Session 3

Life/Health Insurance technical session

Anilraj Pazhety
Life Health Technical Session

ANILRAJ PAZHETY  MS (BUSINESS ANALYTICS), MBA, BE (CS)
Data Innovation Specialist
Asian Markets

Agenda

- Big Data in Life Insurance
- Natural Language Processing (NLP)
  - Convert text to machine readable format
  - Model Framework for a text classifier
- NLP Applications in Life and Health Insurance
Big Data in Life Insurance

Life insurers are lagging behind when it comes to embracing the benefits of big data.

Life insurers collect a substantial amount of data during the application process.

Post Issuance: Limited customer interaction → Limited data during policy life cycle.

Changing dynamics due to increase in customer touch points.

Customers → Application form → Insurer

Data sources being sought and used within advanced analytics applications:

Desirable Attributes: Relevant, Granular, Voluminous, Accurate, Timely, Permissioned.

Potential data sources: Credit bureau, Wearable devices, Bank transactions, Social media, Credit card transactions, Retail transactions, Marketing, Mobile usage, Non-life, Life assurance.
Natural Language Processing (NLP) in Life / Health Insurance

NLP aims to develop algorithms which process human language – Written or Oral

Raw text data is available from a wide range of sources in Life / Health insurance:
- Medical records
- Websites
- Prescriptions
- Customer Care
- Chatbots
- Agent Notes

- BIG potential for insurers to leverage data from these sources to derive information which can drive intelligent data analysis and improved decision making
- For achieving any level of artificial intelligence it is imperative to have machines to process text data
Text Pre-processing

- Text is the most unstructured form of data and hence needs pre-processing to transform it into intelligible format.

**Raw Text** → **Remove Stop words and Punctuations** → **Tokenization** → **Lemmatization / Stemming** → **Clean Text**

- Language: The, of, was, are, is
- Location: Hong Kong, Jakarta
- Time / Numeral: Weekdays, Year
- Domain Specific

**Breaking a sentence into single words (Tokens)**

**Suggested Paracetamol three times a day**


The goal of both **stemming** and **lemmatization** is to reduce inflectional forms and sometimes derivationally related forms of a word to a common base form.

**Stemming**
- Car, Cars, Car's → Car

**Lemmatization**
- Am, are, is → Be

**Document Term Matrix (DTM)**

- Most basic component in text analytics
- Machines understand only numbers. DTM is a numeric representation for a given text after tokenization

<table>
<thead>
<tr>
<th>Documents</th>
<th>Term 1</th>
<th>Term 2</th>
<th>Term 3</th>
<th>Term 4</th>
<th>Term 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Doc 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Doc 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Doc 3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Doc 4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Doc 5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

A two dimensional matrix whose rows are terms and columns represent each document. Hence, each entry \((i, j)\) corresponds to term \(i\) in document \(j\).

**Term Frequency (TF) matrix**

- Simple technique to identify relevance of a word in a given document
- The more frequent the word is the more relevance the word holds in the document

\[
TF(t) = \frac{\text{Number of times word } t \text{ appears in a document}}{\text{Total number of words in the document}}
\]

**Inverse Document Frequency (IDF) matrix**

- Based on the principle that less frequent words are more meaningful

\[
IDF(t) = \log \left( \frac{\text{Total Number of documents}}{\text{Number of documents with word } t \text{ in it}} \right)
\]
Term Frequency (TF) Inverse Document Frequency (IDF) - TFIDF matrix

- Product of TF and IDF
- If a word appears multiple times in a document then it should be more meaningful than other words BUT if a word appears many times in a document but also in many other documents then it may be a stop word or a frequent word in that particular domain

\[ TF = IDF(t) = TF \times IDF \]

Ngrams

It is just a sequence of N words

<table>
<thead>
<tr>
<th>Unigrams</th>
<th>Bigrams</th>
<th>Trigrams</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exercise</td>
<td>Exercises daily</td>
<td>Yoga exercises daily</td>
</tr>
</tbody>
</table>

Increase in Information

Word Embedding – Vector Space Models

The underlying idea is to represent documents as matrices or arrays

This facilitates to represent documents geometrically

<table>
<thead>
<tr>
<th>Prescription</th>
<th>Ativan</th>
<th>Codeine</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>7</td>
<td>9</td>
</tr>
<tr>
<td>2</td>
<td>10</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>9</td>
<td>6</td>
</tr>
</tbody>
</table>

Enables document comparison mathematically
Comparing Documents

Definition of Cosine Similarity

\[
\cos \theta = \frac{A \cdot B}{\sqrt{(X^2 + Y^2)_A} \cdot \sqrt{(X^2 + Y^2)_B}}
\]

\[
\text{Cosine Similarity (P1,P2)} = \frac{(7\times10) + (9\times2)}{\sqrt{7^2 + 9^2} \cdot \sqrt{10^2 + 2^2}} = 0.75
\]

\[
\text{Cosine Similarity (P1,P3)} = \frac{(7\times9) + (9\times6)}{\sqrt{7^2 + 9^2} \cdot \sqrt{9^2 + 6^2}} = 0.94
\]

Statistical models using NLP

- **Text Clustering**: Given a set of text, the model creates clusters of similar words.
- **Topic modeling**: Given a set of documents, identifies the different topics within each document and across documents.
- **Text Summarization**: Given a long sequence of paragraphs, returns a short summary consisting of key points.
- **Sentiment Analysis**: Identifies sentiments based on the context and meaning of words.
Model Development – Overview

Text Classification

- A technique to classify a document into one or more categories
- It can be used to detect presence of certain words, filter documents based on keywords etc.

Natural Language Classifier

- Patient Records, Prescriptions, User Reviews etc.
- Text Preprocessing → Tokenization → Feature Extraction → Machine Learning Algorithm(s)
- Label

NLP Applications in Life / Health Insurance
NLP for Document Digitization

NLP techniques are being used to speed up the process of digitization of medical records. Digitization is crucial for businesses to advance into modern age today.

- Reduce errors and omissions introduced by manual data entry
- Increase accessibility, communication and collaboration, free up a lot of space and more importantly save MONEY!

---

Word Embedding Applications

NLP techniques used to understand biological sequences like DNA and RNA

- Protein structures are similar to human language in terms of composition
- Hence, researchers are treating protein sequences as text and using existing NLP techniques to study them
- These techniques are similar to the approaches used in NLP to identify relationship between words in a given sentence or between sentences in a given document
- Word Embedding (vectors) are used to represent biological sequences over a large set of sequences, and establish physical and chemical interpretations for such representations

Word Embedding Applications

NLP techniques used to understand biological sequences like DNA and RNA

These algorithms accept the whole protein structures (structure alignment) as text and parse the sequence to search for corresponding patterns (sequence alignment). The results of these alignments are traditionally presented in a form of color-coded one-dimensional sequential information.

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5333176/

NLP for Claims Processing

NLP techniques can be used on a real time basis to optimize claims processing

The underlying concept would be similar to those used by virtual assistants like Apple’s Siri, Amazon’s Echo etc.

- Identify client’s speech and fill out relevant fields on the claim application
- Scan email text for relevant content and fill out the claim application form
- Improve customer service levels and enhance customer satisfaction
- Reduce the time required for claims processing by accelerating the time required to gather and analyze information from different sources
NLP for Fraud Detection

NLP techniques can be used to detect fraudulent claims

- Identify common phrases and/or descriptions of incidents from multiple claimants
  - Unstructured data sources include claim forms, applications, notes etc. to flag claims with suspicious text or patterns
- It might be difficult even for a trained human eye to spot such patterns after going through a ton of claim applications

NLP based model would help to eliminate inconsistency and subjectivity and reduce the time required to flag potential fraudulent claims

NLP for Underwriting

NLP techniques for extracting medical information relevant to underwriting

Unstructured Data
- Physician Notes
- Clinical Observations
- Medical History
- Lab Results

- Help automate clinical decisions by taking into account text from various sources
- Identify patients with higher risk at a faster rate
- Enable physicians to derive effective treatment methods based on comprehensive patient data

Understand context, grammar and automate decision making taking into account medical jargons, custom abbreviations, tone etc.
Final Thoughts

Future view on data science in life insurance

Factors for growth
- Increasing volumes of quality data and data products available
- Global demand for personalized offerings and ease of transactions
- Growth of direct to consumer offerings
- Monetization of data assets

Head winds
- Major financial successes yet to be demonstrated
- Effort in data cleaning, manipulation, modelling
- More onerous data protection legislation (explicit consent, profiling)
- Cyber risk – Risks of data being lost, corrupted or stolen
Questions?