

MAP OF FEATURE INDEXES

Features 100 + BERT 768 = 868

Informality (0 to ~~33~~ 32) [~~34~~] [33]

Readability (0 to 8) [9]:

- 0 Kincaid
- 1 ARI
- 2 Coleman-Liau
- 3 FleschReadingEase
- 4 GunningFogIndex
- 5 LIX
- 6 SMOGIndex
- 7 RIX
- 8 DaleChallIndex

Complexity (9 to ~~33~~ 32) [~~25~~ 24]:

(Using [PySemCom Library](#))

9 filename	436.0	(removed)
9 nbWord	1.0	
10 nbConcepts	0.0	
11 nbUniqueConcepts	0.0	
12 conceptsWordsRatio	0.0	
13 uniqueConceptsWordsRatio	0.0	
14 nbNodes	0.0	
15 radius	0.0	
16 diameter	0.0	
17 assortativity	0.0	
18 density	1.0	
19 textDensityMean	0.0	
20 textDensityStd	NaN	
21 nbTypesMean	NaN	
22 nbTypesStd	NaN	
23 nbLinkInMean	NaN	
24 nbLinkInStd	NaN	
25 nbLinkOutMean	NaN	
26 nbLinkOutStd	0.0	
27 nbNodes_DBPedia	1.0	
28 density_DBPedia	0.0	
29 nbNodes_Schema	1.0	
30 density_Schema	0.0	
31 nbNodes_yago	1.0	

32 density_yago

NaN

Note: Same lib that does readability has some syntactic scores

Diversity (33 to 42) [10]

(Using [Lexical-Diversity library](#))

33 Simple TTR

34 Root TTR

35 Log TTR

36 Mass TTR

37 Mean segmental TTR (MSTTR)

38 Moving average TTR (MATTR)

39 Hypergeometric distribution D (HDD)

40 Measure of lexical textual diversity (MTLD)

41 Measure of lexical textual diversity (moving average, wrap)

42 Measure of lexical textual diversity (moving average, bi-directional)

Quantity (43 to 89) [47]

43 #terms

44 #tokens

45 '#'

46 '\$'

47 '"""

48 '('

49 '"

50 "

51 '.'

52 ':'

53 'CC'

54 'CD'

55 'DT'

56 'EX'

57 'FW'

58 'IN'

59 'JJ'

60 'JJR'

61 'JJS'

62 'LS'

63 'MD'

64 'NN'

65 'NNP'

66 'NNPS'

67 'NNS'

68 'PDT'

69 'POS'

70 'PRP'
71 'PRP\$'
72 'RB'
73 'RBR'
74 'RBS'
75 'RP'
76 'SYM'
77 'TO'
78 'UH'
79 'VB'
80 'VBD'
81 'VBG'
82 'VBN'
83 'VBP'
84 'VBZ'
85 'WDT'
86 'WP'
87 'WP\$'
88 'WRB'
89 ''''

Affect (90 to 95) [6]

90 AVG TextBlob's Polarity
91 Positive TextBlob's Polarity
92 Negative TextBlob's Polarity
93 AVG VADER's Polarity
94 Positive VADER's Polarity
95 Negative VADER's Polarity

~~Note: ABL1 showed these two as negative features. Adjusting these features could yield additional performance to the model. [DONE]~~

Subjectivity (96) [1] (MPQA method also available)

96 Sum TextBlob's Subjectivity score over sentences

Specificity (97) [1]

97 [Speciteller](#) scores

Note: the best non-trivial feature used, contributing with more than 1% acc by itself

Pausality (98) [1]

98 Count of '.'-tag tokens

Uncertainty (99) [1]

(Using [Uncertainty](#) library)

99 LUCI: Linguistic Uncertainty Classifier Interface

Passiveness (100) [1]

100 Using Brown Tagger for verbs in gerund form identification preceded to-be. This count is normalized over the number of sentences.