#	Name	Description	Nature	Classification Type	Evolving class labels	No. of classes	Citation	No. of instances	Reference	Link
1	BBC topic detection	A set of 2225 news documents from BBC news, categorized into 5 classes (business, entertainment, politics, sport, tech)	Real	Single Label	No	5	197	2,225	[1]	[1]
2	20 Newsgroups	Set of news documents	Real	Single Label	No	20	2360	~20,000	[2]	[2]
3	Reuters-21578 (RCV1)	News documents appeared on the Reuters financial newswire in 1987	Real	Multi label	No	90	-	10,788	[3]	[3]
4	DBPedia Ontology Classification Dataset	A set of articles obtained from wikipedia.org categorized into 14 classes.	Real	Single Label	No	14	1452	560,000 train & 70,000 test	[4]	[4]
5	Yahoo! Answers Topic Classification	Categorization of answers for questions in Yahoo.com	Real	Single label	No	10	1434	~4.4M & ~ 1.4M	[5]	[5]
6	AG News	News data from 2000 news sources	Real	Single label	No	4	1434	120,000 train and 7,600 test	[6]	[6]
7	Real-Time Classification of Twitter Trends	A set of tweets corresponding to 1036 trending topics	Real	Single Label	Yes	1.036 events or 4 classes	107	~1M	[7]	[7]
8	Twitter News Dataset	Tweets categorized into 5234 news events obtained from twitter	Real	Single label	Yes	5234 news events or 20 clusters	-	~10M	[8]	[8]
9	TREC-6	Question classification into 6 semantic classes.	Real	Single Label	No	6	-	~6,000	[9]	[9]
10	TREC-50	Question classification into 50 semantic classes.	Real	Single Label	No	50	-	~6,000	[10]	[10]

*** Github webpage:** https://github.com/PouyaGhahramanian/Text-Categorization-Datasets/blob/master/Datasets.md

References

[1] Greene, Derek, and Pádraig Cunningham. "Practical solutions to the problem of diagonal dominance in kernel document clustering." Proceedings of the 23rd international conference on Machine learning. ACM, 2006.

[2] Lang, Ken. "Newsweeder: Learning to filter netnews." Machine Learning Proceedings 1995. Morgan Kaufmann, 1995. 331-339.

[3] http://www.daviddlewis.com/resources/testcollections/rcv1/

[4] Lehmann, Jens, Isele, Robert, Jakob, Max, Jentzsch, Anja, Kontokostas, Dimitris, Mendes, Pablo N., Hellmann, Sebastian, Morsey, Mohamed, van Kleef, Patrick, Auer, Soren, and Bizer, Christian. DBpedia - a large-scale, multilingual knowledge base extracted from wikipedia. Semantic Web Journal, 2014.

[5] Used as a text classification dataset in:
Xiang Zhang, Junbo Zhao, Yann LeCun. Character-level Convolutional Networks for Text Classification. Advances in Neural Information Processing Systems 28 (NIPS 2015).

[6] Used as a text classification dataset in:
Xiang Zhang, Junbo Zhao, Yann LeCun. Character-level Convolutional Networks for Text Classification. Advances in Neural Information Processing Systems 28 (NIPS 2015).

[7] A. Zubiaga, D. Spina, V. Fresno, R. Martínez. Real-Time Classification of Twitter Trends, Journal of the American Society for Information Science and Technology (JASIST). In Press.

[8] -

[9] -

[10] -

Links

[1] http://mlg.ucd.ie/datasets/bbc.html

[2] http://qwone.com/~jason/20Newsgroups/ || https://scikit-learn.org/0.19/datasets/twenty_newsgroups.html

[3] http://kdd.ics.uci.edu/databases/reuters21578/reuters21578.html || Available in the nltk.corpus package

[4] https://drive.google.com/drive/folders/0Bz8a_Dbh9Qhbfll6bVpmNUtUcFdjYmF2SEpmZUZUcVNiMUw1TWN6RDV3a0JHT3kxLVhVR2M

[5] https://drive.google.com/drive/folders/0Bz8a_Dbh9Qhbfll6bVpmNUtUcFdjYmF2SEpmZUZUcVNiMUw1TWN6RDV3a0JHT3kxLVhVR2M || https://github.com/LC-John/Yahoo-Answers-Topic-Classification-Dataset

[6] https://github.com/mhjabreel/CharCnn_Keras/tree/master/data/ag_news_csv || http://www.di.unipi.it/~gulli/AG_corpus_of_news_articles.html || http://nlpprogress.com/english/text_classification.html

[7] https://nlp.uned.es/~damiano/datasets/TT-classification.html

[8] https://users.dcc.uchile.cl/~mquezada/breakingnews/

[9] http://nlpprogress.com/english/text_classification.html || https://github.com/PratikBarhate/question-classification/tree/master/dataset

[10] http://nlpprogress.com/english/text_classification.html || https://github.com/AcademiaSinicaNLPLab/sentiment_dataset/tree/master/data