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 Abstract

We decided to analyze different parts of the author's writing. We started with a few questions:

What can we learn about a person from their writing?
What psychological characterics stand out?
What linguistic characterics stand out?
What psychological patterns are interesting?
What linguistic patterns are interesting?

We found from his blog posts that he is overall a more positive person. His word use change over time was also analyzed, and it was found that his clout (confidence) dramatically increased while his authenticity decreased.

2 Introduction

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.

3 Related Work

3.1 Summary of Related Studies

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.

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

Lexical Predictors of Personality Type: Seeking lexical predictors of personality type and other characteristics. Good discussion on why they might choose some types of features vs others.

LIWC Language Manual - The Development and Psychometric Properties of LIWC2015: Many subtle issues to consider when mapping text to psychometrics.

3.2 Differences from Prior Work

- informal given time/resource constraints
- 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

3.2 Similarities to Prior Work

- linguistic cues derived from text
- also derives sentiment and personality

4 Dataset

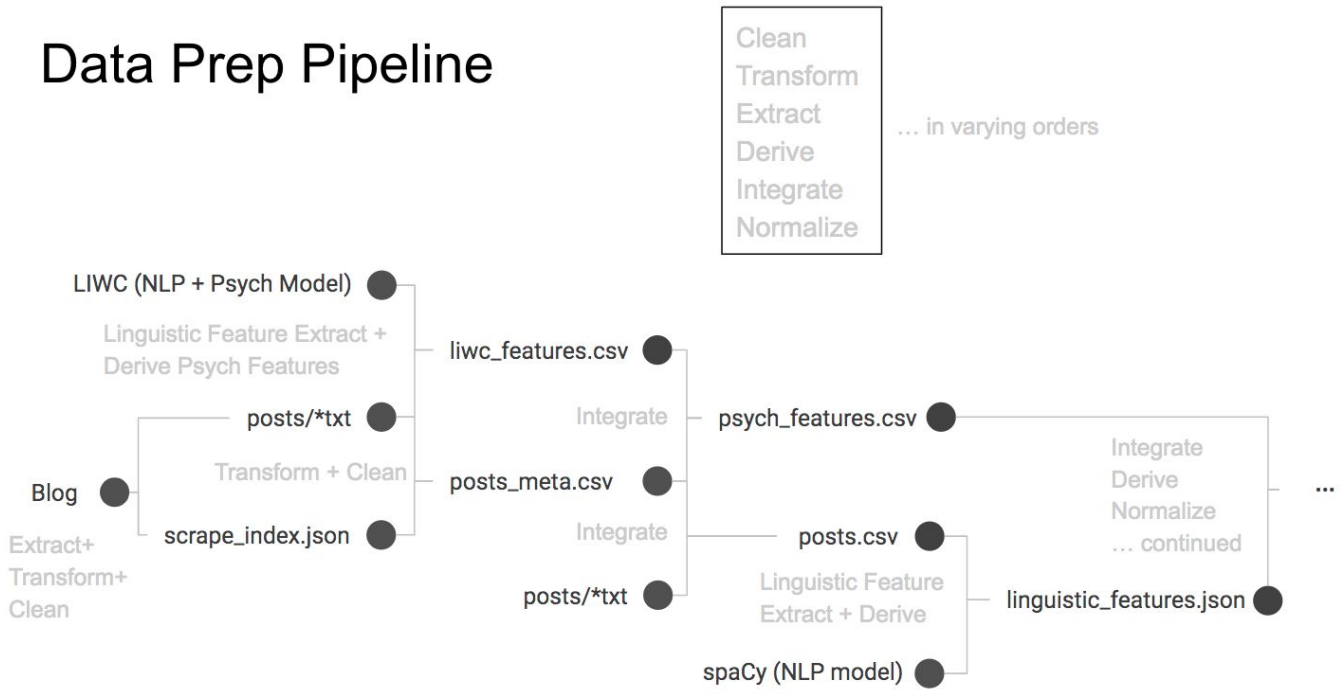
We were unable to find good datasets for longitudinal text analysis, so we created our own by writing a crawler to gather posts from a long-running blog. Content we gathered includes 38,723 blog posts spanning 17 years.

Posts contain text only - no pictures, videos, or other media. Author is mostly one person, though guests did author some posts. JSON metadata file contains year, month, day, title, author, filename.

Posts stored as 212mb worth of individual text files (212mb), zipped to 54mb. Metadata stored as 10mb JSON file. Both combined into a 57mb csv. All available on github.

5 Techniques Applied

5.1 Adam



Those above, plus:
- Dimensionality Reduction: Manual (irrelevant columns, separate year,month,day fields, non-numeric+non-categorical columns)
- Outlier removal (low volume years,<20 words)
- Rescaling (sklearn.StandardScaler())
- Clustering: Manual - by year, month
- Clustering: Manual - by LIWC[1] feature group (e.g., perceptual includes seeing, hearing, feeling words)
- Clustering: Automated - Sklearn OPTICS
- Correlation Analysis (including period-over-period changes)
- Deriving features such as monthly and yearly variability and change
- Visualizations

[1] Linguistic Inquiry and Word Count

5.2 Henry

Read the data through python pandas, archive the article data of the author for about 40 years, and calculate the emotional value for each article, using the library VADER(VADER (Valence Aware Dictionary and sEntiment Reasoner) is a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media, and works well on texts from other domains.), after getting the emotion of each article. By re-sampling the entire data in the monthly dimension, it was found that the author's article was positive for most of the time and the author's frequency of publication was very high, so he was inferred to be a positive and diligent person.

5.3 Kyle

For my analysis I decided to look at if the author of these blog posts is overall a positive or negative person based on what he writes about. LIWC provides us with two statistics pertaining to this: negative and positive emotion words. So, I read in the data using a pandas dataframe and took the average of many different categories that LIWC provided. However, the easiest to clearly see was the positive and negative emotion words.

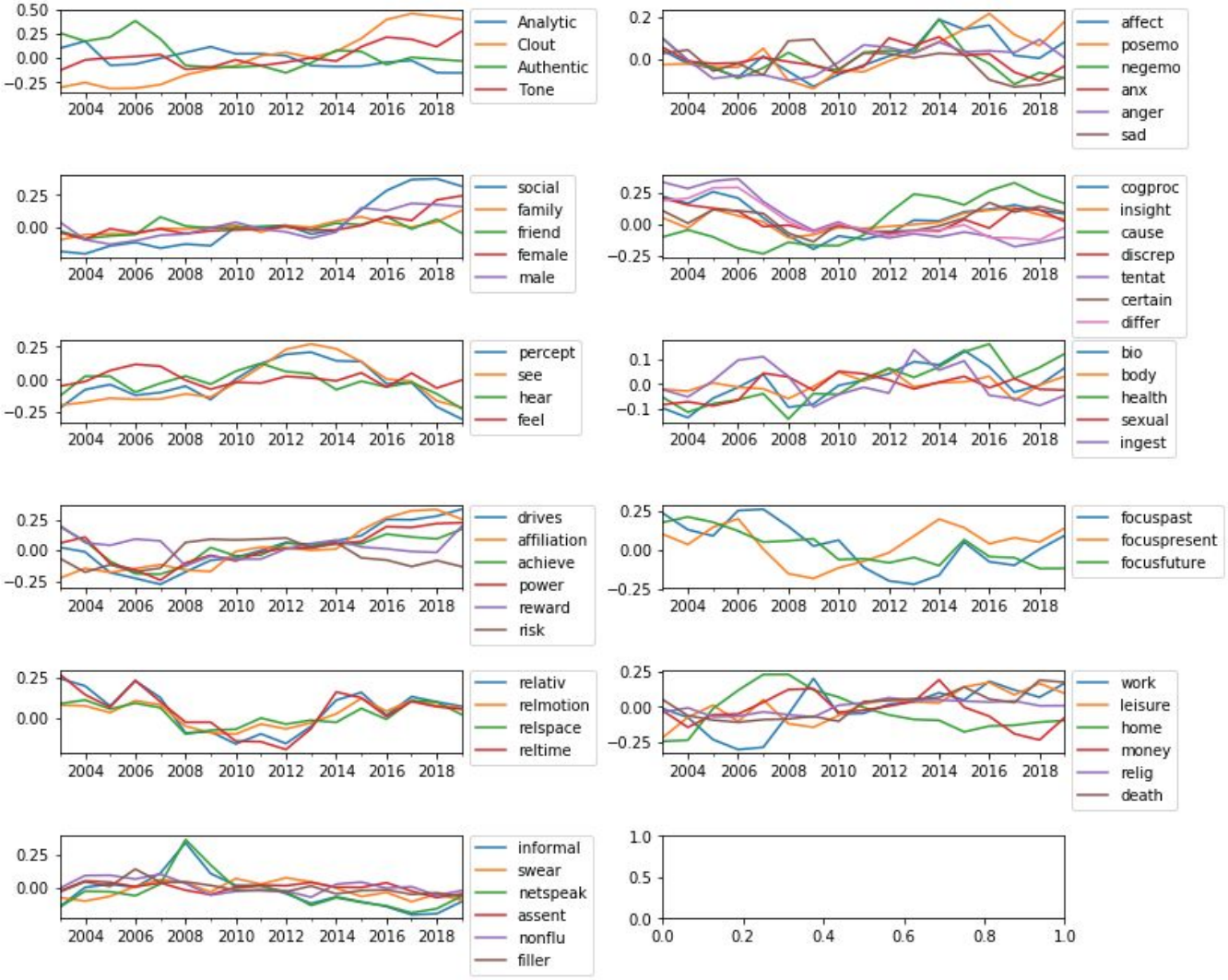
6 Results

6.1 Adam

Plotting the dimensions both separately and in clusters yielded the most useful results. Below are the yearly means for each of LIWC's dimension clusters, minus purely linguistic clusters like grammar and punctuation.

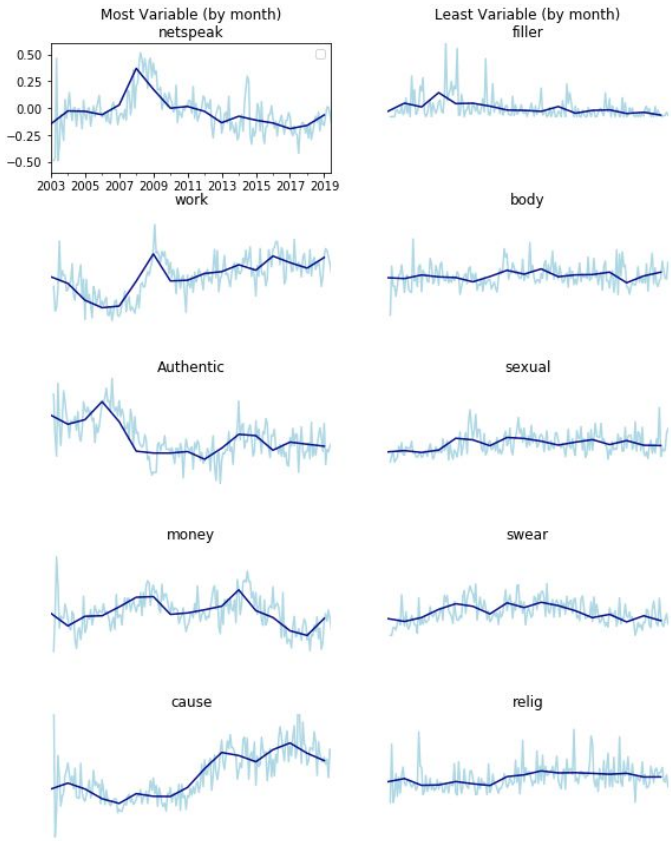
Of note:

- Increased social and affiliation language.
- Increased Clout, decreased Authenticity. My guess is that he's more comfortable stating opinions, and less vulnerable, though both metrics come from other studies I haven't read deeply yet.
- Decreased tentativeness (expected from a longer-term blog)
- Causality language increased, which may reflect experience.
- Consistently low risk language
- Increasing health language (probably correlates with aging in general)

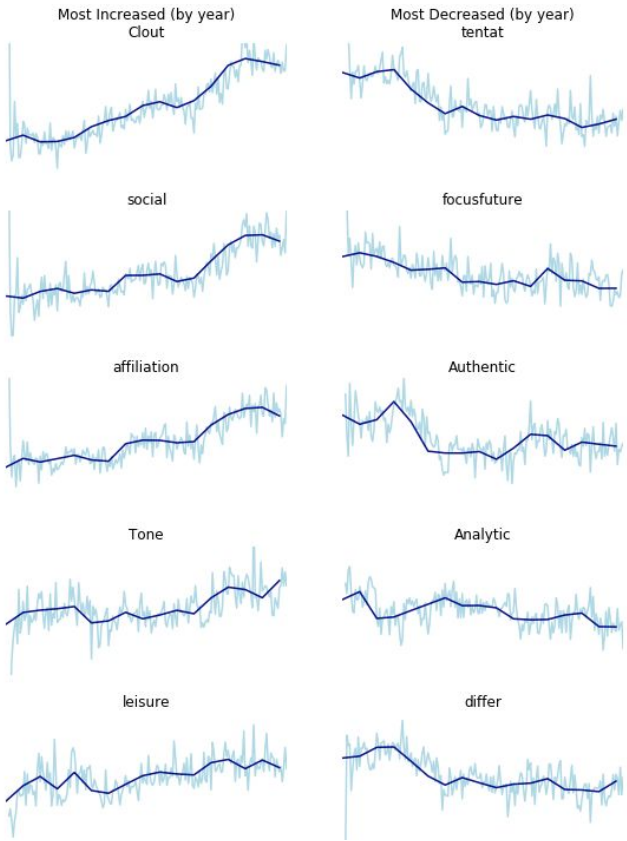


Some patterns jump out more clearly when sorting the dimensions by how they change.

Below, the columns sort all dimensions by different changes. The cells are each dimensions highest in that change metric.



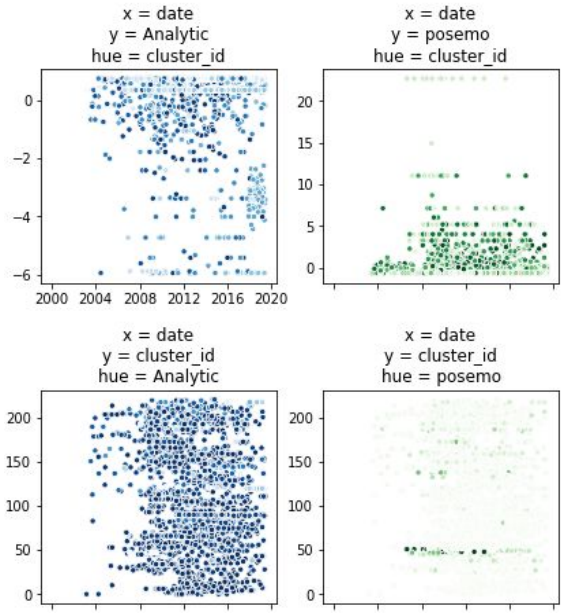
Patterns in Clout, tentativeness, and authenticity are easier to see. Other patterns also stand out, like increased leisure, and decreased future discussion (both make sense with age).



Clustering with sklearn's OPTICS algorithm yielded results I was unsure how best to make sense of.

There were 210 clusters comprised of only 10% of the data. The other 90% was unclustered.

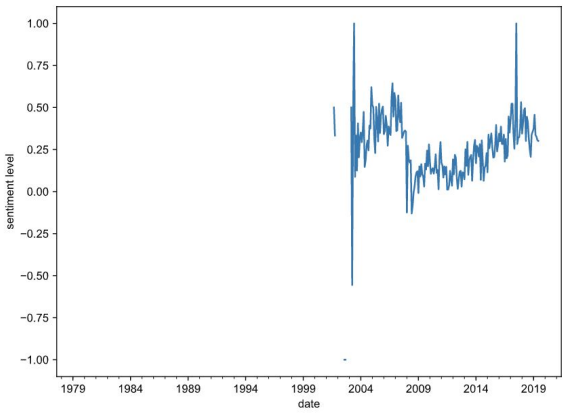
I was unable to find a visualization that made sense of those clusters. For example, here are two different views of the same clustering. One has cluster id on the y axis, and hue is the dimension value. Flipping the two yields different visuals, but neither are intuitive. In the future, I'll likely use a biclustering algorithm, or run clustering on dimension subsets



6.2 Henry

The data shows the author's article was positive for most of the time. So we can think of him as an positive person. At the same time, statistics are made on the author's writing style. Almost all of the author's articles use a lot of scientific data to support their views. Using data to prove the correctness of one's point of view has almost become the author's main rhetorical and writing characteristics. In addition, the author has added pictures or videos to almost all of his articles to help readers understand the views they want to make. Combining articles and pictures to express one's point of view is also a great writing feature of the author.

```
Out[30]: Text(0,0.5,'sentiment level')
```



6.3 Kyle

After taking the averages of many columns in the dataframe, I ended up analyzing the positive and negative emotion words. His negative emotion average was 1.48 and his positive emotion average was 2.67, almost double the negative emotion average. This leads me to believe the author is a more positive person overall.

7 Applications

We can use our results to make assumptions about the author of these blog posts. Using VADAR and analyzing the positive and negative words we can conclude that he is mostly a positive person. There are countless other applications using Timeseries and NLP, such as analyzing the author's psychological and linguistic characteristics.

Appendix A: Work Given Different Starting Models

A.1 Possible Starting Models

- None
- Trained NLP Model
- Trainded NLP + Dataset(s) mapping psych characteristics to writing samples
- Trained NLP + Papers Mapping NLP attributes to psych characteristics
- Trained NLP->Psych Model (LIWC)

A.2 Work starting with Model:

A.2.1 None:

- Train NLP Model on large text dataset (basically create NTLK or spaCy)
- Goto A.2.2

A.2.2 Trained NLP (e.g., Spacy):

- Collect Psych&Text Dataset
- Goto A.2.3

A.2.3 Trained NLP + Dataset(s) mapping writer characteristics to writing samples:

- Extract many text units (words, ngrams)
- Extract many text features per unit (e.g., tense, word count, word length, std devs, other custom, like tentativeness for "I think"...)
- Customize Dictionary (e.g., "kind of")
- Train model to recognize psyc
- characteristics from text
- Isolate most important features per characteristic
- Save as Trained NLP->Psych Model

A.2.4 Trained NLP + Papers Mapping NLP Attributes to writer characteristics

- Decide doc units (word, noun chunk, sentence, paragraph, document, document collection)
- Create map of linguistic features to psych attributes based on those present in papers
- Train NLPsych model that extracts linguistic features, 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)
    - Derive numeric features, e.g.:
    - words per unit
    - sentence (length, std deviation)
- And maps them to psych features, e.g:
    - big 5 personality traits
    - sentiment (valence+intensity)
    - emotion (specific emotion words)
    - optimism
- Tweak until it gets similar results to the papers on the test data
- Goto A.2.5

A.2.5 Trained NLPsych Model

A.2.5.1 Extract Blog Data

- crawl blog
- store site indices' post metadata
- store post content

A.2.5.2 Preprocess (done during crawl)

- pre-cleaning (ignoring empty posts)
- preprocessing (only text, no html or media)

A.2.5.3 Store

- Store posts+meta as files and JSON

A.2.5.4 Transform posts+meta into csvs

- Decide useful linguistic features
    - e.g. most frequent words
- Decide useful psych features
    - e.g., optimism, personality
- Clean (e.g., remove irrelevant columns)
- Normalize (e.g., zscore attributes)

A.2.6.2 Decide useful psych features

A.3 Analyze

A.4 Evaluate Analysis Results

- # of "interesting" patterns found
- Key results learned from this effort