Text Classification on Reddit Data

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

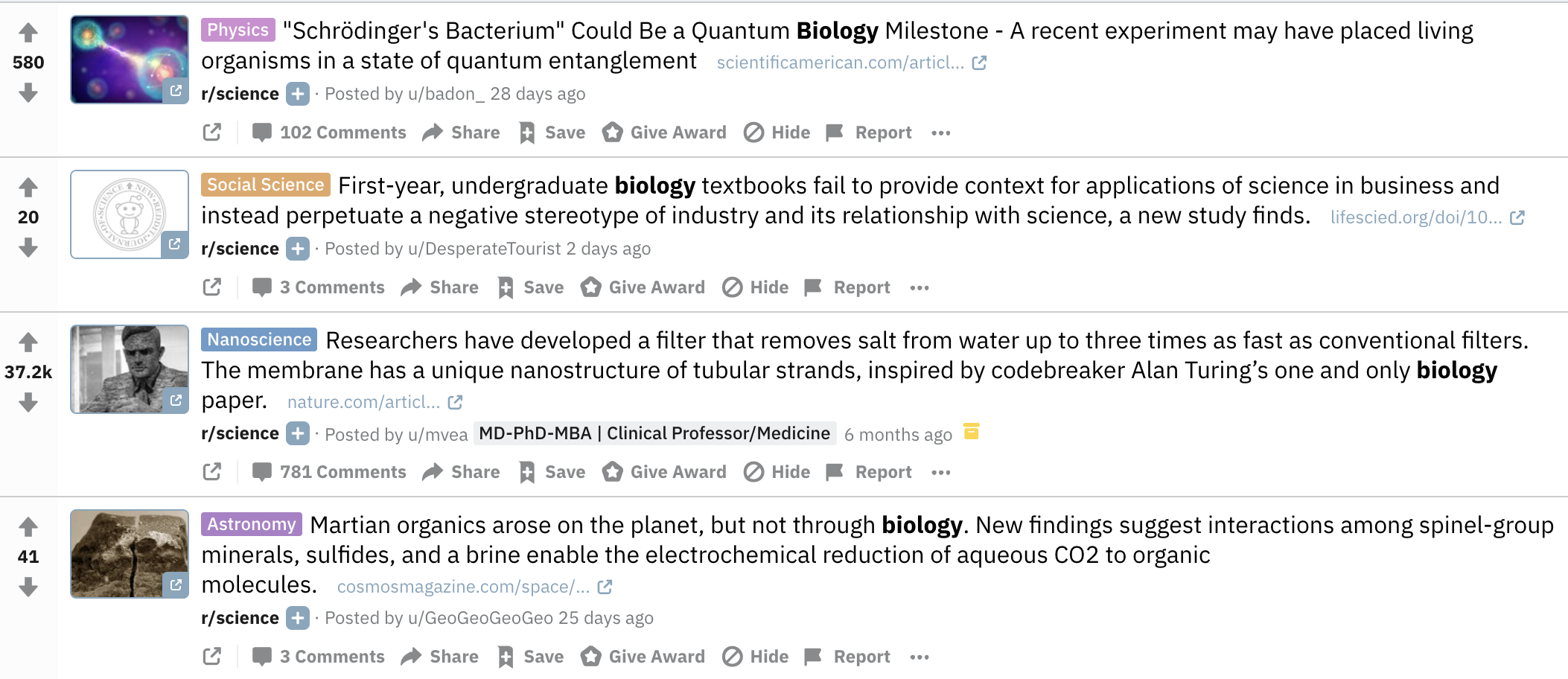
1. Introduction.

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 Rubrics in appendix).

* 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
3. Appendix
   1. Rubric for annotating data

Rubric mainly includes four parts respectively for politics, science, sports, and education. If a title is associated with entity type defined in the table, it will be annotated with the corresponding label.

politics

|  |  |
| --- | --- |
| Entity type | Description |
| Politics | The art or science of government. Ex. Affairs, businesses, policies. |
| Vote | To express one's views in response to a poll especially**:** to exercise a political franchise. Ex, [ballot](https://www.merriam-webster.com/dictionary/ballot), franchise, advance, pose, propose, suggest |
| President | An appointed governor of a subordinate political unit. Ex. Donald trump, Barack Obama, George Walker Bush. |
| Election | The right, power, or privilege of making a choice. Ex. Choice, decision, alternative, selection. |
| Republican | Of, relating to, or having the characteristics of a republic. Ex. nondemocratic, undemocratic. |
| Democratic | Of, relating to, or favoring democracy. Ex. popular, republican, self-governing, self-running. |
| Trump | A US president. Ex. Trumpery, Donald Trump. |
| Government | The act or process of governing, specifically: authoritative direction or control. Ex. Administration, control, direction, governance, management |
| Judge | A public official authorized to decide questions brought before a court. Ex. Adjudicator, arbiter, arbitrator, referee, umpire. |
| White house | A residence of the president of the U.S. Ex. Presidents. |
| Congress | The supreme legislative body of a nation and especially of a republic. Ex. Association, council, consortium, organization, institution. |

Education

|  |  |
| --- | --- |
| Entity type | Description |
| School | A source of knowledge. Ex. Teach, instruct, educate, train, discipline. |
| Teacher | One that teaches especially**:** one whose occupation is to instruct. Ex. educator, instructor, pedagogue, preceptor. |
| High-school | A school especially in the U.S. usually including grades 9–12 or 10–12. Ex. Teenagers. |
| Student | Scholar, learner especially**:** one who attends a school. Ex.  Pupil, scholar. |
| Homework | An assignment given to a student to be completed outside the regular class period. Ex. Assignment, paper, questions. |
| Assignment | A specified task or amount of work assigned or undertaken as if assigned by authority. Ex. Task, work. |
| Class | A body of students meeting regularly to study the same subject. Ex. Major. |
| Education | The action or process of educating or of being educated also**:** a stage of such a process. Ex. Learning. |
| University | An institution of higher (or tertiary) education and research which awards academic degrees in various academic disciplines. Ex. University of Boston, University of Southern California, New York University |
| College | An educational institution or a constituent part of one. Ex. Dartmouth College, The College of William & Mary. |
| GPA | Grading in education. Ex. 4.0, 3.6. |
| PHD | Doctor of Philosophy (PhD), an academic qualification. Ex. Doctor of Science, Doctor of Arts. |
| Bachelor | An undergraduate academic degree awarded by colleges and universities upon completion of a course of study lasting three to seven years. Ex. Bachelor of Science, Bachelor of Arts. |
| School violence | Encompasses physical violence, psychological violence, sexual violence, many forms of bullying. Ex. Rape, verbal abuse. |
| Undergraduate | An entry level university student. Ex. College stage, teenagers. |

Sports

|  |  |
| --- | --- |
| Entity type | Description |
| NBA | National Basketball Association. Ex. Lakers, Golden States Warriors. |
| Basketball | A usually indoor court game between two teams of usually five players each who score by tossing an inflated ball through a raised goal. Ex. Shooting guard, small forward. |
| Football | Any of several games played between two teams on a usually rectangular field having goalposts or goals at each end and whose object is to get the ball over a goal line, into a goal, or between goalposts by running, passing, or kicking. Ex. Soccer, rugby, Real Madrid, FC Barcelona, La Liga. |
| NFL | National Football League. Ex. Los Angeles Chargers, New York Jets. |
| Tennis | An indoor or outdoor game that is played with rackets and a light elastic ball by two players or pairs of players on a level court (as of clay or grass) divided by a low net. Ex. Wimbledon, US Open, French Open, Roger Federer. |
| Player | A person who plays a game. Ex. Participant, gamer. |
| League | An association of nations or other political entities for a common purpose. Ex. Association, union. |
| Olympic | Of or relating to the Olympic Games. Ex. Winter, summer, opening ceremony. |
| Sport | To amuse oneself. Ex. Running, cycling, hiking. |
| Team | A number of persons associated together in work or activity: a group on one side (as in football or a debate). Ex. Crew, party, union, squad. |
| Champion | A winner of first prize or first place in competition. Ex. Championship, winner, gold matel. |

Science

|  |  |
| --- | --- |
| Entity type | Description |
| Biology | The natural science that studies life and living organisms. Ex. Physical structure, chemical process |
| Chemistry | The scientific discipline involved with elements and compounds composed of atoms, molecules and ions. Ex. Composition, structure, reaction |
| Physics | The natural science that studies matter and its motion and behavior through space and time and that studies the related entities of energy and force. Ex. Nuclear physics, electromagnetism. |
| Computer | A device that can be instructed to carry out sequences of arithmetic or logical operations automatically via computer programming. Ex. Hybrid computer, digital computer. |
| Web | An electronic communications network that connects computer networks and organizational computer facilities around the world — used with *the* except when being used attributively. Ex. Internet, Google. |
| Machine Learning | The study of algorithms and mathematical models that computer systems use to progressively improve their performance on a specific task. Ex. Mathematical optimization, unsupervised learning. |
| Robot | A machine—especially one programmable by a computer— capable of carrying out a complex series of actions automatically. Ex. Patient assist robots, dog therapy robots. |
| NLP | Natural language processing. Ex. Speech recognition, natural language understanding. |
| Geometry | A branch of mathematics. Ex. The properties of space, size of figures. |
| Space | The boundless three-dimensional extent in which objects and events have relative position and direction. Ex. Dimension |