



MOVIE-GENRE-MULTI-LABEL-TEXT-CLASSIFICATION

MACHINE LEARNING AND DEEP LEARNING
(650.025, 22W)

PROJECT PRESENTATION

12231704 MUSTAFA TAHİR BİNGÖL

12237686 MUSTAFA TAYYİP BAYRAM



AGENDA

- I. Learning Task
 - i. Types of Classification
 - ii. Mutli-label Classification Applications
 - iii. Performance Measure
- II. Preliminary Data Analysis
- III. Literature Overview
- IV. ML Approach
 - i. Data Preprocessing
 - ii. Modelling
- V. Results

LEARNING TASK

- ❖ Multi-label movie genre classification using movie plot data.
- ❖ Classification Tasks

TYPES OF CLASSIFICATION

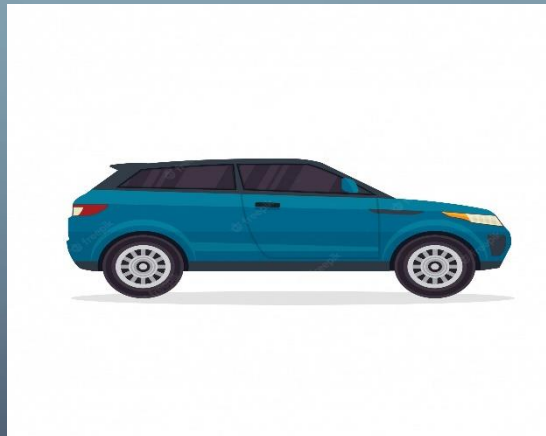
BINARY

Fradulent or Not Fradulent



MULTICLASS

Car or Bike or Bus or Ship



MULTI-LABEL

Crime, Drama, Mystery, Comedy, History



MUTLI-LABEL CLASSIFICATION APPLICATIONS



Image/Text/Music
Classification






Bioinformatics



Recommender
systems

PERFORMANCE MEASURE

Label	True Positive (TP)	False Positive (FP)	False Negative (FN)	Micro-Averaged Values
 Airplane	2	1	1	Precision = $\frac{6}{6+4} = 0.60$ Recall = $\frac{6}{6+4} = 0.60$ F1 Score = $\frac{6}{6 + \frac{1}{2}(4+4)} = 0.60$
 Boat	1	3	0	
 Car	3	0	3	
TOTAL	6	4	4	

MICRO F1 SCORE

- ❖ Global metric calculation for class imbalance.
- ❖ Used for models where accuracy is key and label positivity is similar.
- ❖ F1-score computes label avg, micro F1 computes global avg.

PRELIMINARY DATA ANALYSIS

❖ Obtained from {MPST}: A Corpus of Movie Plot Synopses with Tags

❖ Sources are IMBD and Wikipedia

❖ Number of plot data = 14828

❖ Columns

- imdb_id
- title
- plot_synopsis
- tags
- split
- synopsis_source

imdb_id	title	plot_synopsis	tags	split	synopsis_source
tt0057603	I tre volti della paura	Note: this synopsis is for the original ...	cult, horror, gothic, murder, atmospheric	train	imdb
tt1733125	Dungeons & Dragons: The Book of Vile Da...	Two thousand years ago, Nhagruul the Fo...	violence	train	imdb
tt0033045	The Shop Around the Corner	Matuschek's, a gift store in Budapest, ...	romantic	test	imdb
tt0113862	Mr. Holland's Opus	Glenn Holland, not a morning person by ...	inspiring, romantic, stupid, feel-good	train	imdb
tt0086250	Scarface	In May 1980, a Cuban man named Tony Mon...	cruelty, murder, dramatic, cult, violen...	val	imdb

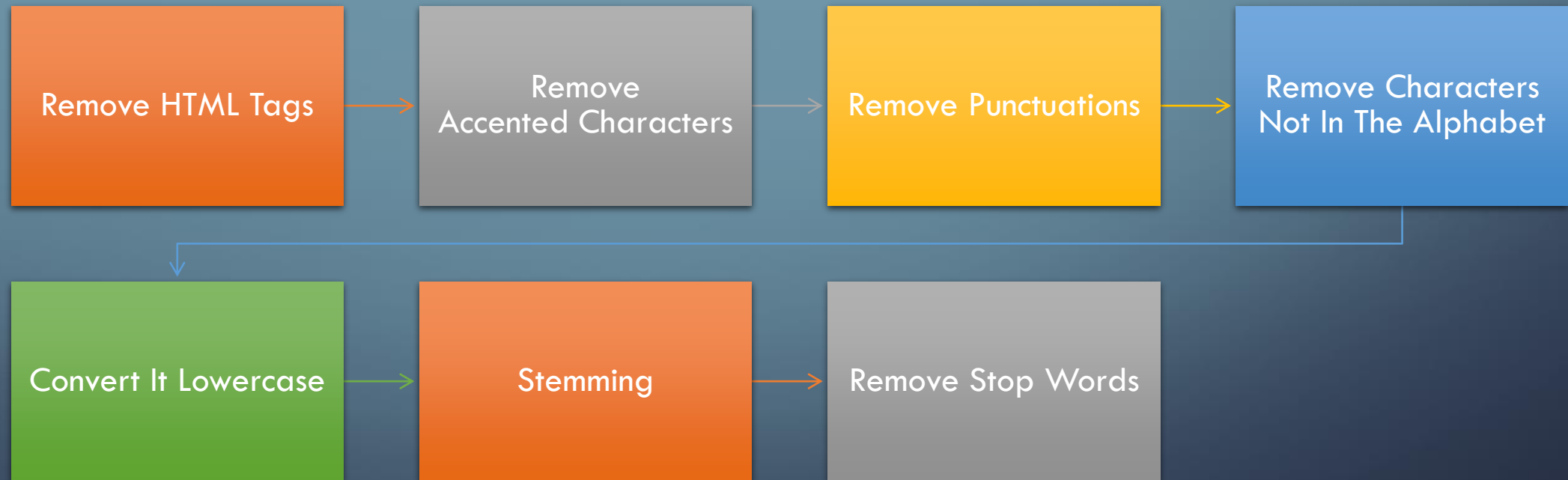
PRELIMINARY DATA ANALYSIS CONTD.

▲ imdb_id	▲ title	▲ plot_synopsis	▲ tags	▲ split	▲ synopsis_source
IMDB id of the movie	Title of the move	Plot Synopsis of the move	Tags assigned to the move	Position of the movie in the standard data split	From where the plot synopsis was collected
14828 unique values	13757 unique values	13848 unique values	murder 7% romantic 5% Other (13093) 88%	train 64% test 20% Other (2373) 16%	wikipedia 72% imdb 28%

LITERATURE

- ✓ MULTIPLE SIMILAR NLP PROJECT.
- ✓ 1 DIRECTLY BENEFITED PROJECT.
 - Logistic Regression
 - Naive Bayes
 - Neural Networks
- ✓ MANY PROJECTS THAT MAKE USE OF VISUALIZATIONS.

DATA PREPROCESSING



MODELLING

TEXT FEATURIZATION (VECTORIZATION)

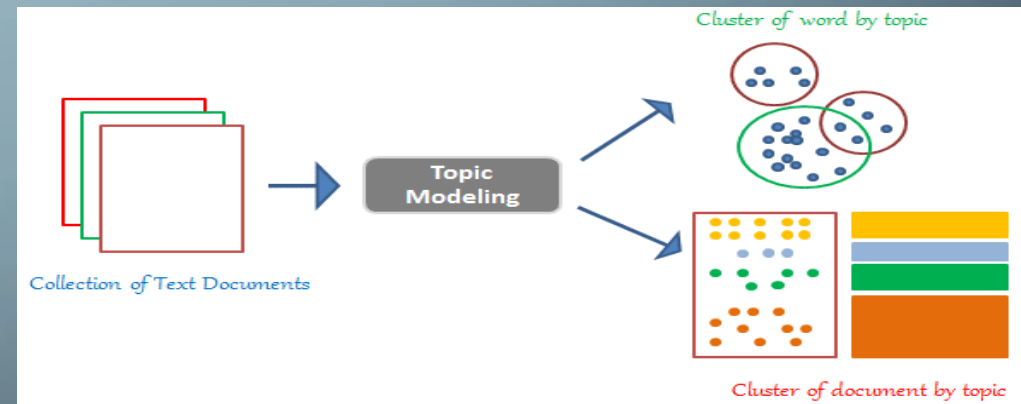
- TF-IDF

$$TF(t, d) = \frac{\text{number of times } t \text{ appears in } d}{\text{total number of terms in } d}$$

$$IDF(t) = \log \frac{N}{1 + df}$$

$$TF - IDF(t, d) = TF(t, d) * IDF(t)$$

- Topic Modelling – LDA (Latent Dirichlet Allocation)



MODELING CONTD. ONE-VS-REST APPROACH USED IN BASED MODELS

- ✓ NAIVE BAYES
- ✓ LOGISTIC REGRESSION ON TF-IDF
- ✓ LOGISTIC REGRESSION ON TF-IDF + LDA

RESULTS

vectorizer	Model	train f1	test f1
tfidf	logistic regression	0.4915	0.3956
tfidf	naïve bayes	0.3391	0.2525
topic modelling+ tfidf	Logistic Regression	0.5539	0.3463



THANK YOU FOR LISTENING
WE WANT YOU TO BE POSITIVE IN
YOUR POLARITY