Big Data Applications Symposium - Spring 2019

Project Name: Author Profiling: Predicting Age, Gender and Personality of Twitter Users

Team: Daniel Rivera Ruiz

Abstract:

Author profiling is an area of text analytics that tries to uncover characteristics of an author based on the style of written text. In this project we will apply author profiling techniques on *tweets* to try to predict the age, gender and strongest personality traits of Twitter users.

Author Profiling On Twitter Users

Motivation

Who are the users of this application?

Companies that want to better understand their customers.

Who will benefit from this application?

Advertising agencies, political and social analysts

Why is this application important?

The information extracted from text plus the metadata (retweets, followers, etc.) helps to make sense of public opinions and reactions to contemporary issues.

Author Profiling On Twitter Users

Goodness

What steps were taken to assess the 'goodness' of the analytic itself?

Test the predictions of the model with unseen data for which the ground truth is known (e.g. celebrities).

Author Profiling On Twitter Users

Remediation

What actuation(s) or remediation actions are performed by the application?

- Top 5 most similar tweets from the training set
- Understand what the models consider similar
- Identify business opportunities based on these similarities

Author Profiling On Twitter Users

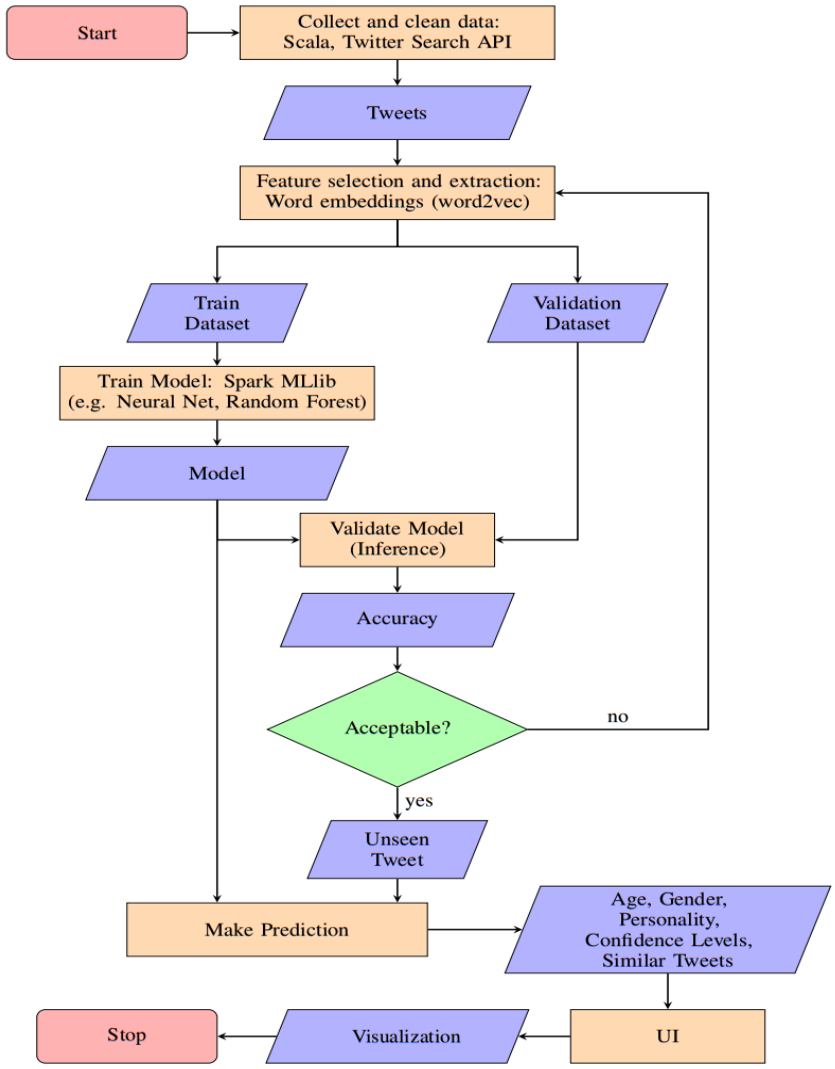
Data Sources

Name: Twitter
Description: Tweets (status + metadata) collected with Twitter API
Size of data: 1G

Name: word2vec
Description: Word embeddings from Google News
Size of data: 10G

Author Profiling On Twitter Users

Design Diagram



Platform(s) on which the application runs:

Dumbo Cluster NYU

Author Profiling On Twitter Users

Insights

1. Skewed data for the age prediction task: more young users.

2. The gender prediction model is the one with more difficulties.

3. The personality trait prediction makes more sense in a per tweet basis rather than per user.

Author Profiling On Twitter Users

Obstacles

1. Retrieving the “right” tweets for the training data (how to determine age/gender/personality a priori).

2. Getting the pre-trained word embeddings in a format that was compatible with Apache Spark.

3. Generating a UI that can interact with the Spark application in the Dumbo Cluster under the hood.

Author Profiling On Twitter Users

Summary

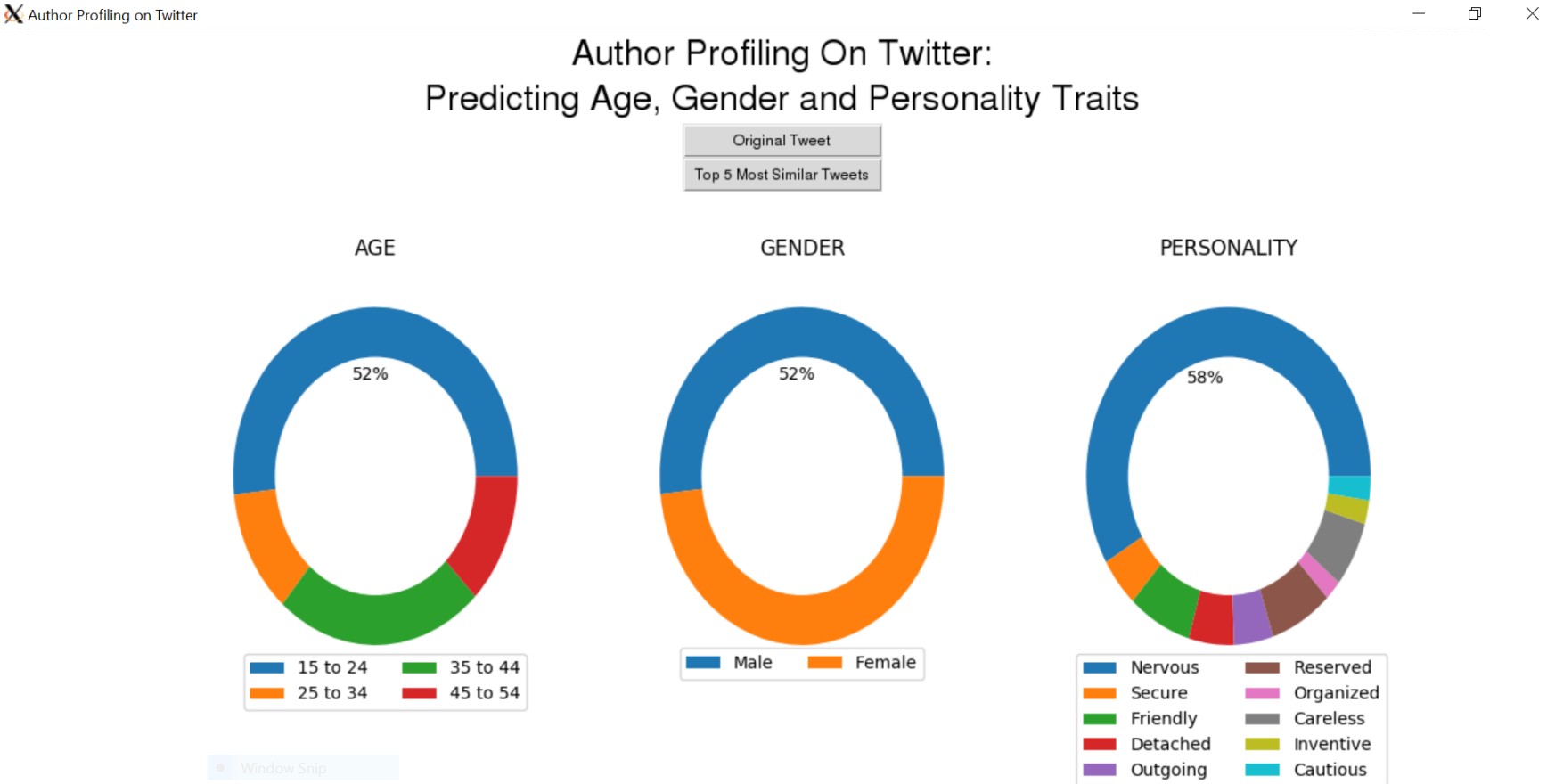
We developed an Author Profiling system that predicts the age, gender and personality traits of Twitter users based on their writing style and metadata from their Twitter accounts.

Acknowledgments

- Professor McIntosh
- HPC@NYU
- Twitter
- Google

Author Profiling On Twitter Users

User Interface



Author Profiling On Twitter Users

References

1. Kursuncu, U., Gaur, M., Lokala, N. U. Ga., Thirunarayan, K., et al. (2019): Predictive Analysis on Twitter: Techniques and Applications.

2. Twitter. (2019): https://developer.twitter.com/en/docs/tweets/search/overview.html

3. Google Code. (2013): Word2vec. https://code.google.com/archive/p/word2vec/

4. Sushree D., Ranjan K. B., Mukesh K., Santanu K. R. (2018): RealTime Sentiment Analysis of Twitter Streaming data for Stock Prediction.

5. Apache Spark 2.3.0. (2019): Machine Learning Library (MLlib) Guide. https://spark.apache.org/docs/2.3.0/ml-guide.html

6. Scala 2.11.12. (2019): Scala Library API. https://www.scalalang.org/api/2.11.12/#package

7. Karau, H., Konwinski, A., Wendell, P., Zaharia, M. (2015): Learning

Spark: Lightning-Fast Big Data Analytics,

8. Alexander, A. (2013): Scala Cookbook: Recipes for Object-Oriented and

Functional Programming,

9. Twitter4J (2019). An unofficial Java library for the Twitter API. http://twitter4j.org/en/

10. Data transfer with Globus (2019). https://www.globus.org/data-transfer

Author Profiling On Twitter Users

Thank you!