**A Linguistic Feature Definitions**

For the following definitions, if the denominator of the ratio is zero the result is treated as zero.

|  |  |
| --- | --- |
| **Feature Name** | **Definition** |
| WRich10\_S | Richness, 100 topics extracted from Wikipedia Dump |
| WClar10\_S | Clarity, 100 topics extracted from Wikipedia Dump |
| WNois10\_S | Noise, 100 topics extracted from Wikipedia Dump |
| Wtopc10\_c | # of topics, 100 topics extracted from Wikipedia Dump |
| WRich20\_S | Richness, 200 topics extracted from Wikipedia Dump |
| WClar20\_S | Clarity, 200 topics extracted from Wikipedia Dump |
| WNois20\_S | Noise, 200 topics extracted from Wikipedia Dump |
| WTopc20\_c | # of topics, 200 topics extracted from Wikipedia Dump |
| BRich10\_S | Richness, 100 topics extracted from WeeBit Corpus |
| BClar10\_S | Clarity, 100 topics extracted from WeeBit Corpus |
| BNois10\_S | Noise, 100 topics extracted from WeeBit Corpus |
| BTopc10\_c | # of topics, 100 topics extracted from OneStop Corpus |
| BRich20\_S | Richness, 200 topics extracted from WeeBit Corpus |
| BClar20\_S | Clarity, 200 topics extracted from WeeBit Corpus |
| BNois20\_S | Noise, 200 topics extracted from WeeBit Corpus |
| BTopc20\_c | # of topics, 200 topics extracted from OneStop Corpus |
| AoA Kuperman | Mean age of acquisition of words (Kuperman database) |
| AoA Kuperman lemmas | Mean age of acquisition of lemmas |
| AoA Bird lemmas | Mean age of acquisition of lemmas, Bird norm |
| AoA Bristol lemmas | Mean age of acquisition of lemmas, Bristol norm |
| AoA Cortese and Khanna lemmas | Mean age of acquisition of lemmas, Cortese and Khanna norm |
| MRC familiarity | Mean word familiarity rating |
| MRC concreteness | Mean word concreteness rating |
| MRC Imageability | Mean word imageability rating |
| MRC Colorado Meaningfulness | mean word Colorado norms meaningfulness rating |
| MRC Pavio Meaningfulness | mean word Pavio norms meaningfulness rating |
| MRC AoA | Mean age of acquisition of words (MRC database) |

Table 1: Existing Semantic Features

|  |  |
| --- | --- |
| **Feature Name** | **Definition** |
| at\_UEnti\_C | average number of unique Entities per token (word) |
| LoCohPA\_S | Local Coherence for PA score from Entity Grid |
| LoCohPW\_S | Local Coherence for PW score from Entity Grid |
| LoCohPU\_S | Local Coherence for PU score from Entity Grid |
| LoCoDPA\_S | Local Coherence dist. for PA score from Entity Grid |
| LoCoDPW\_S | Local Coherence dist. for PW score from Entity Grid |
| LoCoDPU\_S | Local Coherence dist. for PU score from Entity Grid |
| ra\_SSToT\_C | ratio of SS transitions: total, count from Entity Grid |
| ra\_SOToT\_C | ratio of SO transitions: total, count from Entity Grid |
| ra\_SXToT\_C | ratio of SX transitions: total, count from Entity Grid |
| ra\_SNToT\_C | ratio of SN transitions: total, count from Entity Grid |
| ra\_OSToT\_C | ratio of OS transitions: total, count from Entity Grid |
| ra\_OOToT\_C | ratio of OO transitions: total, count from Entity Grid |
| ra\_OXToT\_C | ratio of OS transitions: total, count from Entity Grid |
| ra\_ONToT\_C | ratio of ON transitions: total, count from Entity Grid |
| ra\_XSToT\_C | ratio of XS transitions: total, count from Entity Grid |
| ra\_XOToT\_C | ratio of XO transitions: total, count from Entity Grid |
| ra\_XXToT\_C | ratio of XX transitions: total, count from Entity Grid |
| ra\_XNToT\_C | ratio of XN transitions: total, count from Entity Grid |
| ra\_NSToT\_C | ratio of NS transitions: total, count from Entity Grid |
| to\_EntiM\_C | total number of Entities Mentions |
| as\_EntiM\_C | average number of Entities Mentions per sentence |
| at\_EntiM\_C | average number of Entities Mentions per token (word) |
| to\_UEnti\_C | total number of unique Entities |
| as\_UEnti\_C | average number of unique Entities per sentence |

Table 2: Existing Discourse-based features

|  |  |
| --- | --- |
| **Feature Name** | **Definition** |
| to\_TreeH\_C | total parsed Tree Height of all sentences |
| as\_TreeH\_C | average parsed Tree Height per sentence |
| at\_TreeH\_C | average parsed Tree Height per token |
| to\_FTree\_C | total length of Flattened parsed Trees |
| as\_FTree\_C | average length of Flattened parsed Trees per sentence |
| at\_FTree\_C | average length of Flattened parsed Trees per token |
| nouns per word | number of nouns / number of words |
| proper nouns per word | number of proper nouns / number of words |
| pronouns per word | number of pronouns / number of words |
| conjunctions per word | number of conjunctions / number of words |
| adjectives per word | number of adjectives / number of words |
| verbs per word | number of verbs / number of words |
| adverbs per word | number of adverbs / number of words |
| modal verbs per word | number of modal verbs / number of words |
| prepositions per word | number of prepositions / number of words |
| interjections per word | number of interjections / number of words |
| personal pronouns per word | number of personal pronouns / number of words |
| wh-pronouns per word | number of wh-pronouns / number of words |
| lexical words per word | number of lexical words / number of words |
| function words per word | number of function words / number of words |
| determiners per word | number of determiners / number of words |
| VBDs per word | number of past tense verbs / number of words |
| VBPs per word | number of non-3rd person singular present verbs / number of words |
| VBZs per word | number of 3rd person singular present verbs / number of words |
| adjective variation | number of adjectives / numbers of lexical words |

Table 3: Existing Syntactic features

|  |  |
| --- | --- |
| **Feature Name** | **Definition** |
| TokSenM\_S | total count of tokens \* total count of sentence |
| TokSenS\_S | sqrt(total count of tokens \* total count of sentence) |
| TokSenL\_S | log(total count of tokens)/log(total count of sent) |
| as\_Token\_C | average count of tokens per sentence |
| as\_Sylla\_C | average count of syllables per sentence |
| at\_Sylla\_C | average count of syllables per token |
| as\_Chara\_C | average count of characters per sentence |
| at\_Chara\_C | average count of characters per token |
| number of sentences | number of sentences |
| mean sentence length | number of words / number of sentences |
| number of characters | number of characters |
| number of syllables | number of syllables |
| SmogInd\_S | Smog Index |
| ColeLia\_S | Coleman Liau Readability Score |
| Gunning\_S | Gunning Fog Count Score (New, US Navy Report) |
| AutoRea\_S | Automated Readability Idx (New, US Navy Report) |
| FleschG\_S | Flesch Kincaid Grade Level (New, US Navy Report) |
| LinseaW\_S | Linear Write Formula Score |

Table 4: Existing Lexical Features