Pengembangan Sistem Penghapusan Identitas berbasis Algoritma LSTM = Authorship Obfuscation System Development based on LSTM Algorithm

Hendrik Maulana, author

Abstrak

Stylometry merupakan teknik analisa terhadap kepengarangan menggunakan statistik. Melalui stylometry, identitas kepengarangan dari suatu dokumen dapat dianalisis dengan tingkat akurasi yang tinggi. Hal ini menyebabkan adanya ancaman terhadap privasi penulis. Namun terdapat salah satu jenis metode dari stylometry yaitu penghapusan identitas kepengarangan yang dapat memberikan perlindungan privasi bagi penulis. Penelitian ini menggunakan metode penghapusan identitas kepengarangan yang diterapkan pada korpus Federalist Paper. Federalist Paper merupakan korpus terkenal yang telah banyak diteliti terutama pada metode identifikasi kepengarangan karena di dalam korpus tersebut terdapat 12 artikel yang tidak diketahui identitas penulisnya, salah satu metode identifikasinya adalah menggunakan algoritma Support Vector Machine. Melalui algoritma tersebut didapatkan identitas penulis dari artikel yang tidak diketahui pengarangnya dengan tingkat akurasi sebesar 86%. Tantangan dari metode penghapusan identitas kepengarangan adalah harus mampu mengubah gaya penulisan dengan tetap mempertahankan makna. Long-Short Term Memory (LSTM) merupakan algoritma berbasis Deep Learning yang mampu melakukan prediksi kata secara baik. Melalui model yang dibentuk dari algoritma LSTM, artikel-artikel dalam Federalist Paper diubah gaya penulisannya. Hasilnya, 30% dari artikel yang diklasifikasi dapat diubah identitas kepengarangannya dari satu penulis menjadi penulis lainnya. Tingkat kemiripan dokumen hasil ubahan berkisar antara 40-57% menandakan perubahan makna yang tidak signifikan dari dokumen aslinya. Hasil tersebut menyimpulkan bahwa metode yang diajukan mampu melakukan penghapusan identitas kepengarangan dengan baik.

.....Stylometry is an authorship analysis technique using statistics. Through stylometry, authorship identity of a document can be analyzed with a high degree of accuracy. This causes a threat to the privacy of the author. But there is one type of method of stylometry, namely the elimination of authorship identity which can provide privacy protection for writers. This study uses the authorship method of eliminating the method applied to the Federalist Paper corpus. Federalist Paper is a well-known corpus that has been extensively studied especially in authorship identification methods because there are 12 disputed texts in the corpus, one of the identification

methods is using the Support Vector Machine algorithm. Through this algorithm the author's identity of disputed text is obtained with an accuracy of 86%. The challenge of the authorship identity elimination method is that it must be able to change the writing style while maintaining its meaning. Long-Short Term Memory (LSTM) is a Deep Learning based algorithm that is able to predict words well.

Through a model formed from the LSTM algorithm, the disputed articles in the Federalist Paper are changed in their writing style. As a result, 30% of classified articles can be changed from one author identity to another identity. The level of similarity of the changed documents ranges from 40-57%, which indicates a change in meaning that is not significant from the original document. These results conclude that the proposed method is able to perform authorship identity deletion properly.