Natural Language Processing and its applications in Insurance : A Survey

Courtesy: https://www.einfochips.com/

Data in these times is the real treasure. It is available in numerous forms and our ability to exploit it hugely depends on our capability to easily segregate its elements and then manipulate them in the best possible forms for our use to attain insights and make inferences. Textual data being the most easily and largely available is probably the most exploited form of data. However, manipulation and extraction of textual data are more complicated in comparison to other forms as it is highly unstructured. When we talk about NLP or Text mining it mainly refers to text manipulation techniques or algorithms. A text data point can be a word, a character, sentence, paragraph, or even a document, and text mining works on the extraction of these data points. NLP can be considered as mimicking a normal human text reading process that uses text mining and also captures all the language complexities and intricacies and processes it into a machine learnable form.

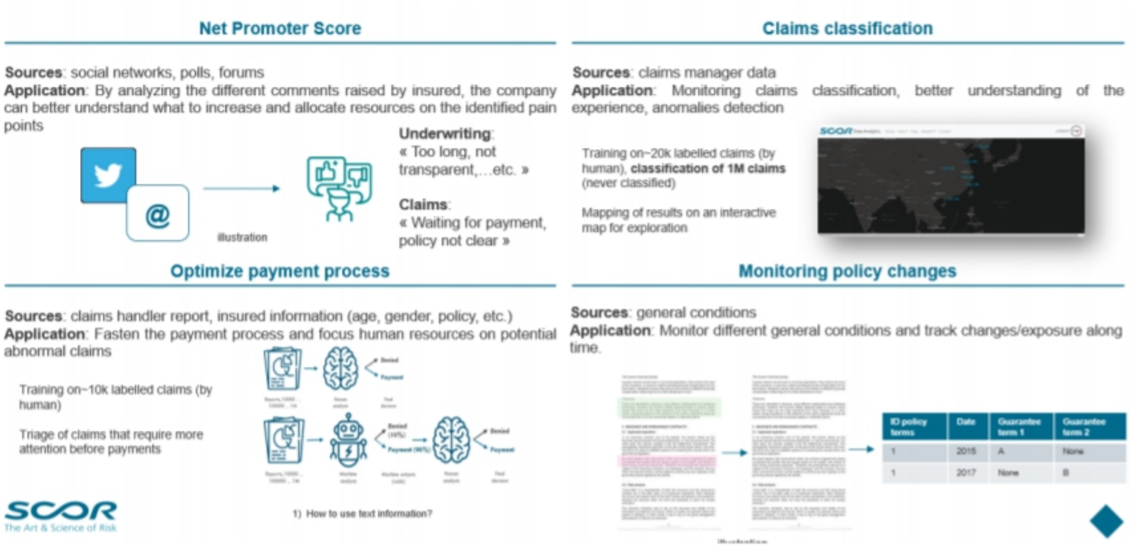
This survey article summarizes all the steps included in Natural Language Processing study and also explains the opportunities NLP is providing to the insurance industry: Marketing, Underwriting, Claims, Reserving, and Prevention. It also highlights the work SCOR teams have been doing to implement state-of-the-art products doing text analysis efficiently and their developed libraries.

Introduction

In this article, we will learn about various areas in the insurance industry that Text analysis or NLP has revolutionized.

In insurance NLP has facilitated many use cases of a policy life cycle which can be summarized as:

- **Marketing**: Marketing is all about reading through the text and understanding the sentiment or emotion behind it. It is also beneficial to extract public feedback to extract the expected trends. So doing sentiment analysis on comments or feedback tweets is a common practice.
- **Underwriting**: When re-insurance happens the underwriters have to analyze every aspect of the policy at renewal periods and digitization of contracts has definitely helped the text mining applications to easily keep track of any changes, thus providing more transparency to the company and policyholder on policy coverage areas. Anonymization techniques like Name Entity Recognition(NER) are used for these tasks.
- **Claims**: Simplifying the area of claim acceptance, reducing operational costs and fraud detection are the main focus points for implementing NLP. With NLP we can classify the claim type and route the claim to the respective department.
- **Reserving**: NLP helps in easing the process of assessing the expert reports that are generated few times annually which helps in anticipating the development of the claim and estimate the expected costs better.
- **Prevention**: NLP is extremely useful in the field of medical diagnostics. If diseases are mapped correctly to their symptoms it increases the chances of patient survival and early treatment.



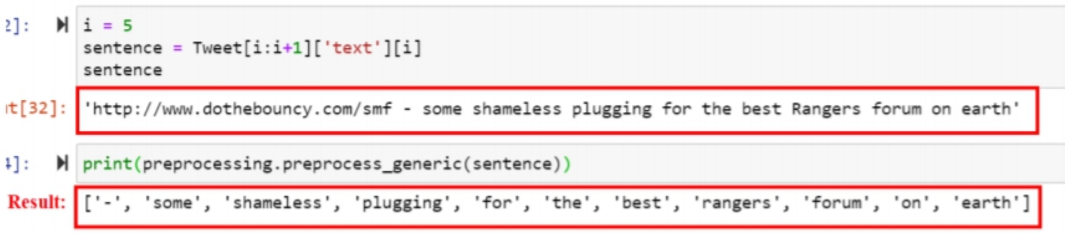
Courtesy: https://arxiv.org/pdf/2010.00462.pdf

Here we look at the SCOR data analytics team’s work with NLP for the insurance industry. We see the models like the Net Promoter Score which takes data from social media, polls, and forums. This model analyses the comments and issues raised by the insured to get a better understanding of areas to allocate more resources. Similarly, they have the Claims Classification model, The Optimize Payment Process model, and the Monitoring Policy Changes models

A Dive into NLP in practice from pre-processing to production

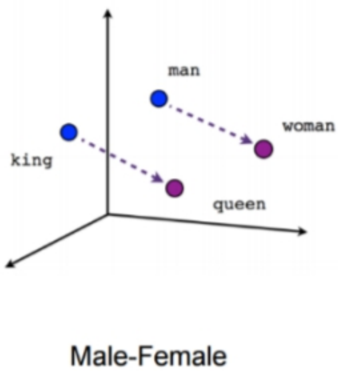
Every text manipulation activity follows a set of steps to generate a working and reusable model. The various steps include:

- **Pre-Processing**: After careful selection and loading of the dataset the following text pre-processing steps are taken to make the data ready for training purposes. Conversion of all letters to **lower case** to remove ambiguity and can be easily achieved in python using .lower(). **Stopwords removal** helps in focussing on the words that really matter thus, removing words like 'the',' and', 'or'. A list of predefined stopwords for multiple languages is available in NLTK lib. **Processing of special characters** is necessary as many writers use special characters like exclamation marks(!) to express their sentiments like excitement. Text is split into smaller sentences and words using a technique called **tokenization** with the help of NLTK and BERT models.



Word level processing

- **Text Embedding**: After data cleaning, data transformation is carried out since a machine can learn through numbers so, by converting textual data into numerical vectors, this is called as text embedding. From the built dictionary of available vocabulary, the rank for each pulled word or sentence is referenced called tokens and making the process as tokenization. After tokenization of words into dictionary indexes, context and semantics are taken into consideration. Word2Vec remains the most popular text embedding lib.



Word2Vec representation

Vocabulary tokenization of a sentence in Chinese

- **Feature Engineering**: Feature engineering practices involving creating new derived features depending on the need and the use case like the length of a phrase, word count, etc. Another technique is data augmentation which enriches the dataset used to train which in turn benefits the prediction. Other techniques involve special character interpretation where a symbol or an emoji can impart greater meaning to the sentiment and specifically for insurance building vocabulary around medical terms is beneficial.

- **Modeling**: To complete the classification or the prediction task as soon as the data is ready in the numerical vector form it can be used to train a model. Models like BERT can be used for both data preparation and modeling when the data is textual. To enrich the training data the numerical vectors are embedded with tabular data for better evaluation.
- **Test and Model Evaluation**: A separate dataset is used to test the model with a similar feature set. And to quantify the model various metrics are used depending on the task achieved:

Metrics	Formula	Meaning	Range
Confusion matrix	$\begin{bmatrix} TP & FN \\ FP & TN \end{bmatrix}$	A confusion matrix is an N×N matrix, where N is the number of classes being predicted. It gives the counts of correct and incorrect classifications for each class.	The higher the diagonal values of the confusion matrix the better. Conversely, the lower the values off the diagonal the better. All values are positive reals.
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$	Percentage of correctly classified instances out of the total predicted instances.	The accuracy is between 0 and 1, where 1 is the perfect score.
Precision	$\frac{TP}{TP+FP}$	Percentage of **positive instances** out of the total predicted positive instances.	As above.
Recall	$\frac{TP}{TP+FN}$	Percentage of positive instances out of the **total actual positive instances**.	As above.
Specificity	$\frac{TN}{TN+FP}$	Percentage of negative instances out of the **total actual negative instances**.	As above.
F-score	$\frac{(1+\beta^2)TP}{(1+\beta^2)TP+\beta^2 FN+FP}$	β is a positive real, and it is chosen such that recall is considered β times as important as precision. In practice β is often set to 1, F1-score is the harmonic mean of precision and recall.	As above.
ROC		The Receiver Operator Characteristic curve represents the tradeoff between Recall and Specificity.	The closer a curve is to the top left corner the better.
AUC		AUC is the area under the ROC curve.	The AUC value lies between 0 and 1 where 1 indicates an excellent classifier and 0.5 the random model.

Metric for Classification

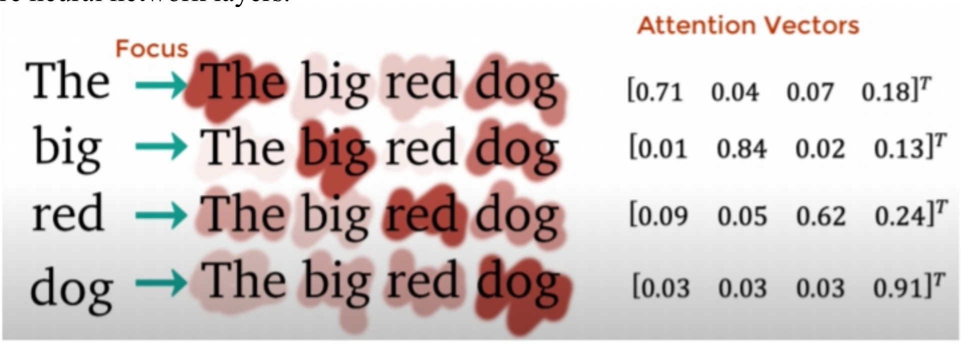
Where:
True Positives (TP) is Predicted positive and actually positive
False Positives (FP) Predicted positives and actually negative
True Negative (TN) Predicted negative and actually negative
False Negative (FN) Predicted negative and actually positive

Metrics	Formula	Meaning	Range
MAE	$\frac{1}{N}\sum_{i=1}^{N}\lvert y_i - \hat{y}_i\rvert$	The Mean Absolute Error is the average of the difference between the actual values and the predicted values.	$[0; +\infty[$
MSE	$\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$	The Mean Squared Error measures the square of the difference between the actual values and the predicted values.	$[0; +\infty[$
RMSE	$\sqrt{\frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{N}}$	The Root Mean Squared Error measures the average magnitude of the error by taking the square root of the average of squared differences between prediction and actual observation.	$[0; +\infty[$
R^2	$1 - \frac{MSE(model)}{MSE(baseline)}$	The Coefficient of Determination or R^2 compares the current model with a constant baseline.	$]-\infty; 1]$

Metric for Regression

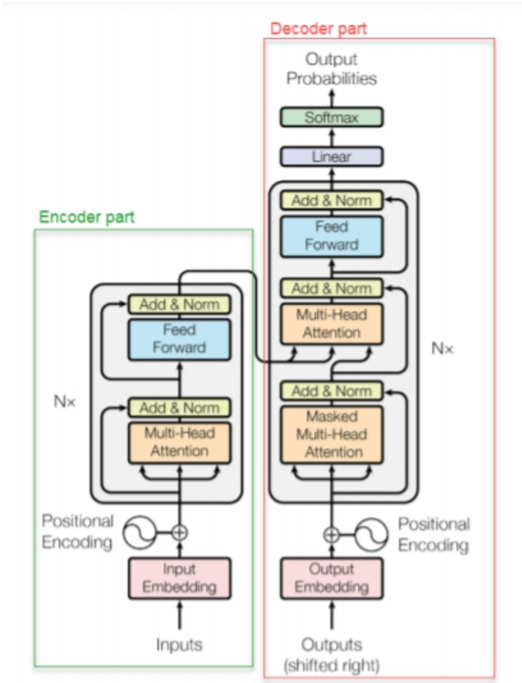
Other approaches in word dependency

- **Attention**: To draw the relationships between the elements of a sequence an attention matrix is drawn to show how much influence one element has over the other one. These are neural network layers.



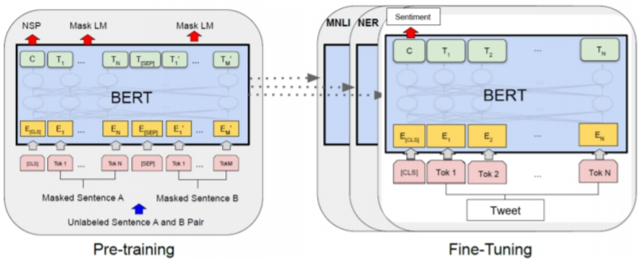
Attention mechanism applied to a sentence

- **Transformer Neural Network**: This has an encoder-decoder architecture. It uses attention and feed-forward neural networks on top of encoder and decoder to improve the efficiency and calculation time for sequential problems.



Architecture of Transformer model

- **BERT:(Bidirectional Encoder Representation for Transformers)** is the most recent model developed to carry out the encoding tasks before the NLP tasks. There are two tasks in the BERT model training pre-training and fine-tuning and they use Transformers Neural Networks.



Conclusions

Although Textual data is hugely available it also corruptable and gathering this data from all possible resources is also a challenge. Since digitization accessing text data has become easier due to the presence of underlying databases, websites, APIs, RSS feeds, etc. Tremendous research has been happening in the NLP areas in the past decade and numerous techniques and models have been revolutionizing the text mining process. BERT remains a state-of-the-art model when it comes to NLP models in these times for its ability to carry multiple NLP tasks.

References

Antoine Ly, Benno Uthayasooriyar & Tingting Wang: A SURVEY ON NATURAL LANGUAGE PROCESSING (NLP) & APPLICATIONS IN INSURANCE