Chair for Mathematical Information Science
Prof. Dr. H. Bölcskei

ETH zürich

Attentive Neural Networks for News Classification

Siddhant Ray
Semester Thesis Presentation
10 June 2021, Zürich

Project Supervisor -
Dmytro Perekrestenko

News classification – Why should we care?

- The world relies on news articles for information
- Personalized news feed for customers
- Faster information retrieval
- Detection of fake news, sentiment analysis etc.

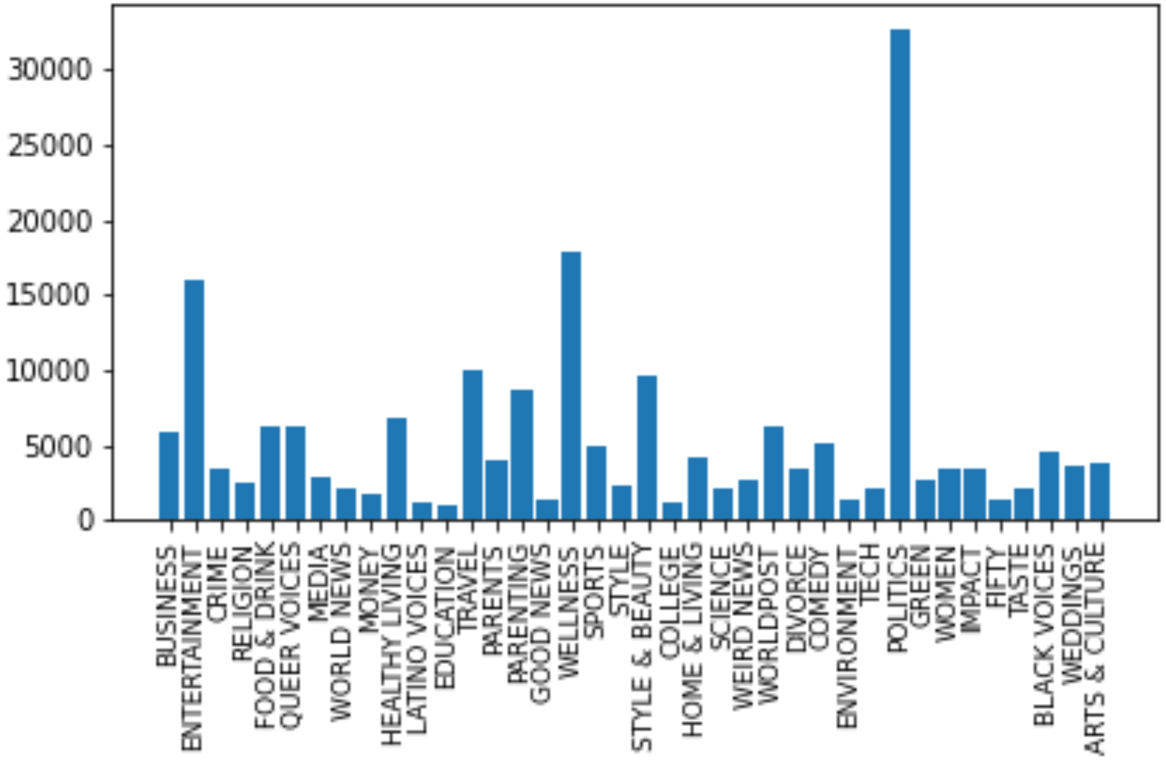
We present :

- A neural network based news classification model to categorize news descriptions
- An algorithm to detect overlap in news categories which reduces redundant labels the dataset
- Improvements in model's performance after reduction using our algorithm

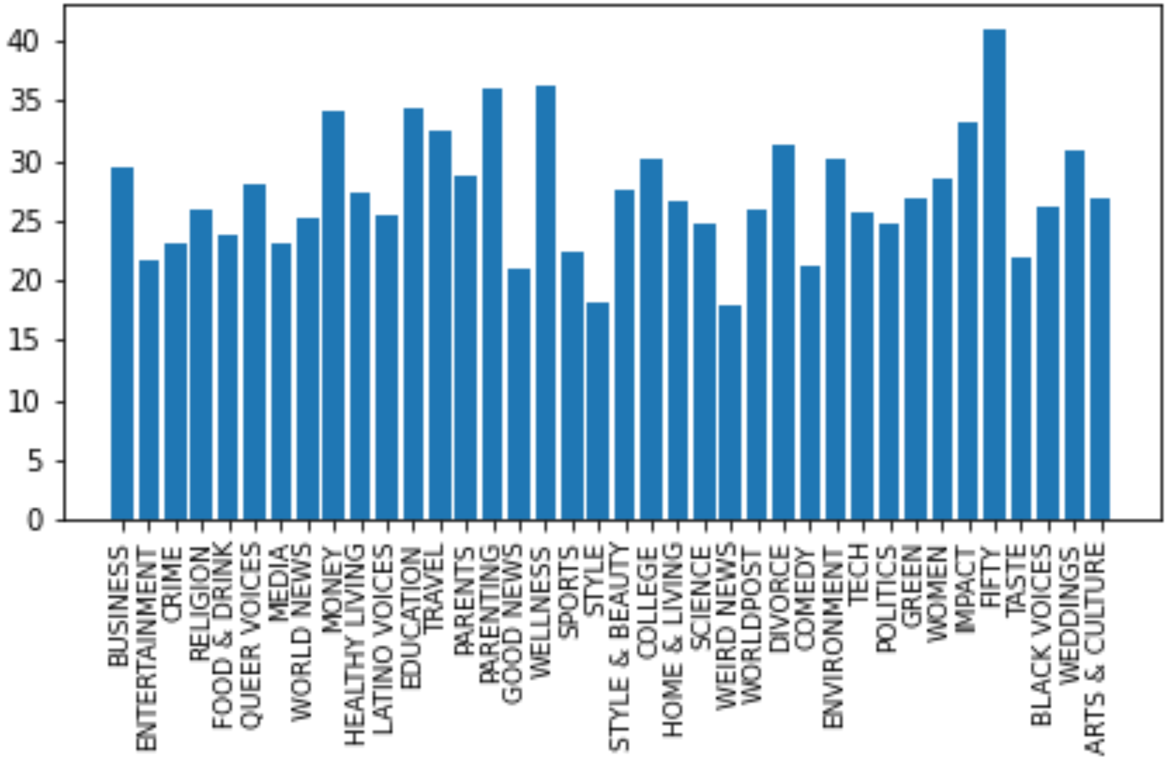
News Dataset

- We use an open-source dataset from Kaggle which is a multi-category news dataset
- Original dataset – 41 news categories
- Each category has :
  - News headline (we use)
  - News description (we use)
  - Other information i.e. author, date etc. (we don't use)

Sample: U.S. Launches Auto Import Probe, China Vows To Defend Its Interests. The investigation could lead to new U.S. tariffs similar to those imposed on imported steel and aluminum in March.

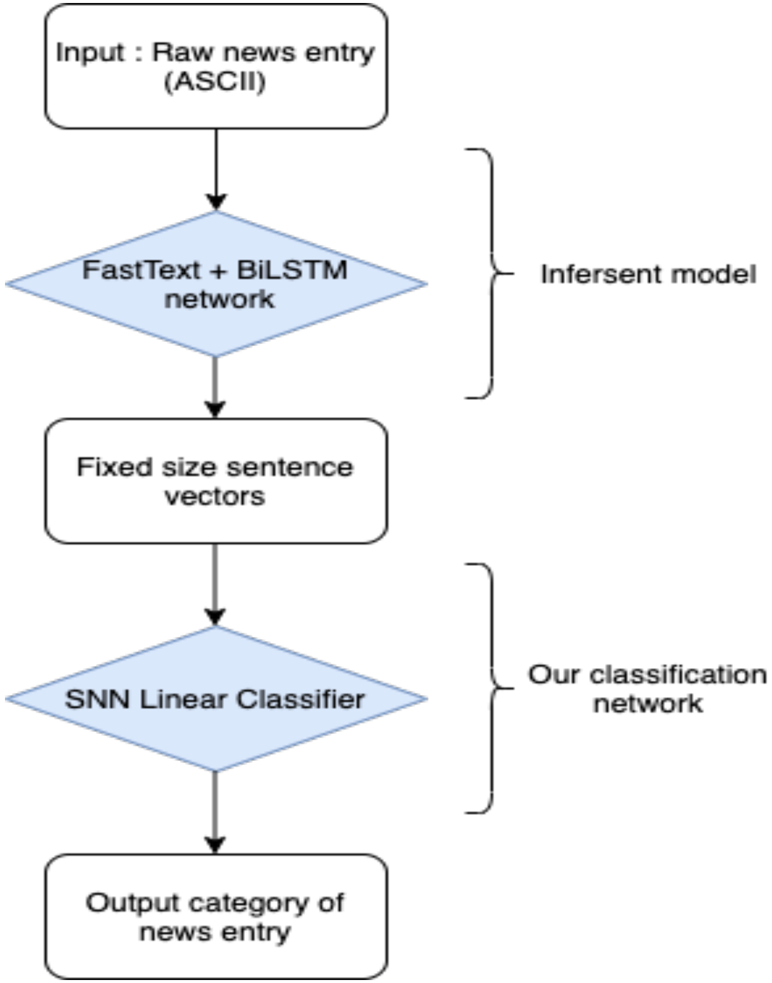


No. of descriptions per news category



Average length of description per category

Preliminary model : Recurrent Neural Network (RNN) based Classifier



The RNN model doesn't work well :

- Word embeddings generated by FastText are limited – i.e. they do not capture contextual meaning
- No robust way to fine-tune the model on our dataset
- RNNs can only process inputs in a sequence, word by word

This makes RNN based text classifiers limited in performance.

Contextual word embeddings are important!!

- *“An apple a day, keeps the doctor away”*
- *”I like using my Apple MacBook”*

- *“I left my phone on the left side of the table.”*

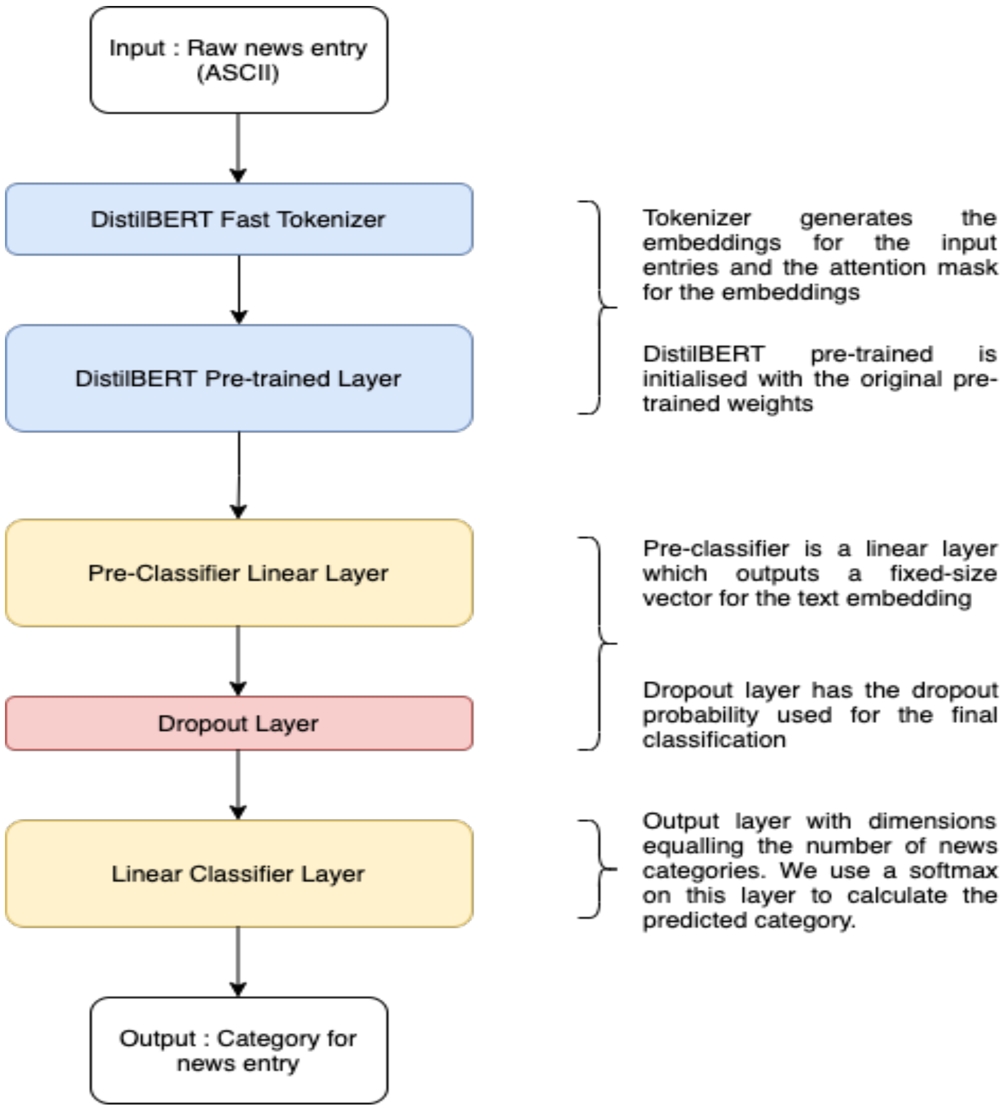
Word embeddings without context, do not capture semantic meaning well.

BERT models give contextual word embeddings

- BERT is a transformer based neural network
- Captures contextual information: uses left context and right context of a given word
- Transformers can process out of order sequences : using attention weights
- Hence, transformers are more parallelizable faster to train than RNNs.

We choose a BERT based model for our news classification task.

Model : BERT based Classifier



Model Evaluation :

- We use 80% of the dataset for training, 20% for evaluation.
- Data for training and validation is sampled randomly.

Performance Metric	Value (BERT)	Value (RNN)
Accuracy (Top prediction)	65.67%	58.82%
Accuracy (Top 3 predictions)	87.75%	-
Mean F1 Score (range [0,1])	0.5920	-
Mean Reciprocal Rank (range [0,1])	0.7574	-

Performance evaluation on the validation dataset

Removing dataset redundancy

- ➢ So, what's the problem? – Inconsistencies in the dataset
- ➢ Dataset has classes which have overlapping items in context
  - example : *"GREEN"* and *"ENVIRONMENT"*
- ➢ Dataset also has huge imbalance in no. of descriptions per category
  - *"POLITICS"* has many more data points

We analyze the evaluation and dataset, to check where the problem lies.

Dataset re-analysis

- ➢ Take category *“GREEN”* and *“ENVIRONMENT”*

"Quarter Of World’s Land Will Be Permanently Drier If Paris Climate Goals Not Met. Countries need to work to prevent the Earth’s temperature from rising more than 1.5 degrees.” (in dataset under GREEN)

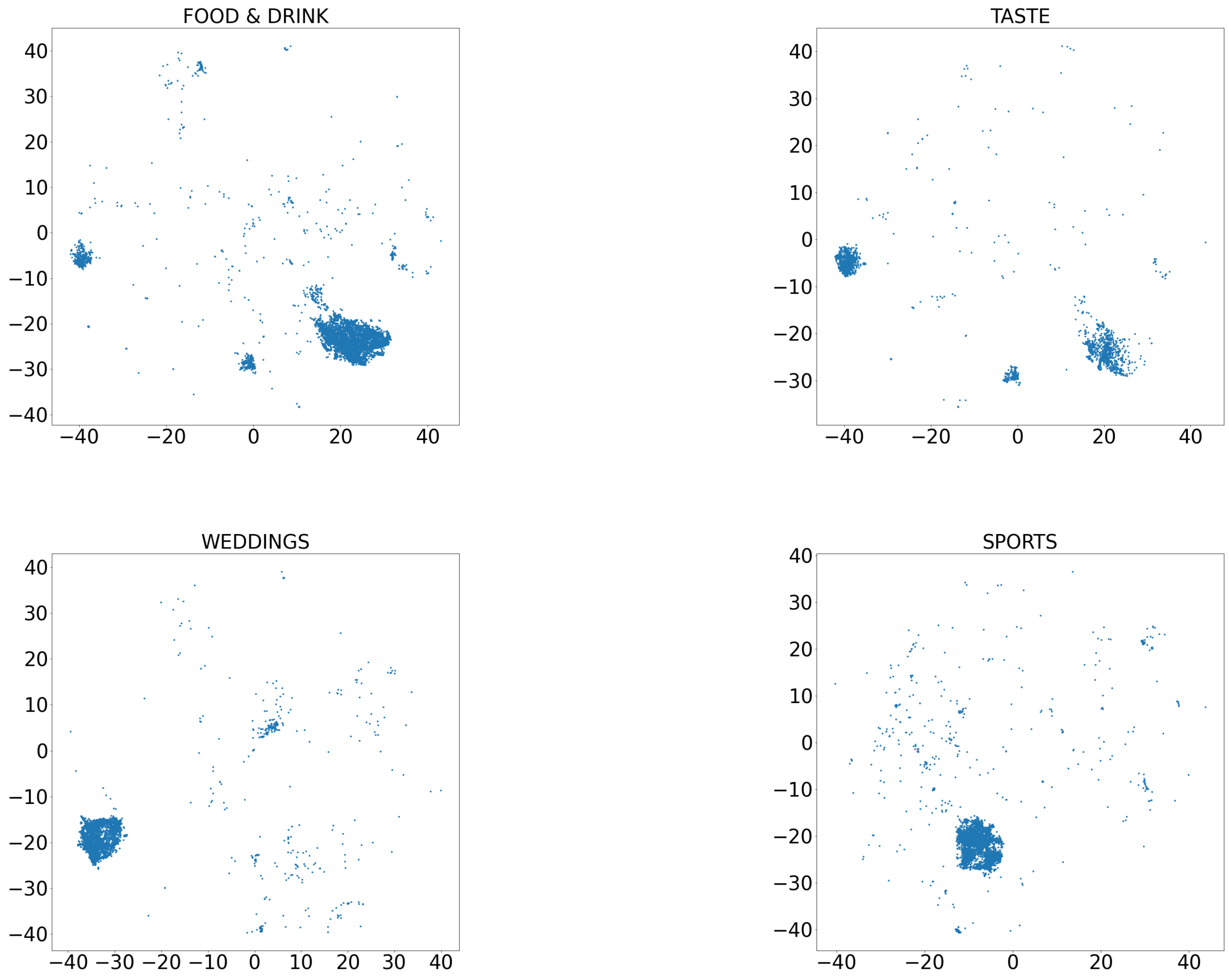
- ➢ Take category *"FOOD & DRINK”* and *“TASTE”*

"Dunkin’ Donuts Coffee Is Being Turned Into A Stout Beer. Dark Roasted Brew is the first beer to be made with the company’s dark roast beans.” (in dataset under TASTE)

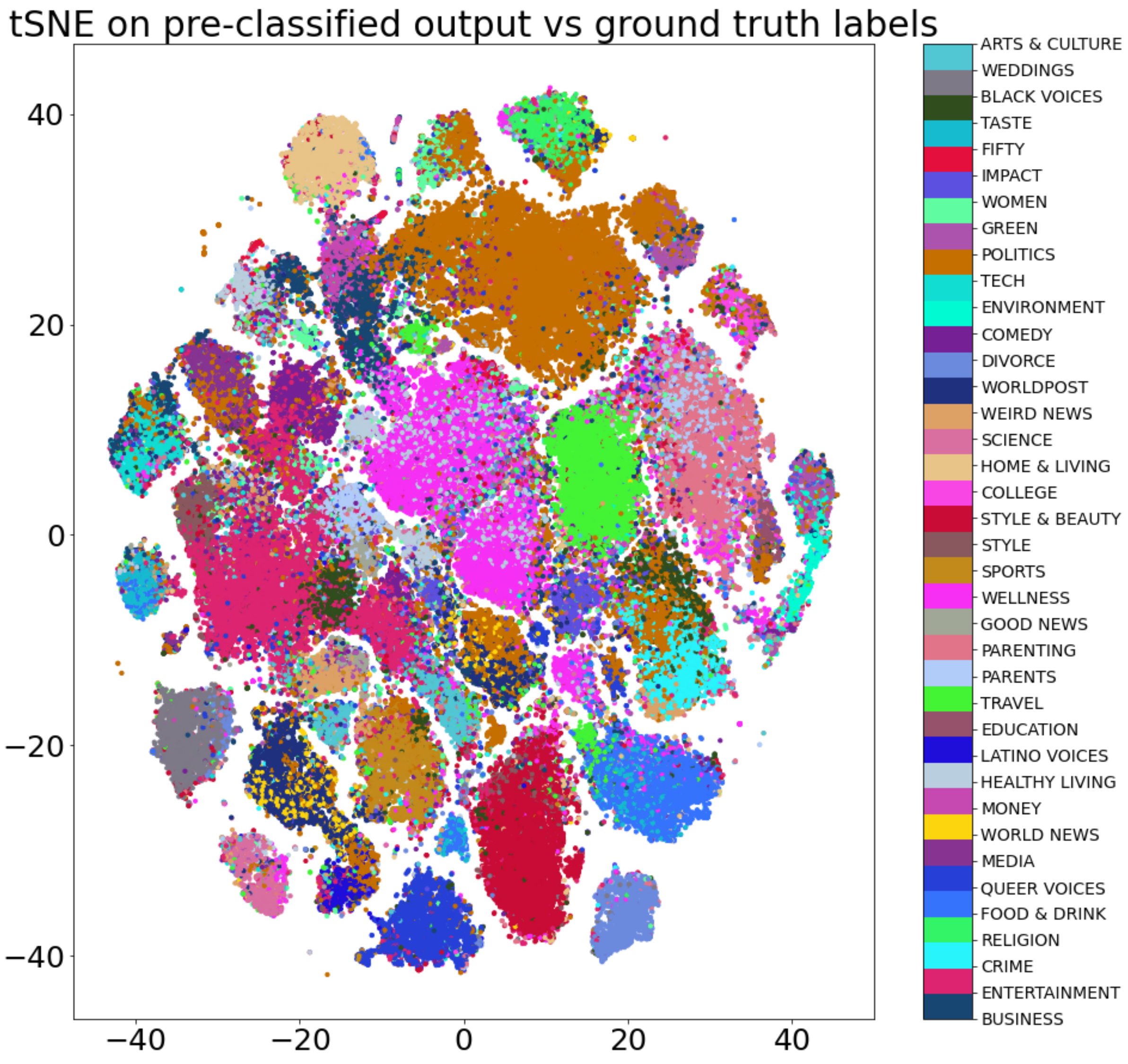
What we do? - Class overlap detection algorithm

First, we run our model on the entire dataset. Our algorithm is :

- PART 1 : Visualization
  - Extract the outputs of the pre-classifier layer of our model.
  - Apply t-SNE to reduce the vectors to 2 dimensions while preserving distances between original vectors with high probability
  - Visualize the individual t-SNE plots per category of news items



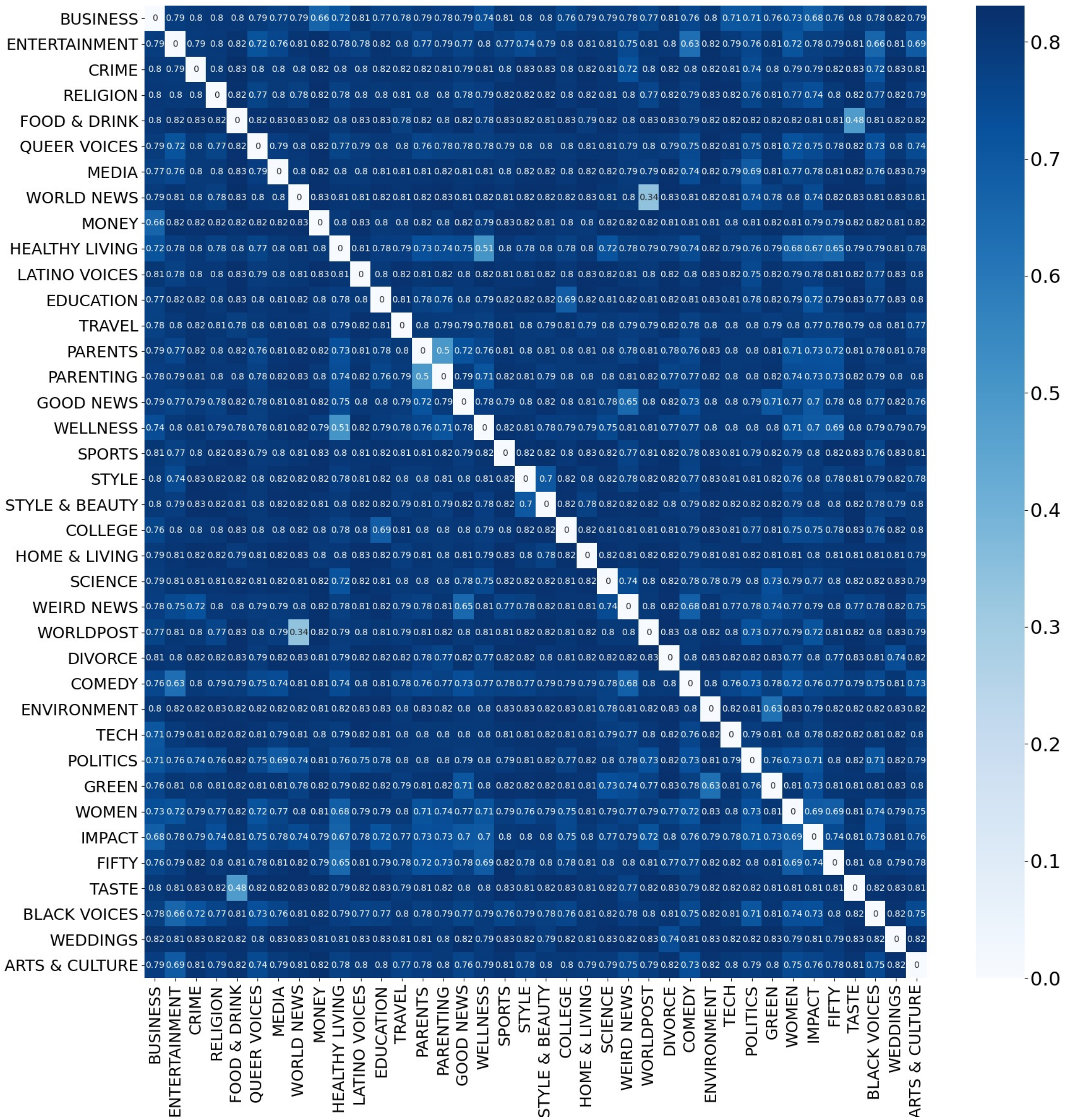
t-SNE plots for individual categories



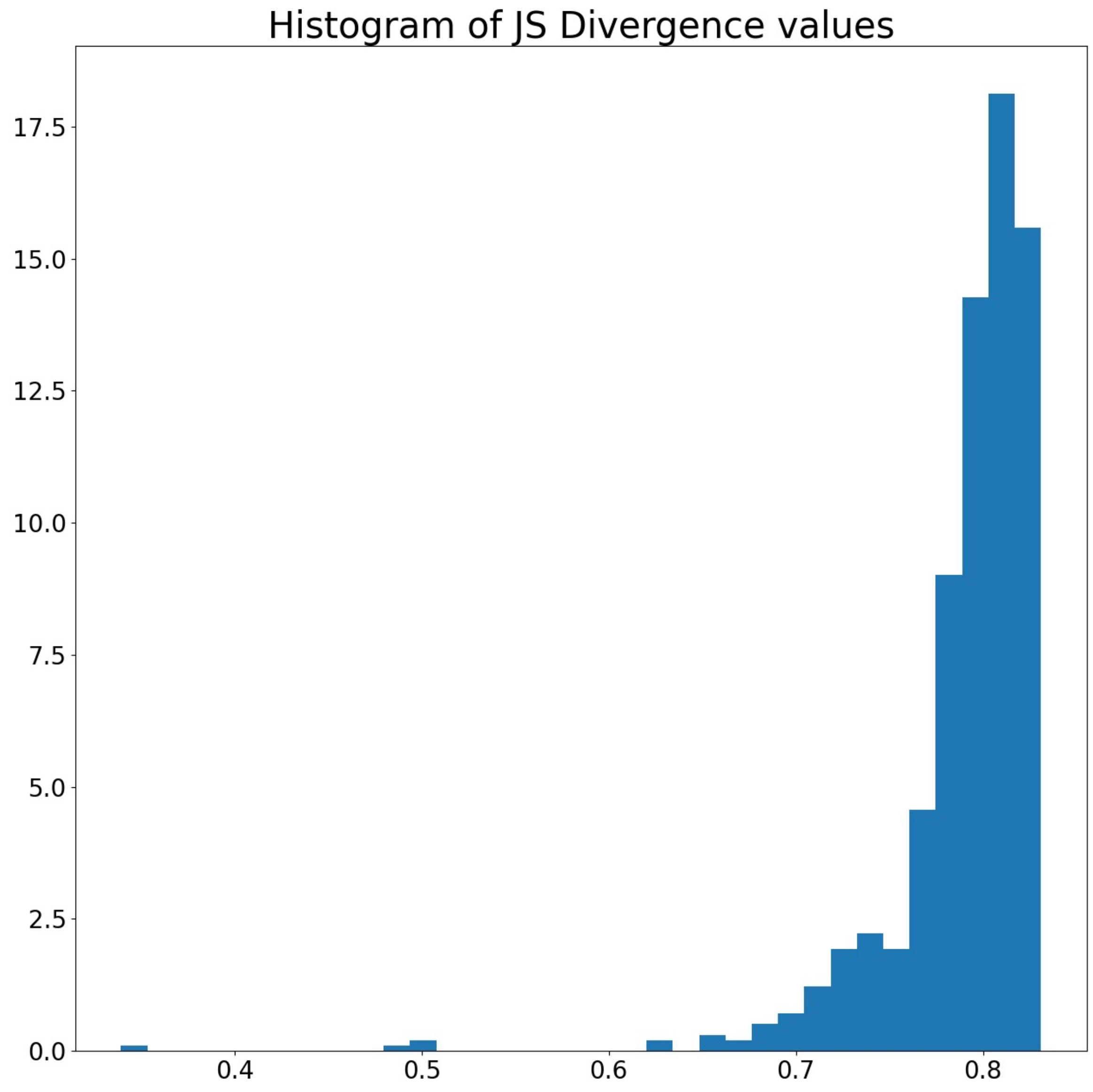
Combined t-SNE plot for all categories

❑PART 2 : Quantification

- Fit 2-D histograms on the t-SNE outputs
- Calculate the Jenson-Shannon divergence between the histogram distributions
- Set thresholds based on divergence values:
  - If too low, then merge to form one category
  - If too widespread (low values with multiple categories), remove the category.
- Returns fewer categories hence, better representation



Jenson-Shannon pairwise divergence heatmap



Histogram of pairwise Jenson-Shannon divergence values

What we get at the end?

Our algorithm solves both problems with the dataset :

- It reduces class overlap in context by merging and removing
- This also solves the imbalance problem : smaller categories become larger

Using our algorithm, we reduce to 28 categories of news descriptions.

Re-Evaluation on Removing Redundancy:

- We still use 80% of the dataset for training, 20% for evaluation.
- Data for training and validation is again sampled randomly.

Performance Metric	Value (On new)	Value(On base)
Accuracy (Top prediction)	73.60%	65.67%
Accuracy (Top 3 predictions)	90.86%	87.75%
Mean F1 Score (range [0,1])	0.6590	0.5920
Mean Reciprocal Rank (range [0,1])	0.8095	0.7574

Performance evaluation on the validation dataset

Class overlap algorithm: Quick review

What’s good :

- We ensure our model doesn’t overfit before we extract the pre-classified outputs
- Our algorithm works automatically, it can merge and remove classes based on similarity threshold

What can be improved :

- Add a way to automate setting the threshold similar to hyperparameter optimization algorithms like grid search.

Concluding remarks:

- In multi-category classification problems, contextual overlap between categories limits learning.
- To find the best model, we first run state-of-art embedding based models on the dataset.
- Following, we discover the extent of category overlap as a problem.
- We build an algorithm to quantify reduction of similar and widespread categories.
- We validate our findings by evaluating the non-redundant dataset against our base model and comparing the results.

Time for questions

Siddhant Ray
Master’s in Electrical Engineering and Information Technology
sidray@student.ethz.ch

ETH Zurich
D-ITET