

Documentation for the Loughran-McDonald Master Dictionary

- **File:** LoughranMcDonald_MasterDictionary_YYYY.xlsx

The Loughran-McDonald Master Dictionary¹

The Loughran-McDonald Master Dictionary was initially developed in conjunction with our paper published in *Journal of Finance* ([“When is a Liability not a Liability?, 2011](#)). The dictionary provides a means of classifying tokens (collections of characters) as actual words, which is important for consistency in word counts, and identifies our sentiment word lists (see below).

Master Dictionary Core Word List

As an artifact of hackers needing word lists to crack passwords, a variety of word lists are available on the internet. Word lists including proper nouns and abbreviations can exceed 600,000 tokens (a token is a collection of characters).

The Master Dictionary we use is based on release 4.0 of the *2of12inf* dictionary documented at: <http://wordlist.sourceforge.net/12dicts-readme.html>. The *2of12inf* dictionary includes word inflections but does not include abbreviations, acronyms, or names. We use inflections instead of stemming because, in our opinion, especially if the focus is on tone, using explicit inflections is less error prone than extending a word using stemming (root morpheme + derivational morphemes). The *2of12inf* word list contains more than 80,000 words.

The one-letter words “A” and “I” are not included in our Master Dictionary for two reasons. First, they are not critical content words, and second they are more likely to indicate headers in financial documents.² All tokens we identify as words contain two or more characters.

Extending the Core Word List

To create the initial Master Dictionary, the core *2of12inf* word list is extended using EDGAR 10-X filings as the basis for business language.³ All tokens—i.e., collections of alphabetic characters—in all 10-X filings are identified that did not appear in the *2of12inf* word list. This collection of orphan tokens was then sorted by frequency of occurrence and each token with a frequency count of 50 or more, or a token that was an inflection of a more common word, was evaluated for inclusion in the Master Dictionary. The only proper noun we have added to the list is “Scholes”, given the importance and frequency of the

¹ In the natural language processing literature the term “word list” is not synonymous with “dictionary”. A word list is a one-dimensional list of words, whereas a dictionary is a word that is associated with additional data, such as a definition or count. For our purposes we will use the terms synonymously.

² Clearly if you are conducting a study that focuses on something such as CEO self-identification the word “I” might become relevant and should, in this case, be appended to our list.

³ We use the term 10-X to reference forms 10-K, 10-K/A, 10-K405, 10-K405/A, 10KSB, 10KSB/A, 10-KSB, 10-KSB/A, 10KSB40, 10KSB40/A.

term Black-Scholes. Beginning with the 2018 version (released in March, 2019), we also include words not in the existing dictionary that appear in earnings calls. Although the list does not generally include abbreviations, in 2020 we added 20 commonly appearing abbreviations to the list.

Sentiment Word Lists

Within the dictionary spreadsheet we also provide flags for the sentiment dictionaries used in the [JF paper](#) and ones developed in subsequent papers ([here](#) and [here](#)). The sentiment categories are *negative*, *positive*, *uncertainty*, *litigious*, *strong modal*, *weak modal*, *constraining*, and *complexity*. Membership in a given classification is flagged in the corresponding column by the year in which the word was added to that sentiment group. A "negative" year indicates the year when a word was removed from the group.

Updating the Master Dictionary

Each year the list is updated using the same process of identifying orphan tokens for the most recent year. All tokens with a frequency count of 50 or more and that are identifiable as words are added to the dictionary.

In earlier versions of the dictionary we did not include additions that were highly industry specific, but starting with the 2012 version we have removed this restriction. Most of these additions are noncommercial pharmacological or chemical terms (e.g., vancomycin or blinatumomab). Some of the additions simply represent the dynamic nature of language (e.g., fracking, bitcoin, covid).

Word and Document Counts

The word counts and document counts included in the Master Dictionary are based on parsing all 10-X documents from 1994 to the current year/version. "Word count" is the simple tabulation of occurrence for the word across all documents/years. "Document Count" indicates the number of 10-X filings containing at least one occurrence of the word.

Additional Data in the Master Dictionary

In addition to the "Word Count" and "Document Count" associated with each word, the file contains the following items:

1. A sequence number. For the Document Dictionaries it is important to have each word associated with a sequence number.
2. The average proportion and standard deviation of the proportion for the occurrence of each word across all documents.
3. Sentiment flags – each sentiment category (previously discussed) has a column. If a word is included in the sentiment category, the year/version when it was included is indicated. Note that if the year is preceded by a negative sign, this flag indicates

- the year when the word was *removed* from the sentiment category.
4. Syllables – the syllable count for each word. About 15,000 of the words were manually identified while the rest were categorized using a syllabification algorithm.
 5. Source – the original source of the words. Most are 12of12Inf, as discussed above, and additions based on 10-X usage updates are identified as “10K_YYYY”. An “A” postscript indicates the token is an abbreviation.

Additions and changes in categorical classifications:

- 2008
 - *Additions:* 871 words
- 2009
 - *Additions:* 14 words
- 2010
 - *Additions:* 1,898 words
- 2011
 - *Additions:* 11 words
 - *Reclassifications*
 - Negative added – {AVERSELY, DELISTS, MISCLASSIFICATION, MISCLASSIFIED, MISDATED, UNDERPERFORM, UNDERPERFORMED, UNDERREPORTING, UNFAVOURABLE, UNFORSEEN, UNMERCHANTABLE, UNPREDICTED, UNPROFITABILITY}
 - Positive added – {INNOVATIVENESS}
 - Uncertainty added – {UNFORECASTED, UNFORSEEN, UNPREDICTED, UNQUANTIFIABLE, UNQUANTIFIED, UNRECONCILED}
 - Litigious added – 140 words were added. See spreadsheet.
 - Constraining added – 28 words were added. See spreadsheet.
- 2012
 - *Additions:* 339 words (added industry specific words-e.g., pharma terms)
 - *Reclassifications*
 - Negative removed – {CASUALTIES, CASUALTY, CONSTRUE, CONSTRUED, CONSTRUES, CONSTRUCTING, DEEPENED, DEEPENING, DEEPENS, DEEPER, DEEPEST, FOREGOING, INAPPLICABLE, REFINANCE, REFINANCED, REFINANCES, REFINANCING, REFINANCINGS, SURRENDER, SURRENDERED, SURRENDERING, SURRENDERS}
 - Positive removed – {OUTSTANDING}
 - Positive added – {BEST}

- Uncertainty added – {VAGUE, VAGUELY, VAGUENESS, VAGUENESSES, VAGUER, VAGUEST}
- Litigious added – {CONSTRUE, CONSTRUED, CONSTRUES, CONSTRUING}
- 2014
 - *Additions:* 462 words
 - *Reclassifications*
 - Negative – added 26 {CYBERATTACK, CYBERATTACKS, CYBERBULLYING, CYBERCRIME, CYBERCRIMES, CYBERCRIMINAL, CYBERCRIMINALS, MISCHARACTERIZATION, MISCLASSIFICATIONS, MISCLASSIFY, MISCOMMUNICATION, MISPRICE, MISPRICING, MISPRICINGS, REDEFAULT, REDEFAULTS, SPAM, SPAMMERS, SPAMMING, UNDERPERFORMS, UNFAVORABILITY, UNMERITORIOUS, UNRECEPTIVE, UNSELLABLE, UNSTABILIZED, UNTRUSTED}
 - Litigious – added 17 {ANTICORRUPTION, BENEFICIATED, CLAIMHOLDER, CLAWBACKS, CONTESTABILITY, COUNTERSUITS, CRIMINALIZE, CRIMINALIZING, DEFEASEMENT, DEFEASES, DEFENDABLE, EXTRAJUDICIAL, HEREFOR, HEREWITHIN, PATENTEE, RECUSES, UNENCUMBER}
- 2016
 - Additions: 85 words
- 2018
 - Additions: 1,265 words (including ones taken from earnings calls).
- 2020
 - Additional:
 - Words – 25.
 - Abbreviations – 20.
 - Removal of select words from sentiment lists based on subsequent contextual analysis. Importantly, in the 2020 Loughran-McDonald Master Dictionary, we drop the following tokens from the negative word list: BREAKING, BRIDGE, BROKEN, CLOSED, CLOSING, CRITICAL, FORCE, FUGITIVE, LATE, SECRECY. For the positive word list in the 2020 version, the following words are dropped, BENEFICIAL, BENEFIT, EFFECTIVE, GREAT, GREATER, HONORABLE, REWARDS.

Updates and Acknowledgments

Language is dynamic. New words enter the business vocabulary and context changes over time. Our classifications are based on an assessment of most likely usage. Thus we will attempt to update the master dictionary and word classifications on a periodic basis. User suggestions are welcomed – please e-mail: mcdonald.1@nd.edu.

The LoughranMcDonald_MasterDictionary has benefited substantially from continuing feedback and those involved in reviewing our initial paper utilizing word lists (JF, 2011). We especially appreciate the contributions made by Cam Harvey in editing the first versions of the lists.

Updated: 202107