. Mohamed Elaraby and Muhammad Abdul-Mageed. Deep models for Arabic dialect identification on benchmarked data. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018)*, pages 263–274, Santa Fe, New Mexico, USA, August 2018. Association for Computational Linguistics.

11. Arabic Language Technologies Group. Farasa is the state-of-the-art full-stack package to deal with arabic language processing, Jul 2020.

12. Nizar Y. Habash. Introduction to arabic natural language processing. *Synthesis Lectures on Human Language Technologies*, 3(1):1–187, January 2010.

13. Daniel Jurafsky and James H Martin. *Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition*. Pearson, 2000.

14. Monkey Learn. Text classification with machine learning & nlp. `https://monkeylearn.com/text-classification`, 2014.

15. Y LeCun, Y Bengio, and G Hinton. Deep learning. 2015.

16. Elizabeth D. Liddy. *Natural Language Processing*. Syracuse University, 2 edition, 2001.

17. Gary Marcus. Deep learning: A critical appraisal, 2018.

18. Tomas Mikolov, Ilya Sutskever, Kai Chen, G.s Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, 26, 10 2013.

19. Hassan Sawaf. Arabic dialect handling in hybrid machine translation. *Proceedings of the Conference of the Association for Machine Translation in the Americas (AMTA)*, 11 2010.

20. Abdulhadi Shoufan and Sumaya Alameri. Natural language processing for dialectical Arabic: A survey. In *Proceedings of the Second Workshop on Arabic Natural Language Processing*, Beijing, China, July 2015. Association for Computational Linguistics.

21. Thoughtvector.io. Subword tokenization - handling misspellings and multilingual data. `https://www.thoughtvector.io/blog/subword-tokenization/`, December 2019.

22. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. 2017.

23. Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October 2020. Association for Computational Linguistics.

24. Omar F. Zaidan and Chris Callison-Burch. The Arabic online commentary dataset: an annotated dataset of informal Arabic with high dialectal content. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 37–41, Portland, Oregon, USA, June 2011. Association for Computational Linguistics.