Predication & summarizing article news

Classification, Clustering, and summarization

Problem Formulation

- Should I take time to read this article?
  - Category? → Classification problem
  - Summary? → Summarization problem

Data preparation

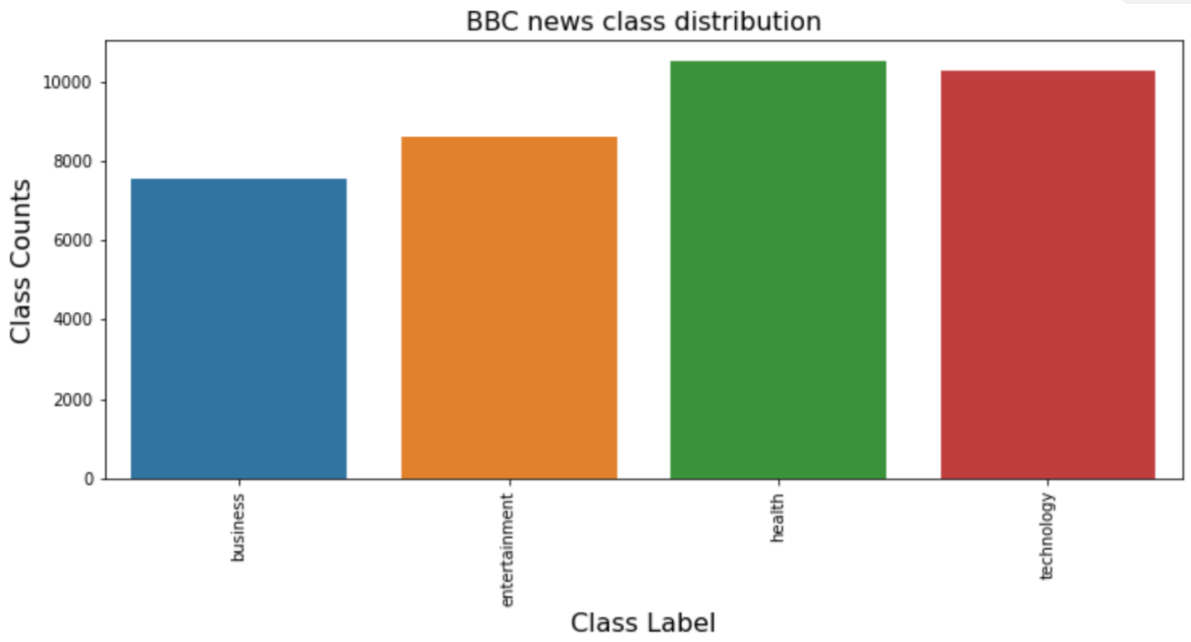
Original Data

Dataframe

	CATEGORY	CONTENT	SUMMARY
0	business	The Federal Reserve approved Ally Financial In...	The Federal Reserve approved Ally Financial In...
1	business	— Major shareholders of Duke Energy Corp. have...	— Major shareholders of Duke Energy Corp. have...
2	business	Photos taken earlier this month show that Nort...	Photos taken earlier this month show that Nort...
3	business	Thanks to dogged reporting by the Associated P...	Thanks to dogged reporting by the Associated P...
4	business	The energy giant says it is committed to clean...	The energy giant says it is committed to clean...
...	...	...	...
36886	technology	Why did this happen?\n\nPlease make sure your ...	Why did this happen?\nPlease make sure your br...
36887	technology	Google Inc (NASDAQ:GOOGL) GOOGL +1.12% (NASDAQ...	Google Inc (NASDAQ:GOOGL) GOOGL +1.12% (NASDAQ...
36888	technology	Google has purchased New Mexico-based unmanned...	Google has purchased New Mexico-based unmanned...
36889	technology	hidden\n\nLooks like Facebook's plans to get I...	Google has beaten the world's largest social n...
36890	technology	Google Has Plans For Titan Drones\n\nTitan Aer...	Google Has Plans For Titan DronesTitan Aerospa...

36891 rows × 3 columns

Histogram



Data preparation

Encoding

	CATEGORY	CONTENT	SUMMARY	category_id
0	business	The Federal Reserve approved Ally Financial In...	The Federal Reserve approved Ally Financial In...	0
1	business	— Major shareholders of Duke Energy Corp. have...	— Major shareholders of Duke Energy Corp. have...	0
2	business	Photos taken earlier this month show that Nort...	Photos taken earlier this month show that Nort...	0
3	business	Thanks to dogged reporting by the Associated P...	Thanks to dogged reporting by the Associated P...	0
4	business	The energy giant says it is committed to clean...	The energy giant says it is committed to clean...	0
...	...	...	...	...
36886	technology	Why did this happen?\n\nPlease make sure your ...	Why did this happen?\nPlease make sure your br...	3
36887	technology	Google Inc (NASDAQ:GOOGL) GOOGL +1.12% (NASDAQ...	Google Inc (NASDAQ:GOOGL) GOOGL +1.12% (NASDAQ...	3
36888	technology	Google has purchased New Mexico-based unmanned...	Google has purchased New Mexico-based unmanned...	3
36889	technology	hidden\n\nLooks like Facebook's plans to get I...	Google has beaten the world's largest social n...	3
36890	technology	Google Has Plans For Titan Drones\n\nTitan Aer...	Google Has Plans For Titan DronesTitan Aerospa...	3

Data preparation

Data cleaning

- Stop words removal
- Converting to lowercase
- Punctuation and links removal
- lemmatization

Data preparation

Data cleaning

	CATEGORY	CONTENT	SUMMARY	category_id	clean_text
0	business	The Federal Reserve approved Ally Financial In...	The Federal Reserve approved Ally Financial In...	0	federal reserve approved ally financial inc ca...
1	business	— Major shareholders of Duke Energy Corp. have...	— Major shareholders of Duke Energy Corp. have...	0	major shareholder duke energy corp called comp...
2	business	Photos taken earlier this month show that Nort...	Photos taken earlier this month show that Nort...	0	photo taken earlier month show north carolina ...
3	business	Thanks to dogged reporting by the Associated P...	Thanks to dogged reporting by the Associated P...	0	thanks dogged reporting associated press know ...
4	business	The energy giant says it is committed to clean...	The energy giant says it is committed to clean...	0	energy giant say committed cleaning dan river ...
...	...	...	...	...	...
36886	technology	Why did this happen?\n\nPlease make sure your ...	Why did this happen?\nPlease make sure your br...	3	happen please make sure browser support javasc...
36887	technology	Google Inc (NASDAQ:GOOGL) GOOGL +1.12% (NASDAQ...	Google Inc (NASDAQ:GOOGL) GOOGL +1.12% (NASDAQ...	3	google inc nasdaq googl googl nasdaq goog goog...
36888	technology	Google has purchased New Mexico-based unmanned...	Google has purchased New Mexico-based unmanned...	3	google purchased new mexico based unmanned aer...
36889	technology	hidden\n\nLooks like Facebook's plans to get I...	Google has beaten the world's largest social n...	3	hidden look like facebook plan get internet de...
36890	technology	Google Has Plans For Titan Drones\n\nTitan Aer...	Google Has Plans For Titan DronesTitan Aerospa...	3	google plan titan drone titan aerospace drone ...

36891 rows × 6 columns

Feature Engineering

- Bag of words (BOW)
- Term frequency- inverse document frequency (TF-IDF)

Prediction Algorithms

Classification

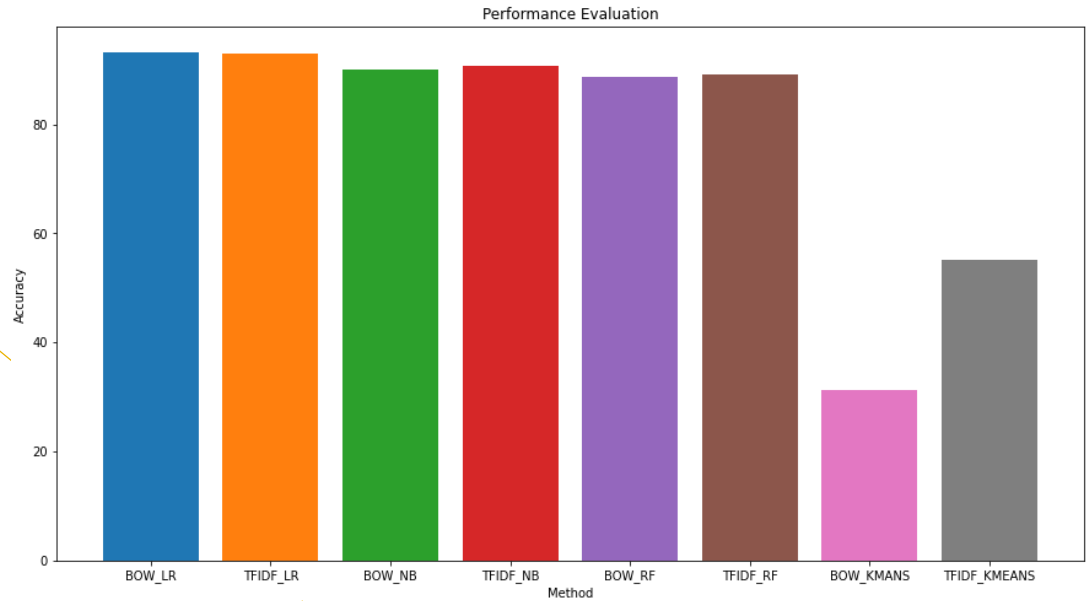
- Logistic regression (LR)
- Naïve Bayes (NV)
- Random Forest (RF)
  - Estimators = 20

Clustering

- K-means

Prediction Algorithms

Evaluation



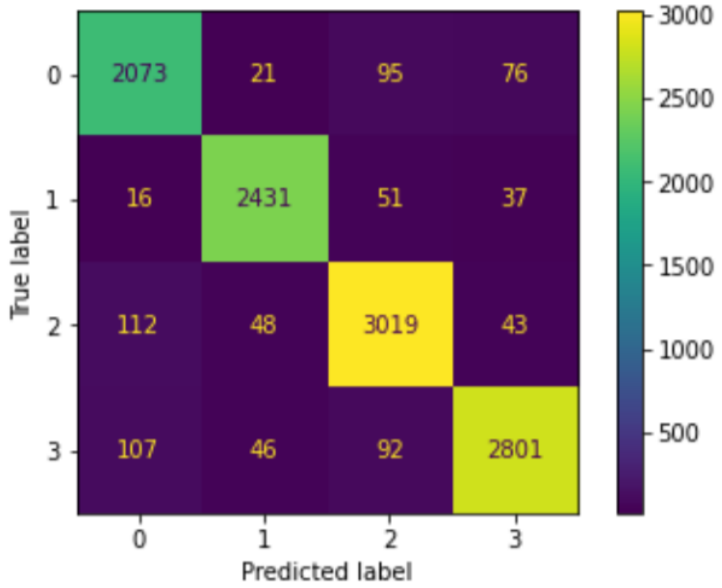
Performance evaluation for the four prediction algorithms.

	BOW	TF-IDF
silhouette score	0.090	0.009
Cohen kappa score	0.075	0.393

Clustering evaluation

Champion Model

Logistic regression with Bag Of Words with accuracy = 93.27% and confusion matrix as shown in the figure.



Error Analysis

Text[35600] misclassified from the best estimator.

True label = ‘technology’ , predicted = ‘health’

CloudWord

Collocations

59	(pilot, program)	7.409391
68	(really, need)	7.409391
67	(rather, excuse)	7.409391
66	(quick, adopt)	7.409391
...	...	...
164	(glass, also)	4.087463
165	(medication, patient)	3.824428
166	(patient, allergic)	3.824428
167	(patient, file)	3.824428
168	(glass, horng)	3.502500

169 rows × 2 columns

45	(pilot, program, instituted)	14.818782
51	(really, need, tweeting)	14.818782
50	(rather, excuse, said)	14.818782
49	(quick, adopt, data)	14.818782
...	...	...
163	(patient, man, allergic)	11.233819
164	(glass, horng, reliance)	10.911891
165	(google, glass, horng)	10.911891
166	(horng, reliance, glass)	10.911891
167	(medication, patient, allergic)	10.233819

168 rows × 2 columns

Summarization

Latent Semantic Analysis (LSA)

Manual Summary

'The Federal Reserve approved Ally Financial Inc.’s capital plan in the bank regulator’s annual review of the industry’s financial health, clearing another potential hurdle to the auto lender’s plans to exit government ownership.\nAlly’s plan was approved after the Federal Reserve found that the bank could keep lending in a severe economic downturn, according to a report Wednesday.'

LSA Summary

<Sentence: Ally’s plan was approved after the Federal Reserve found that the bank could keep lending in a severe economic downturn, according to a report Wednesday.>

Summarization

BERT Summarizer

Manual Summary

'The Federal Reserve approved Ally Financial Inc.’s capital plan in the bank regulator’s annual review of the industry’s financial health, clearing another potential hurdle to the auto lender’s plans to exit government ownership.\nAlly’s plan was approved after the Federal Reserve found that the bank could keep lending in a severe economic downturn, according to a report Wednesday.'

BERT Summary

'The Federal Reserve approved Ally Financial Inc.’s capital plan in the bank regulator’s annual review of the industry’s financial health, clearing another potential hurdle to the auto lender’s plans to exit government ownership.'

Summarization Evaluation

- Similarity using Spacy.
  - LSA Similarity = 99%
  - BERT Similarity = 16%

Summarization Error Analysis

- LSA model covered all the main keywords in the article with more compact size.
- BERT model missed some important information in the final sentences of the article.
- BERT model has more details that don’t affect the meaning of the article.

Innovativeness

- Adding summarization as a service with determining the category.
- Clustering to unlabeled data.

Future Work

- Transfer learning from the pretrained model with different data and tuning the hyper parameters.

References

- https://arxiv.org/abs/1906.04165
- https://github.com/dmmiller612/bert-extractive-summarizer
- http://lia.disi.unibo.it/Events/Confs&Works/URANIA2016/slides/01-urania 2016.pdf
- https://transformersum.readthedocs.io/en/latest/extractive/models-results.html#pretrained-ext

Thank you ☺