



**SOCIETY OF
ACTUARIES®**

SOA Big Data Seminar

13 Nov. 2018 | Jakarta, Indonesia

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

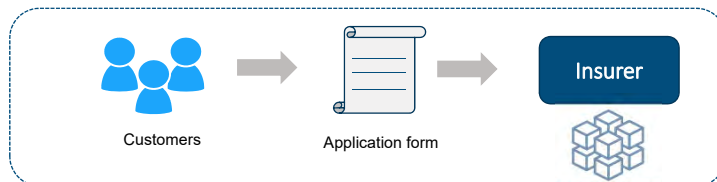


Big Data in Life Insurance



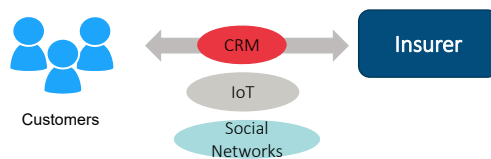
Life insurers are lagging behind when it comes to embracing the benefits of big data

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

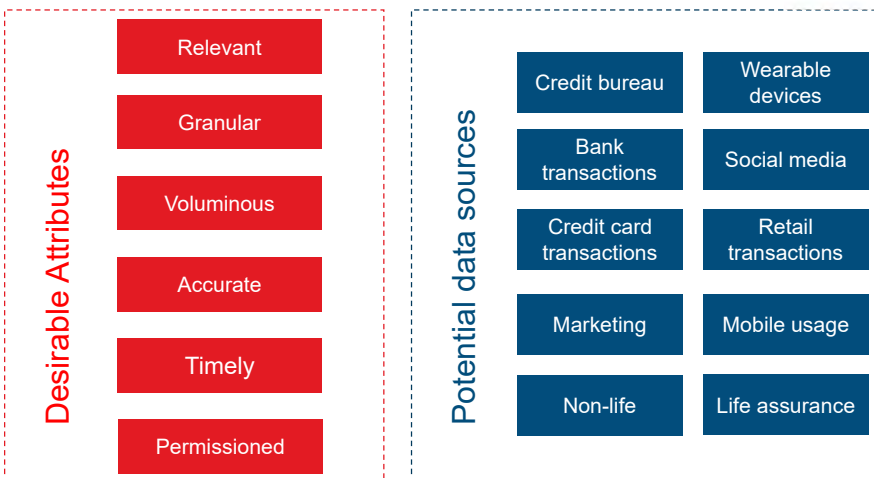


Post Issuance : Limited customer interaction → Limited Data during policy life cycle

Changing dynamics due to increase in customer touch points



Data sources being sought and used within advanced analytics applications



Natural Language Processing(NLP)



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

```

    graph LR
      RawText[Raw Text] --> RemoveStop[Remove Stop words and Punctuations]
      RemoveStop --> Tokenization[Tokenization]
      Tokenization --> Lemmatization[Lemmatization / Stemming]
      Lemmatization --> CleanText[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

Suggested Paracetamol Three Times A Day

SOCIETY OF ACTUARIES

7

Text Pre-processing

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

```

    graph LR
      RawText[Raw Text] --> RemoveStop[Remove Stop words and Punctuations]
      RemoveStop --> Tokenization[Tokenization]
      Tokenization --> Lemmatization[Lemmatization / Stemming]
      Lemmatization --> CleanText[Clean Text]
  
```

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

Source: <https://nlp.stanford.edu/IR-book/html/htmledition/stemming-and-lemmatization-1.htm>

SOCIETY OF ACTUARIES

8

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

Documents	Term1	Term 2	Term 3	Term 4	Term 5
Doc 1					
Doc 2					
Doc 3					
Doc 4					
Doc 5					



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

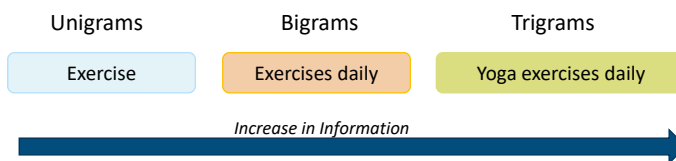


- Product of TF and IDF
- If a word appears multiple time in a document then it should be more meaningful than other words BUT if a word appears many times in a document but also in many other documents then it may be a stop word or a frequent word in that particular domain

$$TF - IDF(t) = TF * IDF$$

Ngrams

It is just a sequence of N words



Word Embedding – Vector Space Models



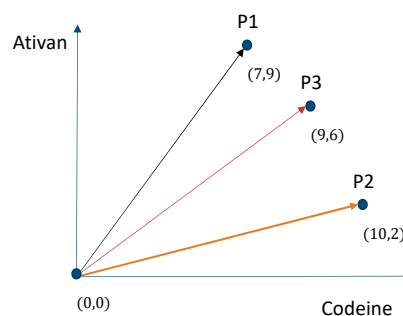
The underlying idea is to represent documents as matrices or arrays

This facilitates to represent documents geometrically

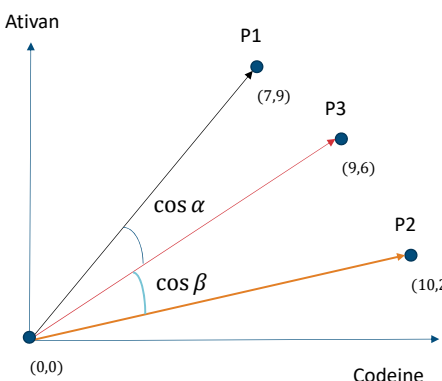
	Ativan	Codeine
Prescription 1	7	9
Prescription 2	10	2
Prescription 3	9	6



Enables document comparison mathematically



Comparing Documents



Definition of Cosine Similarity


$$\text{Cos } \theta = \frac{A \cdot B}{(\sqrt{X^2 + Y^2})_A \cdot (\sqrt{X^2 + Y^2})_B}$$

Cosine Similarity (P1,P2) = $\frac{(7*10) + (9*2)}{\sqrt{7^2 + 9^2} \cdot \sqrt{10^2 + 2^2}}$
= 0.75

Cosine Similarity (P1,P3) = $\frac{(7*9) + (9*6)}{\sqrt{7^2 + 9^2} \cdot \sqrt{9^2 + 6^2}}$
= 0.94

13

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

14

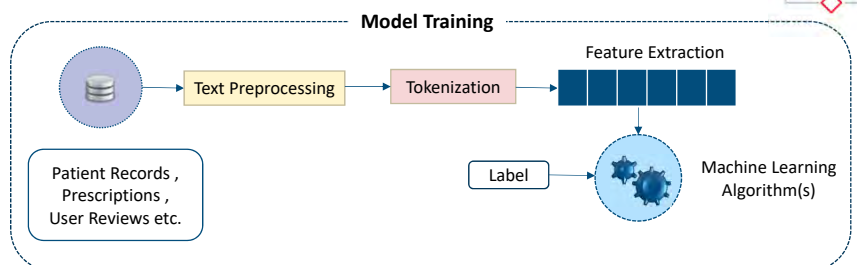
Model Development – Overview



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



NLP Applications in Life / Health Insurance

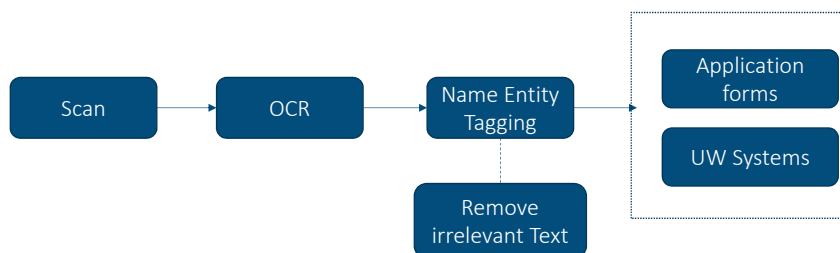


NLP for Document Digitization



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

Citation: Continuous Distributed Representation of Biological Sequences for Deep Proteomics and Genomics Asgari E, Mofrad MRK (2015). PLoS ONE 10(11): e0141287. <https://doi.org/10.1371/journal.pone.0141287>

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

```

KMIGKHKNIINLLGACTQDGPLYVIVEYASKNLEFLRARRPPGMEY
KMIGKHKNIINLLGACTQGGPLYVLVEYAAKGNLREFLRARRPPGLDY
KMIGKHKNIINLLGACTQGGPLYVLVEYAAKGNLREFLRARRPPGLDY
KLIGRHKNIINLLGVCTQEGPLYVIVECAAKGNLREFLRARRPPGPDLL
KLIGRHKNIINLLGVCTQEGPLYVIVECAAKGNLREFLRARRPPGPDLL
KMIGKHKNIINLLGACTQDGPLYVIVEYASKNLEFLRARRPPGLELY
    
```

Each row represents a unique protein structure

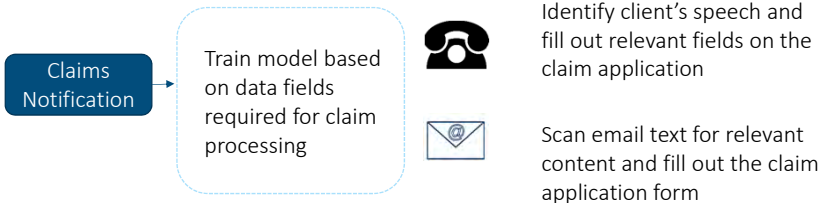
<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5333176/>

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.



- 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 tons 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?

 SOCIETY OF ACTUARIES

25



**SOCIETY OF
ACTUARIES®**

26