Unsupervised Learning for Image Compression

COSC522 (Machine Learning) Project 4

Joseph Teague and Nigel Tan
jteague6@vols.utk.edu and ntan1@vols.utk.edu

December 12, 2019

Abstract

The ability to identify the author of a written work online is important for a number of reasons. While posts are typically tied to a username, individuals can have more than one account on many services and some individuals or organizations have a social media team to handle online posting for them. Author identification can, among other things, assist with tracking down users who are attempting to evadea ban or the author of a post that violates the policies of a celebrity or organization. In this work, we present an attempt to utilize machine learning to identify the individual author of a post. Using a number of machine learning techniques, we show results that have good sensitivity but poor specificity and how, by combining these approaches, we can arrive at a reasonable model

1 Introduction

There is lots of publicly available work for sentiment analysis with

2 Approach

2.1 The Dataset

We obtained archives containing a number of Tweets from prolific celebrity Twitter users from Kaggle [1]. While this archive contains a large number of users, we selected six who we thought would provide a diversity of writing style for the purposes of this examination. Those six are Donald Trump, Hillary Clinton, Richard Dawkins, Neil DeGrasse Tyson, Astronaut Scott Kelly, and Kim Kardashian. The Tweets provided are purely text and have no features extracted, so the dity of identifying features fell to us. The features we selected are:

- Repeated punctuation (e.g., "!!!!!")
- Inclusion of images
- @s directed at other users or organizations
- Typing in all caps (e.g., "NO COLLUSION")
- Inclusion of links
- Use of hashtags (e.g. "#mancrushmonday")
- Quotes

We also, as a comparison, run some tests with highly user-specific features extracted. These typically consist of things that one person is more likely to say than the others. For example, Donald Trump is more likely to say "make America great again," while Scott Kelly is the only selected user who would discuss his time in space. These features target specific words and phrases and do not include any sort of natural language processing. The bulk of our tests do not include these highly-specific features, and the end goal is to develop a user-agnostic approach to author identification.

2.2 Techniques Used

Every technique used in class so far has been employed in this work. We use MPP cases 1, 2, and 3, k-Nearest-Neighbors, support vector machines (linear, poly, and sigmoid), and backpropagating neural networks were used for supervised learning methods. For unsupervised methods, k-Means, winner-take-all, and Kohonen maps were employed. We also employed decision trees, which were not used in class.

Each learning method was run multiple times (with a parameter sweep where appropriate) with m-fold cross validation to examin its effectiveness. In addition to total accuracy, we paid close attention to sensitivity and specificity. Computing performance was not measured - in this case, we are more interested in overall accuracy than anything else.

2.3 Implementation

3 Experiments

4 Results

5 Discussion

6 Appendix

References

[1] J. Littiebrant, "Raw twitter timelines w/ no retweets." https://www.kaggle.com/speckledpingu/RawTwitterFeeds. Accessed:2019-12-12.