Abstract

The complexity of a natural language itself is very challenging as the natural language is not free from ambiguity problem. It is almost impossible to identify that the given text is having sense or not. In today's scenario it becomes even much important to detect that input is given by human or a machine. A valid input with sense is needed everywhere from Social media platforms to Business Intelligence. This Classification algorithm aims to detect whether the given input text is valid, or randomly typed in a keyboard. It returns a percentage value where a lower one means valid text, and a higher value means random text. The approach is based on identifying that the amount of unique chars, amount of vowels of letters, the word/char ratio (in %) are in a usual range. Then it further calculates "deviation score" to compute the accuracy of given input.

References

1. Ms. Ranju Marwaha, Data Mining Techniques and Applications in Telecommunication
Industry, International Journal of advanced research in computer science and software engineering, Volume 4, Issue 9, September 2014

2. Jijy George, Sandhya N., Suja George, "Classification Problem In Text Mining" International Journal of Innovative Research in Advanced Engineering (IJIRAE) ISSN: 2349-2163 Volume 1 Issue 8 (September 2014)

3. Ah-Hwee Tan, "Text Mining: The state of the art and the challenges"


11. D.Y. Sakhare, Dr. Raj Kumar, "Syntactic and Sentence Feature Based Hybrid Approach for Text Summarization, I.J. Information Technology and Computer Science, 2014, 03, 38-46 Published Online February 2014 in MECS

12. Ian Tenney, "A general-purpose sentence-level nonsense etector", December 2014

**Index Terms**

Computer Science, Algorithms

**Keywords**

Data mining; text mining; text classification; sentence validation; pattern learning