Feature Engineering

Die folgenden Features werden für jedes Blogeintrag individuell berechnet.

Feature	Berechnung	Intention
Text length	Anzahl der Zeichen im Text (inklusive Leerzeichen, Sonderzeichen etc.)	Indikator für Textumfang
Number URLs	Anzahl der Wörte die eines oder mehre der folgenden Worte enthalten „urlLink“, „http“, „www“	Indikator dafür, ob auf andere Quellen verlinkt wird.
Number mails	Anzahl der Emails im Text	Sind Ansprechpartner benannt worden?
Uppercase ratio	Anteil der Großbuchstaben an allen Zeichen	Quantifizierung der Zeichenwahl
Lowercase ratio	Anteil der Kleinbuchstaben an allen Zeichen	Quantifizierung der Zeichenwahl
Number ratio	Anteil der Zahlen an allen Zeichen	Quantifizierung der Zeichenwahl
Symbol ratio	Anteil der Sonderzeichen an allen Zeichen	Quantifizierung der Zeichenwahl
Average letters per word	Durchschnitt der Buchstaben pro Wort	Quantifizierung der Wortlänge
Variance of letters per word	Varianz der Wortlänge	Quantifizierung der Heterogenität der Wortlänge
Unique words ratio	Anzahl der verschiedenen verwendeten Worte im Text	Quantifizierung der Heterogenität der Worte
Average letters per sentence	Durchschnitt der Buchstaben pro Satz	Quantifizierung der Satzlänge
Variance of letters per sentence	Varianz der Buchstaben pro Satz	Quantifizierung der Heterogenität der Satzlänge
Average words per sentence	Durchschnitt der Wörter pro Satz	Quantifizierung der Worte pro Satz
Variance of words per sentence	Varianz der Wörter pro Satz	Quantifizierung der Heterogenität der Worte pro Satz
Maximal uppercase ratio per sentence	Anteil der Großbuchstaben in dem Satz mit dem höchsten Anteil an Großbuchstaben	Indikator dafür, ob Sätze mit hohem Anteil an Großbuchstaben vorhande sind.
Length of the maximal uppercase ratio sentence	Anzahl der Zeichen in dem obigen Feature	Indikator dafür wie zuverlässig das obige Feature ist.