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

Arabic Text Dialect Recognition

CSC 496 – Proposal

Prepared by:

Mohand Alrasheed	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

Table of Contents

I. 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.

II. 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.

III. 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 انتمائه.

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. Project Details

The aim of this research is to develop a classifier that predicts the dialect of an Arabic text. In essence, the goal of this research is to train an effective machine learning model using recent advancements in the field of NLP. The research requires a machine learning library and a GPU powerful enough for training a heavy machine learning model.

The objectives of the proposed approach are as follows:
1. Analyze and preprocess the labeled dialects dataset.
2. Use this dataset to train a state of the art supervised classifier on distinguishing Arabic dialects.
3. Assess and fine-tune the performance of the model.
4. Use this system to help researchers in the future.

2. Project Timeline

Milestone	**Due date**
Proposal	26 September
Midterm report	11 November
Final report	9 December
Poster presentation	14 December
Implementation	2nd semester
Documentation and final report	2nd semester