Mowjaz Multi-Topic Labelling Task

Yahya Daqour
Dept. computer science
(undergraduate)
Jordan University of Science and
Technology
Amman, Jordan
yahyadaqour@gmail.com

***Abstract*—This paper focuses on the task of Mowjaz Multi-Topic labelling using Bi-directional Gated Recurrent Unit (Bi-GRU), the model is basically used to classify articles based in their topics that are present within its content. Mowjaz's topic are classified into ten categories and an article can be classified as under as many topics as it covers. In the evaluation, we regard the Mowjaz Multi-Topic labelling task as multi-classification task and use Unigram models to extract features to train a neural network classifier. In the result, the accuracy of my method reached 0.8232, ranking 8th.**

Keywords—Unigram, Mowjaz Multi-Topic labelling, Multi classification, neural network

I. Introduction

Mowjaz is an Arabic topical content aggregation mobile application for news, sport, entertainment and other topics from top publishers that users can follow. With the rapid growth of technology, it is essential to make an enjoyable user experience, our work aims to help users get and display the most relevant news to their interests.

Mawdoo3 company provided a dataset that consists of 9,590 articles classified to 10 topics each topic has 950 **±** articles, the articles are collected from Mowjaz application, I propose an approach to identify the topic of the given article using Bi-GRU model.

The remainder of the paper is organized as follows: Section2 overviews the dataset. Section3 introduces how the text preprocessing and word representation are done. Section4 shows the deep learning model. Finally, section5 provides the conclusion.

II. Dataset

The dataset consists of 9,590 articles split into training, development and testing sets. The following table represents some statistical properties of this dataset.

	Training	Development	Testing
Number of articles	7,681	956	953
Max/Min article length	3,384/1	2,418/1	1,602/2
Avg/Std article length	228.6/218.2	226.2/227.5	235.1/216.0
Number of unique words	228,765	55,256	57,853

III. Text preprocessing and word representation

We cannot go straight from raw text to fitting machine learning or deep learning model, we must clean the text first, which means into words and handling punctuation and case. And since we are dealing with supervised learning there is a need for syntactic and semantic understanding and domain expertise that are not necessarily present in these machine learning approaches. Natural Language Processing (NLP) is important because it helps resolve ambiguity in language and adds useful numeric structure to the data for many downstream applications, such as speech recognition or text analytics.

For text cleaning in NLP, I've removed the following:

- HARAKAT (الحركات) by using strip_tashkeel function from PyArabic library.
- HTML tags using re.
- Twitter usernames and web addresses.
- Arabic and English words with length (1, 2).
- Punctuations and extra spaces.
- Numbers.

After preforming NLP on the text. Now we can go to the word embeddings, Word embedding represent words in array, not in the form of 0's and 1's but continuous vectors, word embedding is a class of approaches for representing words and documents using a dense vector representation, the position of a word in the learned vector space is referred to as is embedding, and there's two popular examples of methods of learning word embeddings from text include: Word2Vec and Glove, for this model I've used Word2Vec with Aravec Unigrams pretrained model (Model : Twitter-Skip Gram with 300 vector size)

IV. Deep learning model

So, for the Deep Learning (DL), I've used the recurrent neural network (RNN) with Bi-GRU since it allows previous outputs to be used as inputs while having hidden states, in the table below shows the properties of the model:

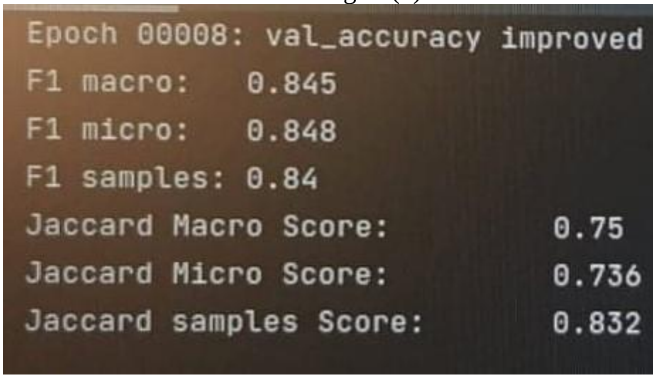
Layer(type)	Output Shape	Param #
Sentence (input layer)	[(None. 256, 300)]	0
Bidirectional	(None, 600)	1083600
Dropout	(None, 600)	0
Dense	(None, 300)	180300
Dense1	(None, 10)	3010

Total params : 1,266,910 **(all of them were trained)**

In the beginning we define a Sequential model, and it takes the sentence (embedded) as input and output with shape of range 256 to 300 neurons (N), next the Bidirectional layer with an input shape range 256 to 300 N and output shape of 600 N followed by a Dropout layer (used for regularization) with a rate of 0.4, followed by two Dense layer the first with input shape 300 N and output shape of 300 N and the last one has an input shape of 300 neurons and output shape of 10 neurons that holds the topic label, and the model compiled with Adam optimizer with a learning rate of 9e-4 with batch size 39 and 8 epochs.

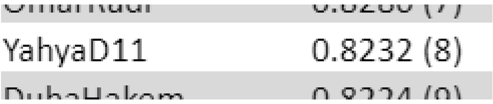
V. Conclusion

The basis of the evaluation was the F1 score and the model got score in the test set as in figure(1),and in Codalab test set as in figure(2)



```
Epoch 00008: val_accuracy improved
F1 macro:   0.845
F1 micro:   0.848
F1 samples: 0.84
Jaccard Macro Score:          0.75
Jaccard Micro Score:          0.736
Jaccard samples Score:        0.832
```

Figure(1)



YahyaD11	0.8232 (8)

Figure(2)

References

M. Al-Ayyoub, H. Selawi, M. Zaghlol, H. Al-Natsheh, S. Suileman, A. Fadel, R. Badawi, A. Morsy, I. Tuffaha, and M. Aljarrah, “Overview of the Mowjaz Multi-Topic Labelling Task,” in The 12th International Conference on Information and Communication Systems (ICICS 2021). IEEE, May 2021.