Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 1318–1326
Marseille, 11–16 May 2020

Habibi - a multi Dialect multi National Arabic Song Lyrics Corpus

Mahmoud El-Haj
School of Computing and Communications
Lancaster University
United Kingdom
m.el-haj@lancaster.ac.uk

Abstract

This paper introduces Habibi the first Arabic Song Lyrics corpus. The corpus comprises more than 30,000 Arabic song lyrics in 6 Arabic dialects for singers from 18 different Arabic countries. The lyrics are segmented into more than 500,000 sentences (song verses) with more than 3.5 million words. I provide the corpus in both comma separated value (csv) and annotated plain text (txt) file formats. In addition, I converted the csv version into JavaScript Object Notation (json) and eXtensible Markup Language (xml) file formats. To experiment with the corpus I run extensive binary and multi-class experiments for dialect and country-of-origin identification. The identification tasks include the use of several classical machine learning and deep learning models utilising different word embeddings. For the binary dialect identification task the best performing classifier achieved a testing accuracy of 93%. This was achieved using a word-based Convolutional Neural Network (CNN) utilising a Continuous Bag of Words (CBOW) word embeddings model. The results overall show all classical and deep learning models to outperform our baseline, which demonstrates the suitability of the corpus for both dialect and country-of-origin identification tasks. I am making the corpus and the trained CBOW word embeddings freely available for research purposes.

Keywords: Corpus, Arabic NLP, Language Identification, Classification, Dialect Identification, Deep Learning, Word Embeddings

1. Introduction

The Middle East and North Africa (MENA) are now commanding attention when it comes to the global music industry[1]. Most importantly, the different territories and the singers emerging from them are already reaping real benefit from that attention. The increasing availability of mobile data has given rise to a rich and diverse global landscape with fans in each territory accessing music through a unique and evolving combination of formats and services. With more than 400 million Arabic speakers worldwide, I believe it is of significant value to study the language used in the lyrics of Arabic songs.

Lyrics are the words that make up a song and are usually consisting of verses and choruses. Unlike western music, Arabic songs are poorly classified and the majority of the songs available online are classified under either Modern Arabic Pop genre or what is now known as Franco-Arabic, which is a blend of western and eastern music but mainly using Arabic lyrics (Soboh et al., 2017). Arabic genres differ from the western ones (e.g. Rock, Pop, Metal, ...etc) and fall into categories that better describe the region (dialect or country of origin) of the singer rather than the type of music. This is more common in modern Arabic music (Touma and Touma, 2003).

Online music streaming giants such as Spotify[2] and Deezer[3] follow a similar categorisation by providing music mix based on the singer's country of origin (e.g. Lebanese songs). Anghami[4], first Arabic streaming platform, adds extra categories describing the mood of a song rather than the genre (e.g. Happy, Sad, Romance ...etc) but they too rely on the singer's country of origin as a genre.

Arabic is widely spoken and is an official language in 25 countries around the globe with a population of more than 400 million speakers. Arabic countries have many styles of music and also many dialects with each country having its own traditional music. Singers are aware of such variety and try to approach their fans by releasing songs fluently singing in various Arabic Dialects. The majority of Online Music Streaming Services categorise Arabic modern music based on the singer's country of origin and that falls into 6 main genres: Egyptian, Gulf, Levantine (Shami), Iraqi, Sudanese and Maghrebi (North African). Such regional categorisation is mainly based on the origin of the singer as indicated earlier regardless of the lyrics' dialect. For example, and despite singing in Egyptian very often, singer Nancy Ajram[5] is always referred to as Lebanese, meaning that her songs are categorised under the Levantine music genre. In this work I rely on this categorisation to identify the dialect of the song lyrics for each of the songs in Habibi corpus.

In addition to the aforementioned genres, there are attempts by Online Music Streaming Services to add Arabic specific sub-genres such as "Dabke", "Chobi", "Shaabi", "Raï" and "Samri" but despite not being explicit those genres are still referring to regions within the Arab world. For example Dabke is a type of Levantine dance native to Lebanon, Jordan, Syria and Palestine as well as Iraq[6]. In Iraq Dabke is referred to as Chobi. Shaabi[7] indicates a type of folk music that is known to Egypt, also referred to as "Mahraganat[8]". Similarly sub genres such as Raï[9] and Samri[10] refer to Al-

[1] www.ifpi.org/
[2] www.spotify.com/
[3] www.deezer.com/
[4] www.anghami.com/
[5] en.wikipedia.org/wiki/Nancy_Ajram
[6] en.wikipedia.org/wiki/Dabke
[7] en.wikipedia.org/wiki/Shaabi
[8] en.wikipedia.org/wiki/Mahraganat
[9] en.wikipedia.org/wiki/Raï
[10] en.wikipedia.org/wiki/Samri

gerian/Moroccan and Gulf music respectively. The use of those sub-genres is limited in comparison to the 6 main regional genres mentioned above.

In this paper I introduce **Habibi Corpus**, an open-source Arabic song lyrics dataset which I called "Habibi" (حبيبي) – a gender-neutral word to mean "my love" in English. Habibi corpus comprises of more than 30,000 Arabic songs segmented into verses (sentences thereafter). The resource provides a rich and diverse venue for researchers working on Dialects Identifications and Authorship Attribution. In this paper I use Habibi corpus to experiment with automatic dialect identification for 6 Arabic dialects: Egyptian, Gulf, Levantine, Iraqi, Sudanese and Maghrebi. In addition the experiments include a task on identifying the country of origin for each singer.

I analysed the song titles in the corpus and found the word "Habibi"[11] to be the most frequent word appearing in more than 35% of Habibi's 30,000 song titles.

Habibi corpus is made of more than 520,000 sentences comprising of more than 3.5 million words. Figures 1 and 2 show the most significant and most frequent words among all 6 dialects of Habibi corpus. The word-cloud shows the word (حبيبي) as one of the most significant words in the corpus.

Figure 1: Dialects Word Cloud

To facilitate researching Arabic lyrics I experiment with the corpus using Deep Learning to automatically identify the dialect of songs in the corpus. The experiments also include the use of Arabic Wiki FastText[12] pre-trained word embeddings as well as Habibi's word embeddings which I built using the song lyrics without the use of any labels or genres. Habibi corpus and Habibi's word embeddings are made freely available for research purposes[13].

2. Related Work

Research on Arabic song lyrics has not received enough attention due to the lack of classified and segmented datasets and also the many conflicting opinions on the suitability of the language used in those songs (Aquil, 2012). The work by Aquil (2012) shows that Arabic songs are rarely viewed as an object of study and analysis or the microscope into the broader culture. The study suggests that Arabic songs need to be viewed differently, they need to be viewed as art, as culture, as history, as philosophy, as group identity, as the way of words, and as the actual voices of people. Despite the common misconception, research found dialectical Arabic Hip Hop songs to use local themes in addition to offering explanation and context to the historical and cultural background of the Arabic region (Terkourafi, 2010).

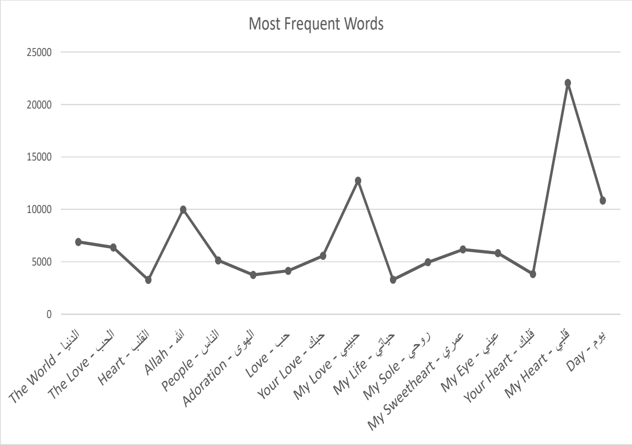


Figure 2: Dialects Word Cloud

As the experiments in this paper are mainly concerned with the automatic identification of Arabic dialects and country of origin, I explore related work on experiments and datasets used for identifying Arabic dialects.

Arabic dialect identification has been an active research topic recently[14]. The majority of Arabic dialectical datasets are collected from Twitter[15] and online fora. Those datasets' instances (text samples) are limited in size due to the enforced restrictions by such platforms (e.g. Twitter only allows 280 characters per tweet). This sometimes results in short tweets and noisy data (Qing Chen et al., 2010).

Zaidan and Callison-Burch (2014) created the Arabic Online Commentary (AOC) corpus. AOC is an Arabic resource of dialect annotation using Mechanical Turk crowdsourcing[16]. The annotators labeled 100,000 sentences defining the Arabic dialect used in writing. The authors trained a simple classifier to identify dialectal Arabic in text harvested from online social media. The dialects used to train their classifier were Egyptian, Gulf, Levant, Iraqi and Maghrebi.

Bouamor et al. (2019) presented MADAR Travel Domain Corpus at the 4th Workshop on Arabic NLP. The dataset comprises of parallel sentences covering 25 Arabic dialects in addition to Modern Standard Arabic (MSA), French and

[11] This includes other variation of Habibi (e.g. حبِيبي and يَاحبيبي)

[12] https://fasttext.cc/

[13] http://ucrel-web.lancaster.ac.uk/habibi/

[14] https://sites.google.com/view/madar-shared-task

[15] www.twitter.com/

[16] www.mturk.com/

English. The dataset is made up of two sub-corpora. First sub-corpus consists of 2,000 parallel sentences translated into 25 Arabic city dialects. The second sub-corpus has an additional 10,000 sentences translated into dialects of 5 major cities in MENA (i.e. Beirut, Cairo, Doha, Tunis and Rabat).

Alsarsour et al. (2018) introduced the Dialectal Arabic Tweets (DARTS)[17] which contains 25,000 Arabic sentences labeled into 5 Arabic dialects (Egyptian, Gulf, Levantine, Iraqi and Maghrebi). The dataset was collected from Twitter and classified into the aforementioned 5 dialects using Crowdsourcing participants from Figure-Eight (known previously as Crowdflower)[18].

The use of crowd-sourcing as in (Zaidan and Callison-Burch, 2014) and (Alsarsour et al., 2018) suffers from a number of quality issues mainly related to the process of filtering spam annotators. Multiple annotators are needed in order to reach a certain degree of agreement among annotators, this remains a challenging task both cost and time -wise. The other problem is the fact that comments and tweets are short and do not contain enough context (Ritter et al., 2011). For example the AOC corpus contains more than 7,000 sentences with less than 10 characters (a maximum of two words), the majority of those sentences are highly overlapping across dialects making automatic identification of dialects using machine learning a challenging task (El-Haj et al., 2018).

3. Data Description

Habibi is the first freely available corpus of Arabic song lyrics. The corpus comprises of more than 30,000 Arabic songs from 18 different Arab countries as shown in Figure 3. Each Song in the corpus comes with a song title along with the singer's full name, country of origin and dialect. I also provide information about the song's writer and composer as shown in Table 1. Songs where a composer or song-writer is missing or nonexistent are simply annotated with "unknown" (e.g. <composer>unknown</composer>)[19]. Each song has a singer, song title, and lyrics with each singer's country of origin and dialect. The corpus was collected using Web as Corpus method (Kilgarriff and Grefenstette, 2001) which was used to collect song lyrics from the web. Based on the structure of the collected lyrics I extract the information shown in Table 1. I used Google[20] and Wikipedia[21] to automatically extract details about each singer's country of origin. I then grouped the countries into the 6 main dialects (Table) 2 based on the dialect spoken by each country.

Figures 4 and 5 show the song count by dialect and singer's country of origin. The corpus is balanced for the Egyptian, Levantine and Gulf dialects with Iraqi just falling short. This is due to the historical popularity of Egyptian, Gulf and Levantine songs. Recently Iraqi songs are gaining a rapid popularity with videos being watched more than 1.5 billion times on YouTube[22].

Songs	30,072
Song Titles	30,072
Sentences	527,870
Singers	1,765
Song writers	3,789
Composers	2,463
Countries	18
Dialects	6

Table 1: Habibi Corpus Stats

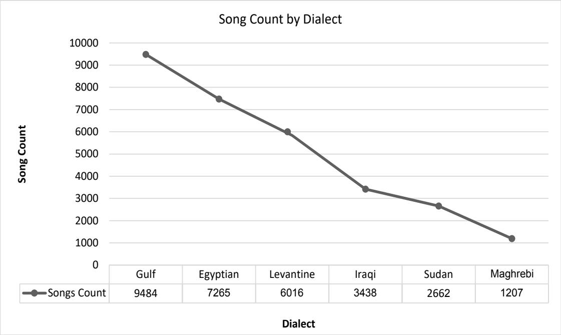


Figure 4: Songs Count by Dialect

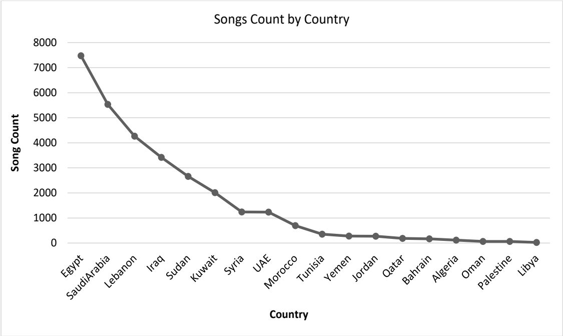


Figure 5: Songs Count by Country

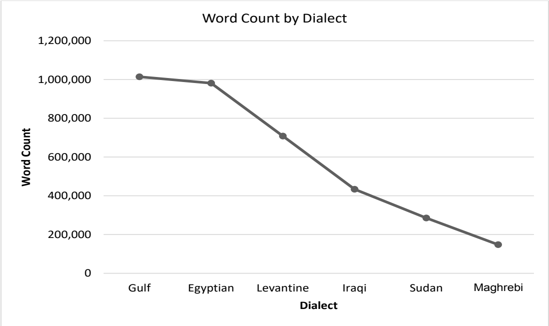


Figure 6: Dialects Word Distribution

[17] http://qufaculty.qu.edu.qa/telsayed/datasets/

[18] www.figure-eight.com/

[19] <composer>غيرمعروف</composer>

[20] www.google.com/

[21] www.wikipedia.org/

[22] www.youtube.com/

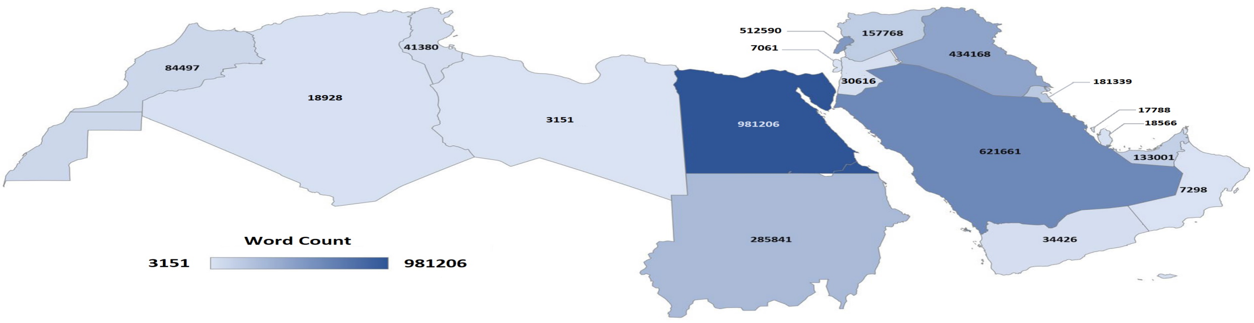


Figure 3: Country Word Count Map

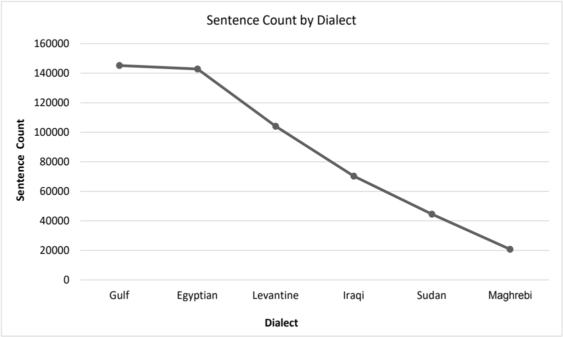


Figure 7: Sentence Count by Dialect

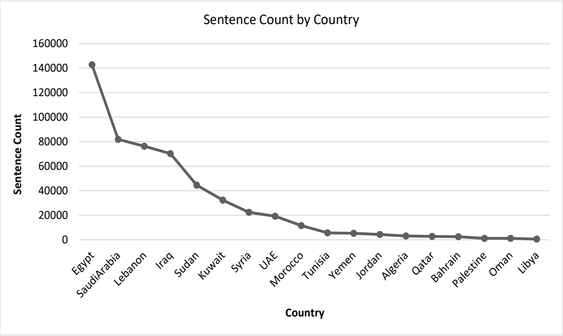


Figure 8: Sentence Count by Country

The corpus contains more than 3.5 million words that are free from spam words, ads, hashtags or emojis resulting in a clean and noise-free dataset. Table 2 and 3 show the total word and song count by dialect and country. Figures 3 and 6 show the word count distribution for the 6 dialects and 18 countries of Habibi corpus.

The corpus is segmented into sentences according to the original song lyrics' verses. Figures 7 and 8 show the sentences count by dialect and singer's country of origin with a total of more than 520,000 sentences.

Habibi is available in comma separated value (csv) format and as annotated plain UTF-8 text files (txt). The plain text (txt) format contains html-like annotation-tags to identify the song and singer details as well as the lyrics of each song. Figure 9 shows a txt annotated sample, the format is consistent across the corpus with a separate file for each song. The csv format follows the same annotations with the tags appearing as column headers. In addition, I converted the csv version into JavaScript Object Notation (json[23]) and eXtensible Markup Language (xml) file formats. As in most of the dialectical datasets, Habibi corpus vocabularies overlap across dialects as shown in Figure 10 heat-map.

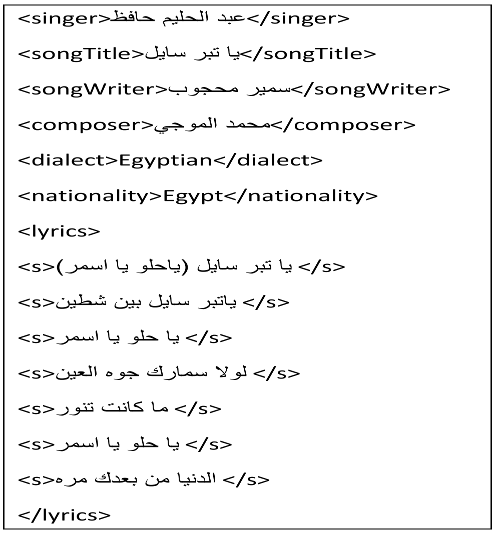
Dialect	Word Count	%	#songs
Gulf	1,014,079	28.40%	9,484
Egyptian	981,206	27.47%	7,265
Levantine	708,035	19.83%	6,016
Iraqi	434,168	12.16%	3,438
Sudan	285,841	8.00%	2,662
Maghrebi	147,956	4.14%	1,207
Total	**3,571,285**	**100%**	**30,072**

Table 2: Habibi Word Count by Dialect

County	Dialect	Word Count	%	#songs
Egypt	Egyptian	981,206	27.47%	7,265
Saudi	Gulf	621,661	17.41%	5,823
Lebanon	Levantine	512,590	14.35%	4,350
Iraq	Iraqi	434,168	12.16%	3,438
Sudan	Sudanese	285,841	8.00%	2,662
Kuwait	Gulf	181,339	5.08%	1,727
Syria	Levantine	157,768	4.42%	1,333
UAE	Gulf	133,001	3.72%	1,237
Morocco	Maghrebi	84,497	2.37%	709
Tunisia	Maghrebi	41,380	1.16%	356
Yemen	Gulf*	34,426	0.96%	279
Jordan	Levantine	30,616	0.86%	271
Algeria	Magherbi	18,928	0.53%	117
Qatar	Gulf	18,566	0.52%	185
Bahrain	Gulf	17,788	0.50%	166
Oman	Gulf	7,298	0.20%	67
Palestine	Levantine	7,061	0.20%	62
Libya	Maghrebi	3,151	0.09%	25
————	**Total**	**3,571,285**	**100%**	**30,072**

Table 3: Word Count by Country
* due to size limitation Yemeni dialect has been labelled as Gulf based on approximation.

[23] The json file displays Arabic text as UTF-8 unicode format.



```
<singer>عبد الحليم حافظ</singer>
<songTitle>يا تبر سايل</songTitle>
<songWriter>سمير محجوب</songWriter>
<composer>محمد الموجي</composer>
<dialect>Egyptian</dialect>
<nationality>Egypt</nationality>
<lyrics>
<s>يا تبر سايل (ياحلو يا اسمر) </s>
<s>ياتبر سايل بين شطين </s>
<s>يا حلو يا اسمر </s>
<s>لولا سمارك جوه العين </s>
<s>ما كانت تنور </s>
<s>يا حلو يا اسمر </s>
<s>الدنيا من بعدك مره </s>
</lyrics>
```

Figure 9: Habibi Corpus Sample

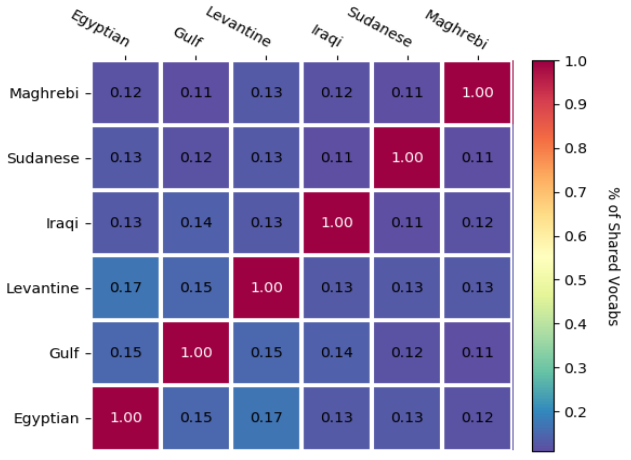


Figure 10: Dialects Shared Vocabs Heatmap

4. Dialect and Country Identification Setup

I divided the experiments into two tasks: Dialect, and Country Identification. All the reported experiments are conducted on a sentence-level where each sentence is a verse from a song, those surrounded with the sentence tags (<s> </s>) as shown in Figure 9. Each sentence is labelled with either the dialect or the singer's country of origin. Both the dialect and the country are associated with the singer rather than the text as explained earlier in the paper. Relabelling the dataset to be associated with the lyrics' dialects will be conducted as future work.

The experiments are setup up to show binary and multi-class classification problems as follows:

1. Dialect Identification. This is done by classifying the lyrics into 6 main dialects: Egyptian, Gulf, Levantine, Iraqi, Sudanese and Maghrebi. The dialect identification experiments are performed on different levels as follows:
   (a) *top_2 dialects*: a binary classification in which the top two dialects are used (i.e. Gulf and Egyptian as in Table 2).
   (b) *top_n dialects*: a three, four and five way classification of the top n dialects in Table 2. This is a multi-class classification where: 2 > n <= 5.
   (c) *all dialects (all_6)*: taking into consideration all 6 dialects in Habibi corpus.
2. Country-of-Origin Identification: classifying the lyrics into 18 different countries as show in Table 3. The experiments are performed as follows:
   (a) *top_2 countries*: This is a binary classification in which the top two countries are used (i.e. Egypt and Saudi Arabia as in Table 3).
   (b) *top_n countries*: a three, four and five way classification of the top n countries in Table 3. This is a multi-class classification where: 2 > n <= 18.
   (c) *all countries (all_18)*: taking into consideration all 18 countries of all singers in Habibi corpus.

5. Classical and Deep Learning Models

To experiment with Habibi corpus I run a number of experiments using classical machine learning and deep learning models.

5.1. Classical Machine Learning

For the purpose of this task I trained different text classifiers using three classical machine learning algorithms: multinomial Naïve Bayes (NB), Logistic Regression (LR) and Support Vector Machine (SVM), following the same setup as in (El-Haj et al., 2018; El-Haj et al., 2016).

5.2. Deep Learning Models

To demonstrate the suitability of Habibi dataset for text classification tasks I apply a number of deep learning neural network models to automatically identify dialects and countries as follows[24]:

5.2.1. Convolutional Neural Network (CNN)

The idea of using CNN to classify text was first described by (Kim, 2014).

CNN is a class of deep, feed-forward, artificial neural networks. Connections between nodes do not form a cycle and use a variation of multilayer perceptrons designed to require minimal pre-processing. CNNs are generally used in computer vision, however they have shown to perform well for text classification tasks. In this paper I apply a word based CNN model similar to (Elaraby and Abdul-Mageed, 2018) and (Kim, 2014). The architecture used to build the model is described by (Elaraby and Abdul-Mageed, 2018).

5.2.2. Long-Short Term Memory (LSTM)

I use LSTM recurrent neural network (RNN) with an architecture consisting of 100 dimensions hidden units (Hochreiter and Schmidhuber, 1997).

[24] The models applied are described in more details in (Elaraby and Abdul-Mageed, 2018), GitHub Repository: https://github.com/ubc-nlp/aoc_id.

5.2.3. Bidirectional LSTM (BiLSTM)

Similar to the LSTM model but processing the data in both directions in two separate hidden layers.

5.2.4. Convolutional Long-Short Term Memory (CLSTM)

CLSTM utilises CNN to extract a sequence of higher-level phrase representations, and are fed into a LSTM to obtain the sentence representation (Zhou et al., 2015; Sainath et al., 2015).
The model used in this paper is based on the work by (Elaraby and Abdul-Mageed, 2018) where a CNN layer is used as a feature extractor by directly feeding the convolution output to a LSTM layer in order to capture long-term dependencies.

5.2.5. Bidirectional Gated Recurrent Units (BiGRU)

The BiGRU model applied in this work is based on the work by (Chung et al., 2014). The architecture used is described in (Elaraby and Abdul-Mageed, 2018) who applied a bidirectional GRU by combining two GRUs each looking at a different direction in a process similar to a BiLSTM network.

5.3. Word Embeddings

For the purpose of this study two Continuous Bag of Words (CBOW) word embeddings models have been used. The first model is the pre-trained FastText word embeddings model which was trained on Wikipedia Arabic articles with 300-dimension vectors (Mikolov et al., 2013). The second model is the in-house word embeddings that I built using Habibi's text without labels, this is referred to as "Habibi's Word Embeddings". To train the word embeddings I used 3.5M words to train a CBOW with 300-dimension vectors. Habibi's word embeddings is made freely available for research purposes[25].

6. Results

In this section I show the experimental results of both the classical machine learning and the deep learning models.
The experiments were conducted on a sentence level using binary and multi-class classification. Working on dialectical level the dataset is balanced for the Egyptian, Levantine and Gulf dialects with Iraqi just falling short. The dataset was randomly split into 70% training and 30% for validation and testing.
As shown in Figures 7 and 8 Gulf and Egypt are the most frequent dialect and country respectively. Gulf is made up of several countries in comparison to Egyptian which is only spoken in one country.
The accuracy of the most frequent class in both dialect and country datasets will be used as a baseline across all experiments conducted in this paper. Baselines accuracies in terms of most frequent class are shown in 4.
Tables 5 and 6 show the results in terms of training and testing accuracy for the classical machine learning experiments. Overall and as expected the algorithms perform better on dialects in comparison to the country of origin, that is simply due to having more classes in the case of country identification. This is also noticeable when comparing the binary (*top_2*) classification against the rest of the multi-class experiments.
The results overall show good training and testing accuracy even when the classes are less balanced. For example the (*all_18*) experiments still perform better than chance and the most-frequent-class baseline combined despite being highly imbalanced.

The drop in accuracy upon increasing the number of classes is not a surprise considering the level of vocabulary overlap between the dialects as shown in Figure 10.

Dataset	Accuracy
Dialect (Gulf)	28.40%
Country (Egypt)	27.47%

Table 4: Most Frequent Class Accuracy

Dataset	Algorithm	Train	Test
top_2	NB	96.8%	92.6%
top_2	LR	95.0%	91.2%
top_2	SVM	95.0%	90.3%
top_3	NB	92.2%	80.1%
top_3	LR	90.7%	76.6%
top_3	SVM	92.4%	75.0%
top_4	NB	90.8%	78.2%
top_4	LR	90.4%	74.5%
top_4	SVM	93.2%	73.2%
top_5	NB	90.5%	75.8%
top_5	LR	93.1%	74.5%
top_5	SVM	92.4%	73.2%
all_6	NB	87.9%	72.6%
all_6	LR	89.9%	71.4%
all_6	SVM	87.6%	69.8%

Table 5: Classical Machine Learning by Dialect

Dataset	Algorithm	Train	Test
top_2	NB	97.0%	92.2%
top_2	LR	96.0%	90.7%
top_2	SVM	95.0%	90.0%
top_3	NB	93.1%	83.0%
top_3	LR	91.7%	79.1%
top_3	SVM	92.2%	77.9%
top_4	NB	92.1%	80.7%
top_4	LR	92.4%	77.6%
top_4	SVM	90.4%	75.9%
top_5	NB	92.5%	78.9%
top_5	LR	92.5%	77.1%
top_5	SVM	93.2%	75.5%
all_18	NB	76.7%	60.9%
all_18	LR	75.4%	59.4%
all_18	SVM	72.3%	58.6%

Table 6: Classical Machine Learning by Country

[25] http://ucrel-web.lancaster.ac.uk/habibi/

Dataset	Model	Train	Val	Test
top_2	CNN	96.5%	90.9%	92.9%
top_2	LSTM	96.3%	90.8%	92.7%
top_2	CLSTM	96.5%	90.5%	92.4%
top_2	BiGRU	96.2%	90.1%	91.8%
top_2	BiLSTM	92.7%	88.8%	89.9%
top_3	CNN	85.6%	73.7%	73.1%
top_3	BiGRU	84.9%	72.5%	72.4%
top_3	CLSTM	86.3%	71.5%	72.0%
top_3	LSTM	85.6%	71.0%	70.3%
top_3	BiLSTM	71.2%	68.6%	68.7%
top_4	CNN	82.5%	72.7%	71.7%
top_4	BiLSTM	71.4%	67.4%	66.8%
top_4	LSTM	74.9%	64.4%	64.5%
top_4	CLSTM	77.8%	61.9%	60.7%
top_4	BiGRU	75.7%	61.9%	59.7%
top_5	CNN	84.6%	68.1%	67.9%
top_5	LSTM	74.0%	64.8%	64.7%
top_5	BiGRU	72.4%	60.3%	59.5%
top_5	CLSTM	73.6%	56.6%	57.1%
top_5	BiLSTM	58.5%	56.3%	55.3%
all_6	CNN	72.1%	59.4%	58.8%
all_6	BiGRU	64.9%	54.0%	54.6%
all_6	CLSTM	68.1%	52.2%	52.5%
all_6	LSTM	61.8%	51.8%	51.9%
all_6	BiLSTM	49.9%	44.0%	46.0%

Table 7: Dialect Identification + FastText Embeddings

Dataset	Model	Train	Val	Test
top_2	CNN	97.6%	91.2%	91.5%
top_2	BiGRU	97.6%	91.3%	91.4%
top_2	CLSTM	97.8%	91.7%	91.2%
top_2	LSTM	97.8%	90.8%	91.0%
top_2	BiLSTM	94.7%	90.7%	90.1%
top_3	CNN	89.9%	77.8%	77.8%
top_3	BiGRU	87.4%	71.4%	72.0%
top_3	CLSTM	84.9%	72.2%	71.8%
top_3	LSTM	86.7%	71.0%	70.5%
top_3	BiLSTM	70.7%	64.4%	65.6%
top_4	CNN	83.2%	71.7%	70.4%
top_4	BiGRU	79.0%	64.0%	64.1%
top_4	BiLSTM	68.7%	64.9%	64.0%
top_4	CLSTM	82.4%	64.6%	63.1%
top_4	LSTM	81.0%	61.7%	62.0%
top_5	CNN	73.1%	63.0%	64.3%
top_5	CLSTM	76.0%	59.7%	60.2%
top_5	BiGRU	76.0%	58.2%	59.7%
top_5	LSTM	78.3%	58.0%	57.9%
top_5	BiLSTM	58.9%	55.5%	56.9%
all_18	CNN	57.0%	48.6%	46.7%
all_18	CLSTM	56.1%	48.0%	45.6%
all_18	BiGRU	51.5%	46.1%	44.1%
all_18	BiLSTM	44.5%	42.9%	41.0%
all_18	LSTM	44.0%	41.9%	40.2%

Table 8: Country Identification + FastText Embeddings

Dataset	Model	Train	Val	Test
top_2	CNN	96.6%	90.9%	93.0%
top_2	CLSTM	96.6%	90.8%	92.5%
top_2	LSTM	96.2%	90.7%	92.0%
top_2	BiGRU	96.0%	90.0%	91.7%
top_2	BiLSTM	93.7%	89.7%	91.1%
top_3	CNN	85.8%	75.4%	75.0%
top_3	BiGRU	85.4%	72.8%	72.9%
top_3	CLSTM	86.4%	71.8%	71.5%
top_3	LSTM	85.6%	71.5%	70.8%
top_3	BiLSTM	70.3%	68.8%	69.0%
top_4	CNN	81.2%	69.9%	70.5%
top_4	LSTM	82.4%	66.0%	66.6%
top_4	CLSTM	80.6%	67.5%	66.6%
top_4	BiLSTM	69.8%	65.6%	65.6%
top_4	BiGRU	67.1%	53.2%	55.0%
top_5	CNN	79.8%	66.9%	67.6%
top_5	BiGRU	73.5%	62.6%	63.1%
top_5	CLSTM	71.1%	59.1%	58.0%
top_5	LSTM	73.2%	55.6%	55.0%
top_5	BiLSTM	54.7%	53.4%	52.9%
all_6	CNN	74.7%	63.0%	63.1%
all_6	BiGRU	66.8%	54.0%	53.9%
all_6	CLSTM	62.4%	52.7%	51.6%
all_6	LSTM	63.7%	50.2%	48.3%
all_6	BiLSTM	51.4%	48.7%	47.9%

Table 9: Dialect Identification + Habibi's Embeddings

Dataset	Model	Train	Val	Test
top_2	CNN	97.6%	91.7%	91.2%
top_2	LSTM	97.3%	91.3%	91.1%
top_2	BiGRU	97.2%	91.4%	90.9%
top_2	BiLSTM	94.6%	91.6%	90.7%
top_2	CLSTM	94.5%	82.3%	83.9%
top_3	CNN	88.8%	76.8%	76.6%
top_3	CLSTM	83.6%	66.6%	66.7%
top_3	BiGRU	73.9%	67.4%	66.4%
top_3	LSTM	65.4%	59.1%	59.3%
top_3	BiLSTM	56.6%	51.5%	52.6%
top_4	CNN	83.8%	72.8%	70.7%
top_4	BiLSTM	72.5%	65.6%	65.8%
top_4	CLSTM	83.8%	65.4%	63.6%
top_4	LSTM	80.9%	61.9%	60.7%
top_4	BiGRU	70.5%	55.8%	55.3%
top_5	CNN	73.8%	60.9%	62.8%
top_5	CLSTM	78.0%	59.6%	59.6%
top_5	LSTM	72.4%	55.9%	57.0%
top_5	BiLSTM	56.1%	53.4%	53.6%
top_5	BiGRU	65.2%	53.3%	53.5%
all_18	CNN	60.1%	51.2%	49.1%
all_18	CLSTM	57.2%	47.7%	45.9%
all_18	BiGRU	51.2%	45.6%	43.1%
all_18	BiLSTM	43.5%	41.5%	40.6%
all_18	LSTM	42.0%	41.3%	40.2%

Table 10: Country Identification + Habibi's Embeddings

On par with the results reported in (Elaraby and Abdul-Mageed, 2018), the testing accuracy results for both dialect and country identification tasks in Tables 5 and 6 show Naïve Bayes to outperform both SVM and LR in all the binary and multi-class experiments, which confirms with the study conducted by McCallum et al. (1998) where they show multinomial Naïve Bayes to work well for text classification.

As explained earlier, the deep learning experiments are conducted using two different word embeddings: a) Arabic Wiki FastText Embeddings and b) Habibi's Word Embeddings. Both word embeddings were created by training a CBOW with 300-dimension vectors.

Tables 7 and 8 show the results using FastText Embeddings to identify dialect and country of origin using the 6 deep learning models mentioned in Section 5.2..

Tables 9 and 10 show the results using Habibi Word Embeddings. Using deep learning models to identify dialects works better using Habibi's word embeddings, but to the contrary using FastText word embeddings works better for country identification.

Examining the results in all four tables we can observe that the word-based CNN model outperforms all other deep models across all binary and multi-class experiments for both dialect and country identification tasks.

Comparing deep learning test accuracy scores to the classical machine learning results in Tables 5 and 6 shows CNN to outperform all classical models only at the binary classification level. This is expected as there is more text for the neural network to learn from in comparison to when there are more classes where the amount of text available for each class drops.

Classical models, especially Naïve Bayes, outperform the deep learning models in all multi-class experiments for both dialect and country identification tasks.

As mentioned earlier, the baseline in both dialect and country identification tasks is the most frequent class as shown in Table 4. The results overall show all classical and deep learning models to outperform the baseline demonstrating the suitability of the corpus for both dialect and country identification tasks.

7. Conclusion and Future Work

In this paper I introduce **Habibi**– a multi-dialect multi-national corpus comprising of more than 30,000 Arabic song lyrics. The paper shows extensive experimental results demonstrating the suitability of the corpus for both dialect and country identification. In addition, I trained a Continuous Bag of Words (CBOW) model to use with the identification tasks.

The experiments include the use of classical machine learning and deep learning neural network models with the use of FastText pre-trained word embeddings in addition to Habibi's in-house word embeddings model. The paper reports results in terms of training and testing accuracy. The results can be used as benchmarks for any future experiments on the corpus. The results find Naïve Bayes to outperform all classical machine learning models across all experiments for both dialect and country identification.

Using deep learning models the results show the word based CNN model to outperform all other deep learning models for both dialect and country identification. CNN performs slightly better using our own Habibi word embedding in comparison to the use of FastText word embeddings.

The experiments also show the word level CNN model to outperform all the deep learning and classical machine learning models used in the dialect identification binary classification task.

For future work I plan to use Habibi corpus for authorship attribution through running experiments that take into consideration the singers as well as the song writers. I also plan to conduct experiments on detecting dialect taking into consideration the song writer's country of origin, which should help in more refining the dialects in order to overcome the problem of singers singing in different dialect than their own. The dataset currently has no information on the songwriters' country of origin. Acquiring such information requires the use of automatic information extraction from a knowledge base in combination with the wisdom of the crowd.

Habibi corpus as txt, csv, json and xml file formats and Habibi's word embeddings are all available as a free online repository for research purposes and can be directly downloaded from: `http://ucrel-web.lancaster.ac.uk/habibi/`.

References

Alsarsour, I., Mohamed, E., Suwaileh, R., and Elsayed, T. (2018). 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 7-12, 2018. European Language Resources Association (ELRA).

Aquil, R. (2012). Revisiting songs in language pedagogy. *Journal of the National Council of Less Commonly Taught Languages*, 11:75–95.

Bouamor, H., Hassan, S., and Habash, N. (2019). The MADAR shared task on Arabic fine-grained dialect identification. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 199–207, Florence, Italy, August. Association for Computational Linguistics.

Chung, J., Gulcehre, C., Cho, K., and Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*.

El-Haj, M., Rayson, P. E., Young, S. E., Walker, M., Moore, A., Athanasakou, V., and Schleicher, T. (2016). Learning tone and attribution for financial text mining.

El-Haj, M., Rayson, P., and Aboelezz, M. (2018). 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. European Language Resources Association (ELRA).

Elaraby, M. and Abdul-Mageed, M. (2018). 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. Association for Computational Linguistics.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Comput.*, 9(8):1735–1780, November.

Kilgarriff, A. and Grefenstette, G. (2001). Web as corpus. In *Proceedings of Corpus Linguistics 2001*, pages 342–344. Corpus Linguistics.

Kim, Y. (2014). Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar, October. Association for Computational Linguistics.

McCallum, A., Nigam, K., et al. (1998). A comparison of event models for naive bayes text classification. In *AAAI-98 workshop on learning for text categorization*, volume 752, pages 41–48. Citeseer.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Qing Chen, Shipper, T., and Khan, L. (2010). Tweets mining using wikipedia and impurity cluster measurement. In *2010 IEEE International Conference on Intelligence and Security Informatics*, pages 141–143, May.

Ritter, A., Clark, S., Etzioni, O., et al. (2011). Named entity recognition in tweets: an experimental study. In *Proceedings of the conference on empirical methods in natural language processing*, pages 1524–1534. Association for Computational Linguistics.

Sainath, T. N., Vinyals, O., Senior, A., and Sak, H. (2015). Convolutional, long short-term memory, fully connected deep neural networks. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4580–4584. IEEE.

Soboh, L., Elkabani, I., and Osman, Z. (2017). Arabic cultural style based music classification. In *2017 International Conference on New Trends in Computing Sciences (ICTCS)*, pages 6–11, Oct.

Terkourafi, M. (2010). *The Languages of Global Hip Hop*. Continuum International Publishing Group, New York, USA.

Touma, H. H. and Touma, H. (2003). *The music of the Arabs*. Hal Leonard Corporation.

Zaidan, O. F. and Callison-Burch, C. (2014). Arabic dialect identification. *Comput. Linguist.*, 40(1):171–202, March.

Zhou, C., Sun, C., Liu, Z., and Lau, F. C. M. (2015). A C-LSTM neural network for text classification. *CoRR*, abs/1511.08630.