Data statement for the LM-SA-2020 Sentiment Word List

Dataset name: LM-SA-2020
Citations: See https://doi.org/10.25403/UPresearchdata.14401178
Dataset developer(s): Michelle Terblanche (michelle.terblanche@gmail.com), Vukosi Marivate (vukosi.marivate@cs.up.ac.za)
Data statement author(s): Michelle Terblanche, Vukosi Marivate
Organisation: Data Science for Social Impact Research Group https://dsfsi.github.io),
Department of Computer Science, University of Pretoria, South Africa

A. CURATION RATIONALE

The *Loughran and McDonald Sentiment Word Lists* were developed using corporate 10-K reports between 1994 and 2008 [1]. These reports are relevant to companies in the United States of America and required by the U.S. Securities and Exchange Commission (SEC)[1].

The motivation for building the ***LM-SA-2020*** word list was based on an experiment using the above-mentioned original lists to detect sentiment-carrying words in South African financial article headlines. A corpus of 808 financial articles (relating to ***Sasol***[2]) were used and only 37% of headlines had words of which the sentiment matched that of the words in the *Loughran and McDonald Sentiment Word Lists* correctly according to ground truth labels. A gap was therefore identified in developing a method for predicting sentiment of financial articles in a South African context.

Due to the size of data set, it was possible to manually examine the headlines to identify sentiment-carrying words to be included in the original word lists. Furthermore, synonyms were added for the existing words in the *Loughran and McDonald Sentiment Word Lists* using *NLTK's WordNet*[3] interface. The sentiment detection/prediction accuracy improved by 29% using the new word list.

This sentiment word list can be further expanded/improved in future by increasing the size of the data set and/or including data from other companies. It highlights the need for not only domain-specific sentiment prediction tools but also region-specific corpora.

B. LANGUAGE VARIETY

The language of this data set is American English. The original lists were developed using reports from American companies and the expansion was done

[1] https://www.investopedia.com/financial-term-dictionary-4769738
[2] www.sasol.com/
[3] https://www.nltk.org/howto/wordnet.html

by adding synonyms using *NLTK's WordNet* interface which was developed by Princeton and hence also American English. The words that were manually added to the list are also considered American English.

C. SPEAKER DEMOGRAPHIC

No specific considerations were made regrading speaker demographic.

The original documents used to develop the sentiment word lists were corporate 10-K reports published by American companies. The specific speaker/author demographic was therefore not available but assumed that the authors are well-versed in the English language.

The financial documents used to expand the sentiment word list were predominantly from South African online news publishers, with 4% of the documents from *Stock Exchange News Reports* published by the ***Johannesburg Stock Exchange***[4]. Again, specific speaker/author demographics are therefore not known but since these articles are forms of formal communication, it is assumed that the authors are fluent in English.

D. ANNOTATOR DEMOGRAPHIC

The original list of words were developed and their sentiment evaluated and annotated by the authors/data set developers, Tim Loughran and Bill McDonald, who are with the University of Notre Dame [1]. No specific information is available regarding their demographics, however from some research, they are white, American males over 50.

Four annotators were used to label the financial articles based on headlines which were then used to expand the word lists. Herewith the demographic information of the annotators:

Table 1. Annotator demographic

	1	**2**	**3**	**4**
Description	Financial understanding	Financial understanding	Financial understanding	Financial understanding
Age	65-70	50-55	45-50	35-40
Gender	Female	Male	Female	Female
Race/ethnicity	Caucasian	Caucasian	Caucasian	Caucasian
First Language(s)	Afrikaans	English	Afrikaans	Afrikaans English
Linguistics training	No	No	No	No

[4] https://www.jse.co.za/services/market-data/market-announcements

E. SPEECH SITUATION

The original documents used to create the *Loughran and McDonald Sentiment Word Lists* were collected between 1994 and 2008. These are official, corporate reports from American companies and available in written format. Since publishing these is a requirement by the SEC, the documents are formal (i.e. edited/scripted). The intended audience are potential investors.

The financial articles used to develop the ***LM-SA-2020*** were published by South African online news platforms from the period mid 2015 - mid 2020. The news articles are in written format and edited before publishing. The intended audience is the general public.

F. TEXT CHARACTERISTICS

The documents/articles used in generating the original word lists as well as for expanding these to develop the ***LM-SA-2020*** list, are formal communications relating to the financial domain. In the South African context, the news articles used are from official online news platforms. These, however, can be subjected to publisher bias.

G. RECORDING QUALITY

N/A

H. OTHER

The ***LM-SA-2020*** sentiment word list can be made available on request.

I. PROVENANCE APPENDIX

The original word lists, the *Loughran and McDonald Sentiment Word Lists*, were used to developed the ***LM-SA-2020*** data set. The original word lists are publicly available[5] [1].
The data statement for these word lists are available at: https://www3.nd.edu/mcdonald/Word_Lists_files/Documentation/Documentation_LoughranMcDonald_MasterDictionary.pdf

References

1. Loughran, T., McDonald, B.: When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. The Journal of Finance **66**(1), 35–65 (2011)

[5] https://sraf.nd.edu/textual-analysis/resources/