IS 584 Course Term Project Proposal

Özkan Tuğberk Kartal
Electrical-Electronics Engineering
METU
Ankara, TURKEY
tugberk.kartal@metu.edu.tr

I. Introduction

The purpose of this document is to give explanations regarding Proposal, EDA (Exploratory Data Analysis) and Quality Checks and Github Repository of IS 584 Course Term Project.

II. Proposal

A. Research Questions

Two research questions are (i) ”How to predict the acceptance/rejection decision of the articles by modeling the hierarchical relationships present in their abstract and (meta)reviews?” and (ii) ”If we use Capsule Neural Networks, can we improve the semantic understanding of abstract texts and (meta)reviews compared to traditional models like CNNs (Convolutional Neural Networks) or LSTMs (Long Short Term Memories)?”.

B. Model Architecture and Framework

It is planned to use Capsule Neural Network architecture in PyTorch framework [1].

C. Performance Metrics

Accuracy, F1-Score, Area Under the ROC (Receiver Operating Characteristic) Curve and Confusion Matrix are used to evaluate the performance of the Capsule Neural Network.

D. Expected Outcomes

By using Capsule Neural Network, hierarchical and semantic relationships in the dataset can be captured in a much advanced manner and due to the dynamic routing mechanism of Capsule Neural Network, unseen review patterns can be handled much better.

E. Potential Benefits

Capsule Neural Networks, in addition to its ability to capture hierarchical and semantic relationships in the dataset and its dynamic routing mechanism, can provide better robustness to the variations in the dataset, can guarantee less information loss and requires fewer training samples.

III. EDA and Quality Checks

The notebook file named ”EDA_and_Quality_Checks.ipynb” was written in order to make EDA and Quality Checks and saved under the folder named ”Notebooks” in [2]. Moreover, by using this notebook file various histograms were plotted and saved in the formats ”eps”, ”jpg” and ”svg” under the directory named ”figures/histograms” in [2]. In this document, we can mention some of these analyses.

The dataset is about ICLR (International Conference on Learning Representations) papers from 2017-2020 and NIPS (Neural Information Processing Systems) papers from 2016-2019 and information related to these papers are located in their corresponding files in ”json” format [3]. All of these files in ”json” format were loaded in order to form pandas data frame variables ”ICLR_data” (only ICLR data) having 5194 rows, ”NIPS_data” (only NIPS data) having 3934 rows and ”data” (ICLR and NIPS data together) having 9128 rows. The column names of the variable ”data”, the format of each column, unique values related to each column can be summarized in Table I and II. It must be emphasized that there exists NaN and None values used for missing data row-column values. Regarding decision column, the histogram can be seen in Figures 1.

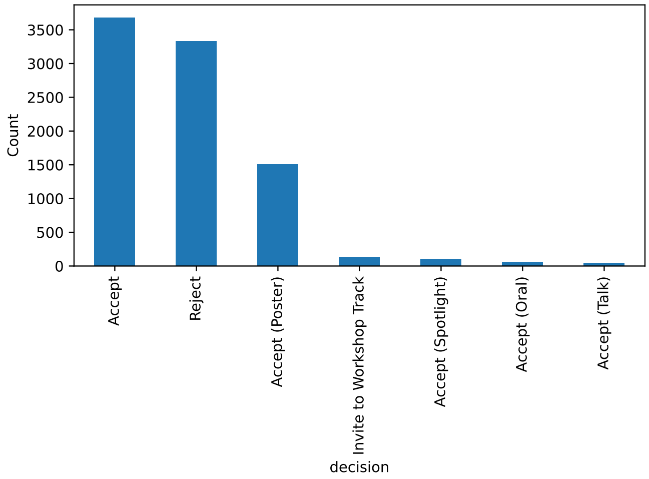

Fig. 1. Decision Histogram

TABLE I
DATA SUMMARY

Column Name	Format	Unique Value
name	string Some Examples ICLR_2017_1.pdf, NIPS_2016_1.pdf	8856 unique values
id	string Some Examples ICLR_2017_1, NIPS_2016_1	each of them
metadata.source	string	CRF, META, NaN
metadata.title	string	5459 unique values (None and NaN for missing 3652 time occurrences, RECURRENT NEURAL NETWORKS 5 time occurrences, DEEP REINFORCEMENT LEARNING 3 times occurrences, some other titles occur more than once, most of the titles occur only once)
metadata.authors	list An Example (Jonathon Cai, Richard Shin, Dawn Song)	not applicable
metadata.emails	list An Example (jonathon@ cs.berkeley.edu, ricshin@ cs.berkeley.edu, dawnsong@ cs.berkeley.edu)	not applicable
metadata.sections	list of dictionaries	not applicable
metadata.references	list of dictionaries	not applicable
metadata. referenceMentions	list of dictionaries	not applicable
metadata.year	int	0, 1969, 2016, 2017, 2018, 2019, 2020, NaN
metadata.abstractText	string	8844 unique values (None and NaN for missing 281 time occurrences, some abstracts occur twice due to either uploading to the same conference twice or to the two distinct conferences, most of the abstracts occur only once)
metadata.creator	string, An Example Microsoft® Word 2016	59 unique values
conference	string	ICLR, NIPS, NaN
decision	string	Accept Accept (Oral) Accept (Poster) Accept (Spotlight) Accept (Talk) Invite to Workshop Track Reject NaN

TABLE II
DATA SUMMARY (CONTINUATION)

Column Name	Format	Unique Value
url	string	8878 unique values
hasContent	boolean	true, false, NaN
hasReview	boolean	true, false, NaN
title	string	8826 unique values (None, NaN, NA and N\A for missing 279 time occurrences, some titles occur twice due to either uploading to the same conference twice or to the two distinct conferences, most of the titles occur only once)
authors	list An Example (Jonathon Cai, Richard Shin, Dawn Song)	not applicable
reviews	list of dictionaries	not applicable
metaReview	string	5876 unique values

IV. GITHUB

The Github link of the repository is given by [2].

REFERENCES

[1] Wikipedia contributors, Capsule neural network — Wikipedia, The Free Encyclopedia. 2024. [Online]. Available: https://en.wikipedia.org/wiki/Capsule_neural_network

[2] Ö. T. Kartal, IS584CourseTermProject. Online. Available: https://github.com/oztuka/IS584CourseTermProject/

[3] Volga Sezen, ”IS 584 Course Term Project Dataset,” 2025. Available: https://drive.google.com/file/d/1nJdljy468roUcKLbVwWUhMs7teirah75/view