1. Text Data in Finance

Manish Jha, Georgia State

General Info

- Features that make text different from other forms of data,
- Relevant statistical methods

Social Science Examples

- In finance, text from financial news, social media, and company filings is used to predict asset price movements and study the causal impact of new information.
- In macroeconomics, text is used to forecast variation in inflation and unemployment, and estimate the effects of policy uncertainty.
- In media economics, text from news and social media is used to study the drivers and effects of political slant.
- In industrial organization and marketing, text from advertisements and product reviews is used to study the drivers of consumer decision making.
- In political economy, text from politicians' speeches is used to study the dynamics of political agendas and debate.

How text differs from other kinds of data

Text is inherently high-dimensional.

Suppose that we have a sample of documents each of which is w words long, and suppose that each word is drawn from a vocabulary of p possible words. Then the unique representation of these documents has dimension $\mathbf{p^w}$.

A sample of 30-word Twitter messages that use only the 1,000 most common words in the English language, for example, has roughly as many dimensions as there are atoms in the universe.

Consequence of High Dimensions

Statistical methods used to analyze text are closely related to those used to analyze high-dimensional data in other domains, such as machine learning and computational biology.

Some methods, such as **lasso and other penalized regressions**, are applied to text more or less exactly as they are in other settings.

Other methods, such as topic models and multinomial inverse regression, are close cousins of more general methods adapted to the specific structure of text data.

Generic Summary

1. Represent raw text D as a numerical array C
2. Map C to predicted values V-hat of unknown outcomes V
3. Use V-hat in subsequent descriptive or causal analysis

1. Represent raw text

Must impose some preliminary restrictions to reduce the di-mensionality of the data to a manageable level.

Even the most cutting-edge high-dimensional techniques can make nothing of 1000^{30}-dimensional raw Twitter data.

This step may involve:

- filtering out very common or uncommon words;
- dropping numbers, punctuation, or proper names; and
- restricting attention to a set of features such as words or phrases that are likely to be especially diagnostic.

2. Map C to predicted values V-hat

In a classic example, the data is the text of emails, and the unknown variable of interest V is an indicator for whether the email is spam. The prediction V-hat determines whether or not to send the email to a spam filter.

Another classic task is sentiment prediction.

Pang et al. 2002: Using movie reviews as data, we find that standard machine learning techniques definitively outperform human-produced baselines.

Some points relevant to Step 2

In the vast majority of settings where text analysis has been applied, the ultimate goal is prediction rather than causal inference.

The interpretation of the mapping from V to V-hat is not usually an object of interest. Why certain words appear more often in spam, or why certain searches are correlated with flu is not important so long as they generate highly accurate predictions.

3. Descriptive or causal analysis

Stephens-Davidowitz (2014) uses Google search data to estimate local areas' racial animus, then studies the causal effect of racial animus on votes for Obama in the 2008 election.

Gentzkow and Shapiro (2010) use congressional and news text to estimate each news outlet's political slant, then study the supply and demand forces that determine slant in equilibrium.

Engelberg and Parsons (2011) measure local news coverage of earnings announcements, then use the relationship between coverage and trading by local investors to separate the causal effect of news from other sources of correlation between news and stock prices.

Representing text as data

Some common ways

Where we are

The field of computational linguistics has made tremendous progress in interpretation. Most of us have mobile phones that are capable of complex speech recognition.

Yet virtually all analysis of text in the social sciences, like much of the text analysis in machine learning more generally, ignores the lion's share of this complexity.

Common simplifications

We typically make three kinds of simplifications:

1. dividing the text into individual documents i,
2. reducing the number of language elements we consider, and
3. limiting the extent to which we encode dependence among elements within documents.

The result is a mapping from raw text D to a numerical array C. A row ci of C is a numerical vector with each element indicating the presence or count of a particular language token in document i.

1. Dividing the text into individual documents

For spam detection, the outcome of interest is defined at the level of individual emails so we want to divide out text that way too.

If V is daily stock price movements which we wish to predict from the prior day's news text, it might make sense to divide the news text by day as well.

If we wish to predict legis- lators' partisanship from their floor speeches (Gentzkow et al. 2016) we could aggregate speech so a document is a speaker-day, a speaker-year, or all speech by a given speaker during the time she is in Congress.

2. Feature selection

1. Strip out elements of the raw text other than words - punctuation, numbers, HTML tags, proper names, and so on.

2. remove a subset of words that are either very common or very rare. Very common words, often called “stop words,” include articles (“the,” “a”), conjunctions (“and,” “or”), forms of the verb “to be,” and so on.

An approach that excludes both common and rare words and has proved very useful in practice is filtering by “term-frequency-inverse-document-frequency” (tf-idf).

3. Stemming: replacing words with their root, such that, e.g., “economic,” “economics,” “economically” are all replaced by the stem “economic.” The Porter stemmer (Porter 1980) is a standard stemming tool for English language text.

tf-idf

For a word or other feature j in document i, term frequency is the count ci j of occurrences of j in i.

Inverse document frequency (idfj) is the log of one over the share of documents containing j.

The object of interest tf-idf is the product tfij ×idfj. Very rare words will have low tf-idf scores because tfij will be low. Very common words that appear in most or all documents will have low tf-idf scores because idfj will be low.

N-grams

Producing a tractable representation also requires that we limit dependence among language elements. A fairly mild step in this direction, for example, might be to parse documents into distinct sentences, and encode features of these sentences while ignoring the order in which they occur.

The simplest and most common way to represent a document is called bag-of-words.

Single words are often insufficient to capture the patterns of interest: “death tax” and “tax break” are phrases with strong partisan overtones that are not evident if we look at the single words “death,” “tax,” and “break”

N-grams contd ...

Unfortunately, the dimension of ci increases exponentially quickly with the order n of the phrases tracked. The majority of text analyses consider n-grams up to three or five at most.

Best practice in many cases is to begin analysis by focusing on single words. Given the accuracy obtained with words alone, one can then evaluate if it is worth the extra time to move on to 2-grams or 3-grams.

References

Gentzkow, Matthew, Bryan Kelly, and Matt Taddy. 2019. "Text as Data." Journal of Economic Literature, 57 (3): 535-74.

DOI: 10.1257/jel.20181020

2. Statistical methods

Four categories

Mapping dtm to predictions

Methods for mapping the document-token matrix C to predictions V-hat.

Could be subdivided into four categories.

1. Dictionary-based methods

Do not involve statistical inference at all: they simply specify vˆi = f (ci) for some known function f (·). This is by far the most common method.

In Tetlock (2007): ci is a bag-of-words representation and the outcome of interest vi is the latent “sentiment” of Wall Street Journal columns, defined along a number of dimensions such as “positive,” “optimisti”.

The author defines the function f () using a dictionary called the General Inquirer, which provides lists of words associated with each of these sentiment categories.

1. Dictionary-based methods

In Baker et al. (2016), ci is the count of articles in a given newspaper-month containing a set of pre-specified terms such as “policy,” “uncertainty,” and “Federal Reserve,” and the outcome of interest vi is the degree of “policy uncertainty” in the economy.

The authors define f () to be the raw count of the pre-specified terms divided by the total number of articles in the newspaper-month, averaged across newspapers.

2. Text regression methods

Generative model: p(vi | ci)

This is intuitive: if we want to predict vi from ci, we would naturally regress the observed values of the former (Vtrain) on the corresponding values of the latter (Ctrain).

Any generic regression technique can be applied, depending upon the nature of vi. However, the high-dimensionality of ci, where p is often as large as or larger than ntrain, requires use of regression techniques appropriate for such a setting, such as L1 regularized linear or logistic regression.

2.1 Linear text regression

This is a regression problem like any other, except that the high-dimensionality of ci makes OLS and other standard techniques infeasible.

Penalized linear models (ridge, lasso, elastic net, log)

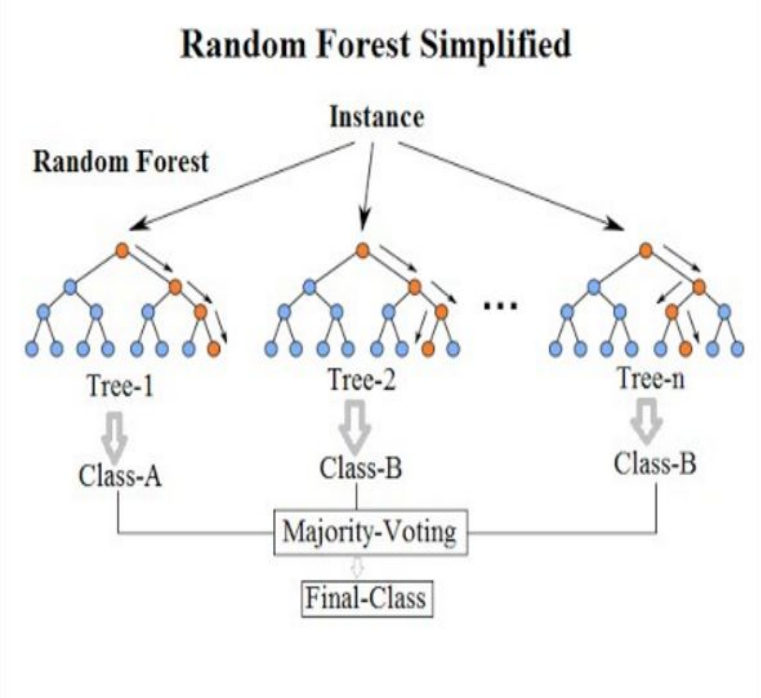
Linear models are intuitive and interpretable

2.2 Nonlinear text regression

May be useful in certain cases where the linear specification is too restrictive.

A reliable class of regression and classification methods are built around decision trees.

For regression tasks, the mean or average prediction of the individual trees is returned.



3. Generative model of p(ci | vi).

The underlying causal relationship runs from outcomes to language rather than the other way around. For example, Google searches about flu do not cause flu cases to occur; rather, people with flu are more likely to produce such searches.

Congress people's ideology is not determined by their use of partisan language; rather, people who are more conservative or liberal to begin with are more likely to use such language.

3. Generative models can be further divided

… by whether the attributes are observed or latent:

1. supervised methods
2. unsupervised methods, we do not observe the true value

   we are willing to impose sufficient structure on it to allow us to infer vi from ci. This class includes principal component analysis (PCA), as well as text-specific methods such as latent Dirichlet allocation (LDA, topic modeling) and its variants.

Finally, in some cases vi includes both observed and latent attributes for a semi-supervised analysis. Ex: Guided LDA

4. Deep learning techniques

These leverage richer representations of the underlying text than the token counts that underlie other methods.

They have seen limited application in economics to date, but their dramatic successes in other machine learning domains suggest they are likely to have high value in the future.

4.1 Word embeddings

Treat textual content in its more full format, as an ordered sequence of transitions between words.

The embeddings are chosen to optimize, perhaps approximately, an objective function defined on the original text such as a likelihood for word occurrences.

Researchers are beginning to connect these vector-space language models with the sorts of document attributes that are of interest in social science.

4.2 Transformers

Transformer - an attention mechanism that learns contextual relations between words (or sub-words) in a text.

In its vanilla form, Transformer includes two separate mechanisms — an encoder that reads the text input and a decoder that produces a prediction for the task.

As opposed to directional models, which read the text input sequentially (left-to-right or right-to-left), the Transformer encoder reads the entire sequence of words at once.

This characteristic allows the model to learn the context of a word based on all of its surroundings (left and right of the word).

Model validation and interpretation

Ex ante criteria for selecting an empirical approach are suggestive at best. In practice, it is also crucial to validate the performance of the estimation approach ex post.

Cross-validation

Whenever one works with complex and high-dimensional data, it is good practice to reserve a testset of data to use in estimation of the true average prediction error.

Looping across multiple test sets, as in CV, is a common way of reducing the variance of these error estimates.

Specific Methods

That I have used, or feel are important for PhD students

Dictionary based method

Often a good starting point

Needs less resources and time

Wordnet package in python

Linear, Ridge and Lasso Regression

Package - sklearn

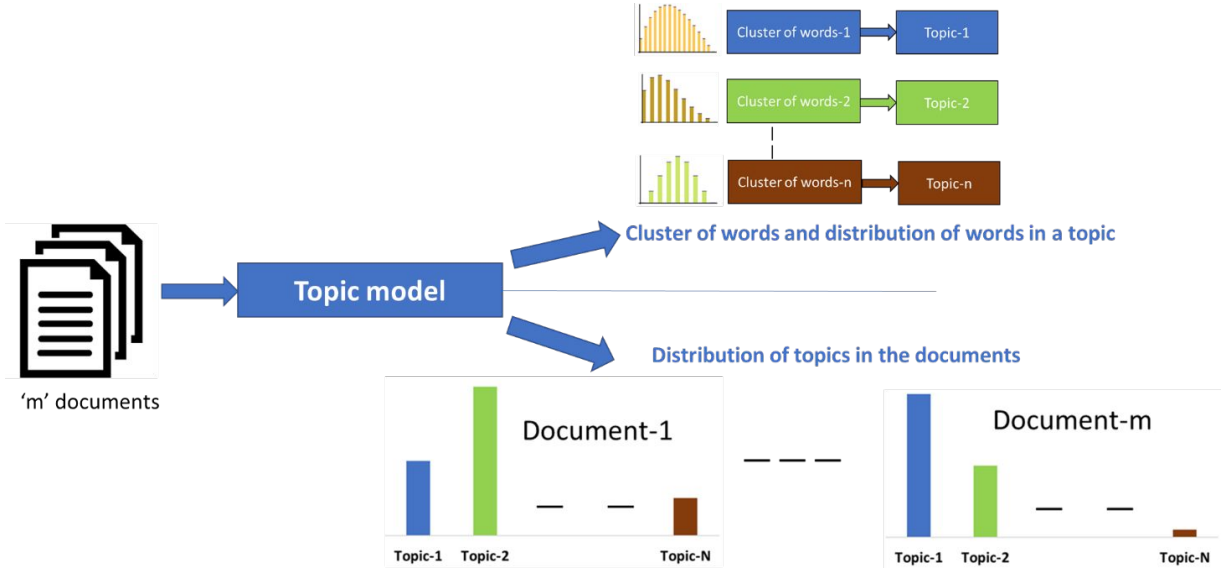
Lasso regression is a type of linear regression that uses shrinkage. Shrinkage is where data values are shrunk towards a central point, like the mean.

The lasso procedure encourages simple, sparse models (i.e. models with fewer parameters).

This particular type of regression is well-suited for models showing high levels of muticollinearity or when you want to automate certain parts of model selection, like variable selection/parameter elimination.

Topic Modelling - LDA, HDP

Latent Dirichlet Allocation (LDA) and Hierarchical Dirichlet Process (HDP) are both topic modeling processes. The major difference is LDA requires the specification of the number of topics, and HDP doesn't.



Word Embeddings

- Word2Vec
- Global Vector for Word Representation (GloVe)
- FastText
- BERT (Transformers)

Text Classification

TensorFlow

- Open-source platform
- Scalable: Almost every operation can be performed using this platform. With its characteristic of being deployed on every machine and graphical representation of a model allows its users to develop any kind of system using TensorFlow.
- Parallelism: TensorFlow finds its use as a hardware acceleration library due to the parallelism of work models. It uses different distribution strategies in GPU and CPU systems.

3. Execution

Manish Jha, Georgia State

Example 1

Catching the Conscience of Kings: How Activists Align with Mutual Funds

Proxy attack players: activist, target, shareholders

Activists
Icahn Capital

In 2013, Michael Dell offered $13.65 per share.
Icahn: privatization, not the best idea

Vanguard (3.9%), Charles Schwab(0.01%)

Targeted firm
Dell Technologies

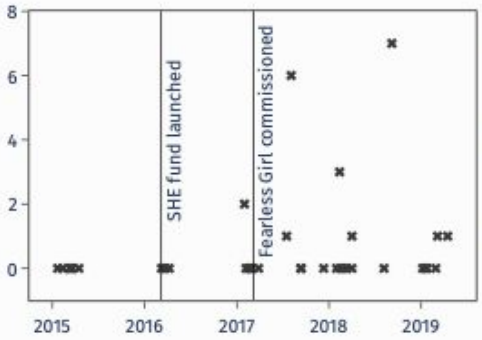
The goals of the confrontational proxy attacks:

- **Shareholder value** - board structure, financing structure, corporate strategy, etc.,
- **Social justice** - climate-friendly policies, women empowerment, etc.

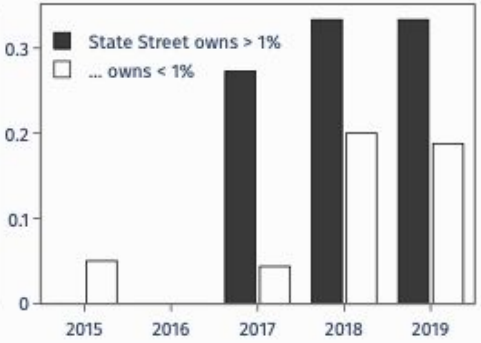
Activists use gender diversity phrases when State Street is a major shareholder

"We will vote against ... incumbent board members if a company does not have **at least one woman on its board**" - State Street, 2020

Gender diversity phrases: "female," "gender," "woman," "women."



(a) Number of times gender diversity phrases were mentioned in attacks where State Street owns > 1%.



(b) Fraction of attacks where gender diversity phrases were used.

Research Question: Do hedge fund activists tailor their campaigns to align with larger mutual fund families? And if so, how does it affect activism?

Main Idea

Do hedge fund activists (X) tailor their campaigns to mutual fund family's preferences (Y)?

We need to:

- Figure out Y
- Whether X's communication matches Y

What do mutual funds want?

Voting reveals mutual fund preferences

Ways mutual funds reveal their preferences (McCahery, Sautner, and Starks 2016):

a. Behind the scenes engagement, executive interviews, websites
b. Proxy voting guidelines
c. Proxy voting - I use **shareholder proposal text**, 2-years prior to the attack.

Relate shareholder proposal text with ***Align*** - the fraction of funds (in a family) that did not follow management recommendation.

$$Align \in [0, 1]$$

If 9 out of 10 invested funds from State Street vote against management, *Align* = 0.9.

Key skill: web scraping

Packages in python:

- Selenium - mimics human
- Beautiful Soup - removes html tag

Also need to know text parsing

- Wordnet - gives you english words
-

Relating proposal text to fund family voting

$$Align_{s,f} = \alpha_f + \beta_f \cdot x_s + \nu_{s,f}$$

Sample proposal: Gender diversity is important to us. To increase gender diversity, we nominate Dr. Rachel Green to the board.

$$x_s = \begin{bmatrix} gen\_div\ (2) \\ rach\_green\ (1) \\ \dots \\ share\_val\ (0) \end{bmatrix};$$

Relating proposal text to fund family voting

$$Align_{s,f} = \alpha_f + \beta_f \cdot x_s + \nu_{s,f}$$

Sample proposal: Gender diversity is important to us. To increase gender diversity, we nominate Dr. Rachel Green to the board.

$$x_s = \begin{bmatrix} gen\_div\ (2) \\ rach\_green\ (1) \\ \dots \\ share\_val\ (0) \end{bmatrix}; \qquad 0.9 = \begin{bmatrix} gen\_div_{coeff} & rach\_green_{coeff} & \dots & share\_val_{coeff} \end{bmatrix} \cdot \begin{bmatrix} 2 \\ 1 \\ \dots \\ 0 \end{bmatrix}$$

- High dimensional input (shareholder proposal text features, ~10,000), with
- Limited observations (~500 shareholder proposals in two years).

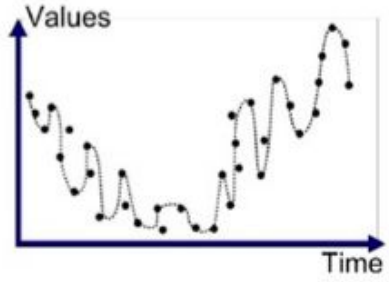
Supervised machine learning > OLS

Over-fitting if we employ Ordinary Least Squares.
Predict perfectly in-sample, fail out-of-sample.

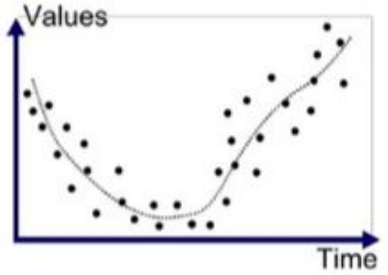
$$0.9 = \begin{bmatrix} gen\_div_{coeff} & \textbf{rach\_green}_{coeff} & \dots & share\_val_{coeff} \end{bmatrix} \cdot \begin{bmatrix} 2 \\ 1 \\ \dots \\ 0 \end{bmatrix}$$

Support Vector Regression (SVR) penalizes non-zero coefficients:

- Benefit: predict well out-of-sample,
- Cost: cannot focus on subspaces, such as "**Paris** Agreement on Climate Change."



Overfitted



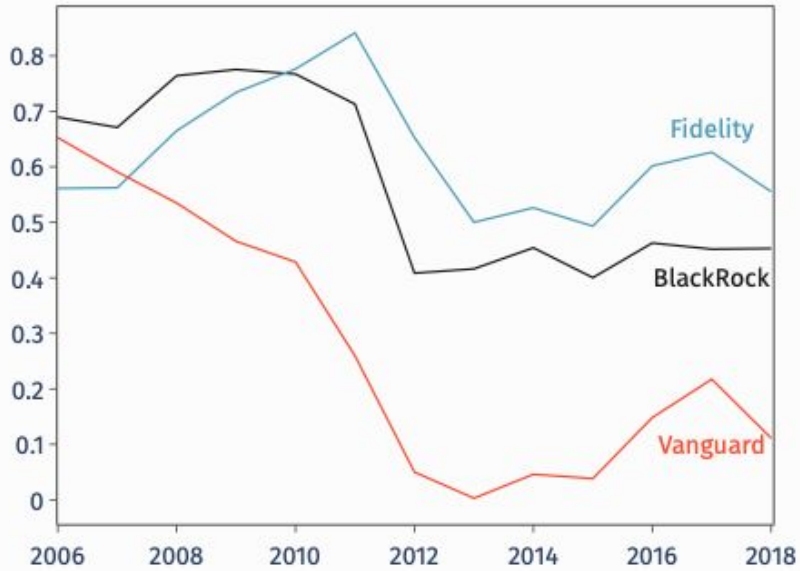
Good Fit/Robust

Key skill: text regression

Packages in python:

- Sklearn
- Count vectorizer

SVR coefficients are interpretable and rooted in proxy voting choices



(a) Fraction of shareholder proposals containing "simple majority vote", where the fund family voted against a management recommendation.

(b) SVR assigned coefficient for "simple majority vote."

Once trained, the model could predict fund family support for the activist

Input

Proxy attack text

Proxy filings (DFAN, DEFC, PREC) by activists to solicit shareholder votes

2013 FrontFour Capital's attack on Ferro Corporation:
"Shareholders request that our board take the steps necessary so that each voting requirement in our charter and bylaws that calls for a greater than **simple majority vote** be eliminated and replaced by a requirement for a majority of the votes cast for and against applicable proposals or a **simple majority** in compliance with applicable laws."
Other Proposals: director nominations, sell solar and pharmaceutical businesses

Trained Model

Output

$\widehat{Align}$ = attack text's alignment with fund family preferences $\in [0, 1]$

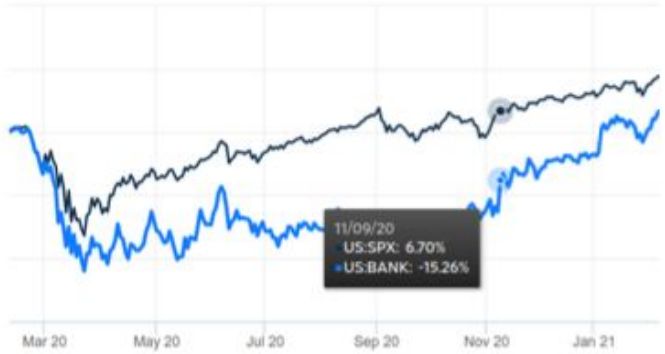
$\cdots$ is more aligned with Fidelity.

Example 2

Natural Disaster Effects on Popular Sentiment Toward Finance

COVID-19

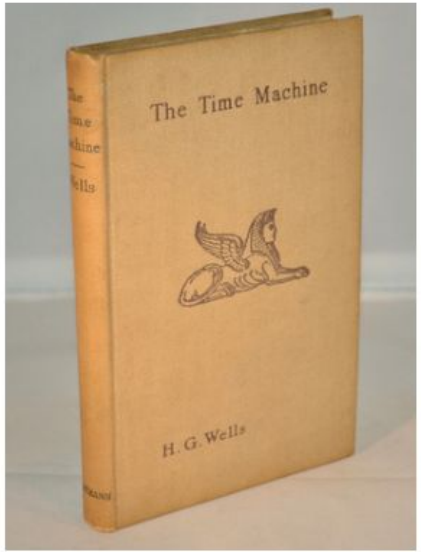
- Financial intermediaries bore most of the blame for 2008 crisis and subsequent recession
- Will the financial sector be perceived more as a *hero* or *villain* after COVID-19?



Source: wsj.com

Our approach

- Measure sentiment toward finance in an annual panel
- 8 large economies matched to languages from 1870–2009
- Computational linguistics approach applied to the text of millions of books



Data

- Text from Google Books corpus
  - Annual sentence (5-gram) counts 1870–2009
  - 8 languages: Chinese, German, French, Italian, Russian, Spanish, UK and US English
  - About 1 billion sentences mentioning "finance"
- Macro data
  - Jorda-Schularick-Taylor macro data for advanced economies
  - Barro-Ursua macro data for Russia and China
- Natural disasters data
  - Emergency Events Database from CRED 1900–2009

Key skill: high level computing

Since the dataset is so big, you need to be able to use resources outside your computer.

Some schools and government agencies provide high level computing resources.

Could also use:

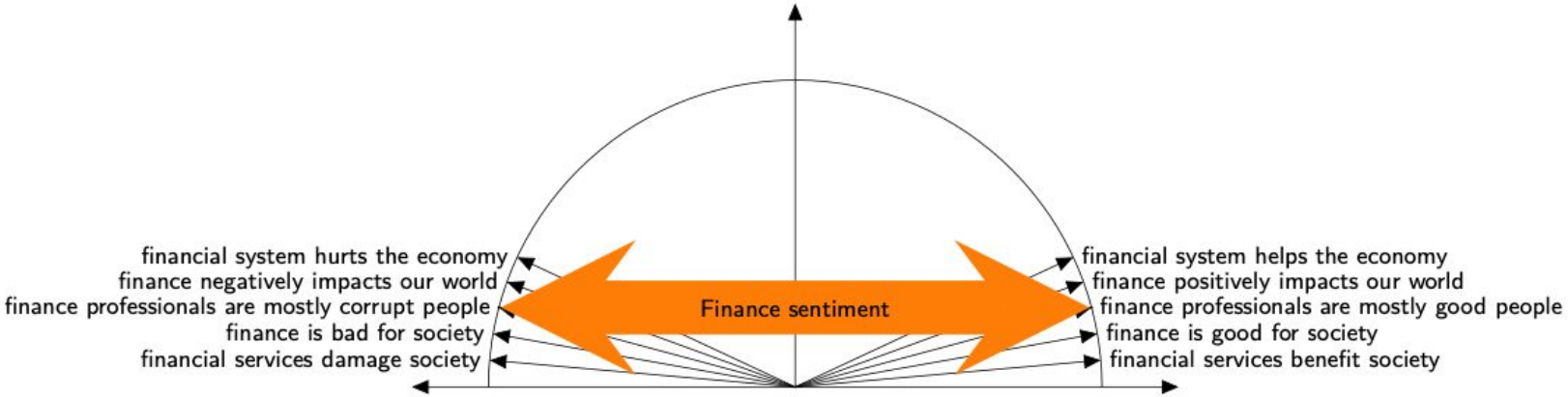
- AWS
- Google Cloud
- Azure

Word embeddings

- We rely on recent language model (BERT, Devlin et al. 2018) to measure if "finance" mentions are on average closer to positive versus negative sentences
- We use BERT to embed sentences in a low dimensional numerical vector (~800d)
- Neural word embeddings produce richer insights into cultural associations than prior methods
  - e.g. $\overrightarrow{king} - \overrightarrow{man} + \overrightarrow{woman} \approx \overrightarrow{queen}$
- BERT is particularly good at distinguishing *context*
- Basic idea
  - e.g. "correcting corruption or financial malpractice"
  - Closer to "finance damages society" than to "finance benefits society"

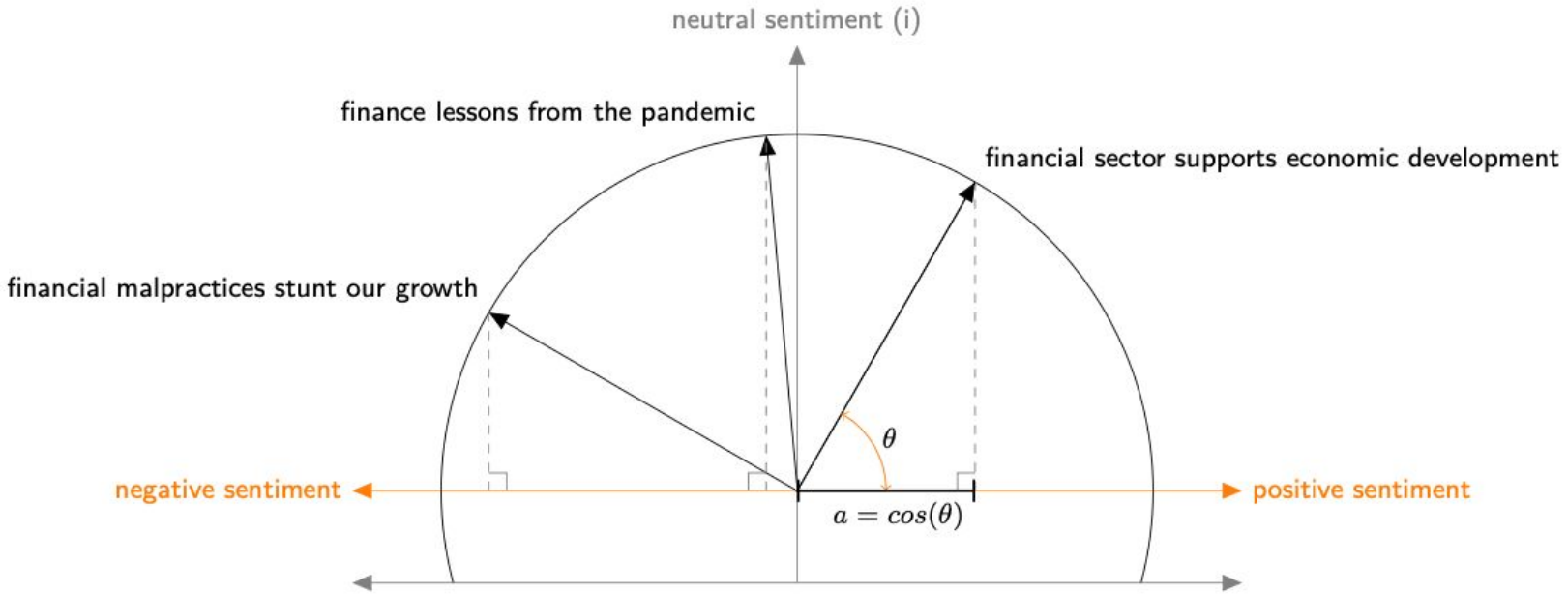
Measuring of finance sentiment

Step 1: Define positive−negative sentiment dimension



Measuring of finance sentiment

Step 2: Project "finance" mentioning sentence j in language i embeddings on the positivity dimension



Finance sentiment for language i in year t is mean cosine similarity across mentions

Key skill: BERT

BERT (Bidirectional Encoder Representations from Transformers) is a recent paper published by researchers at Google AI Language.

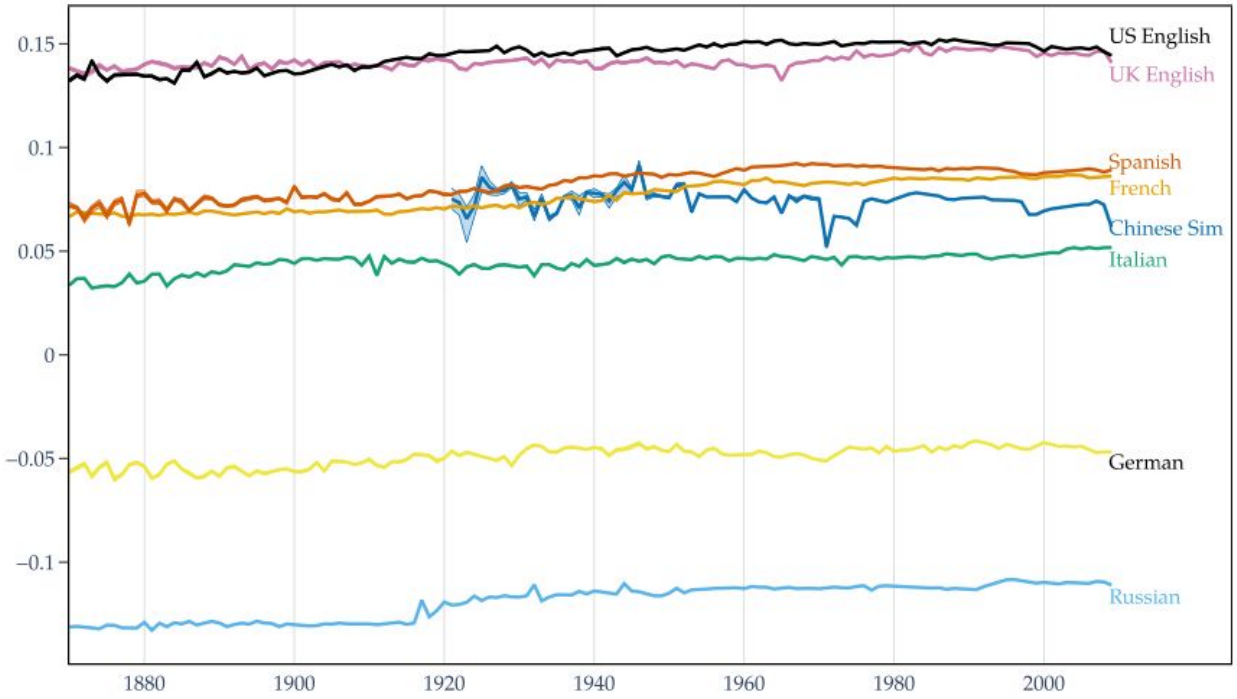
Main advantages of BERT:

- comes pre-trained
- able to account for a word's context
- open-source

The packages for BERT are available on GitHub.

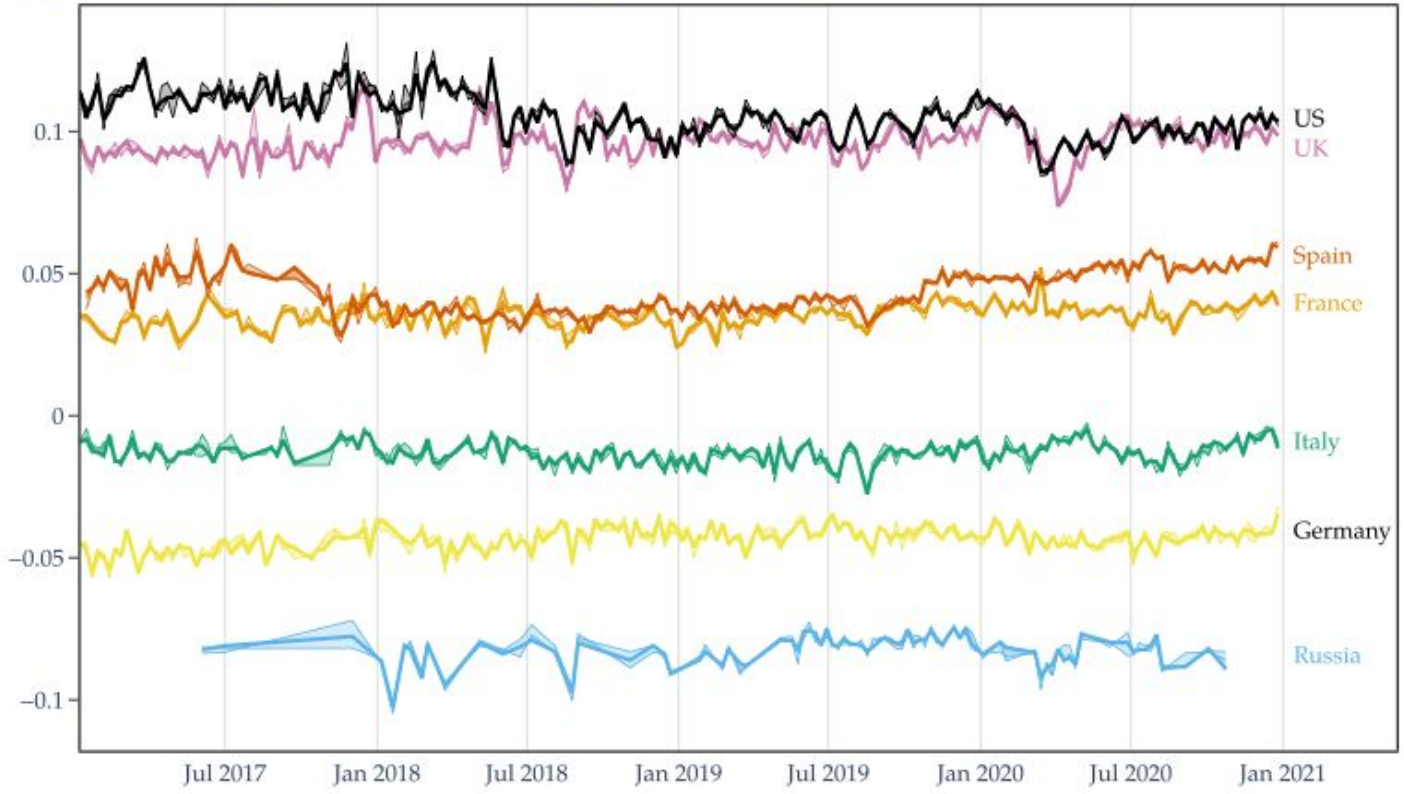
Sentiment toward finance 1870–2009

Persistent differences across languages/countries despite ample time-series variation



Higher frequency finance sentiment for the recent period

Similar ordering as the panel generated from Google Books corpus



Key skill: news-please, common crawl

news-please: recursively follow internal hyperlinks and read RSS feeds to fetch both most recent and also old, archived articles

common crawl: build and maintain an open repository of web crawl data that can be accessed and analyzed by anyone.

Languages you need

You need:

- One from these: Python, R, Julia
- May need one based on your co-author: STATA, LaTeX, SAS

If you are starting from scratch, Python is a good choice.

Thank you

Feel free to reach out to mjha@gsu.edu