Author Detection & Descriptive Analysis on Corpora of Local & International Political Leaders

Farhan Asghar (18030017) — Inam Ullah Taj (18030010)

25 February 2020

Introduction

In today's age of digital media, with the plethora of available textual data, it would be really helpful if there is an automated mechanism of identifying and verifying whether a certain piece of information or text has actually been said by the person it is being associated with. An example of such misuse of information is that of different quotations on digital platforms being wrongly associated with popular scientists and politicians.

Objective and Goals

We aim to devise a mechanism or an application that can correctly determine whether a given text is correctly associated with a given Political Leader. To solve this problem, we will use different techniques of Natural Language Processing (NLP) and Deep Learning (DL) in order to extract and identify characteristics from text corpora that are specific to a respective person (or politician). We will also perform Descriptive Analysis for each Political Leader's corpora in order to determine what mostly used phrases, words and even topics are, with respect to each Political Leader.

Areas of Work

1. Data Acquisition (for 3-4 Politicians) and Dataset Preparation
2. Descriptive Analysis using Language Modeling
3. Authorship Detection using state-of-art Deep Learning Models (like seq2seq)

References

[1] Grigori Sidorov, Francisco Velasquez, Efstathios Stamatatos, Alexander Gelbukh, and Liliana Chanona-Hernández. Syntactic n-grams as machine learning features for natural language processing. *Expert Systems with Applications*, 41(3):853–860, 2014.

[2] Efstathios Stamatatos. A survey of modern authorship attribution methods. *Journal of the American Society for information Science and Technology*, 60(3):538–556, 2009.

[3] Efstathios Stamatatos, Nikos Fakotakis, and George Kokkinakis. Automatic text categorization in terms of genre and author. *Computational linguistics*, 26(4):471–495, 2000.

[4] Wenpeng Yin, Katharina Kann, Mo Yu, and Hinrich Schütze. Comparative study of cnn and rnn for natural language processing. *arXiv preprint arXiv:1702.01923*, 2017.