Time Series NLP

Exploring Author Characteristics from Longitudinal Writing Samples

Adam Laughlin
Computer Science
University of Colorado Boulder
adam.laughlin@colorado.edu

Jiaheng Zhao
Computer Science
University of Colorado Boulder
jizh3194@colorado.com

Kyle Bremont
Computer Science
University of Colorado Boulder
kyle.bremont@colorado.edu

1 Problem Statement / Motivation

What can we learn about a person through their writing? We apply natural language processing (NLP) and other data science methods to a single-author finance blog spanning 17 years. We seek to identify relatively stable characteristics such as introversion/extraversion and more variable characteristics such as affect and optimism. Of particular interest are change patterns (e.g., cycles) in variable characteristics. If possible, we'll incorporate additional data such as market changes and news events to provide context for changing characteristics.

2 Literature Survey (Previous Work)

Using Linguistic Cues for the Automatic Recognition of Personality in Conversation and Text Fantastic exploration of linguistic features used in personality trait prediction. The features are likely useful in non-personality cases.

Linguistic Styles: Language Use As An Individual Difference Another text breaking down linguistic cues that can be used to classify personality characteristics. Between this and the last paper, we have many linguistic cues to work with, though no specific python framework or project tying them together.

After finding these top two today, we may not need the others, but listing them anyway for now.

A Framework for Emotion Mining from Text in Online Social Networks: Excellent resource for thinking about how to derive affect from text, plus well-documented pre-processing, features, and evaluation metrics

Approaches towards Emotion Extraction from Text: Comparison and classification of affect detection methods. Conclusion for our context is finding a hybrid model is likely best for precision, followed by keyword for explainability.

Emotion and Sentiment Analysis: A Practitioner's Guide to NLP: Extracting sentiment via two python tools

3 **Proposed Work**

3.1 EXTRACT

3.1.1 Collect
- crawl
- store site indices' post metadata
- store post content

3.1.2 Preprocess (done during crawl)
- pre-cleaning (ignoring empty posts)
- preprocessing (only text, no html or media)

3.1.3 Store
- Store posts+meta as files and JSON

3.2 TRANSFORM (TBD)

3.2.1 Integrate
- Convert posts+meta to python object

3.2.2 Decide doc units
- word, noun chunk, sentence, paragraph, document, document collection

3.2.3 Derive features/attributes

Derive doc unit metadata, e.g.:
- tokens (e.g., 'I ran' -> ['I', 'ran'])
- semantics (meanings)
- lemmas (e.g., did->do)
- parts of speech (e.g., noun, verb)
- named entities (Denver, My Dog)

Potential conversion to DataFrame

Derive numeric features, e.g.:
- words per unit
- sentence length std deviation

Derive categorical/ordinal features, e.g.:
- big 5 personality traits
- sentiment (valence+intensity)
- emotion (specific emotion words)
- optimism

3.2.4 Map doc units -> features, e.g.:
- document collection {personality}
- document -> {emotion, topic}
- sentence -> {optimism}
- word -> {tense}

3.2.5 Clean features (as necessary)
- remove stop words
- … maybe others?

3.2.6 Normalize features
- e.g., zscoring each feature/attribute
- ensuring consistent time indexing

3.2.7 Store derived features as pickle or csv

3.3 LOAD (TBD)
- load derived features into analysis tools
- decide indexing scheme (e.g., by document, by date?)

3.4 ANALYZE (TBD)

3.4.1 Mine patterns
- find "interesting" patterns
- return to 3.2 when patterns make good features

3.5 EVALUATE (TBD)
- # of "interesting" patterns found
- Key results learned from this effort

4 **Relation to Prior Work**

4.1 Differences
- single-person long-duration (longitudinal) focus vs many-person short duration focus
- focus on attribute change over time
- broader focus than only sentiment or personality
- python for all steps >= 3.2
- if we have time, integrating context such as market changes and news to assess their correlation with other patterns

4.2 Similarities
- linguistic cues derived from text
- may also derive sentiment and personality

5 **Data Set**

5.1 Challenges: We were unable to find good datasets for longitudinal text analysis, so we created our own by writing a crawler to gather posts from a long-running blog.

5.2 Content: 38,723 single-author blog posts spanning 17 years. Posts contain text only. No pictures, videos, or other media. JSON metadata file contains year, month, day, title, author, url.

5.3 Storage: Posts stored as 212mb worth of individual text files (212mb), zipped to 54mb. Metadata stored as 10mb JSON file. Both available on github.

6 **Evaluation Methods**

Evaluating a data mining system is mainly based on five main aspects: accuracy, performance, functionality, availability, and accessibility.

accuracy
The most critical factor in evaluating a data mining system is accuracy. By performing the prediction and classification accuracy of the algorithm on the data mining system, we can judge whether the algorithm in the system is reasonable and whether the data acquisition is comprehensive and the data preprocessing work is perfect.

performance
We need to check if the system can run and the stability of its operation.

Feature
We need to check whether our system can make the user adjust the parameters of the algorithm and algorithm and whether it can randomly extract data from the data set to build a pre-drilling model; whether the mining results can be expressed in different forms.

Availability
We need to check if the visualization of the system is good;

Accessibility
We need to check if the user is allowed to change the error value in the dataset or to clean the data; whether to allow global substitution of values; whether to discretize continuous data; whether to extract subsets from the dataset according to user-defined rules; The null value in the middle is replaced by an appropriate mean or user-specified value; whether the results of one analysis can be fed back into another analysis.

6 Tools

package management:

- python: pip + requirements.txt
- node.js: yarn + package.json

code storage

- github

Tools by Proposed Work Section

3.1 EXTRACT

3.1.1 Collect

- beautifulsoup or selenium or (I used puppeteer for simplicity and performance)
- nodejs
- chromium
- css selectors

3.1.2 Preprocess (done during crawl)

- custom javascript

3.1.3 Store

- nodejs “fs”
- github

3.2 TRANSFORM (TBD)

3.2.1 Integrate

- nodejs “fs” library

3.2.3 Derive features/attributes

Derive doc unit metadata, e.g.:

- spacy (trying first for performance + documentation + simplicity)
- or ntlk
- or stanfordnlp

Derive numeric features, e.g.:

- numpy
- pandas

Derive categorical/ordinal features, e.g.:

- numpy
- pandas
- custom python
- maybe other python libs (tbd)

3.2.4 Map doc units -> features, e.g.:

- custom python
- pandas
- numpy

3.2.5 Clean features (as necessary)

- spacy, or… need to research this

3.2.6 Normalize features

- custom python
- pandas
- numpy
- sklearn

3.2.7 Store derived features as pickle or csv

- pandas and/or pickle and/or spacy (TBD)

3.3 LOAD (TBD)

- pandas

3.4 ANALYZE (TBD)

- pandas
- matplotlib
- seaborn
- lime

7 Milestones

Done: Proposed work 3.1 (Extract)

7/18: Proposed work 3.2-3.5 (first iteration)

8/2: Proposed work 3.2-3.5 (last iteration)

REFERENCES (from original ACM template)

[1] Using in-place links amidst content for now since I don’t know how to auto-link reference numbers.

Formatting reference for our own use:

Example Image

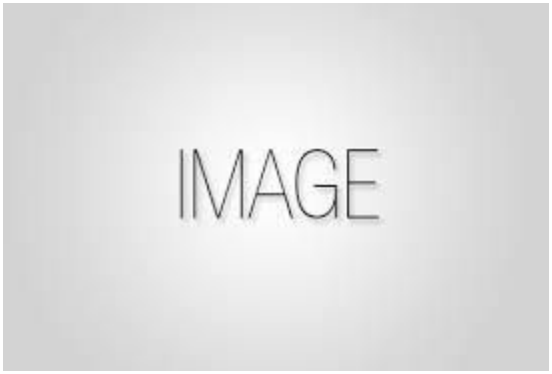

Figure 1: **Figure Caption and Image above the caption**

*Article Title Footnote needs to be captured as Title Note
†Author Footnote to be captured as Author Note

ACM Reference format:

FirstName Surname, FirstName Surname and FirstName Surname. 2018. Insert Your Title Here: Insert Subtitle Here. In *Proceedings of ACM Woodstock conference (WOODSTOCK'18). ACM, New York, NY, USA, 2 pages.* https://doi.org/10.1145/1234567890

FirstName Surname, FirstName Surname and FirstName Surname. 2018. Insert Your Title Here: Insert Subtitle Here. In *Proceedings of ACM Woodstock conference (WOODSTOCK'18). ACM, New York, NY, USA, 2 pages.* https://doi.org/10.1145/1234567890