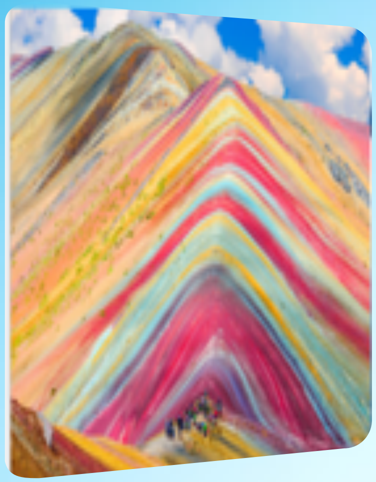
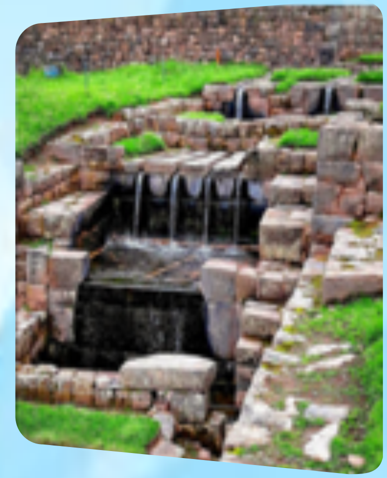






# Generative AI for Consumer Communications: Classification, Summarization, Response Generation

*Nelson Correa, Antonio Correa, Andinum, Inc., USA*
*Wlodek Zadrozny, University of North Carolina at Charlotte, USA*

ANDINUM



IEEE Peru Section

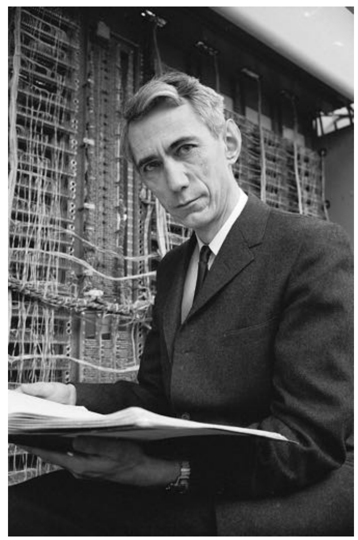
IEEE Andescon
TECHNOLOGY AND INNOVATION FOR ANDEAN INDUSTRY

IEEE

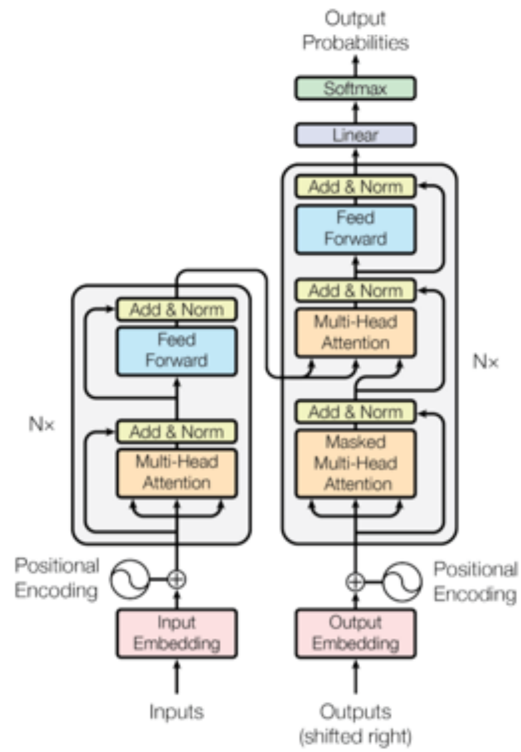
# Generative AI and large language models

*From information theory to the transformer architecture and GPT language models*

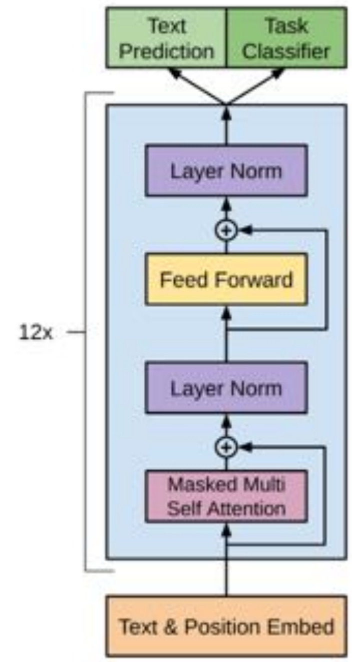
- C. E. Shannon, "*Prediction and entropy of printed English*," Bell System Technical Journal (1951)
- n-gram language models: predict the next character (or word) of a text, based on the previous (n-1) characters (or words)
- Character-based n-gram model from the text of a single book (biography of Thomas Jefferson)
- Google *Transformer* and OpenAI *GPT models* on the right



C. E. Shannon




Google: Transformer Architecture




OpenAI: GPT Model

IEEE Peru Section

IEEE Andescon TECHNOLOGY AND INNOVATION FOR ANDEAN INDUSTRY

IEEE

# Business case

*Automatic handling of customer communications (text, emails, messaging, voice)*

- Businesses must regularly handle large volumes of "*unstructured data*" in the form of text, voice and other modalities
- Large global enterprises: volume of billions of documents per year (social media: trillions)
- Until recently, this volume could be handled only with large scale human input (call centers, customer service representatives)
- Human handling of communications is currently superior, but it is costly, time-consuming, requires training, and suffers from not perfect *inter*- and *intra-evaluator* agreement (i.e., humans performing the task exhibit noteworthy variability)

# Business case

*Example financial consumer communications (U.S. CFPB dataset; 4M+ complaints)*

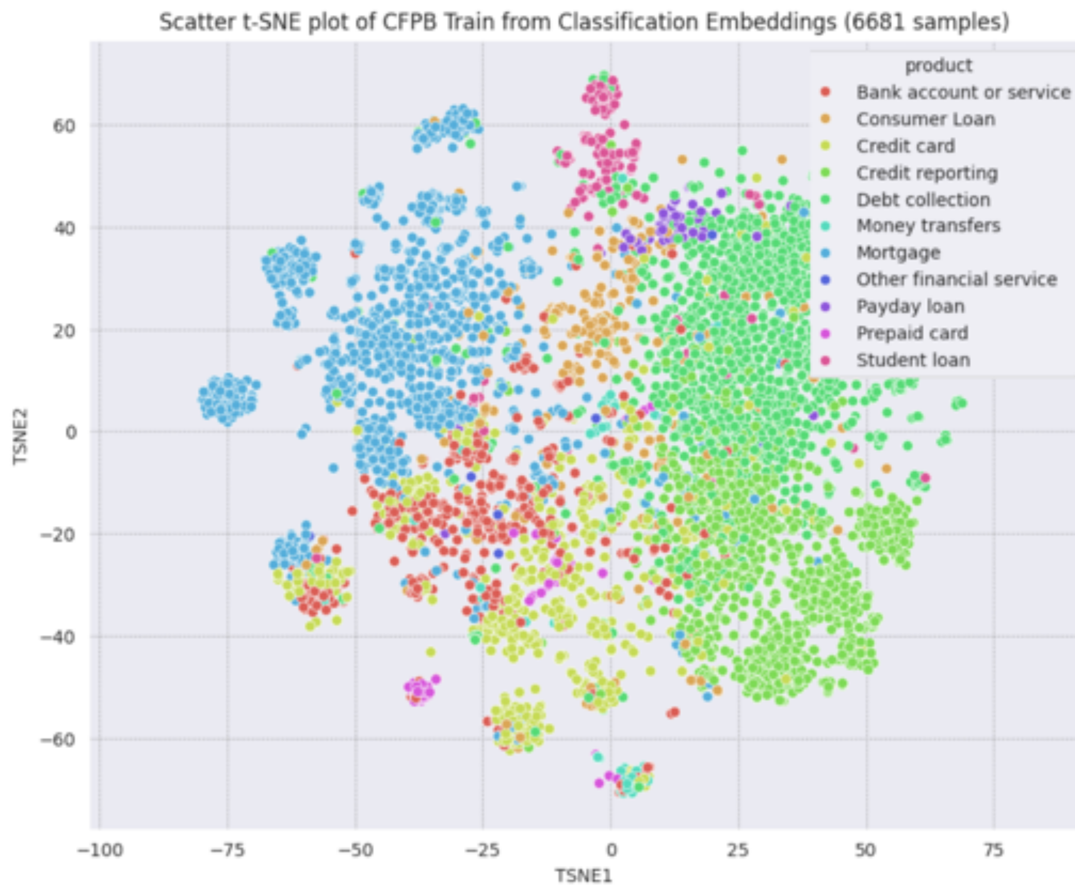
- complaint id: 1290606 (413 words; 2,269 chars)
- Year: 2015 State: New York
- "[Company] used deceptive collection practices while attempting to collect on a purchased debt. Once initial contact had been made and a discussion on solutions to resolve the debt, the attorney office sent out a summons for an appearance in court. [ … redacted] I was hung up on a total of XXXX times while asking to speak to a manager or an attorney within their office. This debt is over XXXX years old and was supposedly incurred by my mother in law who is on a fixed income and under the care of XXXX. [ … redacted] [Company] is using the power and intimidation of the summons and the court system as a scare tactic and collection tool to coerce and discriminate against consumers."

- Complaint classification (for document routing)
  - Product or service: Credit card
  - Issue: Other
- Complaint summary (48 words, 301 chars):
  - [company] sent a summons for court appearance while the consumer was in good faith discussions to resolve a debt, using intimidation tactics to coerce payment. Despite numerous attempts to contact management, the consumer was repeatedly hung up on and denied verification of the debt.
- Complaint response to the customer: …
- CSR instructions: …
- Compliance and regulatory reporting: …

# Data visualization

*Vector embeddings (sparse or dense), dimensionality reduction*

- In modern NLP, texts are represented as dense vector embeddings
- Text vectorization: Sparse vs. dense vector embeddings
  - Keyword indexing (Salton, 1960)
  - SVD, LSA, Word2vec (Mikolov, 2013)
  - LLM vectorization (e.g., 768 with Google Gemini)
- Visualization:
  - Dimensionality reduction, 768 => 2
  - t-SNE, UMAP






IEEE Peru Section

IEEE

# Machine Learning Models: Classification

*Text classification (Gemini LLM 'embedding-001' + multi-layer perceptron)*

- Complaint classification model for CFPB ‘Product’ label, using the Google Gemini 1.0 API without fine-tuning; outputs 768-dimensional embeddings for the input complaints

- Multi-layer perceptron (MLP) with one hidden layer (same dimension as input), a 0.5 ‘dropout’ layer, and a ‘softmax’ output layer with 11 outputs, one per output label class

- MLP model with 599,051 trainable parameters

- CFPB dataset with 80/10/10 train/dev/test split

- MLP trained for 25 epochs; input batch size 32

# Machine Learning Models: Generative tasks

## *Complaint summarization & response: Gemini LLM 'generateContent' model*

- Basic summarization prompt to the LLM:
  - Your task is to summarize the following complaint in {*num_sents*} sentences. Use a maximum {*num_words*} words for the summary. Summarize the complaint succinctly for a customer service representative.

    COMPLAINT: {*complaint*}

    SUMMARY:

- Similar prompt for complaint response generation

- complaint id: 1290606 (413 words)

- SUMMARY 1 (60 words, 392 chars):
  - [company] allegedly used deceptive collection practices by issuing a court summons while negotiating a payment arrangement. The consumer, an advocate for her mother-in-law, claims she was hung up on multiple times and denied verification of the debt. The summons allegedly arrived just days after initial contact, despite prior assurances that no legal action had been filed.

# Model evaluation: classification

*Task 1: Complaint classification*

- Classification Report (Gemini-MLP-Product-11)
  - 88% Gemini classification accuracy
- LLM accuracy is similar to a fine-tuned BERT model with x100 fewer parameters (88.05%)
- Performance competitive with human performance
- See detailed model classification report in conference paper

- Classification model comparison with other models (Correa, 2022)
- Model parameters and classification Accuracy

| Model | Model Parameters | Test Accuracy | Development Accuracy |
|---|---|---|---|
| BoW-MNB | 220,000 | 77.8% | 79.0% |
| BoW-MLP | 2,577,801 | 84.4% | 86.7% |
| Fine-tuned DistilBERT-base | 66,961,931 | 87.05% | 86.86% |
| Fine-tuned ProsusAI FinBERT | 109,490,699 | 88.05% | 87.56% |
| Gemini Pro-MLP no fine-tuning | 7 billion (est.) | 88.00% | 88.47% |

# Model evaluation: summarization

## *Tasks 2 and 3: Complaint summarization and response generation*

- Evaluation based on automatic "text similarity" methods (semantic and pragmatic)
  - Human evaluation with access to reference summary
  - String similarity (Levenshtein or longest common subsequence) too literal
  - n-gram based methods, BLEU in machine translation; ROUGE in summarization
- Character and n-gram based methods (ROUGE, BLEU) have low correlation with human evaluation scores
- New alternative evaluation scores needed: Dense vector similarity; LLMs instructed to evaluate

- Human and ROUGE evaluation
  - Random sample of 50 complaints (15 to 792 words)
  - Human reference summary (target length of 25 words)
  - Human evaluation (33 similar, 14 human, 3 machine)
- Complaint response generation can be evaluated similarly to summarization

| Summarization Scores (F1) | ROUGE-1 | ROUGE-2 | ROUGE-L | Human |
|---|---|---|---|---|
| mean | 0.2700 | 0.0510 | 0.2138 | 0.3939 |
| std | 0.1094 | 0.0680 | 0.0950 | 0.2727 |

# AI Safety, interpretability and explainability

## *Emerging AI regulatory policy and practice standards*

- AI applications bring many potential benefits, along with risks to be considered
- AI models: model cards (Mitchel et al 2019) [45]
- AI technology must be developed and deployed responsibly along a number of dimensions, including: model risk, bias, interpretability and explainability, safety and potential for misuse
- emerging regulatory policy (EU AI Act, 2024) [19] and practice standards (U.S. NIST) [44]

- AI model APIs provide *Safety checking* and *Validation* of *model inputs and outputs*. For Google Gemini [46]
- promptFeedback:
  - Safety checks on model inputs (safetyRatings)
- Outputs candidates: checks along four harm categories, on a categorical probability scale
  - Safety ratings over four categories: Harassment, Hate speech, Sexually explicit, and Dangerous
  - Four discrete probability categories: Negligible, Low, Medium and High

# Conclusion

## *AI has many positive impacts for society, but its deployment poses risks and unknowns*

- Customer communications must be serviced effectively, reliably and cost effectively, at scale, ranging into the billions of documents per year for large organizations
- Presented recent advances in Generative AI and a system for textual complaint classification, summarization and response generation
- Google Gemini large language model
- Presented the CFPB Consumer Complaints Database for evaluation of the machine learning models for the three tasks proposed
- We addressed questions of AI interpretability, explainability, safety and emerging legal AI frameworks

- Presented a system for textual communications in English
- Competitive results
  - Complaint classification, without LLM fine-tuning, at 88% accuracy, competitive with human level of performance
  - Automatic summarization and response generation with LLMs, preliminary study shows competitive performance (72% of machine summaries were similar or better quality than human summaries)
- Future work: Datasets; Multi-modality (voice); multi-linguality; translation