STATISTICAL LEARNING COURSE
FINAL PROJECT

JULY, 2020

FAN FICTION GENRE CLASSIFICATION

DILETTA ABBONATO, LAURA LAURENTI,
SANTO PALAIA, ANNA PRESCIUTTINI

Abstract

Main research aim and framework

Data sources & Data collection

Model & methods

Software /Hardware toolkit

References

Abstract

Text classification is the process of assigning categories to documents according to their content. We'll apply different learning techniques in order to perform a fan fiction genre classification.

In its simplest form, fanfiction is when somebody takes a character, universe, or story from a different scenario to create their own story. These characters and scenarios are commonly pulled from novels, TV shows, movies, and even real life.

We chose for our work 5 genres that are: Fantasy, Romance, Horror, Science Fiction, Humor.

Main research aim and framework

The main goal is to detect the genre of the fanfiction. The idea come from our teenager memories where we were obsessed with famous characters as Harry Potter and Harry Styles and our immagination navigated the ship of our tender heart. To implement our idea we were inspired by the paper made by a team of Nanyang Technological University about to detect the 5 big personal traits from a text.

DATA COLLECTION

SOURCE

Data for this project come from fanfiction.net, a popular hub hosting, a large repository of freely available fan written content.

COLLECTION

We computed the data collection process scraping this fanfiction website through the use of BeautifulSoup, considering just english fan fiction.

FEATURES

The dataset has the following features: 'ID' (unique ID that represent the fanfiction), `canon' ( which topic the fanfiction is related to), `title', `lang' (language -> english), `genres' (different type of genres which we are gonna smoothing with just considering Romance,Adventure,Sci-Fi,Horror,Fantasy), `rated' (audiance age), num_words', `status' (complete or incomplete), `story'.

SIZE

We collected 426 fanfiction for five genres

Data preprocessing steps:

DE-CAPITALIZATION

We performed case-folding reducing everything to lower case. This because, for example, our model might treat a word which is in the beginning of a sentence with a capital letter different from the same word which appears later in the sentence but without any capital latter. This might lead to decline in the accuracy.

PUNTUACTION & NON-TEXTUAL CHARACTERS REMOVAL

We replaced puntuaction and non textual-characters splitting by whitespaces in order to perform a better tokenization

TOKENIZATION & STOP WORDS REMOVAL

We segmented text into sentences and we removed common words that generally do not contribute to the meaning of a sentence for the purposes of information retrieval

PARTS OF SPEECH TAGGING

We marked up a word in a text (corpus) as corresponding to a particular part of speech, based on both its definition and its context.

Scraping

```
[14] 1 data.head()
```

	text	category
0	\tTough\n \n Just cross-posting this from ao3...	Horror
1	\tA Woman Scorned\n \n A Woman Scorned \n \n ...	Horror
2	\tThe Thing That puts Creepy in Creepypasta: V...	Horror
3	\tTorture\n \n The smell of rotting flesh burn...	Horror
4	\tCute Little Kitty\n \n Sometimes, if I used ...	Horror

Not preprocessed/original dataset

Preprocessing

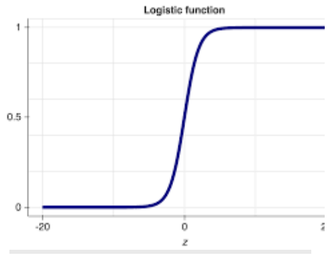
```
[18] 1 data.head()
```

	text	category	text_clean	text_tokens	text_tokens_pos_tagged
0	\tTough\n \n Just cross-posting this from ao3...	Horror	tough crossposting ao im strawberryqueen since...	['tough', 'crosspost', 'ao', 'im', 'strawberry...	[('tough', 'JJ'), ('crosspost', 'NN'), ('ao', ...
1	\tA Woman Scorned\n \n A Woman Scorned \n \n ...	Horror	woman scorned woman scorned said holmes farm t...	['woman', 'scorn', 'woman', 'scorn', 'said', '...	[('woman', 'NN'), ('scorn', 'VBZ'), ('woman', ...
2	\tThe Thing That puts Creepy in Creepypasta: V...	Horror	thing puts creepy creepypasta virus spying coo...	['thing', 'put', 'creepi', 'creepypasta', 'vir...	[('thing', 'NN'), ('put', 'VBD'), ('creepi', '...
3	\tTorture\n \n The smell of rotting flesh burn...	Horror	torture smell rotting flesh burned nostrils bo...	['tortur', 'smell', 'rot', 'flesh', 'burn', 'n...	[('tortur', 'NN'), ('smell', 'NN'), ('rot', 'N...
4	\tCute Little Kitty\n \n Sometimes, if I used ...	Horror	cute little kitty sometimes used imagination f...	['cute', 'littl', 'kitti', 'sometim', 'use', '...	[('cute', 'NN'), ('littl', 'NN'), ('kitti', 'N...

Preprocessed dataset

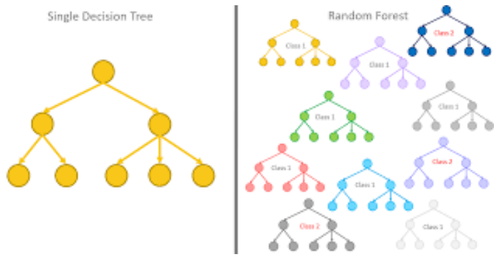
At the end of the preprocessing, we obtained a dataframe with the following columns: Genre, text,text_clean(text without any punctuation),text_tokens(tokenization of thetext_clean)and text_token_POS (tuples of token and tag for each token intext_tokens)

Non- deep learning methods



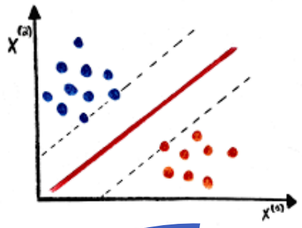
Logistic Regression

Logistic regression is the appropriate regression analysis to conduct when the dependent variable is dichotomous (binary).



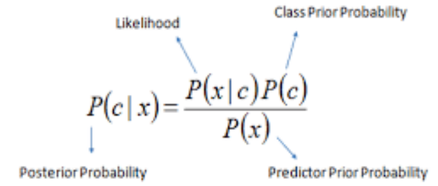
Random forest

Random Forest (RF) is one of the many machine learning algorithms used for supervised learning, this means for learning from labelled data and making predictions based on the learned patterns. RF can be used for both classification and regression tasks.



Support Vector Machine

a support-vector machine constructs a hyperplane or set of hyperplanes in a high- or infinite-dimensional space, which can be used for classification, re gression, or other tasks like outliers detection.

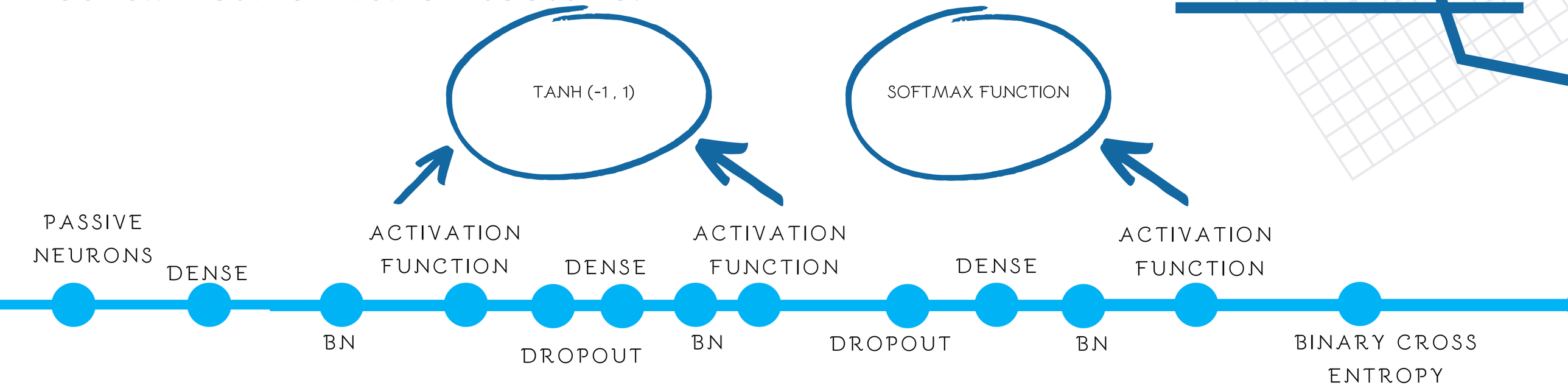


$$P(c \mid \mathbf{X}) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

Naïve Bayes

Naive Bayes is a a family of algorithms based on a common principle: all naive Bayes classifiers assume that the value of a particular feature is independent of the value of any other feature, given the class variable.

Neural Network architecture:



INPUT LAYER

Data input

HIDDEN LAYERS

Data processing

OUTPUT LAYER

Result

Optimization Methods

Batch Normalization

THE BATCH NORMALIZATION IS A SIMPLE HEURISTIC METHOD THAT ALLOWED TO TRAIN DEEP NETWORKS SIGNIFICANTLY BETTER. IT ALLOWS EACH LAYER TO CONTROL THE MEAN AND THE VARIANCE OF ITS OUTPUT, BYAPPROPRIATELY RESCALING THEM WITH A SIMPLE AFFINE TRANSFORMATION

Dropout

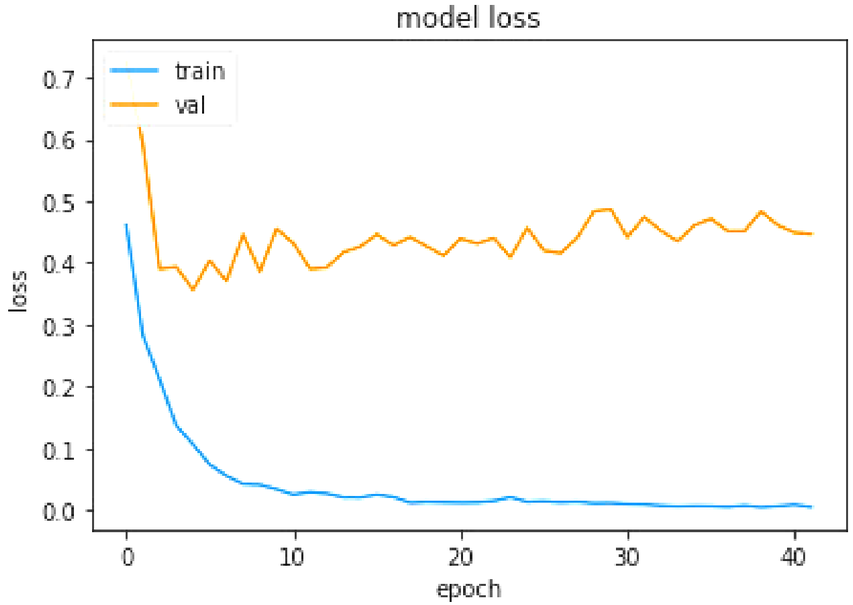
DROPOUT IS A TECHNIQUE WHERE RANDOMLY SELECTED NEURONS ARE IGNORED DURING TRAINING. THEY ARE "DROPPED-OUT" RANDOMLY. THIS MEANS THAT THEIR CONTRIBUTION TO THE ACTIVATION OF DOWNSTREAM NEURONS IS TEMPORALLY REMOVED ON THE FORWARD PASS AND ANY WEIGHT UPDATES ARE NOT APPLIED TO THE NEURON ON THE BACKWARD PASS

Stochastic Gradient Descent

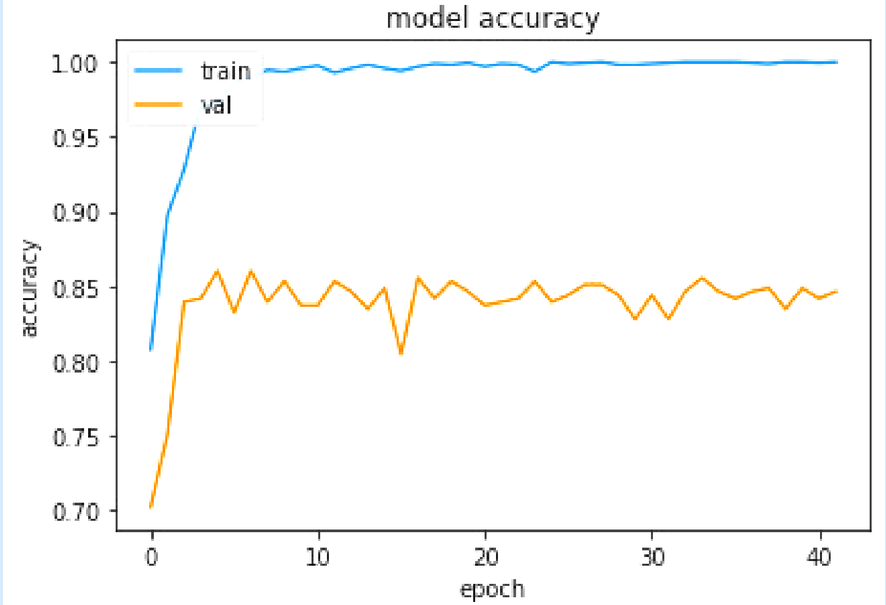
STOCHASTIC GRADIENT DESCENT IS AN OPTIMIZATION ALGORITHM WHICH IMPROVES THE EFFICIENCY OF THE GRADIENT DESCENT ALGORITHM. SIMILAR TO BATCH GRADIENT DESCENT, STOCHASTIC GRADIENT DESCENT PERFORMS A SERIES OF STEPS TO MINIMIZE A COST FUNCTION. UNLIKE BATCH GRADIENT DESCENT, WHICH IS COMPUTATIONALLY EXPENSIVE TO RUN ON LARGE DATA SETS, STOCHASTIC GRADIENT DESCENT IS ABLE TO TAKE SMALLER STEPS TO BE MORE EFFICIENT WHILE ACHIEVING THE SAME RESULT

Results

MODEL LOSS



MODEL ACCURACY



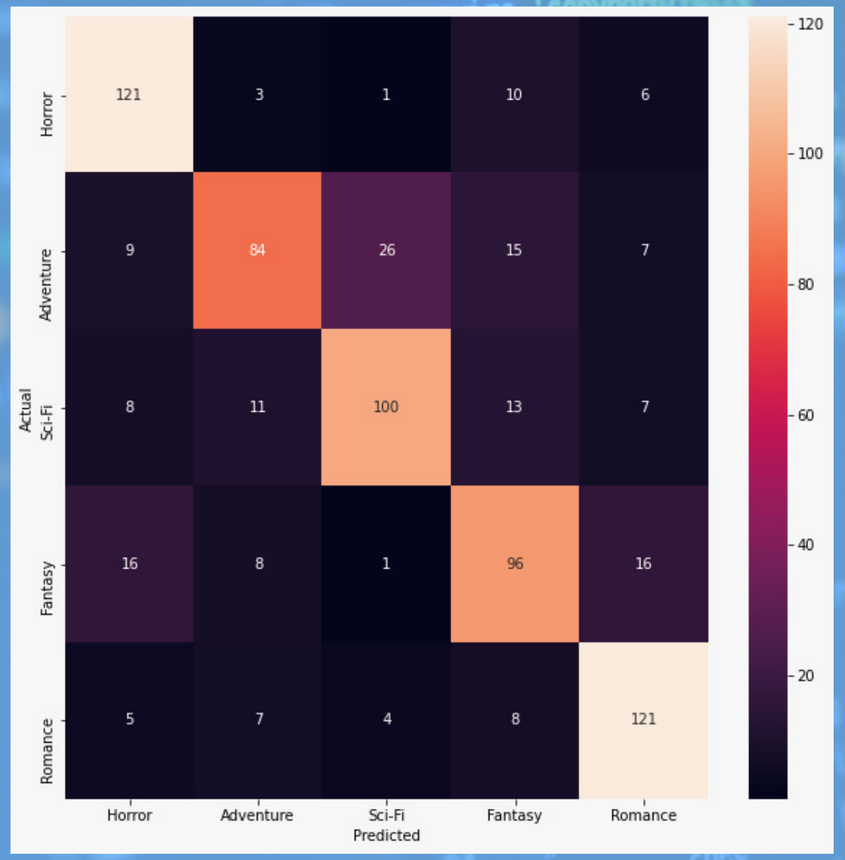
Conclusion

74%

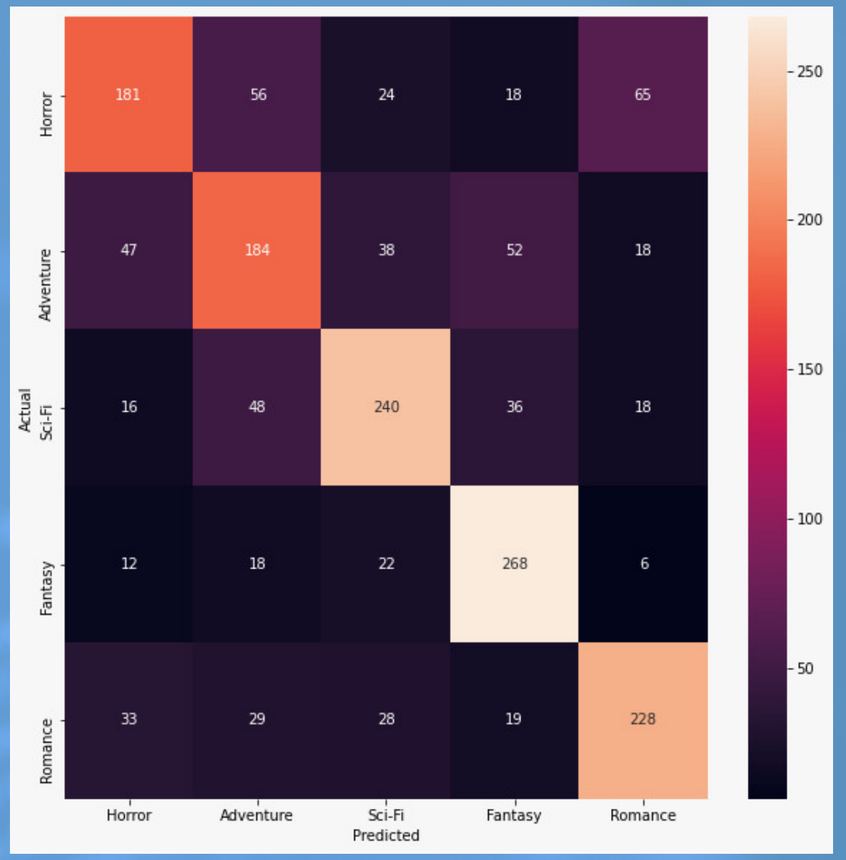
SUPPORT VECTOR MACHINE ACCURACY

86%

FEEDFORWARD NEURAL NETWORK ACCURACY



SVM confusion matrix



NN confusion matrix

Software & Hardware Toolkit

BeautifulSoup

Beautiful Soup is a library that makes it easy to scrape information from web pages. It sits atop an HTML or XML parser, providing Pythonic idioms for iterating, searching, and modifying the parse tree

NLTK

NLTK stands for Natural Language Toolkit. This toolkit is one of the most powerful NLP libraries which contains packages to make machines understand human language and reply to it with an appropriate response.

Pandas

Pandas is a software library for data manipulation and analysis. In particular, it offers data structures and operations for manipulating numerical tables and time series

Keras

Keras is a free open source Python library for developing and evaluating deep learning models. It wraps the efficient numerical computation libraries Theano and TensorFlow. It focuses on being user-friendly, modular, and extensible.

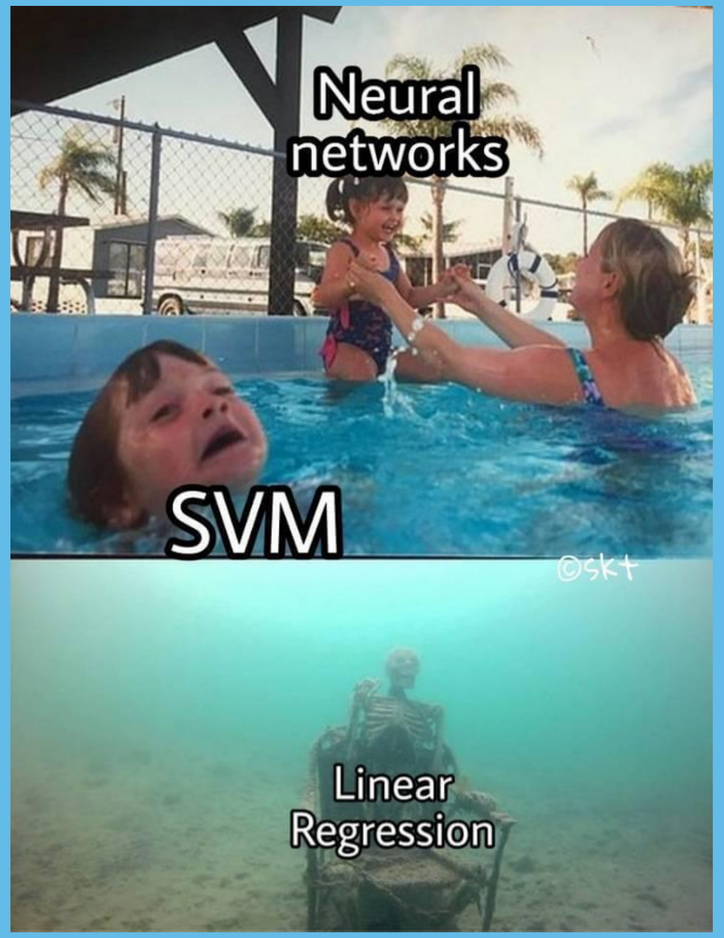
References

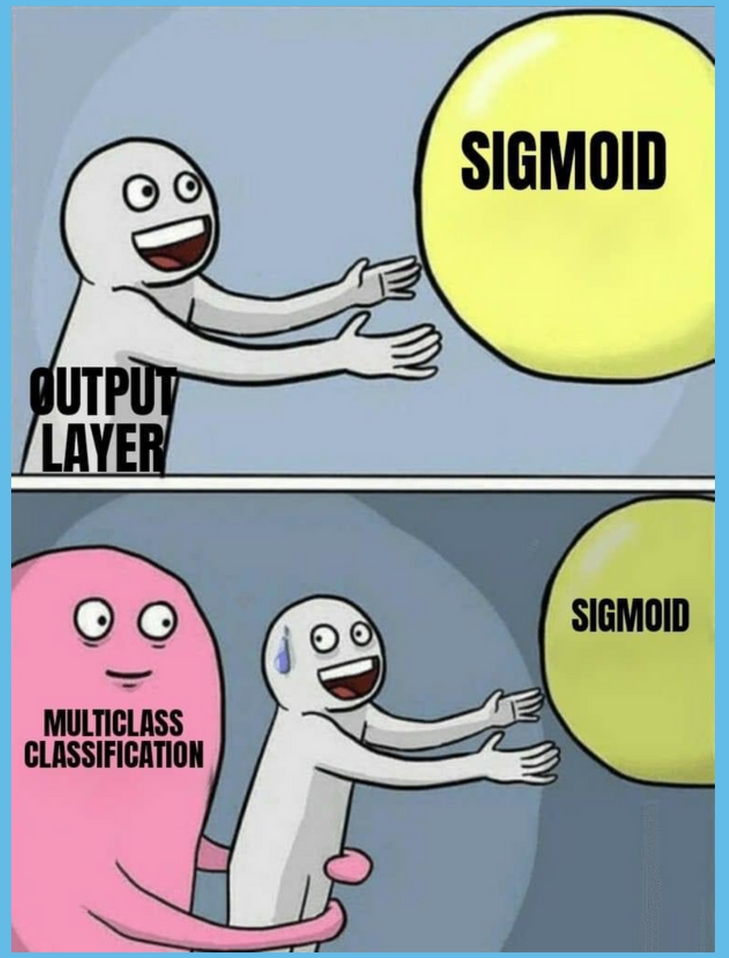
"DEEP LEARNING-BASED DOCUMENT MODELING FOR PERSONALITY DETECTION FROM TEXT"

Navonil Majumder Instituto Politécnico Nacional,
Soujanya Poria, Nanyang Technological University, Alexander Gelbukh, Instituto Politécnico Nacional, Erik Cambria, Nanyang
Technological University

"AN INTRODUCTION TO STATISTICAL LEARNING (WITH APPLICATIONS IN R)"

Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani.





Thank you for your attention!