BBVA Research

Disentangling risks, activity and performance based on US corporate reports

Big Things 2019

Creating Opportunities

BBVA Research

Index

01 Why is the use of NLP important in business and economics?

02 The data and methodology

03 The results and main takeaways

Why is the use of NLP important in business and economics?

Text as a key source of information to enrich business and economic analysis

of the total amount of web pages on the internet is given textual or unstructured data

+

Text mining to extract meaning from strings of letters

=

The potential use of textual information and text sources improves the understanding of business and the economy

Why is the use of NLP important in business and economics?

Focusing on business, corporate reports hint plenty of soft information in form of text with a huge analytical potential

US corporate quarterly and annual reports provide a comprehensive overview on a company's activities and financial performance throughout the year and future plans.

2/3

of the information in corporate quarterly and annual reports is textual or unstructured data.

The workflow to extract meaning from text
From the extraction to the analysis

Extracting the data → **Proccesing the data** → **Analyzing the data**

Information extraction

- EDGAR webpage
- 10K and 10Q reports
- SP500 firms
- Get the CIK and SIC codes by firm

Pre-Processing and text parsing

- Extract words
- Tokenization and multi-word tokens
- Stopword Removal
- Stemming
- Case-folding

Transformation

- Text filtering
- Indexing to quantify text in lists of term counts
- Create the Document-term matrix
- Weighting matrix

Text mining and NPL

- Analysis and Machine learning
- Topics extraction (LDA)
- Neural network (Word2Vec)
- Clustering
- Modelling

Sentiment analysis

- Apply sentiment dictionaries
- Semantic analysis and classification
- Clustering

Results

Analyzing US corporate reports narrative: the data

Exploring the main filling reports by the US companies to the SEC

Part I

Item 1. Financial Statements
Item 2. Management's Discussion and Analysis of Financial Condition and Results of Operations
Item 3. Quantitative and Qualitative Disclosures About Market Risk
Item 4. Controls and Procedures

Part II

Item 1. Legal Proceedings
Item 1A. Risk Factors
Item 2. Unregistered Sales of Equity Securities and Use of Proceeds
Item 3. Defaults Upon Senior Securities
Item 4. Mine Safety Disclosures
Item 5. Other Information
Item 6. Exhibits

Quarterly reports

Company's quarterly report which gives an understanding about the changes that could happen in business even before it reflects in earning figure.

Part I

Item 1. Business
Item 1A. Risk Factors
Item 1B. Unresolved Staff Comments
Item 2. Properties
Item 3. Legal Proceedings
Item 4. Mine Safety Disclosures

Part II

Item 5. Market for Registrant's Common Equity, Related Stockholder Matters and Issuer Purchases of Equity Securities
Item 6. Selected Financial Data
Item 7. Management's Discussion and Analysis of Financial Condition and Results of Operations
Item 7A. Quantitative and Qualitative Disclosures About Market Risk
Item 8. Financial Statements and Supplementary Data
Item 9. Changes in and Disagreements with Accountants on Accounting and Financial Disclosure
Item 9A. Controls and Procedures
Item 9B. Other Information

Part III

Item 10. Directors, Executive Officers and Corporate Governance
Item 11. Executive Compensation
Item 12. Security Ownership of Certain Beneficial Owners and Management and Related Stockholder Matters
Item 13. Certain Relationships and Related Transactions, and Director Independence
Item 14. Principal Accounting Fees and Services

Part IV

Item 15. Exhibits, Financial Statement Schedules
Item 16. Form 10-K Summary

Annual reports

In depth analysis of the company to understand its financial state including corporate history, financial statements, earnings per share and any other relevant data.

Analyzing US corporate reports narrative: the data

Focusing on the analysis of financial conditions and forward-looking statements

Assessing risks and uncertainties

Complementing hard data, it gives hints about the future

ITEM 2 (10Q REPORT) ITEM 7 (10K REPORT)

Management's Discussion and Analysis of Financial Condition and Results of Operations

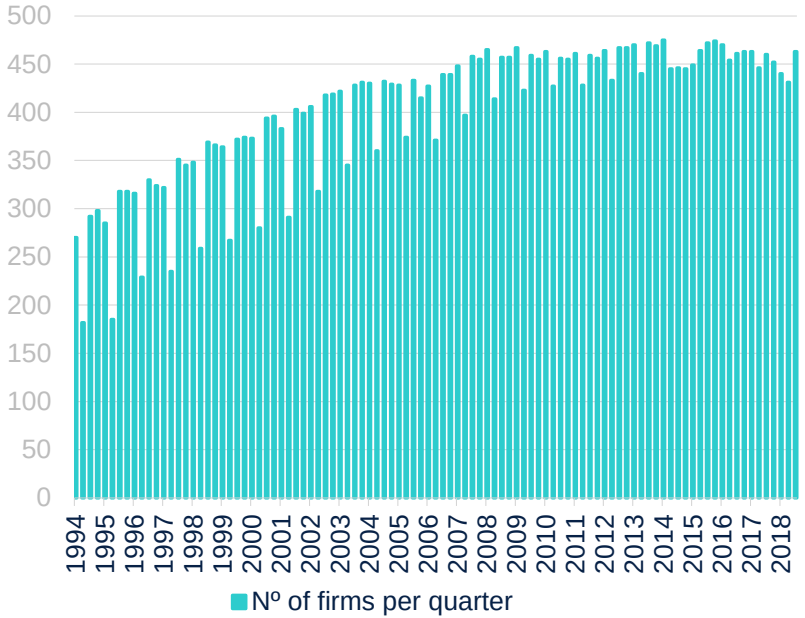
This section and other parts of this Annual Report on Form 10-K ("Form 10-K") contain forward-looking statements, within the meaning of the Private Securities Litigation Reform Act of 1995, that involve risks and uncertainties. Forward-looking statements provide current expectations of future events based on certain assumptions and include any statement that does not directly relate to any historical or current fact. Forward-looking statements can also be identified by words such as "future," "anticipates," "believes," "estimates," "expects," "intends," "plans," "predicts," "will," "would," "could," "can," "may," and similar terms. Forward-looking statements are not guarantees of future performance and the Company's actual results may differ significantly from the results discussed in the forward-looking statements. Factors that might cause such differences include, but are not limited to, those discussed in Part I, Item 1A of this Form 10-K under the heading "Risk Factors," which are incorporated herein by reference. The following discussion should be read in conjunction with the consolidated financial statements and notes thereto included in Part II, Item 8 of this Form 10-K. All information presented herein is based on the Company's fiscal calendar. Unless otherwise stated, references to particular years, quarters, months or periods refer to the Company's fiscal years ended in September and the associated quarters, months and periods of those fiscal years. Each of the terms the "Company" and "Apple" as used herein refers collectively to Apple Inc. and its wholly-owned subsidiaries, unless otherwise stated. The Company assumes no obligation to revise or update any forward-looking statements for any reason, except as required by law.

Analyzing US financial narratives helps to understand the US economic performance, the main opportunities and challenges ahead.

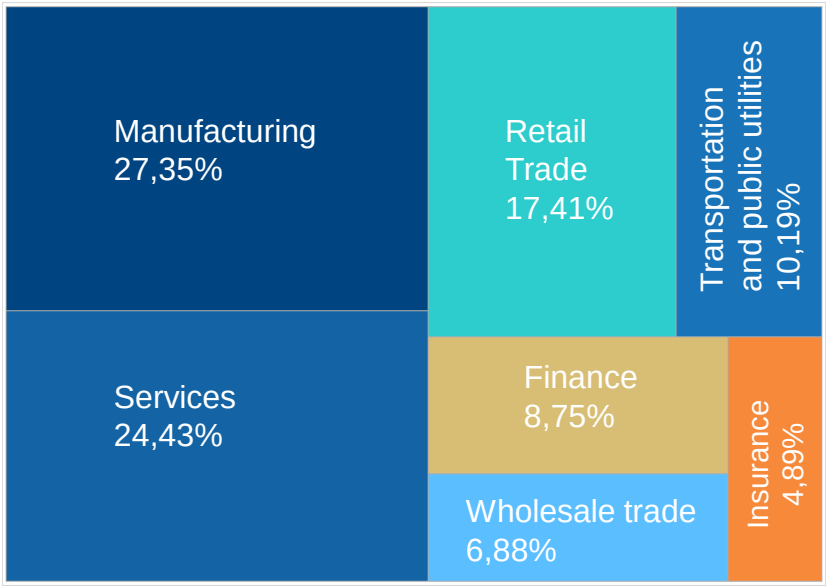
Analyzing US corporate reports narrative: the data

We got more than 21 million corporate files from 1994 to 2019, specially from manufacturing and services sectors

NUMBER OF SP500 FIRMS DISTRIBUITION OVER TIME



NUMBER OF SP500 FIRMS DISTRIBUITION BY SECTOR OF ACTIVITY IN 2000-2019



Analyzing US corporate reports narrative: working with the text
Cleaning and transforming the text

01 Extracting and filtering the text

- Get SP500 10Q and 10K reports
- Get rid of tables, graphs and images
- Focus on the discussion and analysis of financial condition and results

02 Organizing and cleaning the data

- Extract words
- Remove stop words
- Case-folding
- Punctuation

03 Preparing the data for the analysis

- Create the Document-term matrix
- Weighting matrix

04 Converting it into numbers

- Stemming
- Tokenization and multi-word tokens
- Indexing to quantify text in lists of term counts

Analyzing US corporate reports narrative: the methodology
Topic models and Word2Vec

We use two different **approaches** to find hidden semantic structure in documents.

Latent Dirichlet Allocation (LDA)
Unsupervised learning

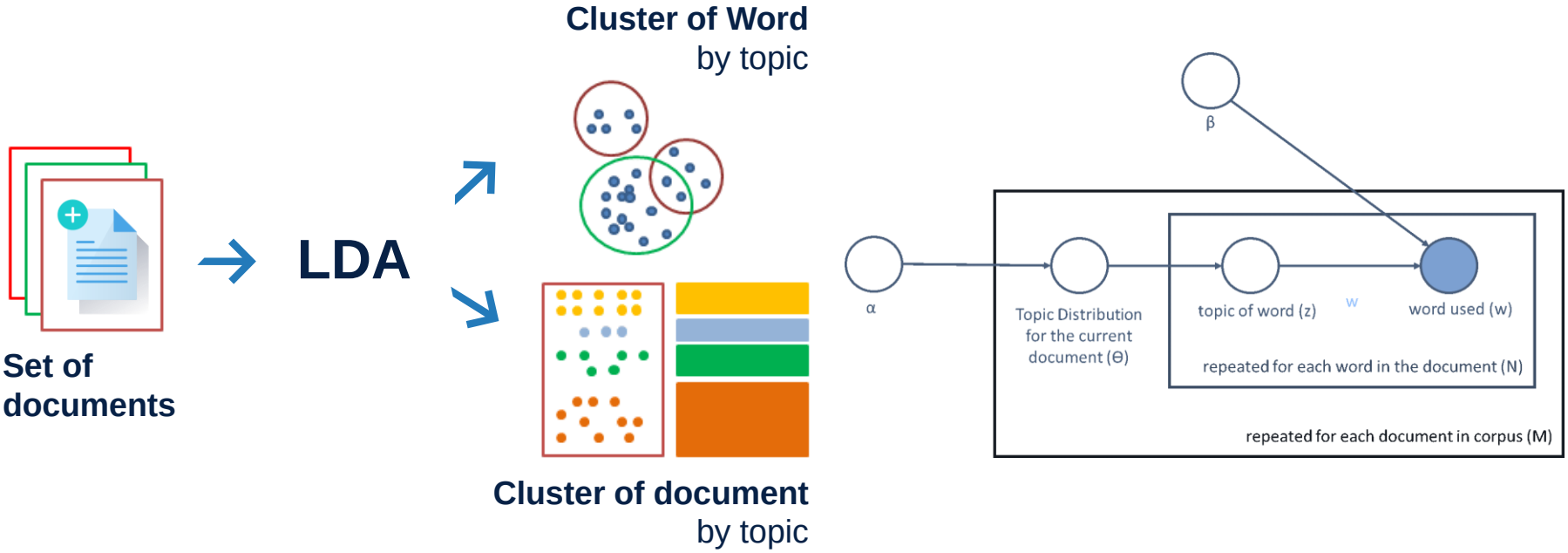
A machine learning algorithm for probabilistic topic modeling that decomposes documents in a set of topics where each **topic** is characterized by a **distribution of words**.

Word2Vec Neural Network
Semi-supervised learning

A neural network to produce word embeddings, trained to reconstruct linguistic contexts of words. It produces a continuous skip-gram architecture where the model uses the current word to predict the surrounding window of **context words**, weighing nearby context words more heavily than more distant context words.

Analyzing US corporate reports narrative: the methodology

Latent Dirichlet Allocation and Dynamic Topic Model



Dynamic Topic Model (DTM)

to introduce time-series dependencies into the data generating process

Analyzing US corporate reports narrative: the methodology
Word2Vec. Skip-gram model

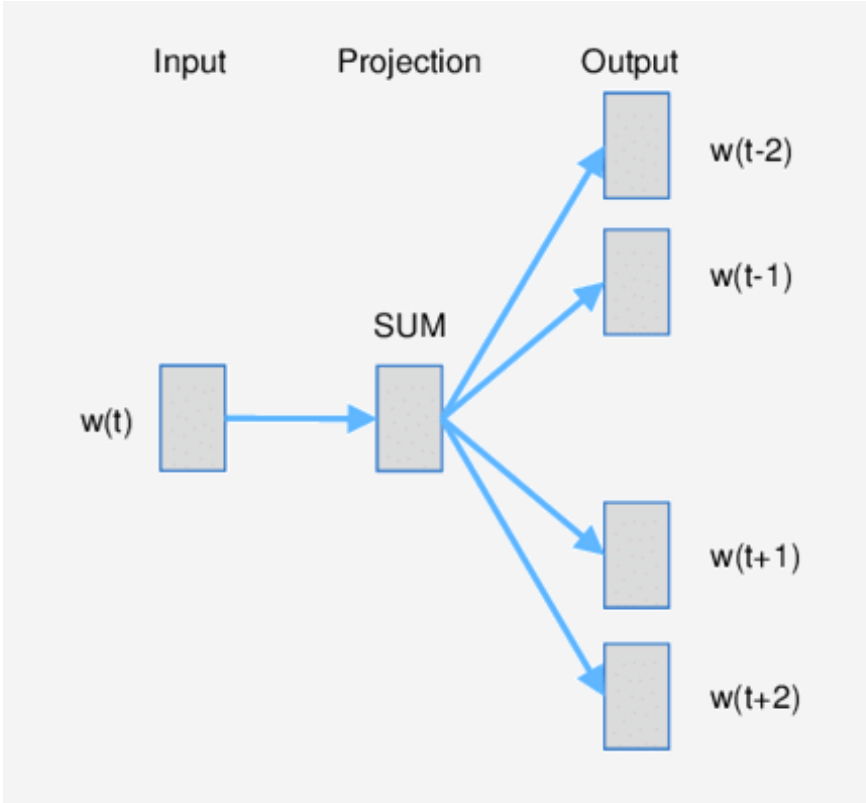
Context word — Target word — Context word

I like natural language processing

I like natural language processing

I like natural language processing

I like natural language processing



Analyzing US corporate reports narrative: the methodology
Sentiment analysis through the lexicon approach

LOUGHRAN AND MCDONALD DICTIONARY
(2011, DEVELOPED TO ANALYZE CORPORATE REPORTS)

Big Data and Data Science techniques allows us to measure and analyze text using natural language processing methodology, also known as text mining or computational linguistics. The information included in the form of text could fully complement and improve our structured databases traditionally used in economic research. This novel approach, which permits to complement and combine traditional economic tools with new emerging tools thanks to the use of Big Data, has plenty of applications with a huge potential for economic research. In this project, we analyze risk interconnectedness, uncertainties and economic performance in US through corporate reports using advanced natural language processing techniques and machine learning. The developed methodologies and tools show a tremendous usefulness to asses risk management and better understand the behavior and driving forces of business performance. There are many signals that precede a collapse, and tracking these trends will help us to anticipate the collapse, becoming an important early warning tool for economic analysis.

Positive words

- ✓ Achieve
- ✓ Benefit
- ✓ Efficiency
- ✓ Outperform
- ✓ Progress
- ✓ Stabilize
- ✓ Strength
- ✓ Versatility

Negative words

- × Bankruptcy
- × Bottleneck
- × Corrupt
- × Downgrade
- × Fallout
- × Imbalance
- × Monopolize
- × Stagnant

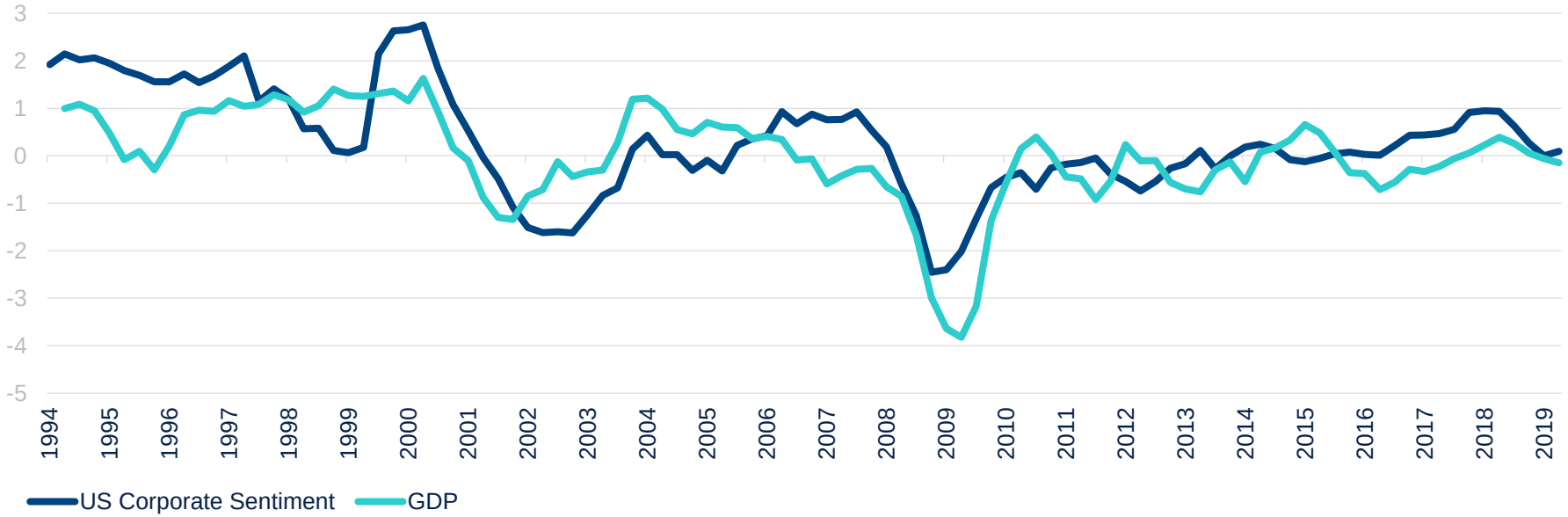
Big Data and Data Science techniques allows us to measure and analyze text using natural language processing methodology, also known as text mining or computational linguistics. The information included in the form of text could fully complement and improve our structured databases traditionally used in economic research. This novel approach, which permits to complement and combine traditional economic tools with new emerging tools thanks to the use of Big Data, has plenty of applications with a huge potential for economic research. In this project, we analyze risk interconnectedness, uncertainties and economic performance in US through corporate reports using advanced natural language processing techniques and machine learning. The developed methodologies and tools show a tremendous usefulness to asses risk management and better understand the behavior and driving forces of business performance. There are many signals that precede a collapse, and tracking these trends will help us to anticipate the collapse, becoming an important early warning tool for economic analysis.

$$\textbf{Average tone} = \frac{\sum Positive\ words - \sum Negative\ words}{\sum Total\ words}$$

Analyzing US corporate reports narrative: sentiment analysis

The sentiment from US corporates narratives mimics the US GDP evolution over time

BBVA RESEARCH US SENTIMENT INDEX BASED ON US CORPORATE REPORTS AND US GDP
(STANDARDIZED, ESTIMATED THROUGH BIG DATA TECHNIQUES FROM 10Q & 10K REPORTS)



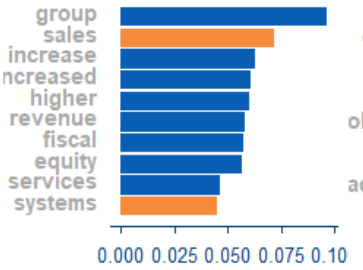
Analyzing US corporate reports narrative: text mining analysis

The US narrative analysis provides a comprehensive image of the main emerging issues through the years

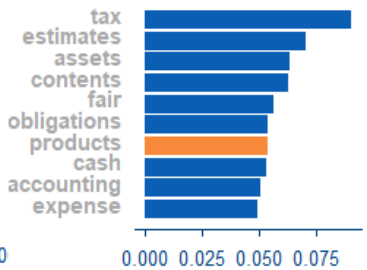
MOST FREQUENT WORDS BY YEAR IN THE ANALYZED DOCUMENTS

(YOY VARIATIONS. WORDS IN BLUE/ORANGE MEANS A POSITIVE/NEGATIVE VARIATION)

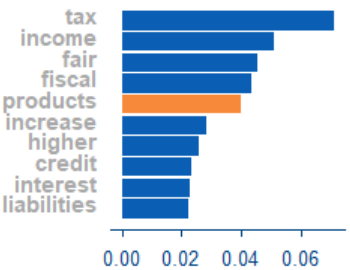
2001



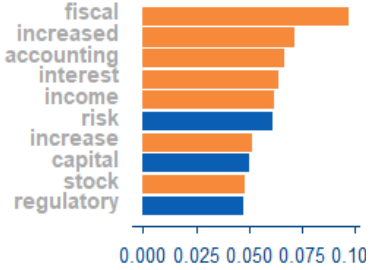
2005



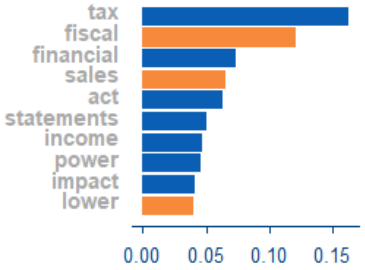
2008



2015

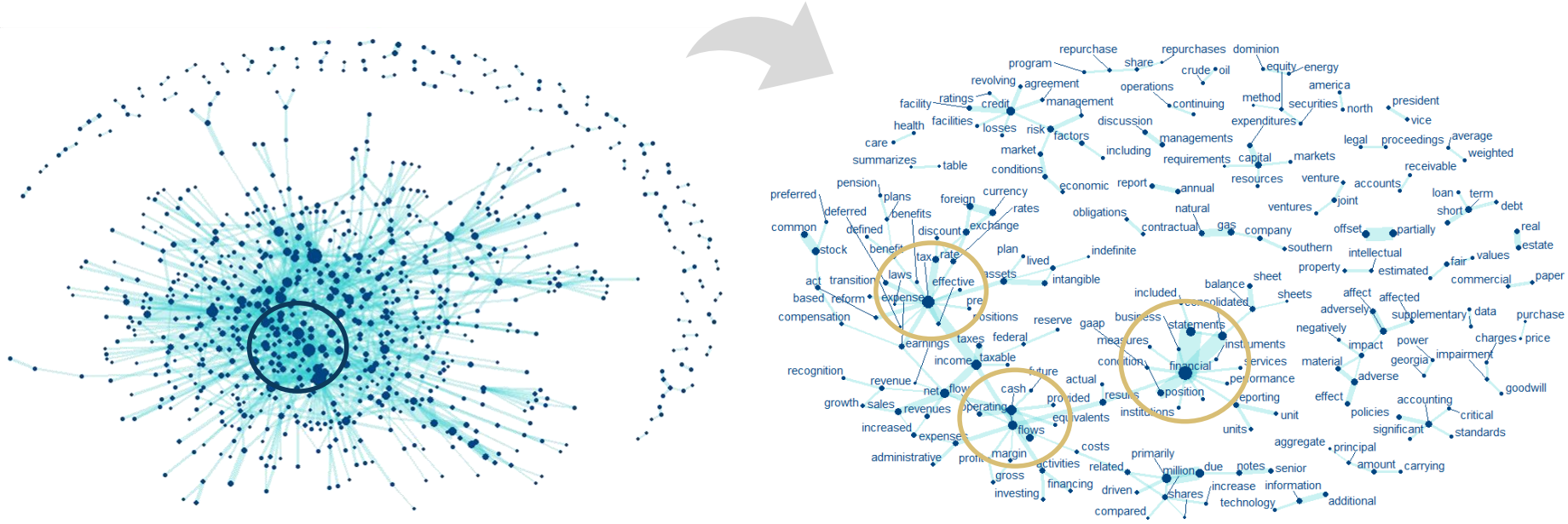


2019



Analyzing US corporate reports narrative: network analysis

More than words: Getting the relation between them…



… to take into account meaningful combination of words …

Analyzing US corporate reports narrative: LDA analysis

… identifying the underlined topics

ENERGY

FINANCE

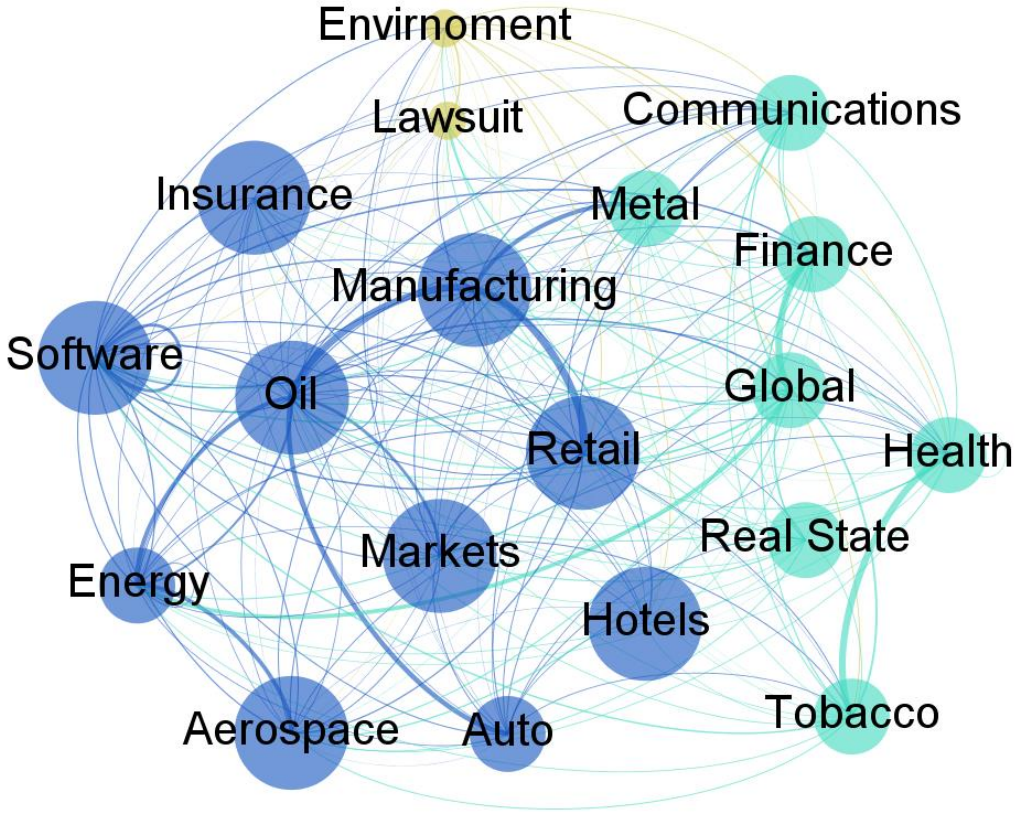
FISCAL

REAL STATE

RETAIL SALES

AUTO

Analyzing US corporate reports narrative: LDA analysis
… and how they interrelate between each other



- The co-occurrence of LDA topics shows **three main clusters** of topics.
- The energy sector, oil and manufacturing are closed related with each other, while global issues is more related with financial sector and the environmental topics with regulation.

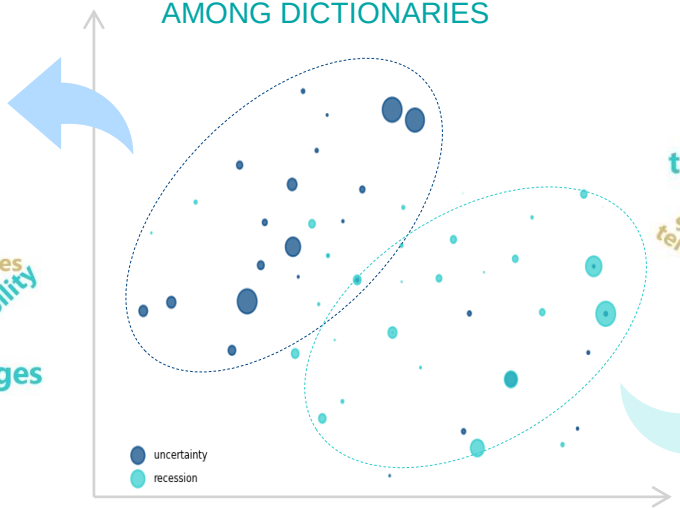
Analyzing US corporate reports narrative: Word2Vec analysis

Focusing on risks and uncertainties, we can also deepen in the context where they are mentioned…

WORD2VEC DICTIONARY FOR RECESSION AND UNCERTAINTY

UNCERTAINTY DICTIONARY

DISTRIBUTION OF WORDS AMONG DICTIONARIES

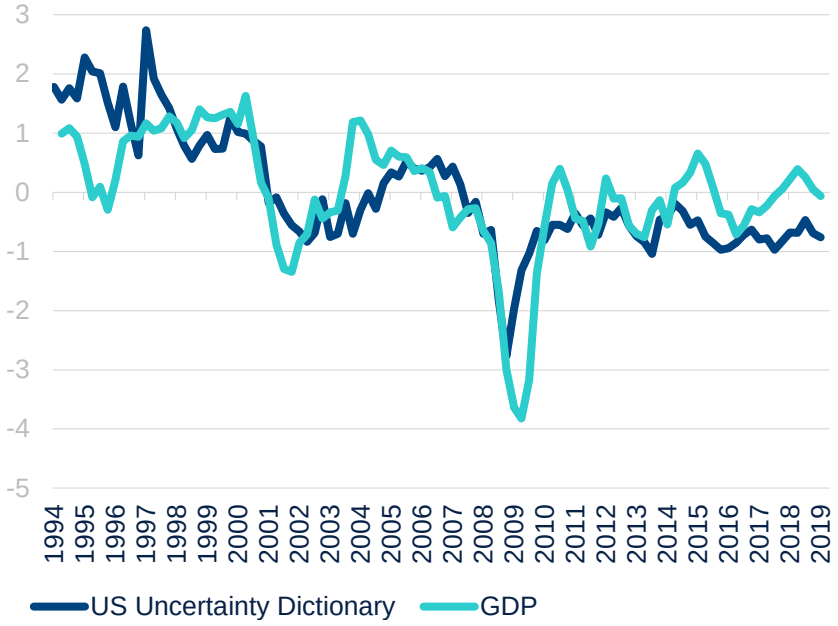


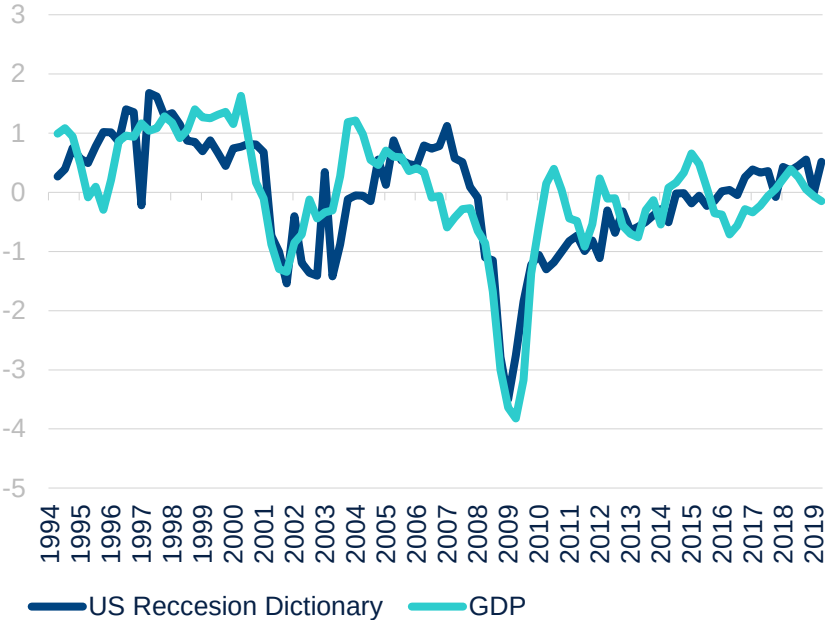
RECESSION DICTIONARY

Uncertainty and recession dictionaries are well differentiated between them, given that the most relevant words in each of them are not common between dictionaries.

Analyzing US corporate reports narrative: Word2Vec analysis
… monitoring its evolution over time

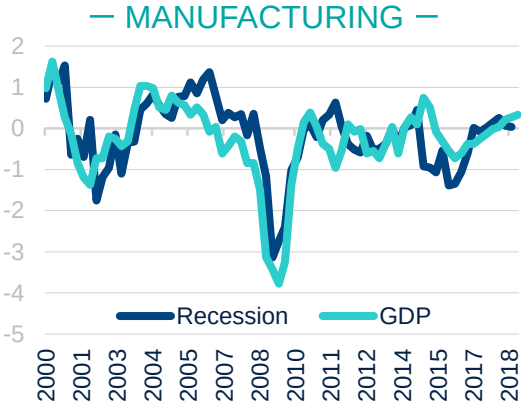
BBVA RESEARCH UNCERTAINTY & RECESSION INDICES BASED ON CORPORATE REPORTS
(STANDARDIZED, ESTIMATED THROUGH BIG DATA TECHNIQUES FROM 10Q & 10K REPORTS)

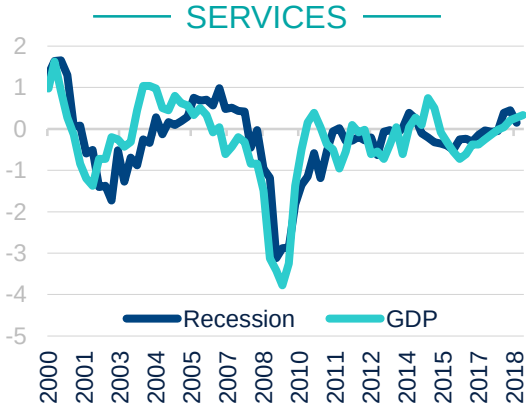


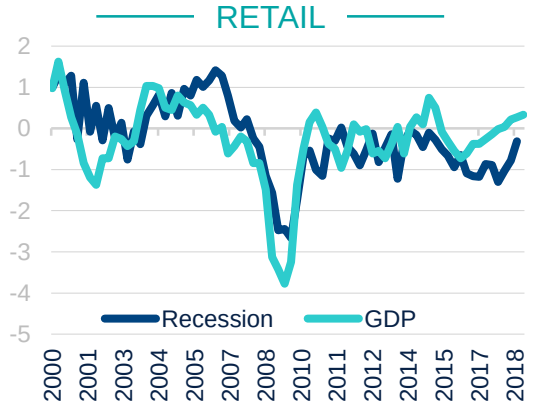


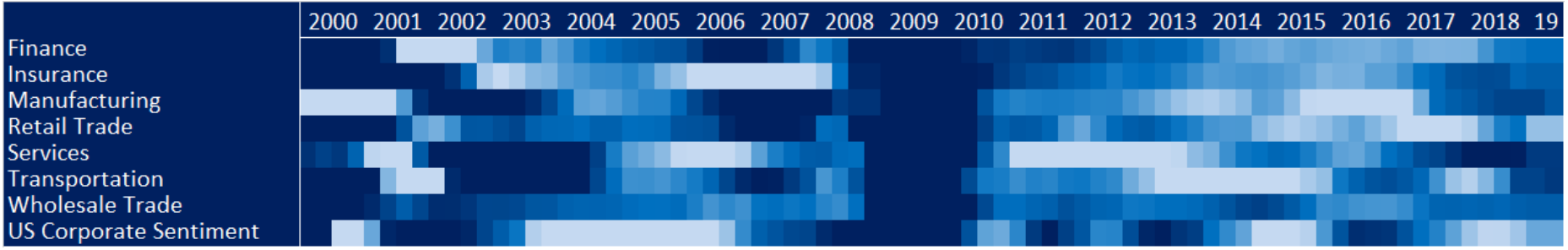
Analyzing US corporate reports narrative: Word2Vec analysis
… disentangling by sector of activity

BBVA RESEARCH RECESSION INDICES BASED ON CORPORATE REPORTS BY SECTOR OF ACTIVITY
(STANDARDIZED, ESTIMATED THROUGH BIG DATA TECHNIQUES FROM 10Q & 10K REPORTS)





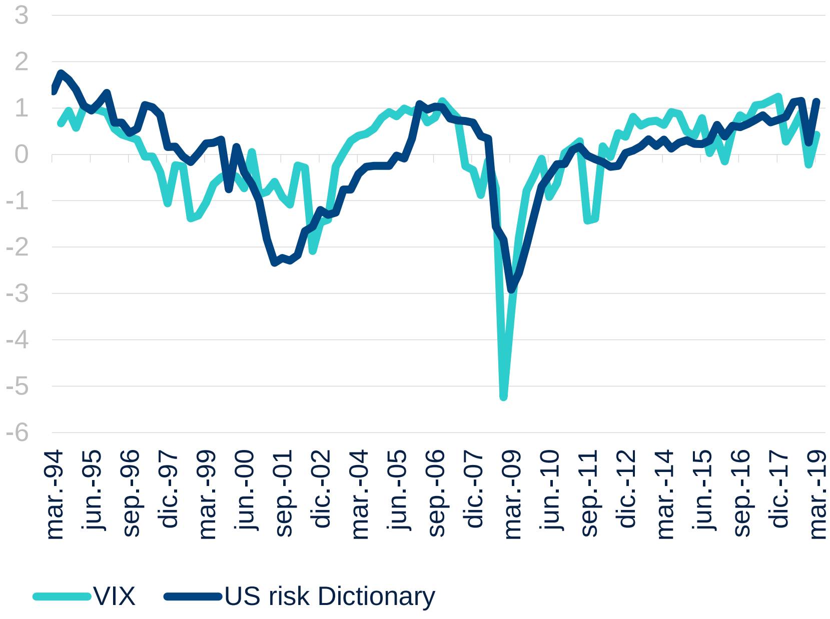




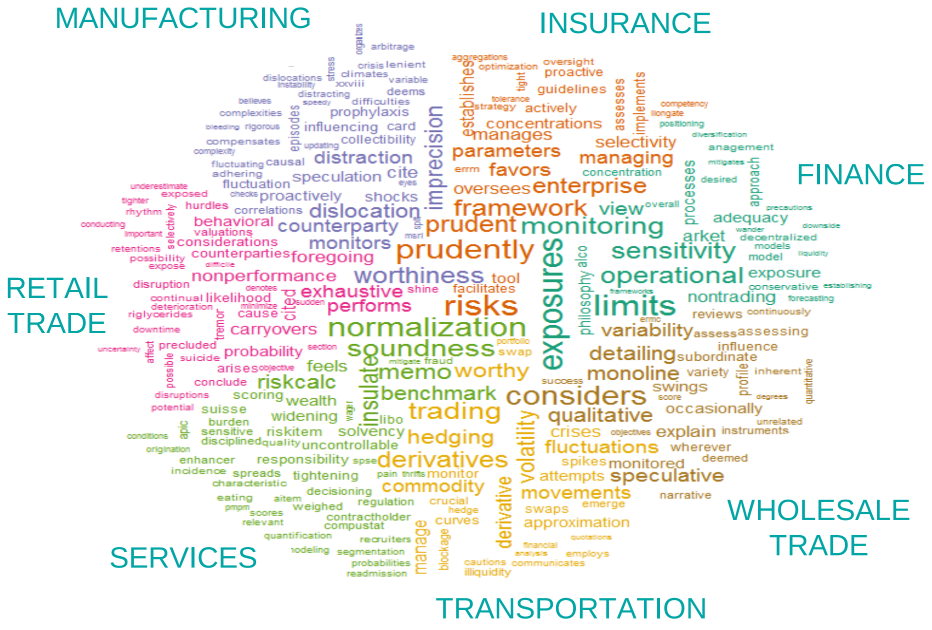
Analyzing US corporate reports narrative: Word2Vec analysis

Risk measures are not equally considered by sector of activity

BBVA RESEARCH RISK INDEX BASED ON US CORPORATE REPORTS AND VIX



BBVA RESEARCH RISK DICTIONARIES BY SECTOR OF ACTIVITY



Analyzing US corporate reports narrative: Word2Vec analysis

Financial sector is more oriented to global issues while the rest of sectors are more related with the economic cycle

MOST RELEVANT WORDS IN RISK DICTIONARIES BY SECTOR OF ACTIVITY

FINANCE

INSURANCE

MANUFACTURING

RETAIL

SERVICES

WHOLESALE

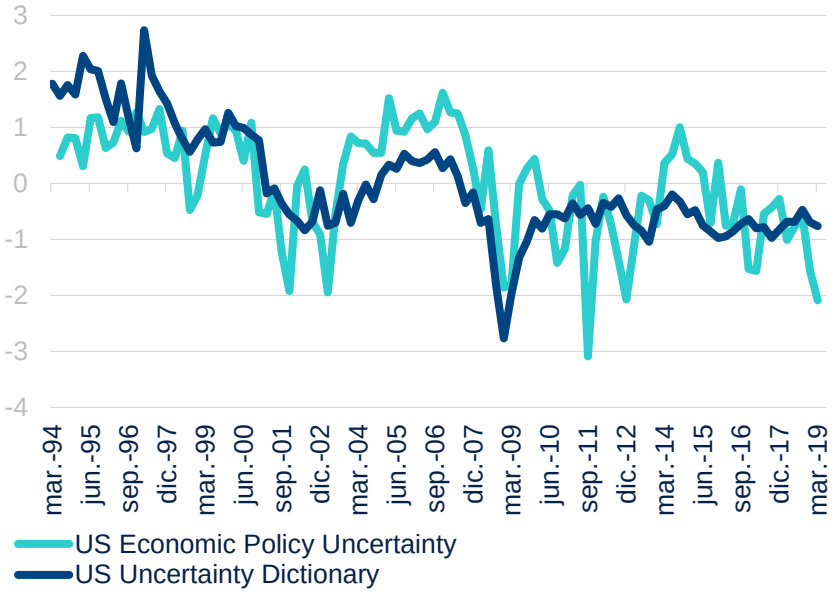
TRANSPORT & PUBLIC UTILITIES

- The analysis by sector of activity of risk dictionaries shows that **finance** and **insurance** narratives are close to each other, while they differ from the rest of sectors.

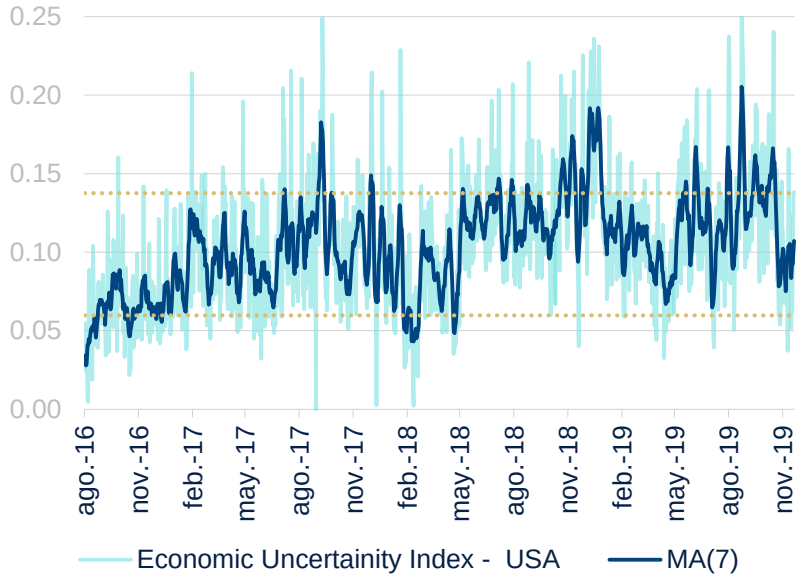
Analyzing US corporate reports narrative: comparing with high frequency data

This historical long-run analysis can be complemented with high frequency data using media information

BBVA RESEARCH US UNCERTAINTY INDEX BASED ON CORPORATE REPORTS AND EPU INDEX
(STANDARIZED, EPU INDEX FROM BAKER ET AL. (2016))



BBVA RESEARCH US UNCERTAINTY INDEX BASED ON NEWS ON DAILY BASIS (NORMALIZED, FROM GDELT DATABASE)



You can find us at:

www.bbvaresearch.com

BBVA Research
Tomasa Rodrigo

@BBVAResearch
@TomasaRodrigo

Thank you!

BBVA Research

Disentangling risks, activity and performance based on US corporate reports

Big Things 2019

Creating Opportunities