Text Classification on Reddit Data

Abstract: (To briefly summarize the accomplishments of our project and figure out keywords.)

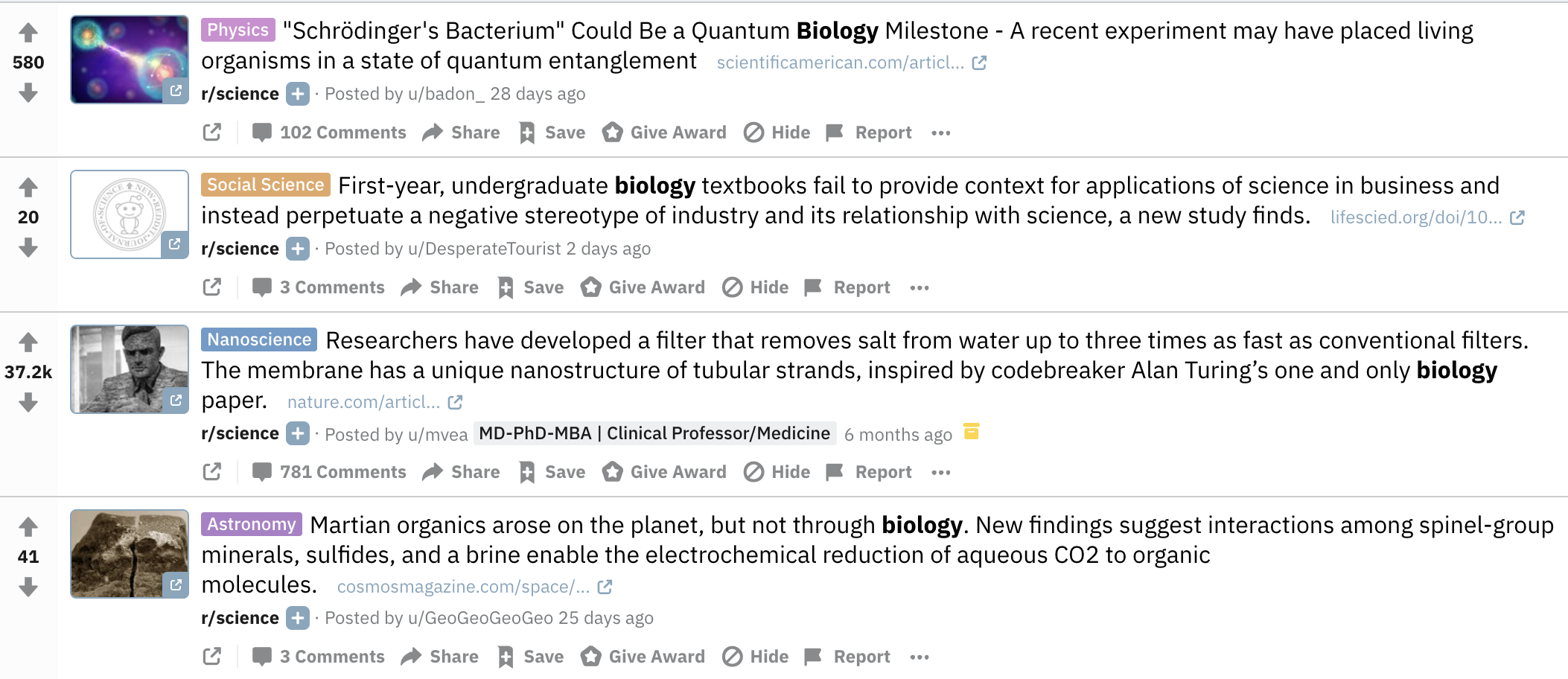
1. Introduction.

To summarize the project in detail, introduce previous study and figure out feasible methods for our project. (You need to include some references.)

1. Data
   1. Analysis of Reddit Data

Reddit is an entertainment, social and news site where registered users can post text or links on the site, making it basically an electronic bulletin board system. By analysis of Reddit, we find that many questions and topics users propose are not classified or even misclassified. Thus, an effective approach for correctly classifying questions can vastly enhance user experience.

We observe that titles contain sufficient information for text classification task (see figure 1). Our experiments show that the average length of titles is around 15, which meets the requirements of the short text classification task. Moreover, the huge dataset on Reddit enables us to retrain word embeddings and train deep learning models.



* 1. Methods of Collecting Data

PRAW library offers python API for retrieving data from Reddit. Data can be scraped by keywords and tags. We first specify four topics used as labels and then come up with related keywords and tags. By applying PRAW library, we collect approximately 20,000 titles on Reddit.

Undoubtedly, the raw data contain noise and even errors. In order to clean raw data, we employ the following strategies. Firstly, titles whose length is less than 4 are filtered. Each sentence is tokenized prior to being fed into our models. Secondly, questions on Reddit might be misclassified individually. It is important to correct labels that is obviously false. Before annotating data, we formulate a rubric which helps us determine which labels sentences should have belonged to (see table 1).

* 1. Dataset

Dataset includes 4 categories: science, politics, education, and sports. After filtering data and correcting their labels, we eventually set up a Reddit dataset containing around 20,000 samples. Each category contains around 5,000 samples.

For the purpose of evaluating models, dataset is splitted into two smaller ones. One is training data with around 16,000 samples while another one is testing data with 4,000 samples.

1. Models
   1. Word Embeddings

Word Embedding can effectively capture similarity among words. Therefore, word embedding instead of one-hot vectors used as input is the key to improve the results. There are many considerations when choosing word embeddings.

Firstly, word2vector invented by Google researchers displays excellent performance. Google Word2Vec and GloVe are two Candidates. Google Word2Vec uses news as training data while the one of GloVe word embeddings uses twitter text as training data. Considering that data on Reddit should be more similar to data on twitter, we use GloVe (for twitter) as our word embeddings.

Secondly, the dimension of word embeddings impacts classification accuracy and time complexity. In terms of trade-off between accuracy and efficiency, we eventually choose word embeddings with 50 dimensions.

Thirdly, we need to retrain word embeddings such that they fit with our data. With the help of Gensim library, we generate new word embeddings with the same dimensions on the basis of GloVe word embeddings.

* 1. RNN
     1. Structure of neural networks
     2. Considerations. (hyper-parameters?, mini-batch?, dropout?, etc.)
     3. Partial results. (loss, confusion matrix, etc.)
  2. CNN
     1. Structure of neural networks
     2. Considerations. (hyper-parameters?, mini-batch?, dropout?, etc.)
     3. Partial results. (loss, confusion matrix, etc.)
  3. LSTM
     1. Structure of neural networks
     2. Considerations. (hyper-parameters?, mini-batch?, dropout?, etc.)
     3. Partial results. (loss, confusion matrix, etc.)
  4. LR
     1. Considerations. (hyper-parameters?, mini-batch?, dropout?, etc.)
     2. Partial results. (loss, confusion matrix, etc.)
  5. Etc.
  6. Comparisons Between Models
     1. Results
     2. Pros and cons

1. Conclusion
   1. Conclusion
   2. Potential improvements in the future.
2. Reference