PERSIAN NEWS CATEGORIZATION

final project

Artificial Intelligence

Instructor: Dr.Pishgoo

TAs: Amirreza Mehrzadian, Alireza Saeednia

Contents

1 Preface

The purpose of this project is to apply the material learned during the semester, on a real issue and examine the different methods in machine learning project on specific datasets. The programming language that can be used in this project is Python and the use of other programming languages is not allowed. You must write all your codes in the Jupyter Notebook space. Given that parts of The project may need a lot of computing volume, it is recommended to use Google's Collab service or Kaggle notebooks.
`https://www.kaggle.com/`

1.1 NLP

Natural language processing, or NLP, combines computational linguistics—rule-based modeling of human language—with statistical and machine learning models to enable computers and digital devices to recognize, understand and generate text and speech.
Natural Language Processing • Word Vectors/Embeddings
`https://youtu.be/CMrHM8a3hqw`
`https://www.youtube.com/playlist?list=PLQY2H8rRoyvzDbLUZkbudP-MFQZwNmU4S`
text-preprocessing-for-nlp

2 Dataset

The dataset contains 3 columns: title, description, and tag.
Each row in the data set corresponds to a news item and the topic of each news item is placed in the tags column. This column has different topics, but your model prediction for the test data set should be one of the following topics:

- social
- economics
- Iran-states
- international
- politics
- scientific-cultural-sports

As a result, the target column (tags) in the initial dataset may require changes. you are allowed to tag news with a maximum of one main category.

For example, according to the topic, the news related to Corona can be placed in the following categories:

- social: quarantines in different countries
- economics: Small business issues and compensation for entrepreneurs
- scientific-cultural-sports: Tokyo Olympics was moved due to the epidemic...
- scientific-cultural-sports: News about symptoms, tips on staying healthy, looking for vaccines, and more

3 Problem description

Your model should predict for each news (line) from the test data set, which of the 6 main categories the content of that news falls into. In the next step, you should encode the obtained strings based on the following table and send the final answer based on numerical values.

number	tag
0	social
1	economics
2	Iran-states
3	International
4	politics
5	scientific-cultural-sports

4 Grading

A time will be set for the presentation to the teaching assistant. important factors in grading:

- How effective are the classifiers on new data
- Good plan for choosing and validating algorithm, parameters, and classifiers
- Clarity and understanding in oral report
- Clean code

F1- score will be used to determine the accuracy of your model. **Accuracy below 70 percent gets no point from the project.**

5 Bonus

for this part you need to build a multi-layer perceptron (MLP) to classify the news. you are not allowed to use MLPClassifier from sklearn. (One way to build mlp is by using keras (Sequential, Dense,...))
to get score of this part your accuracy must be more than 85 percent.

https://youtu.be/TCH_1BHY58I?list=PLAqhIrjkxbuWI23v9cThsA9GvCAUhRvKZ
https://www.kaggle.com/code/fchollet/simple-deep-mlp-with-keras
https://www.datacamp.com/tutorial/multilayer-perceptrons-in-machine-learning