WHITE PAPER

ROLE OF NATURAL LANGUAGE PROCESSING IN FINANCIAL SERVICES ORGANIZATIONS

Abstract

Financial services organizations, much like any organization in the world, generate vast amounts of natural language data, be it documents, emails, wikis, FAQs, queries, chat transcripts, comments, application logs, news articles, social media or blogging forum discussions, statements and reports. Traditional programming techniques have proved insufficient at harnessing value due to the complexities inherent to natural languages. A set of statistical approaches, collectively called Natural Language Processing, have evolved to address this challenge.

This whitepaper introduces key NLP tasks, then elaborates how each of them are being applied by financial services organizations in the spheres of handling customers, predicting stock markets, measuring risks, and in banking operations and information technology. Business relevance of NLP applications is provided for each sphere, followed by a comparison of NLP to traditional approaches, a set of key aspects to be kept under consideration, the current state-of-the-art and the path forward.

We believe that financial services firms will gain immensely by applying NLP techniques to the use cases outlined in this paper.

Infosys®
Navigate your next

Introduction

Natural language processing is the application of statistical algorithms to mine unstructured text data. As per Forrester, 60% of global data and analytics decision makers say their company is sitting on 100 terabytes (TB) or more. Where traditional techniques have failed to harness this vast amount of information, NLP techniques are steadily finding newer grounds of applications in bringing value to organizations.

Task	Description	Algorithms
Topic Modeling	Unsupervised extraction of main topics from text	Bayesian SMM for 20 Newsgroups dataset
Text Classification	Assigning predefined categories to documents	XLNet on AG News dataset; Universal Sentence Encoder on TREC-6 dataset
Information Extraction	Extracting structured information from unstructured text	Automated Concatenation of Embeddings (ACE) on CoNLL 2003 (Engilish) NER dataset
Sentiment Analysis	Detection polarity of subjective statements (opinions)	RoBERTa (large) for SST-2 Binary classification; BERT (large) for Amazon review polarity
Text Summarization	Extracting or generating summaries of large documents	BART-RFX for GigaWord dataset
Information Retrieval	Finding documents that match a user's request	Transformer based Sequence denoising autoencoder on CQADupStack dataset
Relationship Extraction	Predicting attributes and relations for entities in a sentence	RoBERTa with adaptive thresholding and localized contextual pooling on DocRED dataset
Question Answering	Answer questions based on reading comprehension	XLNet on SQUAD 2.0 Dev dataset
Machine Translation	Converting one natural language to another	Transformer+BT (ADMIN init) for WMT2014 English-French dataset
Dialogue Understanding	Understanding chat or spoken conversations	BERT-based tracker on Wizard-of-Oz dataset
Text Generation	Generating human-readable text from a given seed	LeakGAN for COCO Captions dataset

Table 1-A: NLP tasks

Task	Description
Text processing	Language detection, optical character recognition, speech-to-text, tokenization
Morphological analysis	Stemming, lemmatization, part-of-speech tagging
Syntactic analysis	Sentence boundary disambiguation, constituency parsing, dependency parsing
Lexical semantics	Distributional semantics, named entity extraction, word sense disambiguation, sentiment analysis
Relational semantics	Relationship extraction, semantic parsing, semantic role labeling, semantic textual similarity
Discourse analysis	Coreference resolution, topic segmentation, textual entailment

Table 1-B: NLP Sub Tasks

Table 1-A provides a summary of NLP tasks listed in the order of relevance to financial services organizations. A set of sub-tasks are listed in Table 1-B. These are lower level activities built on statistical approaches to model the syntax and semantics of a language. The concern with lower level sub-tasks is largely reduced with the advent of high performing pretrained Transformer models such as BERT, Elmo or GPT. These models have been trained on vast amounts of publicly available text data and claim to perform well on a multitude of the NLP tasks. They can be fine-tuned (through a process called transfer learning) on locally available smaller-sized corpus for customizing to the corporate context.

In this whitepaper we study key use cases of NLP in financial services organizations. The opportunities are endless as almost every financial services business process encounters unstructured text in one form or another. The insights from this paper can be applied to all other NLP use cases. A caveat, some infrequently occurring scenarios, such as multilingual text processing, are not covered in this paper.

Handling Customers

Unstructured text data gets involved right from the point of onboarding a new customer (such as KYC documents) and keeps making frequent appearances during the entire period of servicing the relationship (such as handling queries or complaints through emails or chatbots). Today NLP is being used to augment traditional CRM systems in creating 3600 customer views, tracking engagement levels during onboarding or servicing journey, detecting and servicing complaints, or recommending products with a higher look-to-book ratio.

Functional Task	Description	Applicable NLP Tasks	Data Sources
Onboarding and KYC	Extracting business relevant fields as key-value pairs from scanned bank opening and know-your-client documents	Topic modeling, text classification, information extraction	Account opening and KYC documents
Chatbots and query handling	Understanding customer requirements expressed as a dialogue in a chatbot or a query in search window, and providing the most relevant response	Topic modeling, Dialogue understanding, information extraction, Information retrieval	Utterances, queries, policy documents, FAQs
Complaint detection	Detecting customer dissatisfaction with a product or service by tracking sentiment polarity across all channels of communication	Topic modeling, information extraction	Social media, chatlogs, surveys, emails
Product recommendation	Recommending most relevant products or next-best-action based on assessment of customer interactions and prior decisions	Topic modeling, sentiment analysis, question answering	Chatlogs, Knowledge graph

Table 2: NLP applications for handling customers

Banks allow accounts to be opened online. During that process they collect scanned documents as per mandated KYC requirements. Extracting structured information from these scanned documents is a critical missing piece in straight-through-processing of account opening applications. OCR (Optical Character Recognition) plays a big role which we shall cover in a subsequent paper. Traditionally, this has been done by following a templatized approach. However, this is not scalable as many document types (such as passports) do not follow a standard layout template. Assuming data has been digitized through OCR process, the subsequent challenge for NLP becomes discovering and extracting business relevant fields as key-value pairs. For semi-structured documents such as forms, this can be achieved by understanding table structures. Custom NER (Named Entity Recognition) techniques are applied for completely unstructured documents such as contracts.

The ability to detect negative sentiments in customer communications is essential for executing timely interventions to prevent attrition. Sentiment analysis (or opinion mining) finds useful application in other areas such as adverse media for any negative commentary on an organization or its services. Rudimentary coarse-grained models that can classify a sentence according to a predefined set of sentiments are readily available. The first step is to detect subjective statements (or opinions), and then classify polarity using models trained on supervised data. However, getting the sentiment right with a high degree of accuracy requires custom-crafting fine-grained models that can handle the nuances of double-negation, multipolarity, and irony or sarcasm in speech. Another challenge is when customers voice a positive opinion about one aspect but a negative opinion about another aspect in the same sentence. Weaving out the two threads and accurately understanding the sentiment for each aspect comes under the ambit of aspect-based sentiment analysis. This becomes more relevant for dialogue-oriented systems such as chatbots, as multiple aspects tend to get discussed during a single chat session.

Chatbots (and virtual assistants) work by mapping utterances to a predefined set of intents. For voice interactions, voice-to-text is used to generate transcripts, often a source of many errors. Once the user intent is identified, it triggers a preconfigured action associated with that intent. The actions range from retrieving the right response from an FAQ or executing an event in an underlying system, to essentially anything that can be trigged by invoking an API. These retrieval-based chatbots are simplistic agents that work on simple forms of text classification. Considerably more advanced chatbots based on generative principles (where responses get generated and not fetched) are an active research area and we can expect interesting developments on this front.

Typical recommendation systems are based on content-based or collaborative filtering. These approaches depend on tracking user demographics and preferences. More advanced systems such as candidate generation network delve deeper into user behavior by analyzing likes, comments along with preferences.

BERT based models are top performers. Current state-of-the-art (SOTA) model is "NB-weighted-BON + dv-cosine" – a model that uses document embeddings trained with cosine similarity.

Predicting Stock Markets

The ability to predict stock prices will yield higher portfolio returns. Efficient market hypothesis states that the price of an asset (such as stock, forex, commodity) factors in all available relevant information and, in the absence of new information, mimics the behavior of a random walk, i.e., it cannot be predicted. While this remains a highly debated topic, what it does say is that new information can be used to predict future stock behavior. Traditionally investors have harnessed this 'new information' by consuming breaking news or updates from social media to formulate their investment decisions. Today NLP is taking over this function.

Functional Task	Description	Applicable NLP Tasks	Data Sources
Price movements	Predicting future price of an asset (stock, forex, commodity) based on sentiments expressed by people or in media coverage	Information extraction, sentiment analysis	News, social media feeds, discussion forums
Trade reconstruction	Reconstructing events of a trade by correlating data from all sources (including call or chat logs) as per Dodd-Frank act or MiFID II regulations	Information extraction	Chatlogs, call logs

Table 3: NLP applications for predicting stock markets

The advent of machine learning in this field was with applying time series analysis to predict asset prices based on historical trends, which has not been successful given the non-stationary market behavior. The focus has shifted to combining text mining techniques such as sentiment analysis of opinions found in latest news articles, or combined mood related to an asset on social media sites such as Twitter or based on activity on stock message boards.

News articles are processed by extracting entities (such as individuals, organizations, locations) and identifying the relationships between them. Articles, particularly those sourced from multiple feeds, need to be correlated using techniques such as entity disambiguation. A particular challenge is to determine if two correlated articles contain the same information (and thus should be deduplicated), or, while being largely similar, one of them contains additional information. Sentiment polarity is determined for the asset being referenced in the article (if more than one asset is being referenced, then techniques such as aspect-based sentiment analysis become important). Representing extracted information as a graph is a developing area of interest in news analytics. Social media feeds neglect grammatical rules which makes syntactical and semantic analysis challenging. However, it is possible to extract tone and intent of social media comments. Individual sentiments can be aggregated across all sources (such as news, social media sites or forum discussions) to form an overall opinion. Relative weights of different sources can be learnt using machine learning. Besides sentiments, other market signals such as buzz on bond rates or market indexes can be extracted to provide additional sentiment signals. These signals can be combined with financial features to create more robust prediction models.

Prediction of polarity (buy or sell) has been more successful than predicting the quantum of movement. As new information is becoming available daily, such approaches have been more successful for short-term trades than to make long-term investment strategies. In the related area of trading, NLP techniques are being applied on call and chat logs to help banks reconstruct trades within 72 hours as stipulated by the regulations.

The current SOTA model for stock prediction is MAN-SF that applies deep attentive learning on social media and company correlations.

Measuring Risks

One of the earliest statistical NLP techniques was to apply Zipf's Law (frequency is inversely proportional to the rank in a sorted frequency word list) for detecting source of suspicion in volumes of documents (similar to applying Benford's law for detecting fictitious numerical data). Today NLP is being applied to detect and mitigate across credit, third-party, market, legal and other forms of financial risks.

Functional Task	Description	Applicable NLP Tasks	Data Sources
Credit risk assessment	Assessing credit risk for loan requests based on alternate datasets such as social media profile, relevant in particular for thin-file customers	Text classification	Social media
Anti money laundering	Screening for adverse media coverage for an entity (individual or corporate) involved in a monetary transaction as part of AML compliance	Information extraction, sentiment analysis	News, social media
Default or bankruptcy prediction	Analysis of financial reports and social media trends for detecting signs of potential loan default that can lead to third-party risk	Text classification, information extraction	Financial reports, disclosure statements, social media
Corporate fraud	Recommending most relevant products or next-best-action based on assessment of customer interactions and prior decisions	Information extraction	Disclosure statements, conference call transcripts
Insider fraud	Analysis of employee email communication to detect signs of internal fraud	Text classification, information extraction	Emails

Table 4: NLP for measuring risks

Most banks are unable to extend loans to individuals with no-or-low credit history as the traditional datasets do not contain enough information about them to be able to provide a credit score. This is a huge opportunity lost, particularly in low income countries. Today banks are exploring models that can mine social networks for information about individuals' behaviors and associations to make credit decisions. A challenge particular to these models is the regulatory requirement for transparency in decision-making. Techniques such as Lime are evolving that provide a degree of insight into the factors that led to the neural network's decision. We are also witnessing emergence of decision-tree based models that promise complete transparency and ethical decision-making.

Compliance officers rely on harvesting information from news articles about entities in question to determine if it is indeed on the sanctions list and if a transaction should be flagged as AML. News processing is covered under predicting stock markets section. Corporate reports such as quarterly statements contain hidden subtexts that can be mined. Of particular interest to NLP are elaborate text sections such as management discussion and analysis. Models have been trained using weighted word lists to classify 10-K reports as fraudulent or not-fraudulent, which can be used by regulators to detect corporate fraud. Factors such as risk-related words, readability (e.g. Gunning-Fog index), obfuscation in the form of long-winded narratives, or even file size can predict poor performance. Similar approaches have been tried on textual analysis of earnings calls, with the difference between manager and analyst tones suggesting value uncertainty. Extracting sequence of events from financial reports can help predict likelihood of future events, such as bankruptcy. Employee dissatisfaction is a latent factor for insider fraud. Organizations are applying NLP techniques on intra-office email communication as early-warning systems to detect employee dissatisfaction or predicting financial malaise (even length of senior management mails can be a good indicator). This can be achieved in a zero-revelation mode as contents of emails would not be revealed.

FinBERT was released in 2019 for tacking NLP problems specific to finance domain.

Banking Operations

The business of banking is a vast sea of operational processes with unstructured text flowing across various actors responsible for decision-making. These streams of texts – documents, emails, chats, transcripts, queries, descriptions – are all being mined to extract business relevant information and are being employed in achieving business end objectives such as for reducing overheads, eliminating waste and optimizing costs.

Functional Task	Description	Applicable NLP Tasks	Text Data Sources
Information search	Semantically understanding user queries and retrieving most relevant results from a repository of indexed documents	Information extraction, topic modeling, information retrieval, text summarization	User queries, enterprise databases, document management systems
Email automation	Extracting structured information from emails as key-value pairs, and based on the data automatically triggering actions	Information extraction, text classification, sentiment analysis, text summarization	Emails
Ticket triage	Classifying tickets into predefined categories based on information extracted from ticket descriptions	Text classification, information extraction	Descriptions from ticketing systems
Contract reviews	Reviewing contract documents for possible non-compliance, misrepresentation, or for automatically triggering renewals	Information extraction, question answering	Contracts and other regulatory documents
Project projection	Forecasting a project's ability to meet committed targets based on analysis of current trend reports	Information extraction, sentiment analysis	Project lifecycle systems

Table 5: NLP in banking operations

Information search is an immensely relevant area built on a spectrum of NLP techniques. Documents are transformed into vectors that retain their essence. While this is straightforward for sentences or maybe even for paragraphs, it becomes a challenge for long documents. Text tiling can be applied to split documents into sections that deal with different topics, with each section getting vectorized separately. Understanding the meaning of query keywords in the full context is critical for semantic search. Query expansion is applied for broadening the search scope. Precision (retrieving accurate results) vs. recall (retrieving at least some results) tradeoff must be carefully evaluated to design an information retrieval system optimal for the context.

Emails are interesting as they allow juxtaposition of extracted information with associated email metadata for triggering automated actions such as auto-forwards.

Information Technology

Lastly, we touch upon two interesting areas of application in the domain of information technology – both very relevant to banking organizations.

Functional Task	Description	Applicable NLP Tasks	Data Sources
Rule extraction	Extracting semi-structured rules from documents, typically as part of forward engineering track of a legacy modernization program	Information extraction, text generation	Requirements specification documents
Code translation	Converting code from one programming language to another, typically as part of a legacy modernization program	Text generation	Code repositories
System resource projection	Extracting information from semi-structured system logs for visualization and detecting threshold crossovers	Information extraction	System logs
Operational failure prediction	Extracting information from dynamic application logs and mapping to static program dependency charts for predicting potential downstream failures	Information extraction	Application logs, enterprise knowledge graphs

Table 2: NLP applications for handling customers

Legacy modernization is a vital concern for financial organizations. As most banks look to modernize their legacy platforms, automatically converting legacy code (such as Cobol) to modern languages (such as Java) is an active area of interest. This is typically done using neural machine translation, a technique similar to what is applied for converting one natural language to another (such as English to French). Challenge with this technique has been that it requires a parallel dataset (code in both source and target programming language that does the same thing), and such datasets are not readily available. A monolingual technique called TransCoder has recently been released which claims to give better results than the benchmark of rule-based code translation. Logs are mined for real-time system or application updates. Combining this extracted information with static dependency charts can help predict downstream failures and take preventive actions, a must for creating resilient systems.

Conclusion

This paper studied the importance of NLP for the BFS industry in five major areas: handing customers, predicting stock markets, measuring risks, optimizing banking operations and in information technology. We believe that Financial Services firms will gain immensely by applying NLP techniques to the use cases outlined above. In subsequent papers, we plan to cover the role of machine leaning, speech and computer vision processing in FS organizations.

About the author

Amitab Manu, *Delivery Manager, Leading Innovation for Financial Services*

- Leading innovation for financial services vertical.
- Exploring technology frontiers for readying ourselves for the future.
- Developing intelligent solutions to solve challenging client problems.

References

- Forrester Wave report on AI-based text analytics platforms during Q2, 2018
- https://paperswithcode.com/sota for state-of-the-art NLP models

For more information, contact askus@infosys.com

Infosys®
Navigate your next