

A decorative graphic on the left side of the slide consisting of two overlapping parallelograms. The front one is blue and the back one is a light green. They are positioned diagonally, with the blue one partially covering the green one.

# Impact of Style on New York Times Article Engagement

By Kyelee Fitts and Abram Moats



# Outline

- I. Overview
  - A. What is style? What is engagement?
- II. State of the Field
  - A. NLP feature analysis
  - B. Sources
- III. Methodology
  - A. Getting the data
  - B. SVR (derivation)
- IV. Results and Conclusions
- V. Our futures



# Why do we care?

Kyelee

- Interested in writing/linguistics and wanted to merge that with something math-y
- The Atlantic Data Science team talk

Abram

- Alden Global Capital
  - Own a bunch of local newspapers that they buy as distressed properties
  - This is probably a bad thing
  - Increasing revenue to local/mid-sized newspapers would be an effective buffer against this kind of consolidation



# Sources

- Linguistic correlates of style: authorship classification with deep linguistic analysis features (2004)
  - <https://pdfs.semanticscholar.org/df0e/dbfbc651d6e077a5e70bafc1fec23b174f30.pdf>
- A Corpus-Independent Feature Set for Style-Based Text Categorization (2003)
  - <https://norek.pw/1556992139.pdf>
- A framework for authorship identification of online messages: Writing-style features and classification techniques (2005)
  - <https://onlinelibrary.wiley.com/doi/full/10.1002/asi.20316>
- Automatic Natural Language Style Classification and Transformation (2008)
  - <https://pdfs.semanticscholar.org/acce/7ca7d40ea4c1ce870bd864b4e7f0194ca0ba.pdf>
- A Tutorial on Support Vector Regression (2003)
  - <https://alex.smola.org/papers/2003/SmoSch03b.pdf>




# The Question

Broad Question: Does style matter?

Specific Question: How does the style of New York Times article affect readers' engagement?

- What is 'style'?
- How do we measure engagement?
- What conclusions can we draw from the results?



How do  
we go  
from  
this.....

WASHINGTON — The United States and China have agreed that an [initial trade deal](#) between the two countries would roll back a portion of the tariffs placed on each other's products, a significant step toward defusing tensions between the world's largest economies.

The [agreement](#) has not yet been completed, and a deal could fail to materialize as it has in [previous rounds](#) of negotiations. But if a pact is reached, the Trump administration has committed to cutting some tariffs, according to a United States official and other people with knowledge of the negotiations.

This is the first time the administration has agreed to remove any of the tariffs it has placed on \$360 billion worth of Chinese goods. While President Trump canceled a planned tariff increase in October, he has routinely [dangled the prospect](#) of additional taxes if Beijing does not accede to America's trade demands.

A deal that includes reversing even some tariffs creates a political dilemma for Mr. Trump, a [self-avowed "tariff man"](#) who has used levies to punish China for trade practices that have helped hollow out American manufacturing and to press Beijing to change its



.... To this?

$$A = \begin{bmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 \\ 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 1 \\ 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 1 & 2 \\ 4 & 5 & 6 & 7 & 8 & 9 & 10 & 1 & 2 & 3 \\ 5 & 6 & 7 & 8 & 9 & 10 & 1 & 2 & 3 & 4 \\ 6 & 7 & 8 & 9 & 10 & 1 & 2 & 3 & 4 & 5 \\ 7 & 8 & 9 & 10 & 1 & 2 & 3 & 4 & 5 & 6 \\ 8 & 9 & 10 & 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ 9 & 10 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\ 10 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \end{bmatrix}$$

$$B = \begin{bmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 \\ 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 1 \\ 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 1 & 2 \\ 4 & 5 & 6 & 7 & 8 & 9 & 10 & 1 & 2 & 3 \\ 5 & 6 & 7 & 8 & 9 & 10 & 1 & 2 & 3 & 4 \\ 6 & 7 & 8 & 9 & 10 & 1 & 2 & 3 & 4 & 5 \\ 7 & 8 & 9 & 10 & 1 & 2 & 3 & 4 & 5 & 6 \\ 8 & 9 & 10 & 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ 9 & 10 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\ 10 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \end{bmatrix}$$



# What is 'style'?

- Define **style** as features of an article that could remain constant among disparate article topics
- An emphasis is placed on features of the articles that can be quantified



# Bag-of-words

- Attempt to 'vectorize' sentences or documents
- Makes a lot of sense for word importance analysis
- Unsuitable for style analysis
  - Poor comparison between documents
  - No analysis of specific word variations

## Raw Text

it is a puppy and it  
is extremely cute

## Bag-of-words vector

it	2
they	0
puppy	1
and	1
cat	0
aardvark	0
cute	1
extremely	1
...	...



# TF-IDF

- Term frequency - Inverse document frequency
  - Attempts to determine the importance of a specific word in a text corpus
  - Simply the multiple of the number of times a word appears in a document by the inverse of the documents in which the word appears
- Unsuitable for style analysis due to its reliance on specific words



# What is style? (Parts of Speech)

“The tree sheds its dead leaves in the fall”

“The dog places its dry bone on the ground”

“The seigneur defends his absurd position from the peasants”

“The curmudgeon accosts his conjubilant neighbors in the foyer”

“The noun verb possessive adjective noun preposition the noun”



# What is style? (Components of a text)

- Words
  - Number of words
  - Number of each POS per document
- Sentences
  - Number of sentences
- Document
  - Avg length of words
  - Avg number of each POS per sentence
  - Avg length of sentence



# What style did we not capture?

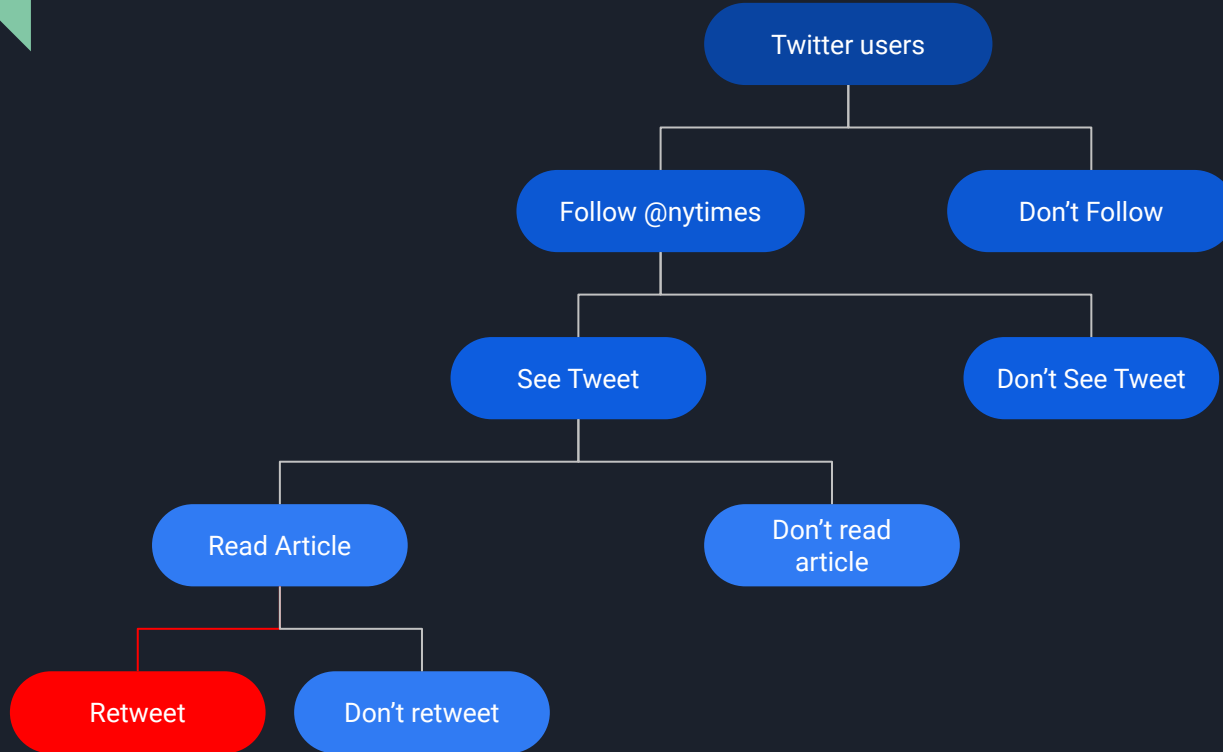
- Sentence structure
- Word bias
- Words that often appear together (n-grams)
- And much more!



# How do we measure engagement?

- Ideal: Access to NYTs internal revenue data, click data, and other user engagement data
- Reality: Twitter engagement data
  - Retweets vs. shares vs. likes
  - Social cost

# Population of Interest- Mental Model



# Methodology





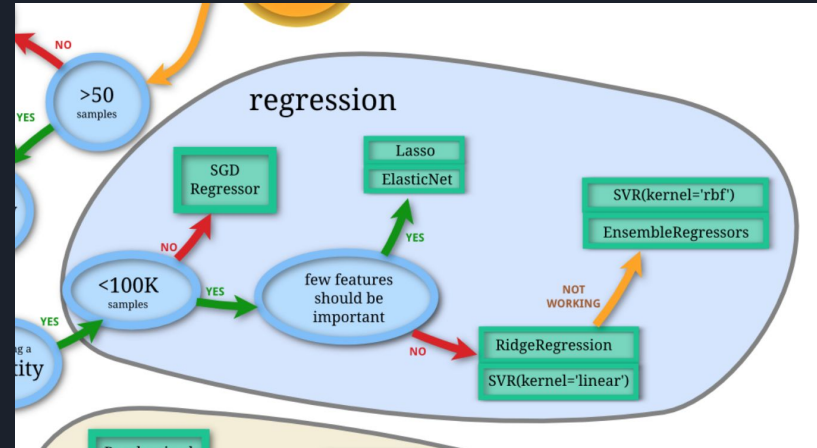
# Data Collection/ Feature Set

- Scrape Twitter for tweet data for NYT articles beginning July 1, 2019
- Using URLs in tweets, scrape article text
- Obtain feature set from article text

CODE:

[https://github.com/ruthlee/NYT-Analysis/blob/master/notebooks/article\\_scraper.py](https://github.com/ruthlee/NYT-Analysis/blob/master/notebooks/article_scraper.py)

[https://github.com/ruthlee/NYT-Analysis/blob/master/notebooks/methods\\_features.py](https://github.com/ruthlee/NYT-Analysis/blob/master/notebooks/methods_features.py)





# Support Vector Regression: Overview

Goal:

- Given a set of feature data ( $x$ ) and targets ( $y$ ), find a linear function  $f(x) = \langle w, x \rangle + b$  which has at most epsilon deviations from the target variables.
  - We want our function to both minimize error and be flat (small  $w$ )

Strategy:

- Linear case: Form primal optimization problem
- Derive Dual problem from Primal Problem
- Non-linear case: Use Dual Problem formulation. Implicit mappings based on kernels

# Technical Nugget: Deriving the Dual Problem from the Primal Problem

## Step 1: Primal Problem + Soft Margins

$$\begin{aligned} \min \quad & \frac{1}{2} \|w\|^2 \\ \text{subj. to} \quad & \|y_i - \langle w, x_i \rangle - b\| \leq \epsilon \end{aligned}$$

$$\begin{aligned} \min \quad & \frac{1}{2} \|w\|^2 + C \sum_{i=1}^l \xi_i \\ \text{subj. to} \quad & \begin{cases} \|y_i - \langle w, x_i \rangle - b\| \leq \epsilon + \xi_i \\ \xi_i \geq 0 \end{cases} \end{aligned}$$

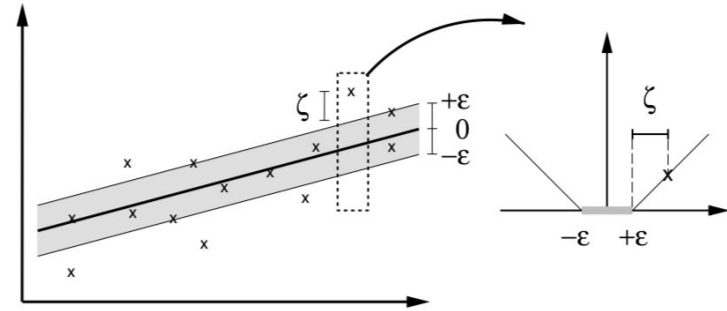


Figure 1: The soft margin loss setting for a linear SVM.

## Step 2: Form the Lagrangian

$$\begin{aligned} L = & \frac{1}{2} \|w\|^2 + C \sum_{i=1}^l (\xi + \xi_i^*) - \sum_{i=1}^l (\eta_i \xi_i + \eta_i^* \xi_i^*) \\ & - \sum_{i=1}^l \alpha_i (\epsilon + \xi_i - y_i + \langle w, x_i \rangle + b) - \sum_{i=1}^l \alpha_i^* (\epsilon + \xi_i^* - y_i + \langle w, x_i \rangle - b) \end{aligned}$$

Saddle points cause partial derivatives to vanish:

$$\delta_b L = \sum_{i=1}^l (\alpha_i^* - \alpha_i) = 0$$

$$\delta_w L = w - \sum_{i=1}^l (\alpha_i^* - \alpha_i) x_i = 0$$

$$\delta_{\xi_i^*} L = C - \alpha_i^* - \eta_i^* = 0$$

$$\delta_{\xi_i} L = C - \alpha_i - \eta_i = 0$$

## Step 3: Simplify

$$\begin{aligned}
&= \frac{1}{2} \sum_{i,j=1}^l (\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*) \langle x_i, x_j \rangle + \sum_{i=1}^l (\alpha_i^* + \eta_i^*)(\xi_i + \xi_i^*) - \sum_{i=1}^l (\eta_i \xi_i + \eta_i^* \xi_i^*) - \\
&\quad - \sum_{i=1}^l \eta_i \xi_i - \sum_{i=1}^l \alpha_i \epsilon - \sum_{i=1}^l \alpha_i \xi_i + \sum_{i=1}^l \alpha_i y_i - \sum_{i=1}^l \alpha_i \langle w, x_i \rangle - \sum_{i=1}^l \alpha_i b - \sum_{i=1}^l \alpha_i^* \epsilon \\
&\quad - \sum_{i=1}^l \xi_i^* \alpha_i^* - \sum_{i=1}^l \alpha_i^* y_i + \sum_{i=1}^l \alpha_i^* \langle w, x_i \rangle + \sum_{i=1}^l b \alpha_i^* \\
&= \frac{1}{2} \sum_{i,j=1}^l (\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*) \langle x_i, x_j \rangle + \sum_{i=1}^l \alpha_i^* \xi_i + \sum_{i=1}^l \alpha_i^* \xi_i^* + \sum_{i=1}^l \eta_i^* \xi_i + \sum_{i=1}^l \eta_i^* \xi_i^* - \sum_{i=1}^l (\eta_i \xi_i + \eta_i^* \xi_i^*) \\
&\quad - \sum_{i=1}^l \eta_i \xi_i - \sum_{i=1}^l \alpha_i \epsilon - \sum_{i=1}^l \alpha_i \xi_i + \sum_{i=1}^l \alpha_i y_i - \sum_{i=1}^l \alpha_i \langle w, x_i \rangle - \sum_{i=1}^l \alpha_i b - \sum_{i=1}^l \alpha_i^* \epsilon \\
&\quad - \sum_{i=1}^l \xi_i^* \alpha_i^* - \sum_{i=1}^l \alpha_i^* y_i + \sum_{i=1}^l \alpha_i^* \langle w, x_i \rangle + \sum_{i=1}^l b \alpha_i^* \\
&= \frac{1}{2} \sum_{i,j=1}^l (\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*) \langle x_i, x_j \rangle + \sum_{i=1}^l \xi_i (\alpha_i^* - \alpha_i) + \sum_{i=1}^l \xi_i^* (\alpha_i^* - \alpha_i) + \sum_{i=1}^l b (\alpha_i^* - \alpha_i) \\
&\quad \sum_{i=1}^l \langle w, x_i \rangle (\alpha_i^* - \alpha_i) - \sum_{i=1}^l \alpha_i \epsilon + \sum_{i=1}^l \alpha_i y_i - \sum_{i=1}^l \alpha_i^* \epsilon - \sum_{i=1}^l \alpha_i^* y_i \\
&= \frac{1}{2} \sum_{i,j=1}^l (\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*) \langle x_i, x_j \rangle - \sum_{i=1}^l \epsilon (\alpha_i + \alpha_i^*) + \sum_{i=1}^l y_i (\alpha_i - \alpha_i^*)
\end{aligned}$$



## Step 4: Get Dual Problem CHECK

We get the following dual optimization problem:

$$\begin{aligned} \max \quad & \frac{1}{2} \sum_{i,j=1}^l (\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*) \langle x_i, x_j \rangle - \epsilon \sum_{i=1}^l (\alpha_i + \alpha_i^*) + \sum_{i=1}^l y_i (\alpha_i - \alpha_i^*) \\ \text{subj. to} \quad & \begin{cases} \sum_{i=1}^l (\alpha_i - \alpha_i^*) = 0 \\ \alpha_i \in [0, C] \end{cases} \end{aligned}$$




# Support Vector Regression: Implicit Kernel Mapping and RBF Kernel

- SVR assumes that data can be fit with a linear regression
- Dual formulation yields a form of  $f(x)$  which only depends on an inner product
- With a proper kernel, can remap input features into a space which can be fit with a linear model *without* specifying that function explicitly

$$w = \sum_{i,j=1}^l (\alpha_i - \alpha_i^*) x_i$$
$$f(x) = \sum_{i,j=1}^l (\alpha_i - \alpha_i^*) \langle x_i, x \rangle + b$$

Support Vector Expansion


$$\begin{aligned} \max & \frac{1}{2} \sum_{i,j=1}^l (\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)k(x_i, x_j) - \epsilon \sum_{i=1}^l (\alpha_i + \alpha_i^*) + \sum_{i=1}^l y_i(\alpha_i - \alpha_i^*) \\ \text{subj. to } & \begin{cases} \sum_{i=1}^l (\alpha_i - \alpha_i^*) = 0 \\ \alpha_i \in [0, C] \end{cases} \end{aligned}$$

Dual Problem generalized to nonlinear input spaces

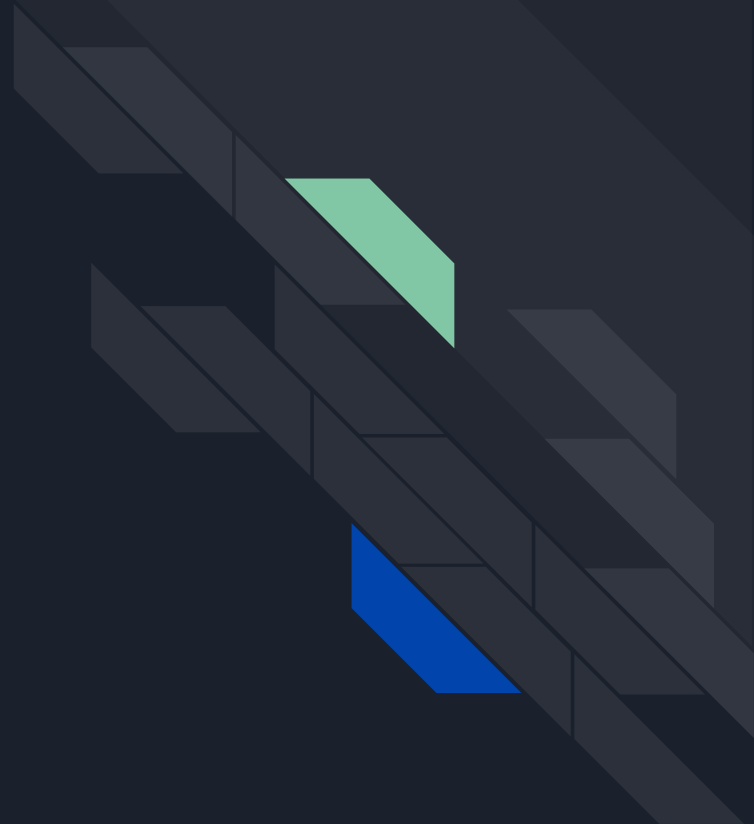
$$k(x_1, x_2) = e^{-\gamma \|x_n - x_m\|_2^2}$$

RBF Kernel

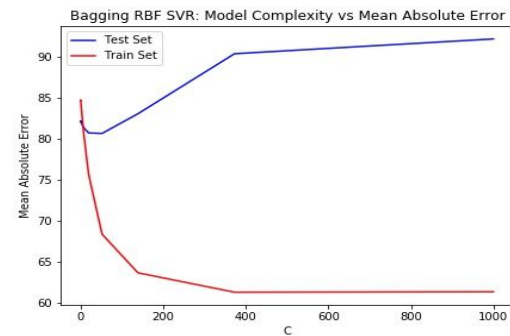
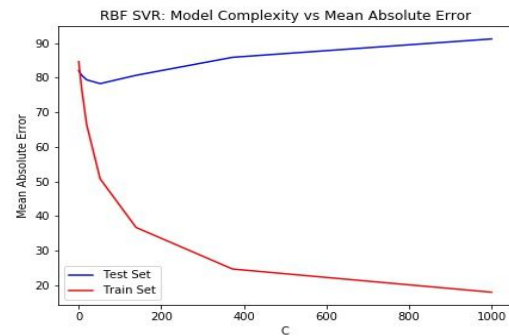
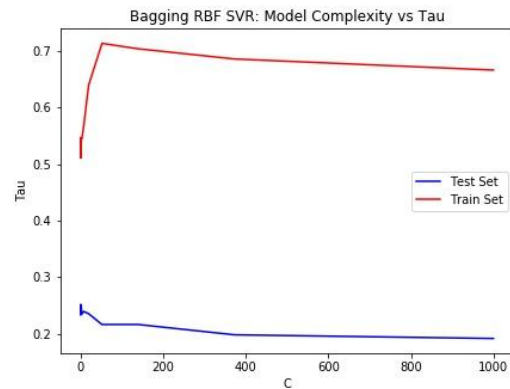
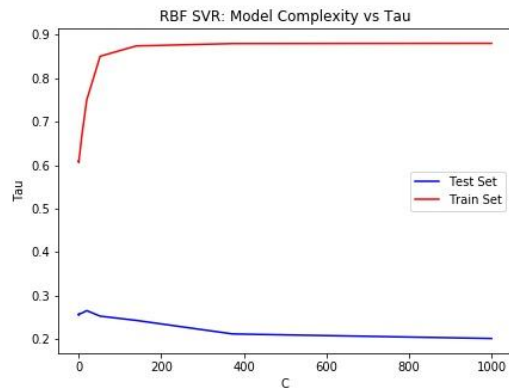


# Results/Conclusions

[https://github.com/ruthlee/NYT\\_Analysis/blob/master/notebooks/rbf\\_kernel.ipynb](https://github.com/ruthlee/NYT_Analysis/blob/master/notebooks/rbf_kernel.ipynb)



# Results





# Future Refinements

- n-grams over parts of speech
- Sentence structure
- Repeat for other news sources and compare (is style less important for the Washington Post?)
- Use better engagement data (@NYT)