Codebook for the Online News Sharing @ Mashable Dataset

Source:
K. Fernandes, P. Vinagre and P. Cortez. A Proactive Intelligent Decision Support System for Predicting the Popularity of Online News. Proceedings of the 17th EPIA 2015 - Portuguese Conference on Artificial Intelligence, September, Coimbra, Portugal. [***minorly adapted for pedagogical reasons***]

Data Set Information:
The dataset contains information of articles published by Mashable (www.mashable.com) and their content as the rights to reproduce it belongs to them. This dataset does not share the original content of the covered articles, but some statistics associated with it. The original content can be publicly accessed and retrieved using the provided urls.

The key outcome variable is "shares", defined at the bottom of the list below.

Variables 21-36 correspond to different statistics used in natural language processing. Broadly, these describe the "sentiment" of each article's text.

Attribute Information:
1. url: URL of the article
2. timedelta: Days between the article publication and the dataset acquisition
3. n_tokens_title: Number of words in the title
4. n_tokens_content: Number of words in the content
5. n_unique_tokens: Rate of unique words in the content
6. n_non_stop_unique_tokens: Rate of unique non-stop words in the content
7. num_hrefs: Number of links
8. num_self_hrefs: Number of links to other articles published by Mashable
9. num_imgs: Number of images
10. num_videos: Number of videos
11. average_token_length: Average length of the words in the content
12. num_keywords: Number of keywords in the metadata
13. category: lifestyle, entertainment, etc. (categorical)
14. weekday: day-of-week of publication (categorical)
15. kwshares_worst: avg shares of the worst-performing included keyword
16. kwshares_avg: avg shares of the average-performing included keyword
17. kwshares_best: avg shares of the best-performing included keyword
18. self_reference_min_shares: Min. shares of referenced articles in Mashable
19. self_reference_max_shares: Max. shares of referenced articles in Mashable
20. self_reference_avg_sharess: Avg. shares of referenced articles in Mashable
21. global_subjectivity: Text subjectivity
22. global_sentiment_polarity: Text sentiment polarity
23. global_rate_positive_words: Rate of positive words in the content

24. global_rate_negative_words: Rate of negative words in the content
25. rate_positive_words: Rate of positive words among non-neutral tokens
26. rate_negative_words: Rate of negative words among non-neutral tokens
27. avg_positive_polarity: Avg. polarity of positive words
28. min_positive_polarity: Min. polarity of positive words
29. max_positive_polarity: Max. polarity of positive words
30. avg_negative_polarity: Avg. polarity of negative words
31. min_negative_polarity: Min. polarity of negative words
32. max_negative_polarity: Max. polarity of negative words
33. title_subjectivity: Title subjectivity
34. title_sentiment_polarity: Title polarity
35. abs_title_subjectivity: Absolute subjectivity level
36. abs_title_sentiment_polarity: Absolute polarity level
37. shares: Number of shares