Assignment 2, Article Classification

Yang Ge, SID:450005028

Baseline System

The baseline system is using bag of words. It first tokenize the 'body' for each row in dataset, and split it to training data and testing data. Then use training data, training target('annotation' or 'class') and Logistic Regression to train, use testing data and testing target ('annotation' or 'class') to predict.
Average accuracy: 0.68079
Macro average precision: 0.57176, Macro average recall: 0.50854, Macro average f1 score: 0.52122
One fold of cross validation

	Precision	Recall	F1_score
Entertainment	0.71	0.75	0.73
Politics	0.75	0.58	0.65
Sports	0.85	0.89	0.87
Business	0.76	0.59	0.67
Other	0	0	0
Society	0.58	0.81	0.68
War	0.50	0.67	0.57
Health	1.00	0.23	0.38
Error	0	0	0
Science and Technology	0.33	0.50	0.40

Most weighted features for each class

class	3 Most Positive	3 Most Negative
Business	'business''market''its'	'old''his''our'
Entertainment	'actor''show''film'	'police''against''said'
Error	'settings''cuomo''log'	'which''year''2014'
Health	'cancer''health''ebola'	'like''police''on'
Other	'your''degrees''pope'	'her''their''they'
Politics	'party''minister''president'	'old''police''university'
Science and Technology	'science''technology''internet'	'not''twitter''president'
Society	'weather''pay''police'	'show''star''film'
Sports	'team''game''games'	'mr''show''by'
War	'islamic''syria''troops'	'with''like''up'

Baseline system did not predict any 'Other' and 'Error' class.
Precisions for 'Society', 'War' and 'Science and Technology' are low. The system classified many other articles to these classes.
Recalls for 'Politics', 'Business', 'Health', 'Science and Technology' are low. The system tend to classify articles which belong to these classes to other classes.
Some meaningless words like "its" in "Business" should not be considered as one of the most weighted features.

Experiment 1

N-gram specification: character n-grams(2-gram)

Comparing with baseline, instead of using unigram, this experiment used bigram(2-gram), pair of words as a feature.
In baseline system, bag of words ignores word ordering, this may lead to wrong understanding. Using bigram word would help the system to deal with word ordering issue. For example, a sentence "A ate B" would be tokenized as ["A", "ate", "B"] in baseline system, in this case, it could be considered as "A ate B", or "B ate A" which is the oppsite of the orginal meaning, while in bigram, it would be tokenized as ["A ate", "ate B"] which can only lead to "A ate B".
Some word may have different meaning and impact when it combines with other words together. For example, "like" and "not like".
2-gram could not cover all possible senarios comparing with 3-gram or 4-gram, it may even cause new misunderstanding.

Average accuracy: 0.61965

Macro average precision: 0.45913, Macro average recall: 0.38627, Macro average f1: 0.39580

One fold of cross validation

	Precision	Recall	F1_score
Entertainment	0.71	0.75	0.73
Politics	0.69	0.58	0.63
Sports	0.87	0.87	0.87
Business	0.77	0.45	0.57
Other	0	0	0
Society	0.50	0.81	0.62
War	0	0	0
Health	0	0	0
Error	0	0	0
Science and Technology	0.33	0.17	0.22

Most weighted features for each class

class	3 Most Positive	3 Most Negative
Business	'20 saturday''phone has''the company'	'year old''on saturday''he was'
Entertainment	'with her''the show''the film'	'the company''said the''according to'
Error	'your video''update your''please log'	'at the''in the''to the'
Health	'the disease''the virus''the hospital'	'on the''in the''to the'
Other	'the city''kay lee''the pope'	'on the''in the''of the'
Politics	'the government''the president''prime minister'	'year old''the company''if you'
Science and Technology	'phone turned''the internet''20 on'	'year old''at the''on the'
Society	'the incident''said the''the police'	'the show''prime minister''the film'
Sports	'to play''the team''the game'	'said the''more than''by the'
War	'in iraq''in syria''islamic state'	'at the''has been''to the'

Experiment 2

Normalisation and weighting: remove stop words

This experiment removed English stopwords.

Some words in English are useless for article classification, e.g. "a" and "the", but they may appear a lot an article. For instance, a meaningless word apears a lot in article1, and the system use this article for training, article2 does not have that word at all, when predicting the class of article2, the result may be affected by that meaningless word.

After removing stopwords, the system would not be affected by these meaningless stopwords.

Average accuracy: 0.68253

Macro average precision: 0.56336, Macro average recall: 0.50579, Macro average f1: 0.51685

One fold of cross validation

	Precision	Recall	F1_score
Entertainment	0.76	0.72	0.74
Politics	0.77	0.65	0.70
Sports	0.88	0.92	0.90
Business	0.92	0.50	0.65
Other	0	0	0
Society	0.60	0.86	0.71
War	0.67	0.67	0.67
Health	0.75	0.23	0.35
Error	0	0	0
Science and Technology	0.36	0.83	0.50

Most weighted features for each class

class	3 Most Positive	3 Most Negative
Business	'industry''business''market'	'old''facebook''police'
Entertainment	'tv''actor''film'	'attack''police''team'
Error	'settings''cuomo''log'	'year''new''2014'
Health	'drug''ebola''cancer'	'police''like''million'
Other	'dreamliner''pope''degrees'	'police''year''old'
Politics	'party''president''minister'	'old''police''doctors'
Science and Technology	'science''internet''technology'	'twitter''police''president'
Society	'pay''police''weather'	'star''film''actor'
Sports	'game''team''games'	'years''like''company'
War	'forces''syria''troops'	'like''just''new'

Experiment 3

Normalisation and weighting: lemmatise words

This experiment lemmatized words.
In baseline system, words like "car" and "cars", "communities" and "community" are different, but they actually refers to the same thing. If the system treat a word like "cars" as symbol of a topic, then article having "car" may not be considered as that topic. After word lemmatizing, the system would treat some words who have same meaning but different form as same word, e.g. "cars" would become to "car". Average accuracy: 0.67511
Macro average precision: 0.56053, Macro average recall: 0.49798, Macro average f1: 0.51215
One fold of cross validation

	Precision	Recall	F1_score
Entertainment	0.77	0.75	0.76
Politics	0.69	0.65	0.67
Sports	0.87	0.89	0.88
Business	0.78	0.64	0.70
Other	0	0	0
Society	0.59	0.80	0.68
War	0.50	0.67	0.57
Health	1.00	0.15	0.27
Error	0	0	0
Science and Technology	0.25	0.50	0.33

Most weighted features for each class

class	3 Most Positive	3 Most Negative
Business	'business''sale''market'	'his''facebook''cm'
Entertainment	'night''actor''show'	'police''company''your'
Error	'setting''log''cuomo'	'which''also''by'
Health	'cancer''ebola''health'	'on''like''police'
Other	'university''what''pope'	'her''their''about'
Politics	'election''president''minister'	'university''police''post'
Science and Technology	'science''internet''technology'	'not''twitter''president'
Society	'girl''pay''police'	'star''game''show'
Sports	'league''game''team'	'year''actor''show'
War	'syria''troop''force'	'with''like''new'

Experiment 4

N-gram specification: words only from title

This experiment use words only from title text instead of the whole body text. This could be an improvement because for some articles, title contains the most meaningfull words, in this case, the system could easily get the most usefull features for training and predicting without read every words and be affected by some ambiguous and meaningless words in body text. For example, the article "Basketball: Tall Blacks progress in thriller - Sport - NZ

Herald News", it has words "Basketball" and "Sport" in its title, the system may directly link it to "Sports".
However, some title text may not have enough information to represent a article. Title does not have as much information to classify the documents as the full body tetxt.
Accuracy: 0.52052
Macro average precision: 0.44328, Macro average recall: 0.32286, Macro average f1: 0.33828 One fold of cross validation

	Precision	Recall	F1_score
Entertainment	0.43	0.40	0.42
Politics	0.43	0.48	0.45
Sports	0.70	0.42	0.52
Business	0.80	0.18	0.30
Other	0	0	0
Society	0.42	0.80	0.55
War	0.50	0.33	0.40
Health	0.33	0.08	0.12
Error	0	0	0
Science and Technology	0	0	0

Most weighted features for each class

class	3 Most Positive	3 Most Negative
Business	'retail''price''business'	'man''over''from'
Entertainment	'star''shows''film'	'obama''police''things'
Error	'flight''despair''tell'	'india''the''and'
Health	'medical''cancer''ebola'	'as''after''by'
Other	'your''things''tips'	'with''be''at'
Politics	'president''election''obama'	'man''best''your'
Science and Technology	'shower''smartphone''facebook'	'for''as''her'
Society	'students''man''found'	'china''star''goes'
Sports	'football''league''arsenal'	'who''that''woman'
War	'isis''ukraine''iraq'	'news''from''be'

Experiment 5

Syntax and stylometry: use POS tags

This experiement use POS tags, changed features like 'word' in baseline to ('word', POS).
In baseline system, some words which have more than one meaning, like "cook" which have different meaning when it used as noun and verb, would be treated as same word. After use POS tags, the system can distinguish these meaning, so it would get better understanding of words and articles. It can help the system deal with polysemy issue.
Average accuracy: 0.67336
Macro average precision: 0.55909, Macro average recall: 0.49582, Macro average f1: 0.50930
One fold of cross validation

	Precision	Recall	F1_score
Entertainment	0.71	0.75	0.73
Politics	0.76	0.61	0.68
Sports	0.85	0.87	0.86
Business	0.72	0.59	0.65
Other	0	0	0
Society	0.56	0.78	0.65
War	0.67	0.67	0.67
Health	1.00	0.23	0.38
Error	0	0	0
Science and Technology	0.33	0.50	0.40

Most weighted features for each class

class	3 Most Positive	3 Most Negative
Business	'('business', 'NN')''('market', 'NN')''('its', 'PRP$')'	'('his', 'PRP$')''('her', 'PRP$')''('free', 'JJ')'
Entertainment	'('music', 'NN')''('film', 'NN')''('show', 'NN')'	'('against', 'IN')''('said', 'VBD')''('your', 'PRP$')'
Error	'('photo', 'NN')''('log', 'NN')''('cuomo', 'NN')'	'('year', 'NN')''('also', 'RB')''('which', 'WDT')'
Health	'('cancer', 'NN')''('ebola', 'NN')''('health', 'NN')'	'("'s", 'POS')''('like', 'IN')''('on', 'IN')'
Other	'('your', 'PRP$')''('pope', 'NN')''('what', 'WP')'	'('in', 'IN')''('their', 'PRP$')''('her', 'PRP$')'
Politics	'('party', 'NN')''('minister', 'NN')''('president', 'NN')'	'('police', 'NN')''('islamic', 'JJ')''('university', 'NN')'
Science and Technology	'('science', 'NN')''('internet', 'NN')''('technology', 'NN')'	'('president', 'NN')''('twitter', 'NN')''('not', 'RB')'
Society	'('weather', 'NN')''('day', 'NN')''('police', 'NN')'	'('show', 'NN')''('star', 'NN')''('film', 'NN')'
Sports	'('team', 'NN')''('games', 'NNS')''('game', 'NN')'	'('show', 'NN')''('minister', 'NN')''('actor', 'NN')'
War	'('forces', 'NNS')''('islamic', 'JJ')''('troops', 'NNS')'	'('with', 'IN')''('like', 'IN')''('new', 'JJ')'

Best-performing experiment

Based on average accuracy, the best-performing experiment is experiment 2, removing stop words, but only a few improvements comparing with baseline system. Experiment 2 removed many meaningless words, so that the system would not be affected by these words.
Comparing with baseline system in the one fold of cross validation, the precisions for "Society", "War" and "Science and Technology" increased, the recall for "Politics" and "Science and Technology" increased.
In baseline system, "its" is one of the most weighted feature in class "Business" but it actually cannot help the system to classify article, "its" was removed in experiment 2, other most weighted words in baseline system, "his", "our", "agains", "which", "on", "your", "her", "their", "they", "by", "up" were also removed.
The article "Problem in Space Station Said to Pose No Danger" which should be "Science and Technology" was predicted as "Society" by baseline system. Lots of "the", "in" and other stopwords in this article. After removing stopwords, the experiment 2 predicted it correctly.
However, the macro average precision, recall and f1 score slightly decreased. This may be because the annotations of some articles are not perfectly correct, and the system need more improvments.

Virality Dataset

Baseline System: Accuracy: 0.57948, Macro average precision: 0.57681, Macro average recall: 0.57695, Macro average f1: 0.57537
Experiment 1: Accuracy: 0.60961, Macro average precision: 0.60684, Macro average recall: 0.60423, Macro average f1: 0.60213
Experiment 2: Accuracy: 0.57817, Macro average precision: 0.57557, Macro average recall: 0.57557, Macro average f1: 0.57385
Experiment 3: Accuracy: 0.58559, Macro average precision: 0.58344, Macro average recall: 0.58278, Macro average f1: 0.58079
Experiment 4: Accuracy: 0.55022, Macro average precision: 0.54648, Macro average recall: 0.54573, Macro average f1: 0.54341
Experiment 5: Accuracy: 0.59956, Macro average precision: 0.59712, Macro average recall: 0.59671, Macro average f1: 0.59519

The relative performances of each experiment do not exactly match those on the 'topic' dataset. Experiment 2 which have the best perfomance in 'topic' dataset got slightly lower accuracy, precision and recall comparing with baseline system. Experiment 4 did not have good performance, similar as using topic dataset. Other experiments got improvements comparing with baseline system.
Experiment 1, using bigram words, have the highest average accuracy and macro average f1 score.