NDSC Supplementary Material

Overview

This report aims to outline the approaches taken to tackle the classification problem presented by NDSC and the reasons behind the success or failure of each approach. First, several solution pipelines will be described and the choice of pre-processing methods and model architectures explained. Next, the performance of each pipeline will be discussed and the reasons for their success or failure will be explored. In particular, we hope to rationalize the underwhelming performance of solution pipelines that utilized ensemble models or reduced the problem into smaller components. Finally, the report will highlight areas to build on for solution pipelines that performed well and areas for improvement for solution pipelines that fell short of expectations.

Data-driven pipelines

We believe that unique datasets require unique pipelines and our solutions are designed with the data in mind. The NDSC dataset is unique as it already contains broad categories for each record as well as a mix of image and text data. These features of the dataset lay the foundation for the choices that characterize our solution pipelines.

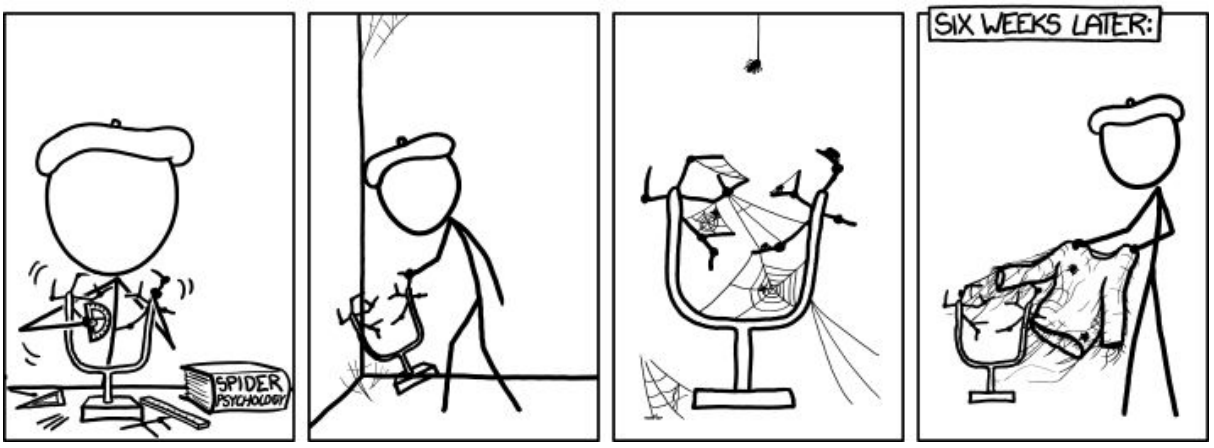

Fig. 1: With a good understanding of the problem, good solutions can come from anywhere

Firstly, the dataset was already sliced into three larger groups: Beauty, Fashion and Mobile. This led us to consider including **problem reduction** as a possible feature of our pipelines. For these pipelines, the dataset would be sliced by their groups and a different model fitted to each group. Also, the dataset was unique in offering both text and image data. This lends itself well to **ensembling**, which would be able to capitalize on the strengths of different model architectures in interpreting different forms of information.

The following sections describe the flow of two pipelines that were designed using problem reduction and ensembling in different combinations to best utilize the data.

Data preprocessing

In both pipelines, each variable in the dataset was examined and the title and image_path were selected as relevant variables. We then set out to convert the titles into word vectors and extract then resize the images from the provided archives.

Extracted and resized images were obtained from the dataset shared to the NDSC discussion section for fashion and mobile categories. However, due to hardware limitations, we were unable to obtain this data for the beauty group.

Word vectors are finite representation of the semantics of their respective words and can be used to capture the degree of similarity between seperate words. We used the standard english model from the spacy nlp library to convert each title to word vectors and used the mean word vectors to represent the title.

Ensemble model with problem reduction

The first pipeline was developed to take advantage of the groupings in the data and capitalize on both text and image data. Each group is independent, reducing room for misclassification, and uses an ensemble to combine the results of text and image models.

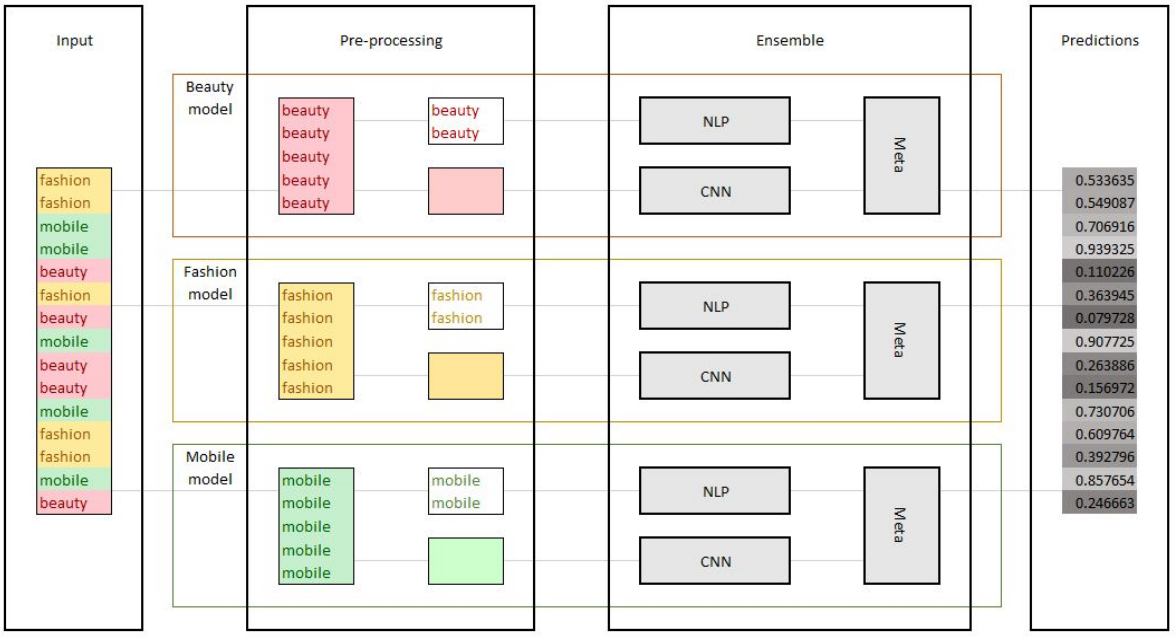

Fig. 2: Pipeline flow (ensemble model with problem reduction)

The base models consist of a NLP model and a CNN while the meta classifier is a simple softmax classifier that takes the predictions of the base models as input. The dataset is first labelled with then split by the groups extracted from the “image_path” variable. Next, image (colored cells) and text data (text cells) from each group is used train the respective base models. Finally, the last layer of these models are then concatenated and used as input for

the meta classifier. The weights for the base models are locked and the meta classifier is trained.

Single model without problem reduction

A second model pipeline was developed as a baseline model to evaluate how well our previous pipeline met its goals. This pipeline uses a single softmax classifier with the same architecture as the meta classifier in the previous pipeline and was only trained on the word vectors from the entire dataset.

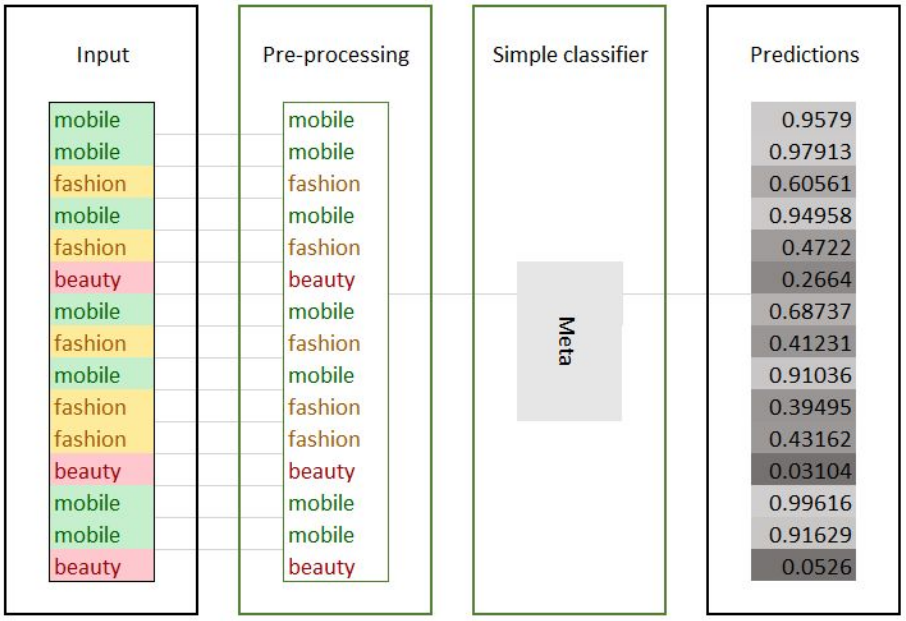

Fig 3: Pipeline flow (Single model without problem reduction)

Methodology

To train both pipelines, the train data was first pre-processed accordingly, then split into train and validation batches. Each model was fitted on the train batch and evaluated on the validation batch. Model weights were saved and validation accuracy was recorded after each epoch. Model fitting was stopped after there was no improvement for 5 epochs or just before the kernel time limit (all models were trained on kaggle). The best model weights were then exported for use in other parts of the pipeline or to resume training if training had been interrupted by kernel time constraint.

All models were constructed in tensorflow on a kaggle kernel and several public resources were used in our pipelines. Namely, our CNN architecture was based off the Xception model which was trained using transfer learning. Also, spacy's standard english nlp model was used for word embedding.

Results

During training, the individual meta models in the ensemble pipeline with problem reduction obtained validation accuracies ranging from 75 - 80% while the single model pipeline without problem reduction obtained validation accuracies of only around 70%. Surprisingly, the ensemble pipeline with problem reduction performed much worse than the single model pipeline without problem reduction on the test set. The former attained a test accuracy of 41.316% while the latter attained a test accuracy of 69.764%.

Additionally, it was noted that all CNN models achieved far lower accuracies than NLP models in all situations.

Discussion

We feel that the surprising results from the test set were definitely due to overfitting of the models in the first pipeline. Although these results are discouraging, it does indicate that the approach of problem reduction is indeed effective.

However, every cloud has a silver lining and we were pleasantly surprised with the results achieved by the second pipeline, which took only word vectors as input. One of the key features of this pipeline that only emerged on hindsight was it’s simplicity.

Summary

On the whole, the project has been a valuable learning process that has improved our knowledge in the field of machine learning and data science. While our results are not spectacular and are far from what we initially hypothesized, we feel that the shortcomings in our first pipeline can be addressed with more careful regularization methods and cross-validation.

Additionally, we gleaned the insight that simple models making use of relevant and meaningful data may perform better than complex models. Again, this reiterates our philosophy of finding a solution that fits the data. Perhaps future work on this dataset could incorporate more analysis and feature selection of the image data.