Text Classification on Reddit Data

Zheng Cao, Guangfan Chen, Dongrun Lu, Shuang Yuan

November 30, 2018

Abstract

This report is mainly concerned about the strategies we apply to collect and annotate data and the models for addressing the issue of text classification. Data is scraped from Reddit and annotated by predefined rubric. GloVe,which is word embedding, is fed into our models. Through the experiments, LSTM with self-attention technique achieves 0.96 f1 score on our dataset.

1 Introduction

Among classification tasks, text classification is important in our daily life. The purpose of text classification is to classify the text into an appropriate categories based on its context. Text classification is mainly used for information retrieval, machine translation, automatic summarization, information filtering, mail classification, and etc. Our experiments mainly focus on text classification on questions posted on Reddit.

2 Dataset

2.1 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.

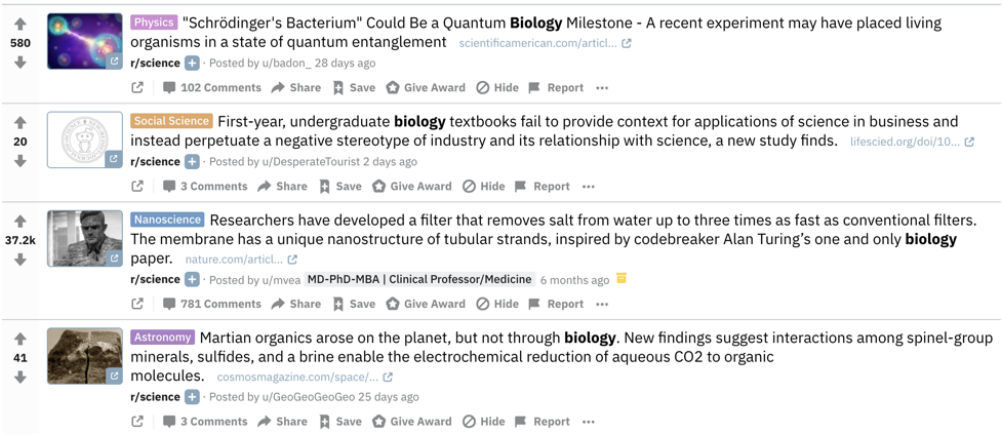


Figure 1: partial search results from Reddit

2.2 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. 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.(see Rubrics in appendix).

2.3 Parameters of 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.

3 Models

3.1 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.

3.2 Linear Regression Model

3.2.1 Motivation

Multinomial Logistic Regression is the model that we first try to deal with the problem of our project. It's the simplest and most intuitive way to handle the multi-classification problem. Our model is using a linear function to predict the topic of a vector of problem text and using cross entropy loss to optimize our model.

3.2.2 Hyperparameters setting

To find optimal hyperparameters, we alter one hyperparameter while keeping others fixed. Since other parameters in different models can be seen in the code we submit, only parts of important hyperparameters are shown as below (see Table 1):

Hyperparameters	**value**
Epochs	5
Batch size	16

Table 1: Hyparameters for LR

3.2.3 Pseudocode for Training Process

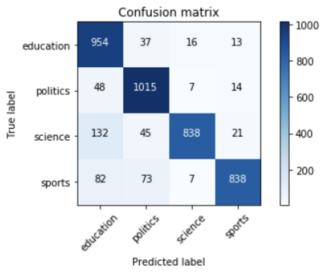
Algorithm framework is almost same when applying different models. So we just display the training process of LR.

Algorithm 1 LR training process

Initialization: Initialize $model, optimizer, lossfunction$;
repeat
  for $X\_batch, y\_batch \in dataset$ **do**
    $output = LogisticRegression(X\_batch)$;
    use $CrossEntropyLoss()$ to calculate $loss$;
    use $Adamoptimizer$ to update parameters
  end for
until maximum epoch is reached

3.2.4 Results

The result is shown in Figure 2.



	precision	recall	f1-score	support
education	0.78	0.94	0.85	1020
politics	0.87	0.94	0.90	1084
science	0.97	0.81	0.88	1036
sports	0.95	0.84	0.89	1000
avg / total	0.89	0.88	0.88	4140

Figure 2: confusion matrix, precision, recall and f1-score for LR

3.3 CNN

3.3.1 Motivation

We notice that using Logistic Regression to handle the text classification problem is not an optimal solution because it ignores the relation between the words which is an important feature of a sentence. Therefore, we use a multi-layer CNN to obtain the local relationship between words of a sentence and see whether it can improve our model.

3.3.2 Model Structure

Model structure is shown in Figure 3 .

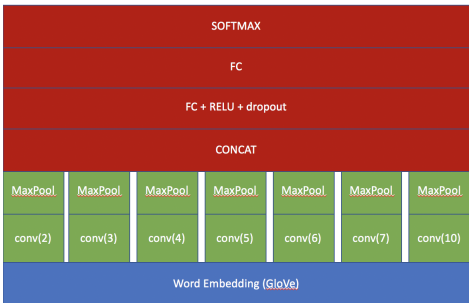
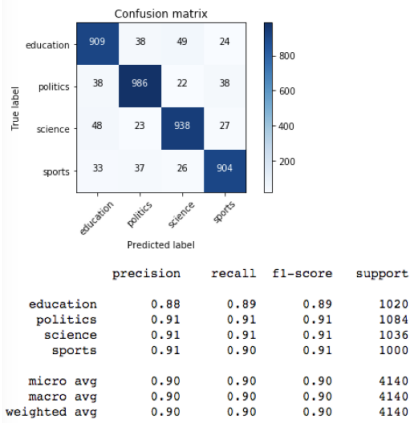


Figure 3: CNN model structure

3.3.3 Results

The result is shown in Figure 4 .



	precision	recall	f1-score	support
education	0.88	0.89	0.89	1020
politics	0.91	0.91	0.91	1084
science	0.91	0.91	0.91	1036
sports	0.91	0.90	0.91	1000
micro avg	0.90	0.90	0.90	4140
macro avg	0.90	0.90	0.90	4140
weighted avg	0.90	0.90	0.90	4140

Figure 4: confusion matrix, precision, recall and f1-score for CNN

3.4 RNN

3.4.1 Motivation

Although the CNN model can get some local relation features of a sentence, the information is constrained by the filter size of convolutional layer. In order to express the context features better, we use RNN as our improved model.

3.4.2 Model Structure

Rnn model structure is shown in Figure 5 .

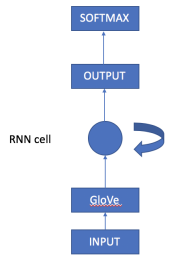
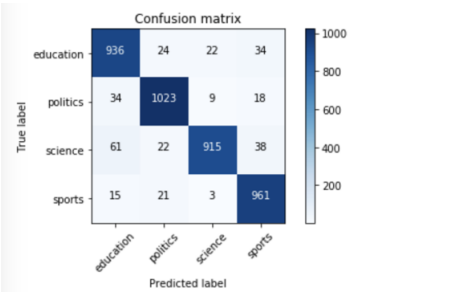


Figure 5: RNN structure

3.4.3 Results

The result is shown in Figure 6



	precision	recall	f1-score	support
education	0.89	0.92	0.91	1016
politics	0.94	0.94	0.94	1084
science	0.96	0.88	0.92	1036
sports	0.91	0.96	0.94	1000
avg / total	0.93	0.93	0.93	4136

Figure 6: confusion matrix, precision, recall and f1-score for RNN

3.5 LSTM

3.5.1 Motivation

After training our model by using RNN, we have already got pretty good results in our test dataset. However, we want to find whether we can get a better result by using a more complicated model. Therefore, we use the LSTM to replace basic RNN model.

3.5.2 Model Structure

Lstm model structure is shown in Figure 7

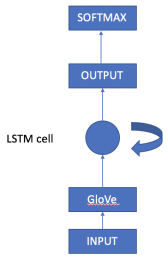
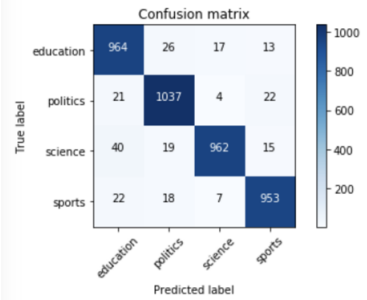


Figure 7: LSTM model structure

3.5.3 Results

The result is shown in Figure 8



	precision	recall	f1-score	support
education	0.92	0.95	0.93	1020
politics	0.94	0.96	0.95	1084
science	0.97	0.93	0.95	1036
sports	0.95	0.95	0.95	1000
avg / total	0.95	0.95	0.95	4140

Figure 8: confusion matrix, precision, recall and f1-score for LSTM

3.6 BI-LSTM with Self-attention

3.6.1 Motivation

In our project, we also use the self-attention framework to solve the problem of short text classification (https://arxiv.org/pdf/1703.03130.pdf).

3.6.2 Results and Discussion

In this algorithm, we have several hyper-parameters. We refer to a open source code (https://github.com/kaushalshetty/Structured-Self-Attention): $batch\_size = 512, h = 150, epoch = 2, max\_length = 40, a\_hops = 10, a\_h = 150$

Results can be seen in Figure 9.

	precision	recall	f1-score	support
education	0.94	0.96	0.95	1020
sports	0.97	0.96	0.96	1000
science	0.96	0.97	0.96	1036
politics	0.98	0.96	0.97	1084
avg / total	0.96	0.96	0.96	4140

Figure 9: Precision, recall and f1-score for LSTM with attention

3.7 Comparison among Models

For solving the short text classification problem, especially for our project of classifying reddit titles, the main task is extracting the keywords of a sentence. In our project, we first use Logistic Regression model as a baseline. Then, we tried CNN and RNN to improve the performance of our model. In order to get the further improvement, we tried self-attention mechanism. Compared with RNN and LSTM, self-attention contains a Bi-LSTM layer, which can further extract the contextual features in both sequential and reverse order. And it uses attention to extract some keywords at the same time. Attention mechanism can be considered as a kind of "exquisite Bag-of-words model". It can give higher weights to some keywords with strong effects on classification task in a sentence and thus can contribute to a better classification result.

Model	**F1 Score**	**Time**	**Pros**	**cons**
LR	0.88	5s	simple and fast	it ignores relations among words.
CNN	0.90	59s	it expresses local relation features among words.	it is unable to establish long term dependency.
RNN	0.93	29s	it makes full use of context of the whole sentence	it is unable to omit useless information.
LSTM	0.95	80s	it deals with the problem of long term dependency.	it is more complicated and requires more training time.
LSTM with self-attention	0.96	-		

Table 2: Comparison among models

4 Conclusion

I am the conclusion

5 Appendix

5.1 Rubric for annotation

Rubric for 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, 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.

Rubric for label "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.

Rubric for label "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.

Rubric for label "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