Dataset Appendix – Similarity Evidence Representation in Synthetic Documents forMeta-Feature Generation in Text Classification

May 2019

Table 1 provides basic information about the topic datasets, reporting the number of features, documents, and the number of classes.

Dataset	**Domain**	**#Documents**	**#Classes**	**Density**
20NewsGroup (20NG)	News	18805	20	130
WebKB (4UNI)	Webpages	8277	7	140
ACM-DL (ACM)	Digital Library	24897	11	29
Reuters90 (REUT)	News	13327	90	78
Medline (MED)	Digital Library	861454	7	31

Table 1: Topic Dataset Description

For sentiment categorization, we use eighteen recent and publicly available real-world textual datasets gathered from different works for the categorization of the text sentiment as one of the three categories: positive, negative or neutral. These data provide a heterogeneous benchmark with varying length of documents, number of documents and unbalancement. They are named aisopos_ntua [2], debate [1], en_dailabor [5], nikolaos_ted [7], pang_movie [6], sanders_tw[1], ss_bbc [8], ss_digg [8], ss_myspace [8], ss_rev [8], ss_twitter [8], ss_youtube [8], stanford_tw [3], semeval_tw[2], vader_amzn [4], vader_movie [4], vader_nyt [4], and yelp_rev[3]

[1] http://www.sananalytics.com/lab/twitter-sentiment

[2] https://www.cs.york.ac.uk/semeval-2013/task2

[3] http://www.yelp.com/dataset_challenge

References

[1] Nicholas A Diakopoulos and David A Shamma. Characterizing debate performance via aggregated twitter sentiment. In *SIGCHI'10*, pages 1195–1198. ACM, 2010.

[2] Fotis Aisopos. Manually annotated sentiment analysis twitter dataset ntua., 2014.

[3] Alec Go, Richa Bhayani, and Lei Huang. Twitter sentiment classification using distant supervision. Technical report, Stanford, December 2009.

[4] Clayton J. Hutto and Eric Gilbert. VADER: A parsimonious rule-based model for sentiment analysis of social media text. In *ICWSM'14*, 2014.

[5] Sascha Narr, Michael Hulfenhaus, and Sahin Albayrak. Language-independent twitter sentiment analysis. *Knowledge Discovery and Machine Learning (KDML)*, pages 12–14, 2012.

[6] Bo Pang and Lillian Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *ACL'04*, Stroudsburg, PA, USA, 2004.

[7] Nikolaos Pappas and Andrei Popescu-Belis. Sentiment analysis of user comments for one-class collaborative filtering over ted talks. In *SIGIR'13*, pages 773–776. ACM, 2013.

[8] Mike Thelwall. Heart and soul: Sentiment strength detection in the social web with sentistrength, 2013.