Dissertation thesis review

The dissertation thesis titled Entropy Utilization in the Text Similarity, submitted by Ing. Michal Prilepok, describes author’s original research in the field of text analysis and processing with a focus on practical applications in the areas of Spam and plagiarism detection and EEG signal analysis. The author proposes, theoretically analyzes, and evaluates several algorithms, designed specifically for the aforementioned application domains. The unifying idea of the proposed approaches is the utilization of lossless data compression (i.e. estimated algorithmic complexity of data) to evaluate the similarity of text-like documents.

The thesis is logically structured and presents in 6 chapters an introduction to research topics, brief theoretical introduction to the concept of Entropy and algorithmic complexity, and then target application fields and proposed approaches. The algorithms are experimentally evaluated on well-known test data sets and compared to some baseline methods. The experimental comparison is sound and clearly demonstrates timeliness of the conducted research and usefulness of the developed concepts.

The thesis is well written and nicely typeset. Although the use of the English language in the manuscript is not flawless, the author commits only a bare minimum of grammatical mistakes and avoids most typos. The text reads well and clearly describes to interested readers the main points of the conducted research as well as appropriate technical details. The content is free of major technical errors and the presented information is correct. However, some minor comments can be still made: on p. 12, the author states that a distance is subject to several requirements; however these requirements define a metric; the concept of conditional Kolmogorov complexity, $K(x|y)$, on p. 15 is not formally defined; the last 2 sentences on p. 16 are almost identical to the first 2 sentences of par. 4 on the same page; table 2 on p. would be better if it contained also information on False Positive/False Negative (or a clear explanation that no FP’s were obtained); the meaning of figure 9 is not clear; on p. 61, 2 subsection titles called ‘summary’ after each other are not nice; etc. These remarks, however, are of rather a formal nature.

Experimental results, presented in the thesis, show that the proposed methods have real application potential for spam and plagiarism detection and EEG signal classification. Author’s publication portfolio demonstrates that the described concepts and other results of his research have been already published in the form of peer-reviewed research articles in 2 international journals with Impact Factor indexed by the Web of Science, 4 international journals indexed by Scopus, and 12 papers published in proceedings of international scientific conferences. Therefore, it is my pleasure to recommend the thesis for defence.

(doc. Ing. Pavel Krömer, Ph.D.
Hlučín, 30. 5. 2016)