Statistical Learning for Unstructured Data

FA590 Statistical Learning in Finance

Dr. Zonghao Yang

Stevens Institute of Technology

Previous Lecture

- Understand the differences between supervised and unsupervised learning
- Apply clustering techniques such as hierarchical clustering, K-means, and DBSCAN to segment data into meaningful groups
- Explore Principal Components Analysis (PCA) and its extensions in the context of asset pricing to create low-dimensional representation of data

Learning Objectives

- Describe the unique features and challenge of analyzing textual data
- Summarize the process of textual analysis
- Explain the analytical tools for textual analysis

Part I

Unstructured Data

Data

- “Data is the new oil" - Clive Humby (Tesco)
  - Data is an important asset in the digital economy
  - Data is useful only if it’s mined/analysed
- **Data** Factual information used as a basis for reasoning, discussion, or calculation
- **Structured Data** Data that has been predefined and formatted to a set structure
  - Tabular format
  - Easy use by machine learning algorithms

Date	Open	High	Low	Close	Volume
Nov 29, 2024	6,003.98	6,044.17	6,003.98	6,032.38	2,444,420,000
Nov 27, 2024	6,014.11	6,020.16	5,984.87	5,998.74	3,363,340,000
Nov 26, 2024	6,000.03	6,025.42	5,992.27	6,021.63	3,835,170,000
Nov 25, 2024	5,992.28	6,020.75	5,963.91	5,987.37	5,633,150,000
Nov 22, 2024	5,972.90	5,944.36	5,944.36	5,969.34	4,141,420,000

Table. S&P 500 Index, Yahoo Finance

Unstructured Data

- **Unstructured Data** Data stored in its native format
- Text, audio, images, and videos
- International Data Corporation (2022): Unstructured data will comprise 80% of enterprise data in 2025[1]

[1] International Data Corporation (2022). "Worldwide Global DataSphere and Global StorageSphere Structured and Unstructured Data Forecast, 2022–2026."

How Many Forms of Data Can You Find?

Figure. An 𝕏 Post from Elon Musk

How Many Forms of Data Can You Find?

Figure. An 𝕏 Post from Elon Musk

How Many Forms of Data Can You Find?

Figure. An 𝕏 Post from Elon Musk

1. Text
2. Video
3. Audio
4. Image
5. Tabular data

TEXT

- Daily: Social media, messages, textbooks, web content
- Finance: 10-K reports, transcripts of earning calls, news articles
- Business: Contracts, patent descriptions, resumes

Text as Data

- **Finance** Predict asset price (news, social media, and company fillings)
- **Macroeconomics** Nowcast the macroeconomic indicators, such as inflation and unemployment (news)
- **Media economics** Study the drivers and effects of political slant (news and social media)
- **Marketing** Study the drivers of consumer decision making (advertisements and product reviews)
- **Political economy** Study the dynamics of political agendas and debate (politicians' speeches)

Features and Challenge

- Raw text consists of an ordered sequence of language elements
  - Length of a sentence varies
  - An ordered sequence
  - Context
- Challenge for analyzing text: **High-dimensionality**
  - The unique representation of a thirty-word $\mathbb{X}$ message, using one thousand common English words, has dimension 1000^{30}

Examples

- “Alice loves Bob." vs. “Bob loves Alice."
- “My favorite fruit is an **apple**." vs. “The CEO of **Apple** is Tim Cook."

Why Today?

- **Data:** An ever-increasing share of human interaction, communication, and culture is recorded as digital text
- **Methodology:** Natural language processing and understanding
  - Semantic meaning: *Word2Vec*[2]
  - Length and sequence: *Recurrent Neural Networks* (RNN)[3]
  - Context: *Transformer* architecture based on the attention mechanism[4]
- **Technology:** Computing power and resources

- ChatGPT of OpenAI; Search and translation functions of Google

[2] Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space.

[3] Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. NeurIPS.

[4] Vaswani, A., Shazeer, N., ... & Polosukhin, I. (2017). Attention is all you need. NeurIPS.

Part II

TEXTUAL ANALYSIS

Textual Analysis Pipeline

1. Represent raw text $\mathcal{D}$ as a numerical array $\mathbf{C}$
2. Map $\mathbf{C}$ to predicted values $\hat{\mathbf{V}}$ of attributes $\mathbf{V}$
3. Use $\hat{\mathbf{V}}$ in subsequent descriptive or causal analysis

Representing Text as Data

1. Represent raw text $\mathcal{D}$ as a numerical array **C**: **Bag-of-Words** (BoW)

- Divide raw text $\mathcal{D}$ into individual documents $\{\mathcal{D}_i\}$
- Build the vocabulary
  - Strip out elements like punctuation and numbers
  - Remove stop words: e.g., 'and', 'the', 'we'
  - Stemming: e.g., 'economic', 'economics', 'economically' $\rightarrow$ 'economic'
- Create **C** by counting the occurrence of words in each $\mathcal{D}_i$

Example of BoW Representation

- Raw text $\mathcal{D}$
  - $\mathcal{D}_1$: How time flies! It was the best of times.
  - $\mathcal{D}_2$: How time flies! It was the worst of times.
  - $\mathcal{D}_3$: It was the age of wisdom.
  - $\mathcal{D}_4$: It was the age of foolishness.
- Step 1: Build vocabulary
- Step 2: Count the number of occurrences
- This word count table is also referred toas term-document matrix **C**

	best	worst	time	age	wisdom	foolish	fly
$\mathcal{D}_1$	1	0	2	0	0	0	1
$\mathcal{D}_2$	0	1	2	0	0	0	1
$\mathcal{D}_3$	0	0	0	1	1	0	0
$\mathcal{D}_4$	0	0	0	1	0	1	0

Table. Document-Token Matrix

TF-IDF Score

$$\text{TF-IDF Score} = tf_{ij} \times idf_j$$

- i - document, j - word
- **Term Frequency** (tf_{ij}): The count of occurrences of j in i
- **Inverse Document Frequency** (idf_j): The log of one over the share of documents containing j, i.e., $\log(n/d_j)$
  - Number of documents containing word j, $d_j = \sum_i \mathbf{1}_{[tf_{ij}>0]}$
  - Intuition: Word j is discounted if it appears in many documents, such as stop words (e.g., 'and', 'the', 'we')
- By excluding the words with tf-idf scores below some cutoff, this approach excludes both common and rare words

BoW Representation using TF-IDF Score

- Raw text $\mathcal{D}$
  - $\mathcal{D}_1$: How time flies! It was the best of times.
  - $\mathcal{D}_2$: How time flies! It was the worst of times.
  - $\mathcal{D}_3$: It was the age of wisdom.
  - $\mathcal{D}_4$: It was the age of foolishness.
- Step 1: Build vocabulary
- Step 2: Calculate tf-idf score for each word in each document

$$\text{TF-IDF Score} = tf_{ij} \times idf_j = tf_{ij} \times \log(n/d_j)$$

	best	worst	time	age	wisdom	foolish	fly
$\mathcal{D}_1$			$2 \times \log(4/2)$				
$\mathcal{D}_2$			$2 \times \log(4/2)$				
$\mathcal{D}_3$			$0 \times \log(4/2)$				
$\mathcal{D}_4$			$0 \times \log(4/2)$				

Table. Document-Token Matrix

Comments on BoW Representation

- Limitations
  - Loss of word order and context
  - BoW creates high-dimensional, sparse vectors, leading to inefficiency in storage and computation
  - Semantic ignorance

- Positive side
  - Simplicity and ease of implementation
  - Often surprisingly effective for simple text classification tasks (like spam detection, sentiment analysis, etc.)

WORD2VEC

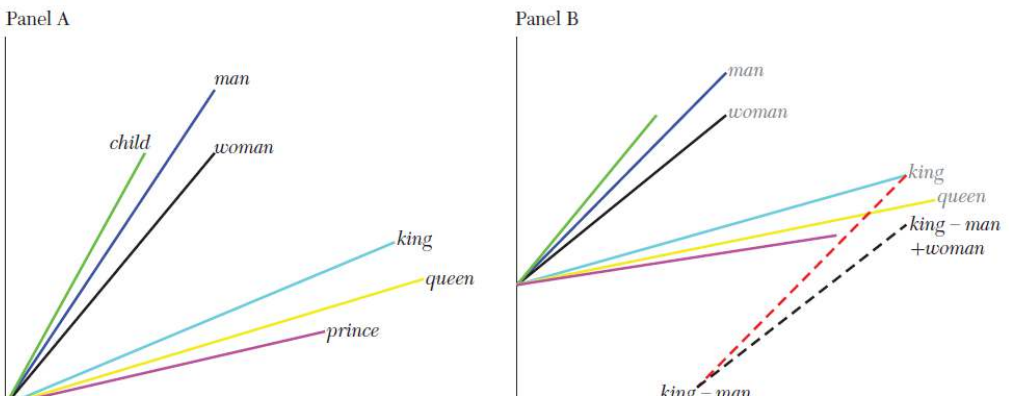


Figure. A Graphical Example of Word Embedding

Statistical Methods

1. Represent raw text $\mathcal{D}$ as a numerical array $\mathbf{C}$
2. Map $\mathbf{C}$ to predicted values $\hat{\mathbf{V}}$ of attributes $\mathbf{V}$

- Examples of attributes
  - Observable: Frequency of flu cases, the positive or negative rating of movie reviews, or the unemployment rate
  - Latent: Topics being discussed
- Statistical methods
  - Dictionary-based methods: Predefined dictionary[5]
  - Text regression methods: $p(\mathbf{v}_i|\mathbf{c}_i)$
  - Generative model: $p(\mathbf{c}_i|\mathbf{v}_i)$

[5] Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. Journal of Finance.

Penalized Regression Model

Penalized linear model in the simplest form,

$$E[\mathbf{v}_i|\mathbf{c}_i] = \alpha + \mathbf{c}_i'\beta \tag{1}$$

The penalized estimator is then the solution to,

$$\min\{l(\alpha,\beta) + n\lambda \sum_{j=1}^{p} k_j(|\beta_j|)\}, \tag{2}$$

where $l(\alpha,\beta)$ is the OLS loss and $k(\cdot)$ is a predefined function that penalizes deviations of β_j from 0 (such as Lasso and Ridge)

Remark Linear models are intuitive and interpretable but can produce suboptimal forecasts

Principal Components Regression (PCR)

PCR consists of two steps:

- PCA on the original input array $\mathbf{C}$ to get K principal components
- The K components are used in standard predictive regression

Remark Loss of interpretability

Nonlinear Text Regression

To consider the nonlinear dependencies between $\mathbf{c}_i$ and $\mathbf{v}_i$:

- Generalized linear models: Expand the linear model to include nonlinear functions of $\mathbf{c}_i$ such as polynomials or interactions
- Support vector machines: Text classification problems (e.g. spam)
- Regression trees: Random forests and boosted trees

Remark Tree structure is an effective way to accommodate rich interactions and nonlinear dependencies.

Generative Language Models

- Motivation Text regression treats the token counts as generic high-dimensional input variables without any attempt to model structure that is specific to language data

- **Generative Model** The words in a document are viewed as the realization of a generative process defined through a probability model for $p(\mathbf{c}_i|\mathbf{v}_i)$
  - Unsupervised: Topic model
  - Supervised: Naïve Bayes Classifier

Topic Modeling

- Unsupervised machine learning algorithm
- Latent Dirichlet Allocation (LDA) method[6]
- Intuition for topic modeling
  - A document is a distribution of topics
  - A topic is a distribution of words
- Input the document-token matrix ($\mathbf{C}$); Output the **document-topic distribution** ($\hat{\mathbf{V}}$) and topic-token distribution ($\hat{\Theta}$)

[6] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. Journal of Machine Learning Research.

Assumptions and Model Specifications

- Assume k topics across all documents, with a vocabulary of p words
- Document i is represented as token probability vectors $\mathbf{q}_i$ with multinomial distribution:

$$\mathbf{c}_i \sim MN(\mathbf{q}_i, m_i), \quad m_i = \sum_{j=1}^{p} c_{ij}$$

- Topic l is a probability distribution θ_l over tokens:

$$\sum_{j=1}^{p} \theta_{lj} = 1, \quad l = 1, \ldots, k$$

- Document i-specific language composition:

$$\mathbf{q}_i = \Theta \mathbf{v}_i, \quad \text{with} \sum_{l=1}^{k} v_{il} = 1$$

- Parameter estimation via EM algorithm: $\hat{\mathbf{V}}$ and $\hat{\Theta}$

Naive Bayes Classifier

- Supervised generative model: Attributes $\mathbf{v}_i$ are observable
  - Example of $\mathbf{v}_i$: The author of articles
- 'Naïve': Conditionally independence between tokens j

$$p(\mathbf{c}_i|\mathbf{v}_i) = \prod_j p_j(c_{ij}|v_i) \tag{3}$$

- Prediction: With certain prior probability π_a for class a, we have

$$p(\mathbf{V}|\mathbf{c}_i) = \frac{p(\mathbf{c}_i|\mathbf{V})\pi_v}{\sum_a p(\mathbf{c}_i|a)\pi_a} \tag{4}$$

Model Selection

- Dictionary-based methods heavily weight prior information about the function mapping features $\mathbf{c}_i$ to outcomes $\mathbf{v}_i$. They are appropriate in cases where priors are strong and reliable information in the data is weak.
- Text regression is a good choice for predicting a single attribute
- Generative models are appropriate when 1) multiple attributes, and 2) the interdependence of attributes and language are of interest

Part III

APPLICATIONS

NOWCASTING MACROECONOMIC CONDITIONS

Raw text: *Wall Street Journal* articles

1. Text representation $\mathbf{C}$: Document-token matrix via Bag-of-Words
2. Map $\mathbf{C}$ to predicted attributes $\hat{\mathbf{V}}$: The topic distribution of the day using topic modeling
3. Use $\hat{\mathbf{V}}$ to predict nowcast the macroeconomic indicators, such as industrial productivity, GDP, and unemployment

- Forthcoming in *Journal of Finance*[7]
- The Structure of Economics News[8]

[7] Bybee et al. (Forthcoming). Business News and Business Cycles. Journal of Finance.

[8] http://www.structureofnews.com/

Application (Cont.)

- **State-of-the-art** of textual analysis in Finance
- Motivation: Macroeconomic indicators are published with delays and in low frequency, whereas news articles are in real-time
- Methodological contribution: Topic modeling - Interpretability and predictability

Group Discussion

- What improvements would you make if you were the authors?

Group Discussion

- What improvements would you make if you were the authors?
- Hints (1/2): Length, sequence, and context of text

Group Discussion

- What improvements would you make if you were the authors?

- Hints (1/2): Length, sequence, and context of text
- Hints (2/2): Embedding and deep learning

Word Embedding

- Vectors in the latent embedding space
- Estimated by optimizing an objective function (such as a likelihood for word occurrences and co-occurrences)
- Advantages compared to Bag-of-Words:
  - Semantic meaning
  - Combined with deep learning, we can consider the order of word sequence and the context
- Word2Vec, Global Vector for Word Representation, and BERT

Deep Learning

- Deep learning is applied with word embedding
- Recurrent Neural Networks (RNN, such as GRU and LSTM) and Transformer (attention mechanism)
- Proven success in language tasks like machine translation, chat bots, sentiment analysis, voice assistant, etc.

Revisit Text Analysis Pipeline

1. Represent raw text $\mathcal{D}$ as a numerical array $\mathbf{C} \Leftarrow$ Word embedding
2. Map $\mathbf{C}$ to predicted values $\hat{\mathbf{V}}$ of attributes $\mathbf{V} \Leftarrow$ DL/ML
3. Use $\hat{\mathbf{V}}$ in subsequent descriptive or causal analysis

Stock Prices

Three examples that study equity return prediction using text

- Preexisting dictionary: Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. Journal of Finance.
- Regression: Jegadeesh, N., & Wu, D. (2013). Word power: A new approach for content analysis. Journal of Financial Economics.
- Generative Models: Bybee, L., Kelly, B., Manela, A., & Xiu, D. (2024). Business news and business cycles. Journal of Finance.

AUTHORSHIP[9]

- Task: Infer the authorship of the disputed Federalist Papers that had alternatively been attributed to either Alexander Hamilton or James Madison
- Data: Features are counts of function words such as ‘an,’ ‘of,’ and ‘upon’; The outcome is the indicator for the identity (Binary)
  - Why function words: The key feature of these words is that their use by a given author tends to be stable regardless of the topic, tone, or intent of the piece of writing
- Algorithm: Naive Bayes Classifier; Train on a sample of undisputed Federalist Papers authored by either Madison or Hamilton
- Result: Disputed papers were all authored by Madison

[9] Mosteller, F., and Wallace, D. L. (1963). Inference in an authorship problem: A comparative study of discrimination methods applied to the authorship of the disputed Federalist Papers. Journal of the American Statistical Association.

Central Bank Communication[10]

Federal Open Market Committee (FOMC) began disclosing their meeting transcripts after 1993.

- Task: Researchers study how FOMC transparency affects debate during meetings by studying a change in disclosure policy
- Data: FOMC meeting transcripts
- Algorithm: Topic model and difference-in-differences (DiD) regression
- Result: Increased accountability and conformity

[10] Hansen, S., McMahon, M., & Prat, A. (2018). Transparency and deliberation within the FOMC: A computational linguistics approach. Quarterly Journal of Economics.

NOWCASTING

Important variables such as unemployment, GDP, and racial prejudice are measured at low frequency and estimates are released with a significant lag.

- Task: Construct alternative real-time estimates of the current values of these variables.
- Data: Text produced online, such as search queries, social media posts
- Algorithm: Regression (Google Flu Trends project[11]) and Summary Statistics (Corruption estimate[12])
- Result: Google Flu Trends project produces accurate flu rate estimates for all regions approximately 1-2 weeks ahead of the CDC's regular report publication dates

[11] Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., & Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. Nature.

[12] Saiz, A., & Simonsohn, U. (2013). Proxying for unobservable variables with internet document-frequency. Journal of the European Economic Association.

Policy Uncertainty[13]

- Task: High-frequency measure of economic policy uncertainty (EPU) and then estimate its economic effects
- Data: Text from news outlets; An element of the input data c_{ij} is a count of the number of articles in newspaper j in month i containing at least one keyword from each of three categories defined by hand: economy, policy, and uncertainty; The outcome variable v_i is a simple average of counts across newspapers
- Result:
  - The authors validate $\hat{v}_i$ using a human audit of 12,000 articles
  - The resulting human-coded index has a high correlation with $\hat{v}_i$

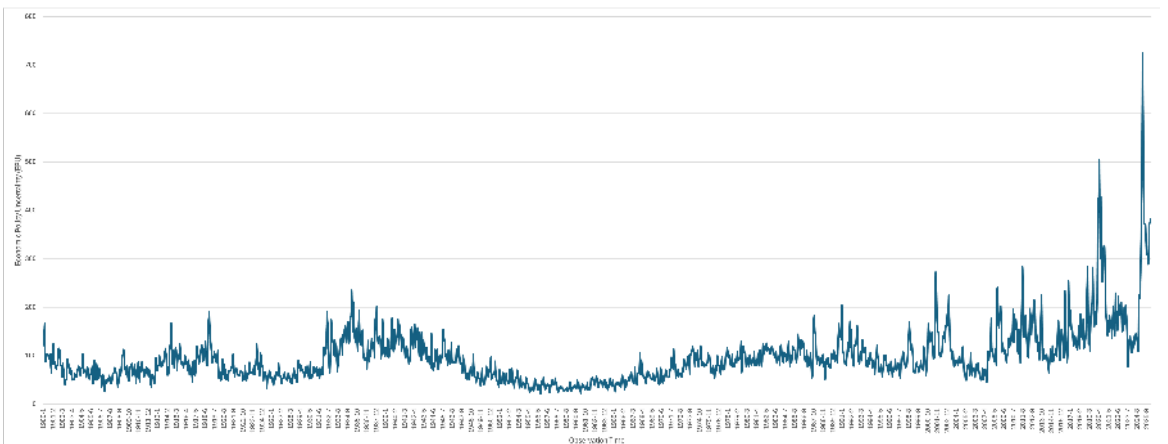
[13] Baker, S. R., Bloom, N., & Davis, S. J. (2016). Measuring economic policy uncertainty. Quarterly Journal of Economics

Effects of Policy Uncertainty

With the estimated $\hat{v}_i$, the authors analyze the micro- and macro-level effects of EPU.

- Firm-level regressions: Measure how firms respond to the uncertainty
  - EPU leads to reduced employment, investment and greater asset price volatility for the firm
- US and international panel VAR models: The increased $\hat{v}_i$ is a strong predictor of lower investment, employment, and production.

Economic Policy Uncertainty



Source: Economic Policy Uncertainty

Market Definition[14]

- Task: A flexible industry classifications that may vary over time as firms and economies evolve
- Data: Product description in the annual 10-K report
- Algorithm: Clustering
  - Distance measure: The cosine scores of the token counts vector
- Results:
  - Three-hundred time-evolving industries
  - Then study the effect of shocks on competition and product offerings

[14] Hoberg, G., & Phillips, G. (2016). Text-based network industries and endogenous product differentiation. Journal of Political Economy.

More on Unstructured Data

- Images: Convolutional Neural Networks (CNN)
- Textual data: Recurrent Neural Networks (RNN), Long-Short Term Memory (LSTM), Transformer
- The foundation and application of Large Language Models (LLMs)

- **FA690 Machine Learning in Finance**