

Authorship identification in text documents using BERT and POS features

Daniel Díaz-Ocampo, Wesam Al-Nabki, Andrés Vasco-Carofilis, Laura Fernández-Robles,
Eduardo Fidalgo, Enrique Alegre

Department of Electrical, Systems and Automation Engineering, Universidad de León, León, ES
{ddiao, wesam.alnabki, andres.vasco, l.fernandez, eduardo.fidalgo, enrique.alegre}@unileon.es

Digital communication makes it easy for anyone to write and publish texts all over the world. In some situations, this text is written with non-legitimate intentions, attempting to manipulate or mislead the reader.

Usually, authorship identification is approached as a text classification task by assigning a document to an author. However, this approach does not fit the real-world setting when the model receives a new document produced by a new author of unknown identity.

In this work, we reproduce a state-of-the-art solution proposed by Fabien et al. to analyze the impact of adding a new category called “new author”; that captures the identity of different authors never seen in the training set. Additionally, we evaluate the benefit of using part-of-speech (POS) as features in the task of authorship identification. The POS tag features used consist of sequences of POS tag n-grams extracted using TF-IDF.

We evaluate the algorithm on three datasets with different levels of text quality, one with high-quality (ENRON professional emails) and two with low quality (Amazon customer reviews and slang chat messages from ChatCoder). We observe a decline in the model performance after introducing the “new author” class. Moreover, we observe a relationship between the text quality and the part-of-speech feature. For high-quality text, the F1 score increases by 1% up to 95,30%, and with low-quality text, the performance drops by 3% to 91,30%. This result shows that the POS tag features are useful when the text has high- quality but when the text has low quality those features act as noise for the classifier. Also, the category “new author” decreases the performance in each of the datasets compared to the results obtained before adding this class.