#

|  |  |  |
| --- | --- | --- |
| **Pruning Method** | **Rationale** | **Criteria** |
| Remove Stopwords | Stopwords are common words in a language that are frequently filtered out during text processing as they typically do not add high meaning to sentences, such as prepositions. There is not a single, universally accepted stopword list; the stopword list used in this work was RapidMiner’s English stopword list. | Words from the following list:"a", "abaft", "aboard", "about", "above", "across", "afore", "aforesaid", "after", "again", "against", "agin", "ago", "aint", "albeit", "all", "almost", "alone", "along", "alongside", "already", "also", "although", "always", "am", "american", "amid", "amidst", "among", "amongst", "an", "and", "anent", "another", "any", "anybody", "anyone", "anything", "are", "aren", "around", "as", "aslant", "astride", "at", "athwart", "away", "back", "bar", "barring", "be", "because", "been", "before", "behind", "being", "below", "beneath", "beside", "besides", "best", "better", "between", "betwixt", "beyond", "both", "but", "by", "can", "cannot", "certain", "circa", "close", "concerning", "considering", "cos", "could", "couldn", "couldst", "dare", "dared", "daren", "dares", "daring", "despite", "did", "didn", "different", "directly", "do", "does", "doesn", "doing", "done", "don", "dost", "doth", "down", "during", "durst", "each", "early", "either", "em", "english", "enough", "ere", "even", "ever", "every", "everybody", "everyone", "everything", "except", "excepting", "failing", "far", "few", "first", "five", "following", "for", "four", "from", "gonna", "gotta", "had", "hadn", "hard", "has", "hasn", "hast", "hath", "have", "haven", "having", "he", "her", "here", "hers", "herself", "high", "him", "himself", "his", "home", "how", "howbeit", "however", "id", "if", "ill", "immediately", "important", "in", "inside", "instantly", "into", "is", "isn", "it", "its", "itself", "ve", "just", "large", "last", "later", "least", "left", "less", "lest", "let", "like", "likewise", "little", "living", "long", "many", "may", "mayn", "me", "mid", "midst", "might", "mightn", "mine", "minus", "more", "most", "much", "must", "mustn", "my", "myself", "near", "neath", "need", "needed", "needing", "needn", "needs", "neither", "never", "nevertheless", "new", "next", "nigh", "nigher", "nighest", "nisi", "no", "one", "nobody", "none", "nor", "not", "nothing", "notwithstanding", "now", "er", "of", "off", "often", "on", "once", "oneself", "only", "onto", "open", "or", "other", "otherwise", "ought", "oughtn", "our", "ours", "ourselves", "out", "outside", "over", "own", "past", "pending", "per", "perhaps", "plus", "possible", "present", "probably", "provided", "providing", "public", "qua", "quite", "rather", "re", "real", "really", "respecting", "right", "round", "same", "sans", "save", "saving", "second", "several", "shall", "shalt", "shan", "she", "shed", "shell", "short", "should", "shouldn", "since", "six", "small", "so", "some", "somebody", "someone", "something", "sometimes", "soon", "special", "still", "such", "summat", "supposing", "sure", "than", "that", "the", "thee", "their", "theirs", "them", "themselves", "then", "there", "these", "they", "thine", "this", "tho", "those", "thou", "though", "three", "thro", "through", "throughout", "thru", "thyself", "till", "to", "today", "together", "too", "touching", "toward", "towards", "true", "twas", "tween", "twere", "twill", "twixt", "two", "twould", "under", "underneath", "unless", "unlike", "until", "unto", "up", "upon", "us", "used", "usually", "versus", "very", "via", "vice", "vis-a-vis", "wanna", "wanting", "was", "wasn", "way", "we", "well", "were", "weren", "wert", "what", "whatever", "when", "whencesoever", "whenever", "whereas", "where", "whether", "which", "whichever", "whichsoever", "while", "whilst", "who", "whoever", "whole", "whom", "whore", "whose", "whoso", "whosoever", "will", "with", "within", "without", "wont", "would", "wouldn", "wouldst", "ye", "yet", "you", "your", "yours", "yourself", "yourselves" |
| Word Length | Words that were extremely short or extremely long were most likely to be too uncommon to be meaningful in text processing, and/or proper nouns which also would not be meaningful in text processing. | Words which were less than 3 characters or more than 25 characters. |
| Frequency | Words which appear extremely frequently would not provide meaningful term frequency-inverse document frequency (TF-IDF) scores due to scores being offset by how frequent words appear across the entire dataset. Rare words would not be encountered frequently enough to consistently influence predictions. | Words which appeared in greater than 99% of all comments or less than 2% of all comments. |

#

# **Table, Supplementary Digital Content 3**. Description of types of words removed during text processing.