Jakka Shashank
IMSc. Applied Mathematics
IIT Roorkee
7895189165
jakkashashank1@gmail.com

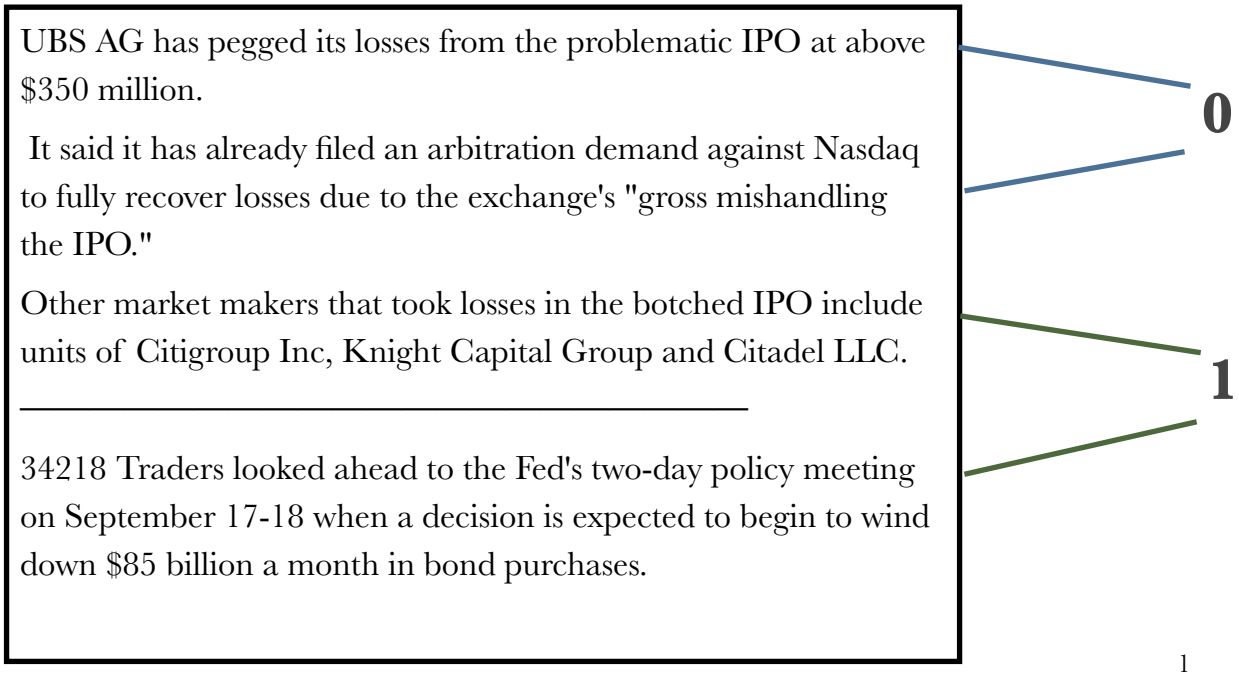
Problem 1 Report

Topical Segmentation of Financial News Documents

Approach

- The idea is to train the document vector space model on a corpus of given train and val sets using gensim's Doc2Vec.
- Then split each text file into sentences and then pair two consecutive sentences.

1. If the pair of sentences are from the same topic then label it as 0.
2. If the pair of sentences are from the different topic then label it as 1.



- Now for building the corpora we need to first investigate training on different lengths of text.
- For example combinations of all full text files, all text files broken down into their sections. Since we are also interested at a sentence level, we also want to try combinations with sentences too.
- The final corpora looks like this.

    1. Whole text file as a document {T}
    2. Each topic in a text file as a document {P}
    3. Consecutive pair of sentences as a document {S}

- The whole final corpora **{F}** = {T} U {P} U {S}
- Once we have this model, we are looking to classify sentences as members of binary classes.

Challenges and Short Comings

- The entire approach depends upon how well we split the text files into sentences.
- Presence of different forms of acronyms made the pre-trained tokenisers like Punkt work very bad.
- Had to depend heavily on regex, considering multiple cases which were found by manually looking up. (could still be missing so many edge cases)
- After converting the pairs of sentences into vectors, the usual similarity measures have been explored, these all required manual thresholding.(not so efficient)
- An XGBoost model has been trained for binary classification on two sentences vector dataset.

- Combining this vector pair with the usual addition, concatenation losses information.
- So instead could have trained an Embedding layer with shared weights in Keras.