**A Combined CNN and LSTM Model for**

**Arabic Sentiment Analysis**

Saif Rahal

**Introduction:**

In the past decade, social media networks have become a valuable resource for data of different types, such as texts, photos, videos, voices, GPS reading, etc.

Sentiment analysis refers to the use of natural language processing, text analysis to systematically identify, extract, quantify, and study affective states and subjective information.

Sentiment analysis is widely applied to voice of the customer materials such as reviews and survey responses, online and social media, and healthcare materials for applications that range from marketing to customer service to clinical medicine.

Sentiment analysis deals with the texts or the reviews of people that include opinions, sentiments, attitudes, emotions, statements about products, services, foods, films, etc.

There will be three different levels of sentiment analysis for each proposed dataset. The reason of using different levels is to try to expand the number of features in short tweets and to deal with many forms of a single word in Arabic.

**The levels are as follows:**

1. Character Level (Char-level)

by converting the sentence into characters instead of words, At this level, the number of features is increased

1. Character N -Gram Level (Ch5gram-level)

where we measure the length of all the words in each dataset and we calculate the average length of words (which is five characters for all the different datasets). Then, we split any word that, has more than the average number into several sub-words. Whereas, any word that consist of the same average number of characters or less will be kept as it is. The average word’s length for each dataset is five characters. This level can be useful in order to deal with many forms of Arabic words, especially for words with more than five letters.

1. Word Level (Wordlevel)

where the sentence is divided into words using the space as splitter, this level is the most commonly chosen option in the field of sentiment analysis.

**Study Knowledge:**

1. The connection between LSTM and Neural Network.
2. Learning about the CNN model.
3. Learning how to get better accuracy of the output.
4. Learning the reLU function.
5. Learning about the sigmoid function.

**The Project Flow:**

Load datasets:

* AAQ – 4000+ Tweets
* AJDG - 1800+ Jordanian spelling tweets
* ArTweets- 3000+ Tweets with different spelling

Preprocess the data:

We used help functions to clean and normalize the data sets.

We build 3 main functions for the data split levels.

Building the model:

In our project the architecture we have used is a Combined CNN and LSTM layers:

* Conv1D - (32 Filters , size of kernel 3 ,strides 1)
* Maxpool1 - (pool size 2 , strides 2)
* Conv1D - (32 Filters , size of kernel :3 ,strides 1)
* Maxpool2- (pool size 2 , strides 2)
* LSTM – (50 cells)
* LSTM – (50 cells)
* LSTM – (2 cells)
* FullyConnected1 – (Dense = 1, activation = sigmoid)
* Loss = binary\_crossentropy
* Optimizer = nadam

Train and evaluate the model:

Training the model with 3 different datasets and each set is split at 3 ways,

at this table we can see the results of each data set with the relative methods.

|  |  |  |  |
| --- | --- | --- | --- |
| **Sentiment Level** | **AAQ** | **AJGT** | **ArTwitter** |
| Char-level | 0.61 | 0.58 | 0.70 |
| Ch5gram-level | 0.75 | 0.82 | 0.86 |
| Word-level | 0.75 | 0.83 | 0.85 |