Stylistic Author Attribution Methods for Identifying Parody Trump Tweets

Abstract

Style-sensitive methods for authorial attribution have been proposed and developed in recent years in order to assess the authorship of documents with content unrelated to the rest of an author's body of work. We aim to train a classifier with a style-sensitive and syntactically deep feature space to examine how vital stylistic analysis is to a similar task – parody detection. The distinct style of Donald Trump's tweets has spun off a litany of parody tweets and parody accounts: we will train the classifier on Trump tweets and use it to determine whether tweets by other users are parodying Trump. Using feature ablation, we will determine the contribution of stylistic features to successful parody identification.

Background

Authorial style, as divorced from content, is the heart of pastiche. An author with an immediately recognizable style generates immediately recognizable pastiches, or, if the intents of the pastiches are sarcastic, immediately recognizable parodies. Authorship attribution is a natural language processing task that can use both the style and the content of an author's (attributed) body of work to determine if an anonymous document was, in fact, written by the author [1]. An attempt at authorship attribution needs to examine non-content features in the case that the content of the training corpus is disjunct or near-disjunct from the content of the anonymous document [2]. In the case of parody, an author's style is emulated while content – settings, characters, even themes – can change.

An increasingly large body of authorial data is being produced currently by the sitting President of the United States, who, since the beginning of his presidential campaign, has used a distinct writing style on Twitter to lambast his opponents and garner support from his followers. Donald Trump's tweets have given rise to a literature of parody tweets and several parody accounts; people familiar with Trump's extremely idiosyncratic Twitter style can identify Trump parody tweets by their style. The altered content of the tweet, displayed alongside Trump's style and juxtaposed with his status, seems to form the punchline.

This project will attempt to define a feature space that includes both content-based features (distribution of word or sentence length, $n$-gram probability) [3] and style-based features (part-of-speech trigram probabilities, production rule frequencies and $pq$-gram probabilities for
syntax, clause depth) [4, 5] across a corpus of Trump’s tweets. The features of those tweets and a group of control tweets will be extracted and used to train a support vector machine (SVM) for use as a classifier. The frequency with which the classifier can successfully identify Trump parody tweets from non-parody tweets will be measured. Some feature ablation will help to identify which features are most significant to parodying the distinct Trump Twitter voice; a comparison of a classifier trained on style and content features with one trained content-agnostically should reveal whether a Trump parody is identifiable by style alone.

**Deliverables**

- Corpus of Trump and control tweets for training, annotated for part-of-speech and syntactic category, as well as a ‘test set’ formed of Trump parodies and non-parody tweets.
- Implementation and training of the SVM
- Code using the trained classifier for a style-and-content-sensitive Trump parody tweet detector, almost definitely in Python
- Final report with analysis of the performance of a content-sensitive classifier versus a style-only classifier, obtained by comparing results from models trained with isolated feature sets

**References**


