



NLP and GenAI in FinTech

Sumedha Rai

New York University | Acorns Grow

Senior Data Scientist

What is Natural Language Processing (NLP) in Artificial Intelligence (AI)?

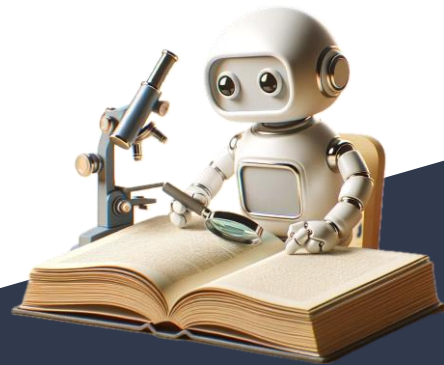
Human Language Understanding

- Teach machines to understand + interpret + generate human language
- Seamless communication between humans and computers
- Analyze VAST amounts of textual data

Two main use cases

TEXT ANALYSIS

TEXT GENERATION



Diverse Applications

Text Summarization

Sentiment Detection

Language Translation

Virtual Assistants/
Chatbots

Some common problems in Fintech



Understanding your customers



Ingesting long documents



Customization:
product/
communication



Important themes
to work on



Fraud detection
and prevention

Sources of text in a fintech firm?

Regulatory and legal
data (data laws/ fintech
laws)

Customer feedback
(complaints/ surveys/
emails)

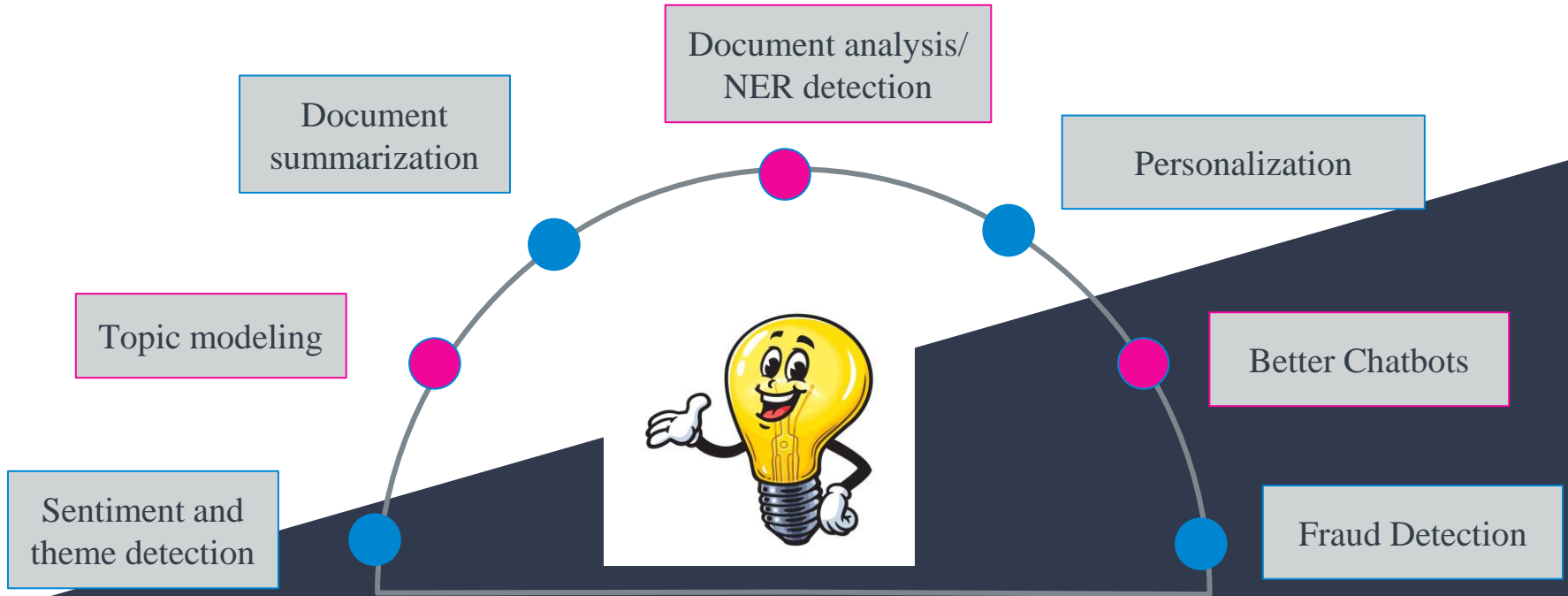
Transactional data
(descriptions)

Company's data
(website/ blogs/
learning platform/
internal)

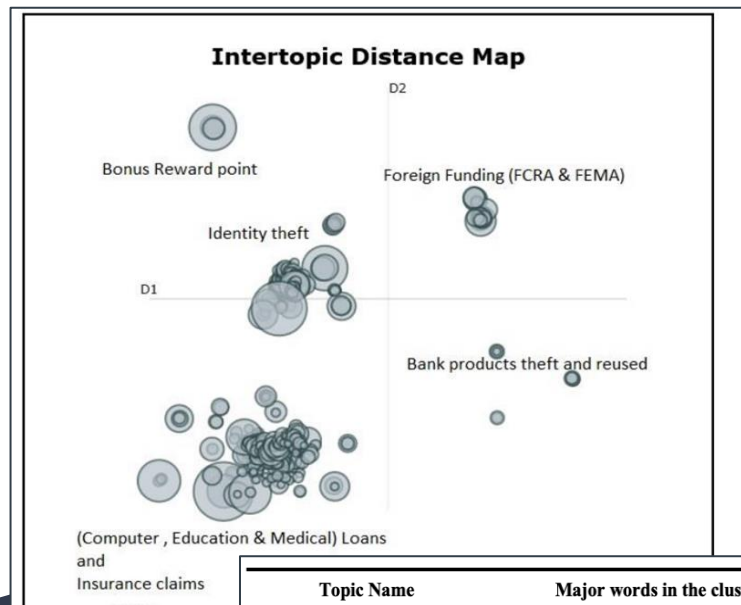
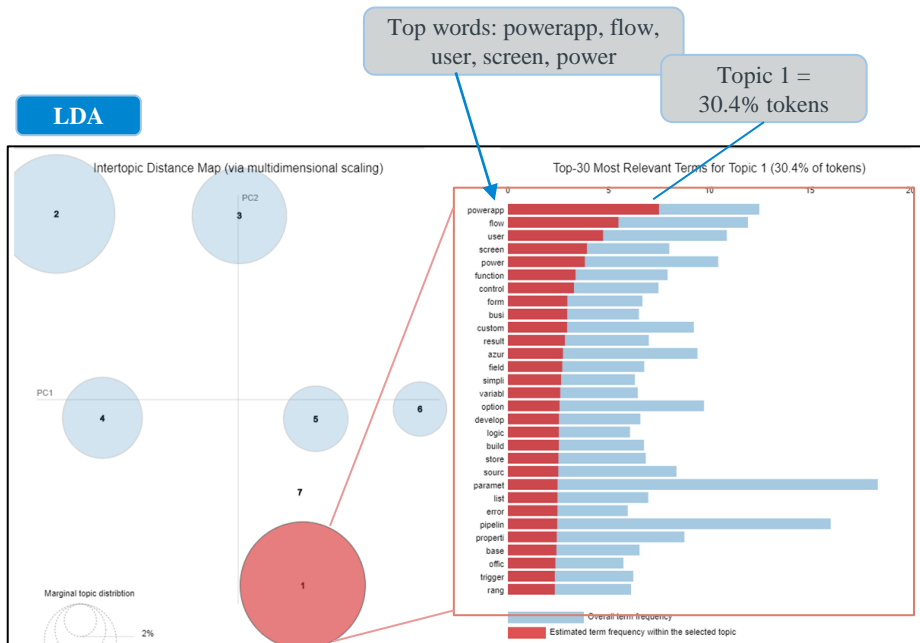


Claims calls (agent
feedback + customer
data)

What are some NLP use cases in fintech firms?



Example: Topic Modeling



BERTopic

Topic Name	Major words in the cluster
Identification Theft	victim, modify, identify, theft, trust, employ, voice, fraudulent
Foreign Funding (FCRA & FEMA)	FCRA, thank, notify, shortage, fay, fema
Bonus Reward points	reward, membership, spend, bonus, promotion, advantage, consecutive
(Computer, Education & Medical) Loans or Insurance claims	rehabilitation, student, course, interest, mortgage, roadlaon, visa, suntrust, pma, college, costco, ira, atm
Bank products theft and reused	reinsert, grave, suffer, inquiry, reopen, grace, theft

* Publicly available dataset of customer complaints about financial firms compiled directly by Consumer Financial Protection Bureau

** Graphs: <https://adatis.co.uk/building-an-lda-topic-model-with-azure-databricks/> (left)

Sangaraju et al [2022], Topic Modelling on Consumer Financial Protection Bureau Data: An Approach Using BERT Based Embeddings (right)

Example: Topic Modeling

Latent Semantic Analysis (LSA)

Uses Singular Value Decomposition (SVD), a linear algebra framework. Reduces the dimensionality of the term-document matrix to capture the underlying structure, helping to identify patterns

Latent Dirichlet Allocation (LDA)

A generative probabilistic model with Bayesian statistics. It assumes that a document is generated by first choosing a distribution over topics and then, for each word in the document, choosing a topic from this distribution and selecting a word from that topic.

LDA2vec

An algorithm that combines Word2Vec's word embeddings with LDA's topic modeling, capturing both local context of words and their topic memberships.

BERTopic (different variations)

A topic modeling technique that leverages the BERT embeddings and c-TF-IDF to create dense clusters of topics, offering contextually relevant topic identification

Example: Entity Recognition

input

& AMZN Mktp US*RZXXXXXXXXX Amzn.com/bill WA

⚡ Inference API ⓘ

🏷️ Token Classification

Examples ▾

AMZN Mktp US*RZXXXXXXXXX Amzn.com/bill WA

Compute

output

in time on cpu: 0.048 s

AMZN Mk **ORG** tp US*RZXXXXXXXXX Amzn.com/bill **WA** **LOC**

</> JSON Output

Maximize



bert-base-NER

Model description

bert-base-NER is a fine-tuned BERT model that is ready to use for **Named Entity Recognition** and achieves **state-of-the-art performance** for the NER task. It has been trained to recognize four types of entities: location (LOC), organizations (ORG), person (PER) and Miscellaneous (MISC).

input

```
sentence = (  
    "Last Friday, John congratulated his friend for getting a job at Google."  
    "They celebrated with a bottle of champagne that cost over 300 dollars."  
    "Later they grabbed some dessert at the Magnolia Bakery in Upper West. "  
)
```

✓ 0.0s

```
doc = nlp(sentence)
```

✓ 0.0s

```
for ent in doc.ents:  
    print(ent.text, ent.label_)
```

✓ 0.0s

Last Friday DATE
John PERSON
Google ORG
over 300 dollars MONEY
the Magnolia Bakery FAC
Upper West GPE

DATE: Dates/ periods of time
PERSON: Individuals (real/fictional characters)
MONEY: currencies/ amounts
FAC: facilities/ buildings
GPE: geopolitical entity (countries, cities, states)

output

spaCy

- Spacy has ~15* entities that it can recognize
- You can customize/ add your own entities on top of the existing entities

* the number may vary depending on the model

Example: Document/ Text Summarization

(a) Extractive Summarization

Source Text: Peter and Elizabeth took a taxi to attend the night party in the city.

While in the party, Elizabeth collapsed and was rushed to the hospital.

Summary: Peter and Elizabeth attend party city. Elizabeth rushed hospital.

(b) Abstractive Summarization

Source Text: [redacted] and [redacted] took a taxi to [redacted] the night [redacted] in the city.

While in the party, [redacted] collapsed and was rushed to the [redacted].

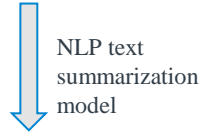
Summary: Elizabeth was hospitalized after attending a party with Peter.

- Extractive can be easier and more accurate
- Abstractive is how humans would summarize (english sentences)

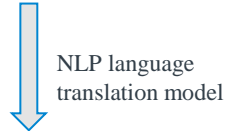
*<https://devopedia.org/text-summarization>

Potential flow

Data Law/ Filing



Summary



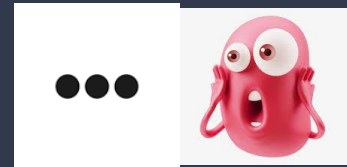
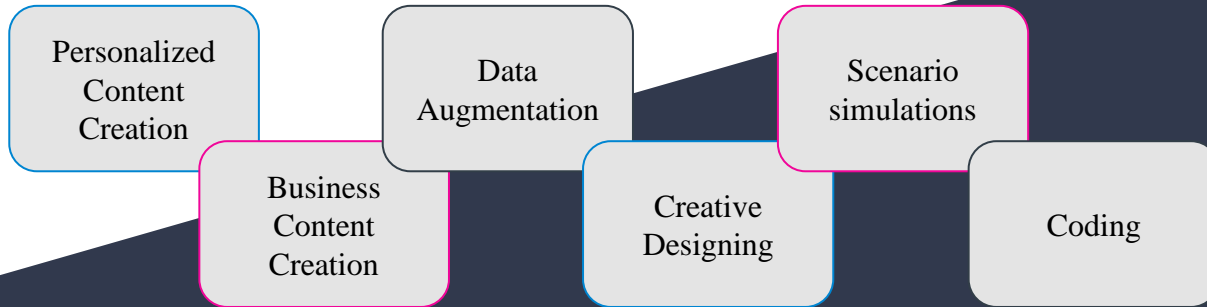
Translated
Summary



- Quarterly Earnings/ Annual Reports
- Data Laws from different countries
- Investment prospectus
- Legal and compliance documents
- Financial news analysis (daily topics!)
- Claims documents
- Industry Reports
- Internal docs

What is Generative AI?

- Type of AI that can generate new content
- It can be text, images, music, speech to text, text to image
- Process:
 1. Learn patterns, styles, sequences + knowledge from very large datasets
 2. Use this learned information to generate new creations that mimic the original data in some way



GenAI in Action

Example: Sentiment Analysis

SR You

Give the sentiments of these statements, with reasons, on a scale of 1-5:

1. I really like the products of this company.
2. I love the services here! They are totally worth it!
3. Totally not worth getting the so called great subscri
4. I feel cheated!
5. I DO NOT LIKE THIS FIRM!

- Out of the box options are quite good for use cases that are not very niche - understanding customer feedback and sentiment
- If numerical values are to be more indicative/ comparative in nature, this can be powerful quick tool for analysis
- Similar analysis can be done for themes and topics

1. **"I really like the products of this company."** - This sentence expresses a clear positive sentiment, but it's not as strong as an expression of love or adoration. On a scale of 1-5, a more fitting rating might be **4** (Positive).
2. **"I love the services here! They are totally worth it!"** - This sentence expresses a very strong positive sentiment. The use of "love" and the emphasis on worth indicate high satisfaction. Therefore, a rating of **5** (Highly Positive) seems appropriate.
3. **"Totally not worth getting the so called great subscription!"** - This is a clear negative sentiment, expressing disappointment and suggesting that expectations were not met. However, the phrasing does not convey as strong an emotion as "I feel cheated!" Thus, a rating of **2** (Negative) might be more suitable.
4. **"I feel cheated!"** - This statement strongly conveys negative emotions, implying a breach of trust or expectations. Such a sentiment is likely to be more intense and warrants a **1** (Highly Negative) rating.
5. **"I DO NOT LIKE THIS FIRM!"** - The use of capital letters emphasizes the negative sentiment, indicating strong dislike. However, it might not carry the same level of personal feeling of being wronged as "I feel cheated!" Therefore, a rating of **2** (Negative) could be more accurate.

Example: Topic Modeling



You

I have the following set of 6 sentences. Classify them in two major themes:



ChatGPT

Of course, please provide the sentences and I'll help you classify them into two major themes.



You

One of the goals of the Act was to set and achieve NAAQS in every state by 1975 in order to address the public health and welfare risks posed by certain widespread air pollutants.

These emission standards are commonly referred to as "maximum achievable control technology" or "MACT" standards.

Online investing comprises wealth management and "roboadvisers", payment-dependent notes and feeder fund investments.

With potential SEC enforcement actions and resulting fines a growing reality, smart fintech companies are seeking guidance with legal compliance, licensing, implementing internal controls, and ensuring their platforms are in line with regulatory standards.

The Office of Water (OW) ensures drinking water is safe, and restores and maintains oceans, watersheds, and their aquatic ecosystems to protect human health, support economic and recreational activities, and provide healthy habitat for fish, plants, and wildlife.

There are a variety of funding opportunities for new and growing businesses in the U.S. market, including equity, debt, convertible notes, and future issuances of digital assets (e.g. simple agreements for future tokens) as a means for growing companies to raise capital.



Example: Topic Modeling



ChatGPT

Based on the content of the sentences, I'd classify them into two major themes as follows:

1. Environmental Regulations and Compliance:

- "One of the goals of the Act was to set and achieve NAAQS in every state by 1975 in order to address the public health and welfare risks posed by certain widespread air pollutants."
- "These emission standards are commonly referred to as 'maximum achievable control technology' or 'MACT' standards."
- "The Office of Water (OW) ensures drinking water is safe, and restores and maintains oceans, watersheds, and their aquatic ecosystems to protect human health, support economic and recreational activities, and provide healthy habitat for fish, plants, and wildlife."

2. Financial Services and Investment Opportunities:

- "Online investing comprises wealth management and 'roboadvisers', payment-dependent notes and feeder fund investments."
- "With potential SEC enforcement actions and resulting fines a growing reality, smart fintech companies are seeking guidance with legal compliance, licensing, implementing internal controls, and ensuring their platforms are in line with regulatory standards."
- "There are a variety of funding opportunities for new and growing businesses in the U.S. market, including equity, debt, convertible notes, and future issuances of digital assets (e.g. simple agreements for future tokens) as a means for growing companies to raise capital."

Give me
two
subthemes

Environmental Regulations and Compliance:

1. Air Quality Regulations:

2. Water Management and Conservation:

Financial Services and Investment Opportunities:

1. Fintech Compliance and Regulation:

2. Diverse Funding Options for Businesses:

Give me two
broad themes
but only 2
words each



ChatGPT

Sure:

1. Environmental Regulation
2. Financial Investment

Zero Shot Prompting

Zero examples of how the task should be solved - LLM relies on the existing knowledge and understanding of language

Few Shot Prompting

Few examples given for model to learn but no updation to model weights; model uses provided context to generate response

You
[clothes, labels, designers]: retail,
[claims, accidents, policies]: insurance,
[money, equity, trading]: finance

what will be the category for the word 'bank'?

ChatGPT
Given the provided categories for the keywords, the word "bank" would most likely fall under the category of "finance."

You
[high, peak, rugged]: mountains,
[green, leaves, bark]: trees,
[sand, water, pebbles]: river

what will be the category for the word 'bank'?

ChatGPT
Based on the provided categories for the keywords:

- "Mountains" for words like "high," "peak," and "rugged."
- "Trees" for words like "green," "leaves," and "bark."
- "River" for words like "sand," "water," and "pebbles."

The word "bank" would likely fall under the category of "river," referring to the sloping land beside a body of water like a river.

Model retains context within a conversation

"Prompting" can be different from "learning"

Context lasts for a session

Example: Named Entity Recognition: LLM + Langchain

Langchain: Framework for developing applications that are powered by LLMs

```
from langchain.chat_models import ChatOpenAI
from langchain.chains import create_extraction_chain
```

✓ 0.0s

```
schema = {
  "properties": {
    "name": {"type": "string"}, "organization": {"type": "string"}, "money": {"type": "integer"},
    "height": {"type": "integer"}, "sentiment": {"type": "string"}
  }
}
```

✓ 0.0s

```
llm = ChatOpenAI(openai_api_key = api_key)
chain = create_extraction_chain(schema, llm)
```

✓ 0.0s

```
input_sent = "Alex, who is a data scientist at Amazon, was recently reading about a million dollar AI deal and was happy to see the progress in AI"
chain.run(input_sent)
```

✓ 0.9s

```
[{'name': 'Alex',
  'organization': 'Amazon',
  'money': 1000000,
  'sentiment': 'happy'}]
```

```
input_sent = "Sarah and John are tired of the long project that they have been working on. \
It's been running for 14 months and could go on for another 2 months. It has cost them five grand already."
```

```
chain.run(input_sent)
```

✓ 1.0s

```
[{'name': 'Sarah'},
 {'name': 'John'},
 {'sentiment': 'tired'},
 {'money': 5000},
 {'money': 14000}]
```

```
input_sent = "X is 5 feet tall and Y, his brother, is 1 feet taller than him."
chain.run(input_sent)
```

✓ 0.8s

```
[{'name': 'X', 'height': 5}, {'name': 'Y', 'height': 6}]
```

Entrepreneurial GenAI in action

1

You

I'm starting a company that provides NLP solutions in fintech. Design a logo for me using my initials S and R. Make sure that the NLP aspect of the company is highlighted in the logo. Also, give me a tagline for my company



For a tagline, how about: "Transforming Finance with Words and Wisdom"? This conveys the essence of leveraging natural language processing to innovate and enhance financial services.

2

You

I need to get a press release out for the launch of my new fintech product that uses AI and NLP to help banks in decision making. Craft two 6 sentence press releases: one with a formal tone and the other with a tone of excitement

ChatGPT

Formal Tone Press Release:

[Your Fintech Company] Launches Cutting-Edge AI-Powered Banking Decision Tool

[City, State, Date] – [Your Fintech Company] announces the release of its innovative AI and NLP-driven platform, revolutionizing banking decision-making processes globally. Leveraging advanced technologies, the platform provides banks with actionable insights, real-time analytics, and predictive modeling capabilities. This transformative tool streamlines operations, ensuring compliance and risk management. CEO [Your Name] notes, "Our AI-powered platform represents a significant leap forward in decision-making technology." For further details, visit [Your Fintech Company's website] or contact our press office at [Press Contact Information].

Excited Tone Press Release:

Revolutionize Banking Decisions with [Your Fintech Company]'s AI Marvel!

[City, State, Date] – Brace yourselves, bankers! [Your Fintech Company] just dropped the mic with its AI-powered decision-making tool! Featuring lightning-fast insights and predictive powers, this bad boy (or girl) is set to turbocharge banking operations. CEO [Your Name] declares, "We're not just disrupting the status quo – we're reinventing it!" Ready for the future? Head over to [Your Fintech Company's website] or contact us at [Press Contact Information] to get on board!

- Everything that we have seen in the LLM chatbot environment can be integrated into a company's code/ internal infrastructure using APIs
- Some of them can also be downloaded freely such as the LLAMA models by Meta

Ethical use of NLP and Gen AI

Bias and
discrimination

Data Security

Data Privacy

Transparency

Hallucinations

Respectful
content
creation

Thank you!
Questions?