Neural Text Classification for Digital Transformation in the Financial Regulatory Domain

ANDESCON 2022

Nelson Correa, Ph.D.*, Antonio Correa, MBA

Andinum, Inc., New York / West Palm Beach
*ncorrea@ieee.org

IEEE

Who we are

Nelson Correa, Ph.D.

- Electrical engineer / Ingeniero Electrónico
- Computational linguist, software developer
- AI/ML/NLP, Entrepreneur
- ex-IBM, ex-Universidad de los Andes

Antonio Correa, MBA

- MBA UniAndes / Ingeniero Electrónico
- Business development 5G Mobile networks
- 25 years in the Telecom sector: ex-Radiar, Nortel, Alcatel/Lucent, Oracle, Mavenir

IEEE Andescon
TECHNOLOGY AND INNOVATION

AGENDA

Artificial intelligence, Machine learning, NLP and finance applications

- AI/ML/NLP and digital business transformation
- Neural NLP: Attention and Transformers
- Financial NLP: CFPB consumer financial complaints
- Text classification: Baselines and Large Language Models (LLMs)
    - CFPB baselines: Bayes, Perceptron and LSTM classification
    - CFPB LLMs: Transformer models (DistilBERT, FinBERT)
- Model evaluation and responsible AI
- Conclusion & questions

Digital transformation of knowledge work

Text classification use case: Scale and applications

- Knowledge-based businesses: information, know-how, documents and media
  - Essential services must be reliable, economical and accessible to everyone (need human-augmentation through automation)
  - Global business-to-consumer businesses (retail, technology, finance, entertainment, education, etc.) must be delivered to millions and billions of customers ($8*10^9$)
  - Large financial institutions: contact centers can receive 10^7 *or more calls/month*, that must be categorized and routed
- Economic transformation from exploitation of natural resources, agriculture, energy, manufacture and services, to knowledge and intangible assets (often in text documents)
- Digital transformation since the 2000s
  - Logistics and manufacture; big data; IoT
  - Transformation of knowledge work in all verticals. Key to competitiveness
- Document workflows in the modern office and the enterprise
- Most knowledge work is currently done manually by people, assisted by IT solutions

Neural NLP: Attention and Transformers

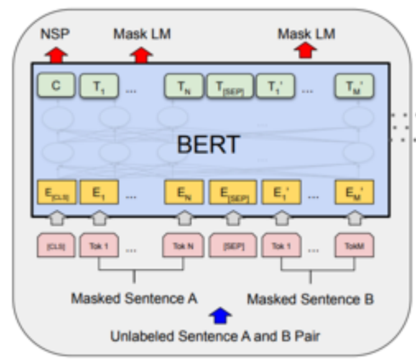
BERT, GPT and Large Language Models (LLMs)

BERT: Devlin et al., 2018

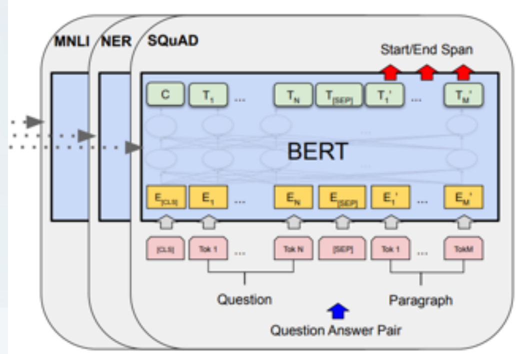
- Encoder of Transformer - Universal encoder
- Masked LM; Next sentence
- Pre-train; Fine-tune; size base, large
- BERT-large: 340M parameters

GPT - GPT-3: Radford, 2018 - 2020

- Decoder of Transformer
- Task: Predict next word; Language prompt
- Zero-shot, Few-shot training
- GPT-2 large: 1.5B parameters
- GPT-3: 175B parameters



Pre-training

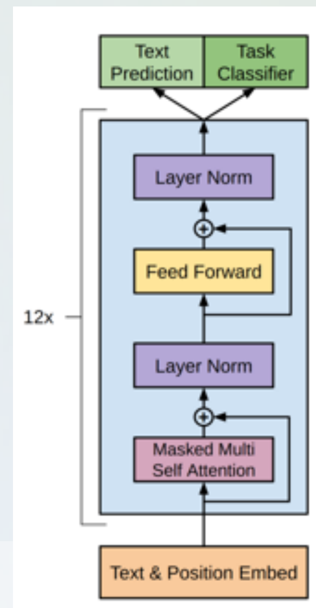


Fine-Tuning

GPT-2 (2019)

Parameters	Layers	d_{model}
117M	12	768
345M	24	1024
762M	36	1280
1542M	48	1600

Table 2. Architecture hyperparameters for the 4 model sizes.



GPT (2018)

CFPB consumer complaints dataset

U.S. Consumer Financial Protection Bureau

CFPB consumer complaints database (CCD), 2011-2022

- Collected since 2011, over 3,000,000 complaints (994,000 in 2021)
- 18 data fields. Structured data fields: Date, State, ZIP, Company, Product, Issue
- “complaint_what_happend” (narrative text)
- 1,000,000 complaints with narrative text

Key data fields

- complaint_id, complaint_what_happend, date_received, company, state, zip_code, product, sub-product, issue, sub-issue

Results here: CFPB dataset from Kaggle, 2016

- 555,957 records; 66,806 non-empty complaint narratives

```
# CSV or JSON files
cfpb_base = "/home/nelson/dataset/regulator/cfpb/"                  # Ubuntu
cfpb_base = "/Users/nelson/dev/datasets/nlp/cfpb/cfpb_kaggle/"      # MBP macOS,
cfpb_kaggle_csv_fn = "cfpbk-consumer_complaints.csv"

print(f"READ path: {cfpb_base}")
print(f"READING ... cfpb_kaggle_csv_fn: {cfpb_kaggle_csv_fn}")
ccd_df = pd.read_csv(cfpb_base+cfpb_kaggle_csv_fn)

print(f"ccd_df.shape: {ccd_df.shape}")
print(f"ccd_df.columns: {ccd_df.columns.to_list()}\n")
```

	date_received	product	sub_product	issue	sub_issue	consumer_complaint_narrative	company_public_response
553086	02/11/2016	Payday loan	Payday loan	Charged fees or interest I didn't expect	Charged fees or interest I didn't expect	I have been paying ($180.00) a month through d...	NaN
553090	03/30/2016	Mortgage	Conventional fixed mortgage	Application, originator, mortgage broker	NaN	I recently became aware that Amerisave Mortgag...	Company believes it acted appropriately as aut...
553096	02/12/2016	Mortgage	Conventional fixed mortgage	Application, originator, mortgage broker	NaN	Bank of America has demonstrated an on-going l...	Company has responded to the consumer and the ...

Sample CFPB complaint

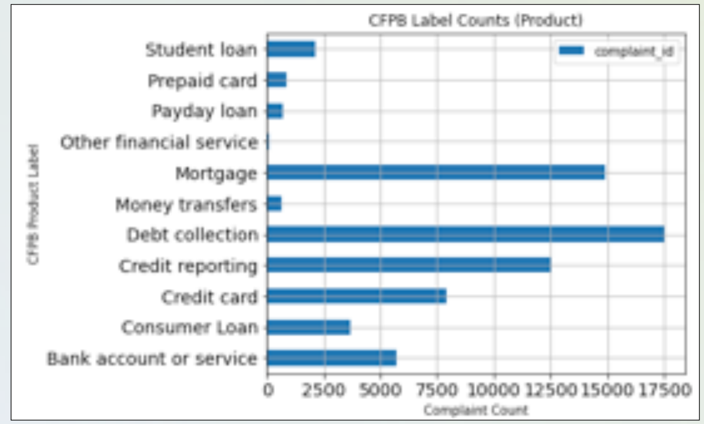
- *“I became a victim of identity theft couple of years ago. After that incident I have noticed many incorrect and unauthorized items appeared in my report. So, I am requesting that you delete all the following ACCOUNTS AND INQUIRIES from my credit report immediately.”*

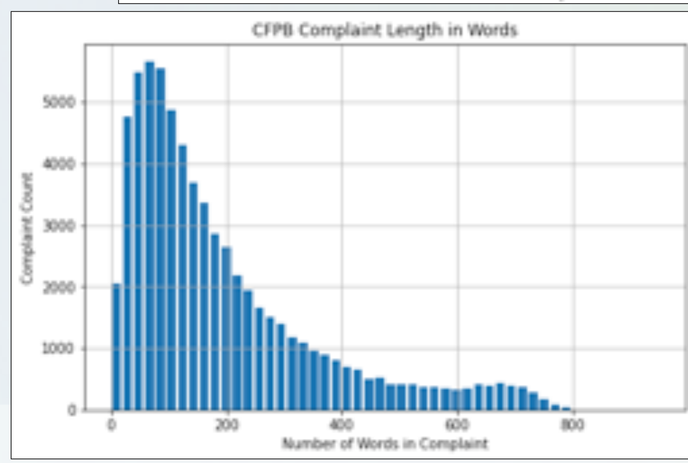
CFPB consumer complaints

Class distribution and text

- "product": 13 consumer financial product and service categories (collectively "product")
- "issue": 90 complaint issue categories
- The class distributions are highly skewed (typical). For "product":
  - "Debt collection" 17,552 (26.27%)
  - "Money transfers" 666 (1%)
  - "Other financial service" 110 (0.15%)

We limit ourselves to "product" below.





Text Classification

Baseline MNB, MLP with BoW-TF-IDF

Vectorizer: Bag-of-Words TfidfVectorizer (BoW-TF-IDF)

- Scikit-Learn library (sklearn)
- ngram = (1, 3); max_features = 20000

Multinomial Naive Bayes (MNB)

- sklearn.MultinomialNB()
- Model Parameters: 220,000; 77.8% accuracy

Multi-Layer Perceptron (MLP)

- Tensorflow-Keras API, 3-layer MLP
- Input (2000x128), Hidden (128), Output (128x11)
- Parameters: 2,577,801; 84.4% accuracy

```
# Model TfidfVectorizer
tokenizer = '(?u)\\b\\w\\w+\\b'
ngram = (1, 3)
max_features = 20000
vectorizer = TfidfVectorizer(
   token_pattern=tokenizer,
   ngram_range=ngram,
   max_features=max_features)

# Model: BoW-MNB
tfidf_mnb_model = make_pipeline(
   vectorizer, MultinomialNB())
```

```
# BoW-MLP model
output_classes = 11 # dataset classes
dense_nodes = 128   # hyperparam
dropout = 0.5       # hyperparam

model = Sequential()
model.add(Dense(dense_nodes,
   activation='relu', input_shape=(20000,)))
model.add(Dropout(dropout))
model.add(Dense(dense_nodes, activation='relu'))
model.add(Dense(
   output_classes, activation='softmax'))
```

Neural Text Classification

Bi-LSTM: Popular 1990s neural model

Long Short-Term Memory (LSTM)

- Word-based, |V| = 20,000
- Input tokens dense vector embedding

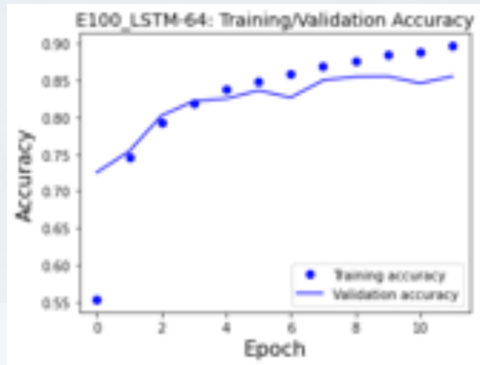
Bidirectional model (TensorFlow)

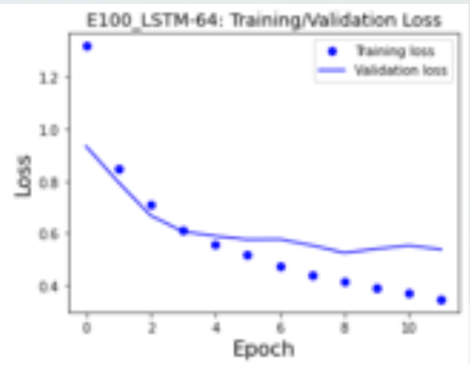
- Input embedding dim-100
- Hidden dim-64
- Model training: 10 epochs

Parameters: 2,093,451; 83.1% accuracy

```
# E100-Bi-LSTM model
lstm_size = 64          # LSTM hidden nodes
embed_dim = 100         # hyperparam
dropout = 0.3           # hyperparam

model = tf.keras.models.Sequential([
    tf.keras.layers.Embedding(
  max_features, embed_dim),
    tf.keras.layers.Bidirectional(
        tf.keras.layers.LSTM(
  lstm_size, dropout=dropout,
  recurrent_dropout=dropout)),
    tf.keras.layers.Dense(
  dense_nodes, activation='relu'),
    tf.keras.layers.Dropout(dropout),
    tf.keras.layers.Dense(
  output_classes, activation='softmax')])
```





Neural Text Classification

Large Language Models - Transformers

DistilBERT, distilled version of BERT-base

- 40% fewer parameters, comparable performance
- Parameters: 66,961,931; 87.05% accuracy

FinBERT Large Language Model

- BERT-base fine-tuned to financial news
- Parameters: 109,490,699; 88.05% accuracy

HuggingFace transformers library

- Models fine-tuned for two epochs
- Batch size 32 on CFPB training data

```
# DistilBERT transformer model
model_ckpt = "distilbert-base-uncased"
num_labels = len(cfpb_product_names)
batch_size = 32

tf_model = hf.TFAutoModelForSequenceClassification.
from_pretrained(model_ckpt, num_labels=num_labels)

tf_model.compile(optimizer=tf.keras.optimizers.
Adam(learning_rate=5e-5),
loss=tf.keras.losses.
SparseCategoricalCrossentropy(from_logits=True),
metrics=tf.metrics.SparseCategoricalAccuracy())
```

```
# FinBERT transformer model
model_ckpt = "ProsusAI/finbert"
num_labels = len(cfpb_product_names)
batch_size = 32

tf_model = hf.TFAutoModelForSequenceClassification.
from_pretrained(model_ckpt, num_labels=num_labels)

tf_model.compile(optimizer=tf.keras.optimizers.
Adam(learning_rate=5e-5),
loss=tf.keras.losses.
SparseCategoricalCrossentropy(from_logits=True),
metrics=tf.metrics.SparseCategoricalAccuracy())
```

Model Evaluation

Models, metrics and performance

Model quality is measured *probabilistically*, in terms of *loss* or in terms of *accuracy* of classification decisions. For text classification and information retrieval, *Precision*, *Recall* and *F1–measure* are the common measures.

- CFPB classification results on 11 categories
- Model accuracy results: 77.8% to 88.05%
  - Baseline 26.27% (17,552/66,806), most likely class
  - Human (IAA): 70%-85% typical for classification
- Model parameters and hyper-parameters
  - Five models presented: 220,000 to 109 million param
  - Hyper-parameters: custom to each model

Model	Model Parameters	Test Accuracy	Validation Accuracy
BoW-MNB	220,000	77.8%	79.0%
BoW-MLP	2,577,801	84.4%	86.7%
E100-Bi-LSTM	2,093,451	83.1%	85.6%
Fine-tuned DistilBERT-base	66,961,931	87.05%	86.86%
Fine-tuned ProsusAI/FinBERT	109,490,699	88.05%	87.56%

```
# FinBERT model hyper-parameters
“attention_probs_dropout_prob": 0.1,
"hidden_dropout_prob": 0.1,
"hidden_size": 768,
"layer_norm_eps": 1e-12,
"max_position_embeddings": 512,
"num_attention_heads": 12,
"num_hidden_layers": 12,
"vocab_size": 30522
```

Model Evaluation

Classification report and Confusion Matrix

Additional model performance analysis tools

- Classification report: Model metrics per classification class
- Confusion Matrix: True vs. predicted class, for each class

	precision	recall	f1-score	support
Bank account or service	0.91	0.81	0.85	193
Consumer Loan	0.90	0.81	0.85	173
Credit card	0.86	0.89	0.87	285
Credit reporting	0.88	0.96	0.92	363
Debt collection	0.87	0.79	0.83	400
Money transfers	0.70	0.54	0.61	13
Mortgage	0.92	0.97	0.94	474
Other financial service	0.00	0.00	0.00	5
Payday loan	0.69	0.44	0.54	25
Prepaid card	0.61	0.94	0.74	18
Student loan	0.83	0.96	0.89	51
accuracy			0.88	2000
macro avg	0.74	0.74	0.73	2000
weighted avg	0.88	0.88	0.88	2000

Fig. 5. FinBERT Classification Report on CFPB Data

FIG. 6. FinBERT Test Confusion Matrix

Responsible AI

Data and model interpretability

Data quality and risks

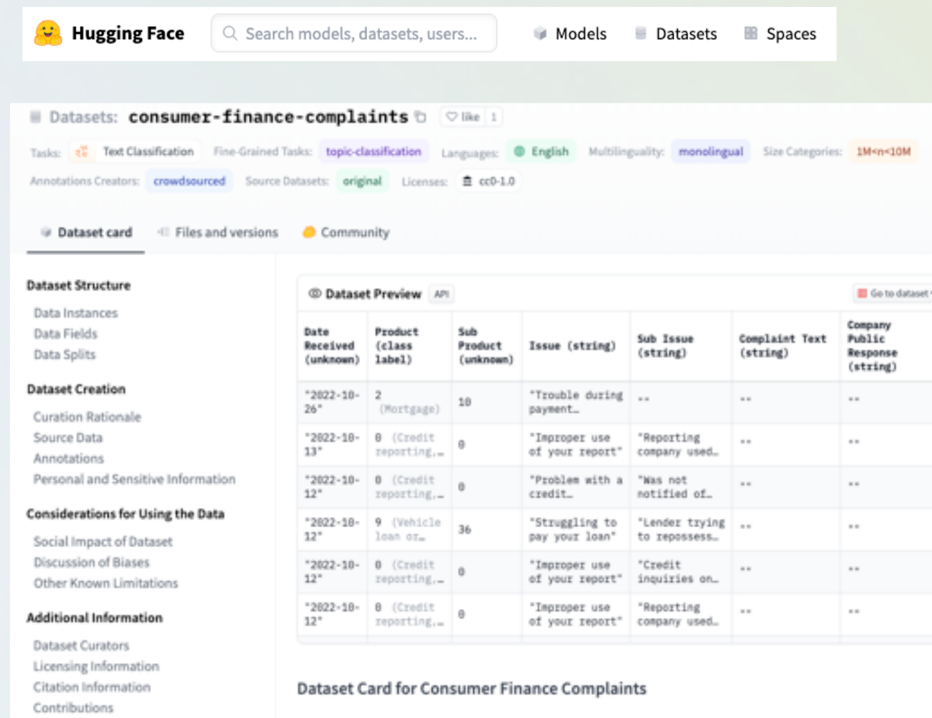
- Data provenance, representativeness, documentation
- Issues: Bias and fairness

Interpretability

- Semantics of dense vector spaces (clusters and labels)
- Model transparency and explainability: e.g., BM25 (transparent) vs. dense models (opaque)
- Interpretabiliy and explainability of AI/ML models are critical and required by emerging regulation

Data and model cards

- Model details, Intended use, Factors, Metrics, Evaluation data, Training data, Analysis, Ethics, Caveats
- https://arxiv.org/abs/1810.03993 (Mitchell et al., 2019)



https://huggingface.co/datasets/consumer-finance-complaints

Conclusion

CFPB Neural Text Classification

We presented text classification, an increasingly important business use case for process automation

- Contrasted traditional feature representations (TF-IDF) and ML models (MNB) to dense vector representations with large neural networks (LLMs) for text classification
- Models: Naive Bayes, perceptron, LSTM models; FinBERT, DistilBERT transformer models
- Used the CFPB consumer complaints database and the Python HuggingFace transformers library
- 88.05% classification accuracy with FinBERT model
- Model risk and ethics considerations, including use of model cards and AI/ML governance

GitHub code and slides

- https://nelscorrea.github.io/andescon2022

ANDINUM

- AI / ML / Natural Language Processing
- Business Process Automation
- Regulatory compliance

Contact: *ncorrea@ieee.org*

Twitter: *@nelscorrea*