Natural Language Processing and its applications in Insurance : A Survey

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

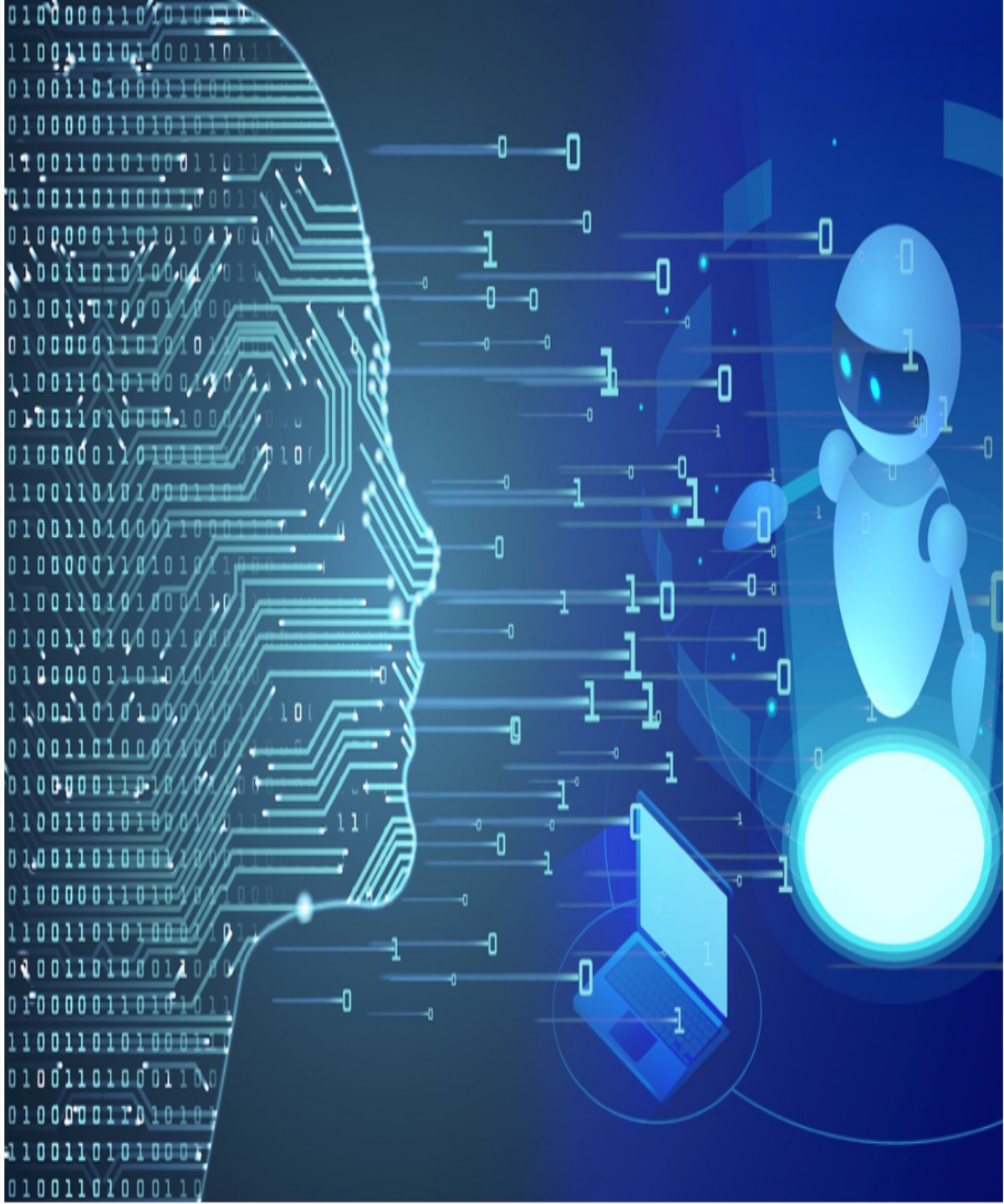
Stuti Agarwal
Student ID: 015229994
email: stuti.agarwal@sjsu.edu

Data forms

Data is present in various forms :

- Text data
- Tabular data
- Visual Data

Textual data is most abundantly available data and is most unstructured.



NLP(Natural Language Processing) and Text Mining

- 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

- Text mining mainly refers to text manipulation techniques or algorithms

NLP in Insurance industry

- Marketing: 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
- Claims: With NLP we can classify the claim type and route the claim to the respective department.

NLP in Insurance industry

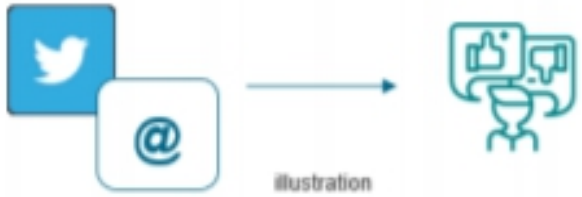
- Reserving: NLP helps in easing the process of assessing the expert reports that are generated few times annually which helps in anticipating the development of the claim and estimate the expected costs better.

- 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.

NLP in Insurance industry

Net Promoter Score

Sources: social networks, polls, forums
Application: By analyzing the different comments raised by insured, the company can better understand what to increase and allocate resources on the identified pain points



illustration

Underwriting:
« Too long, not transparent,…etc. »

Claims:
« Waiting for payment, policy not clear »

Claims classification

Sources: claims manager data
Application: Monitoring claims classification, better understanding of the experience, anomalies detection

Training on~20k labelled claims (by human), **classification of 1M claims** (never classified)

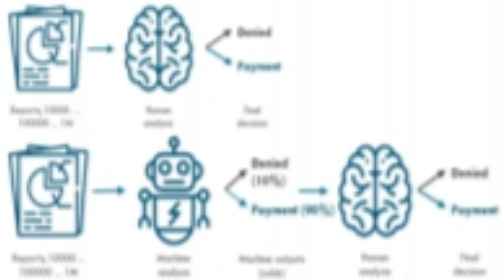
Mapping of results on an interactive map for exploration

Optimize payment process

Sources: claims handler report, insured information (age, gender, policy, etc.)
Application: Fasten the payment process and focus human resources on potential abnormal claims

Training on~10k labelled claims (by human)

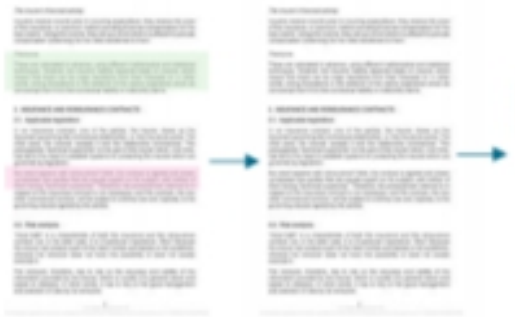
Triage of claims that require more attention before payments



1) How to use text information?

Monitoring policy changes

Sources: general conditions
Application: Monitor different general conditions and track changes/exposure along time.



ID policy terms	Date	Guarantee term 1	Guarantee term 2
1	2015	A	None
1	2017	None	B

SCOR
The Art & Science of Risk

NLP in practice from pre-processing to production

- Pre-Processing: Steps involved
- Conversion to lower case to remove ambiguity
- Stopwords removal: helps in focussing words that really matter
- Processing of special characters
- Tokenization

NLP in practice from pre-processing to production

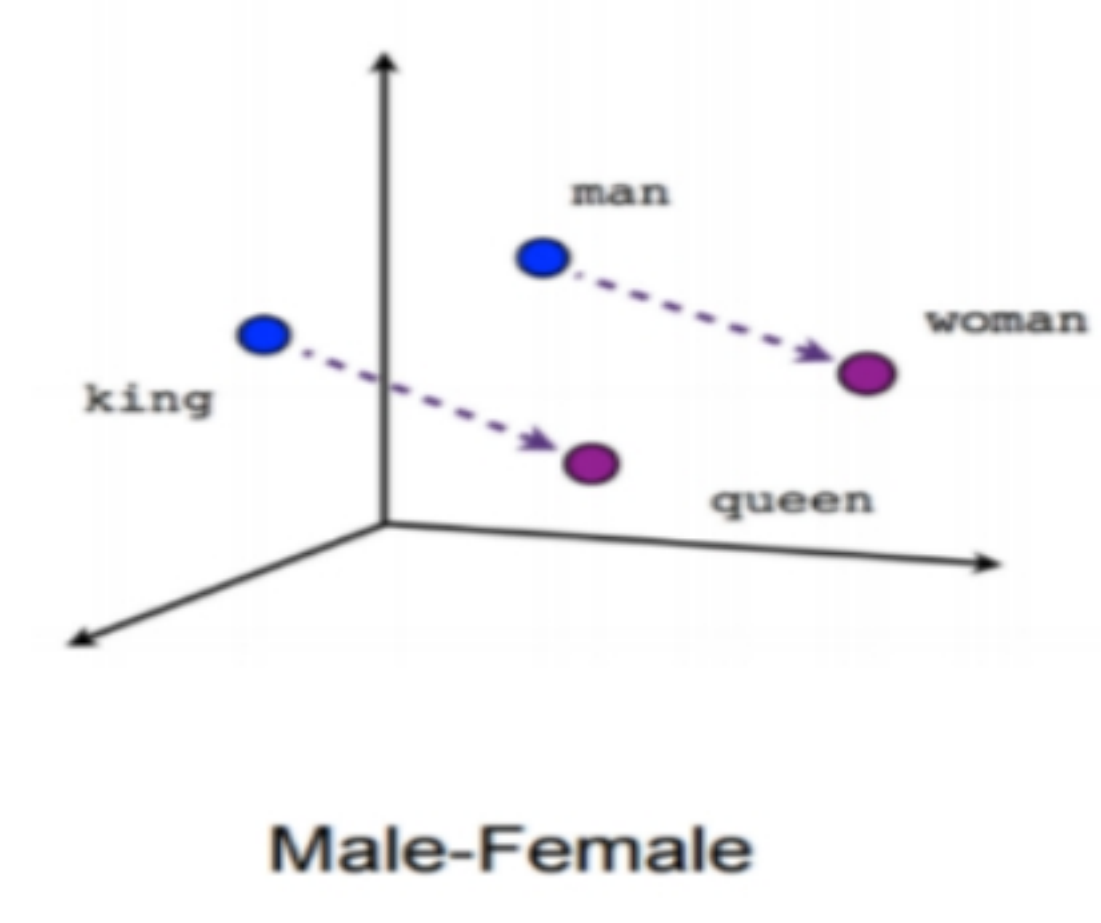
```
2]:  i = 5
     sentence = Tweet[i:i+1]['text'][i]
     sentence

t[32]: 'http://www.dothebouncy.com/smf - some shameless plugging for the best Rangers forum on earth'

4]:  print(preprocessing.preprocess_generic(sentence))

Result: ['-', 'some', 'shameless', 'plugging', 'for', 'the', 'best', 'rangers', 'forum', 'on', 'earth']
```

word level processing

NLP in practice from pre-processing to production

- Text Embedding:
- Machines understand and learn through numbers . Method of converting words into numerical vectors is called text embedding
- From the built dictionary of available vocabulary, the rank for each pulled word or sentence is referenced called tokens
- After tokenization of words into dictionary indexes, context and semantics are taken into consideration.
- Word2Vec remains the most popular text embedding lib

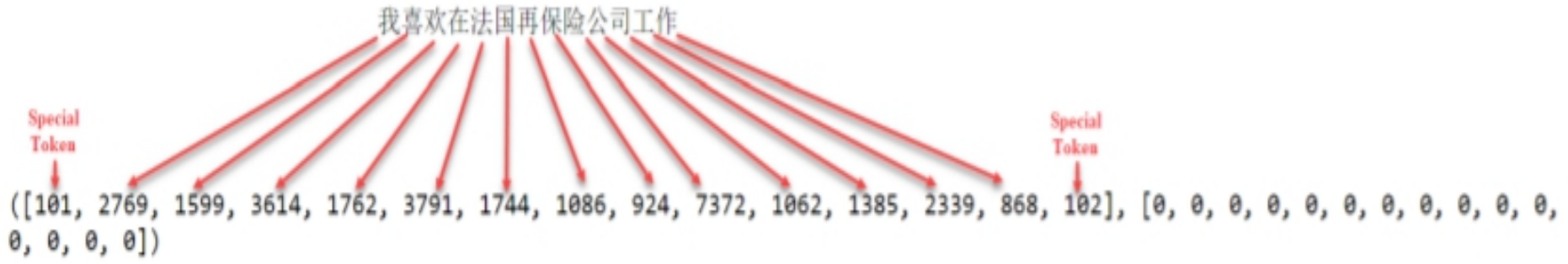
NLP in practice from pre-processing to production



Word2Vec

NLP in practice from pre-processing to production

```
print('                              ', "我喜欢在法国再保险公司工作")
print('\n\n\n')
print(processor.Bert_Tokenizer.encode("我喜欢在法国再保险公司工作"))
```



Tokenization

NLP in practice from pre-processing to production

- **Feature Engineering**:

- Creation of new features called derived features
- Technique of data augmentation
- Character interpretation: like symbols and emojis

NLP in practice from pre-processing to production

- **Modelling**:

- As soon as the numerical vectors are ready we can use the data for training the model

- BERT can be used for both data preparation and modelling both

- **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

Other approaches in word dependency

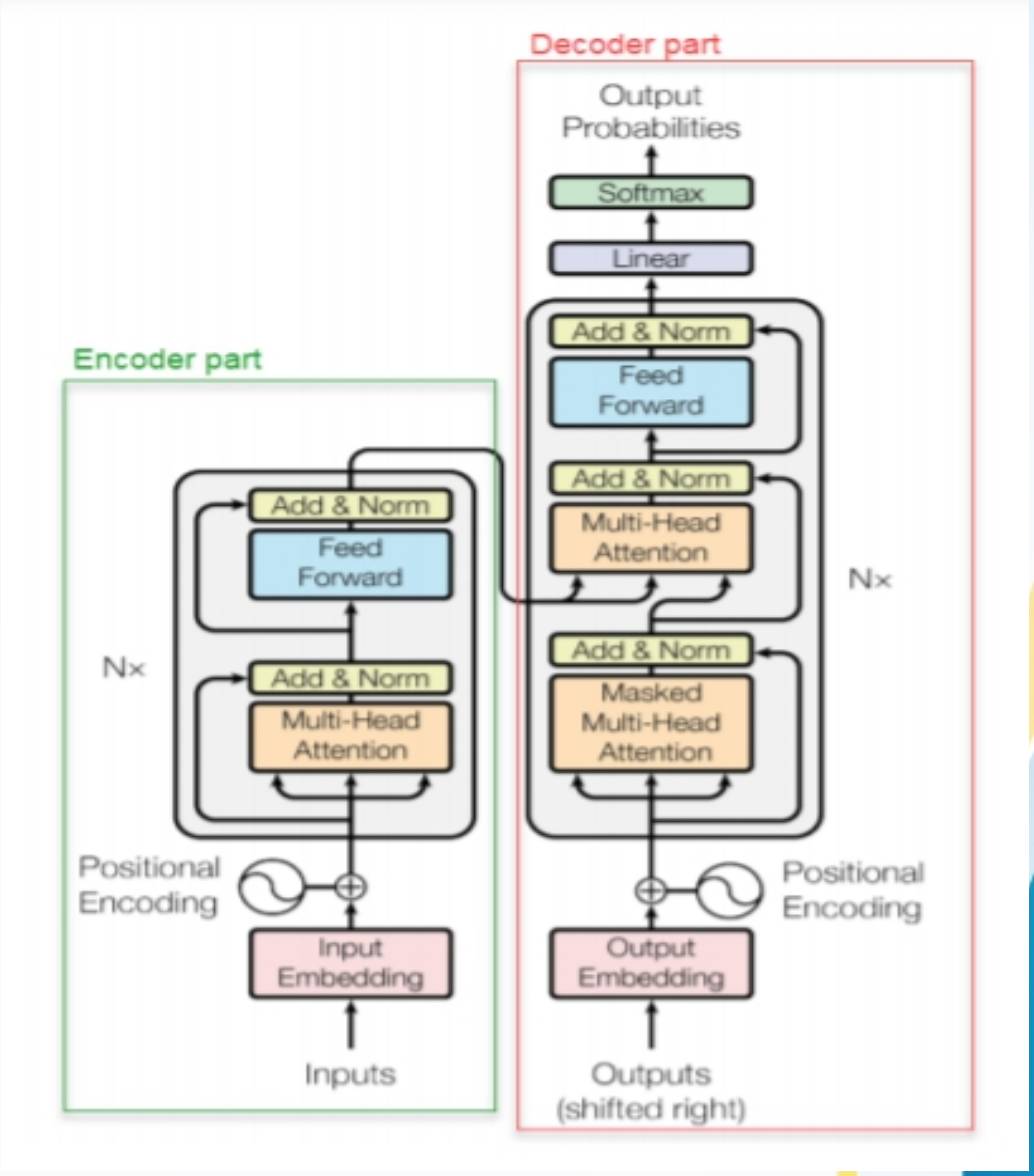
- **Attention**:

- 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.

Other approaches in word dependency

- **Transformer Neural Network**:

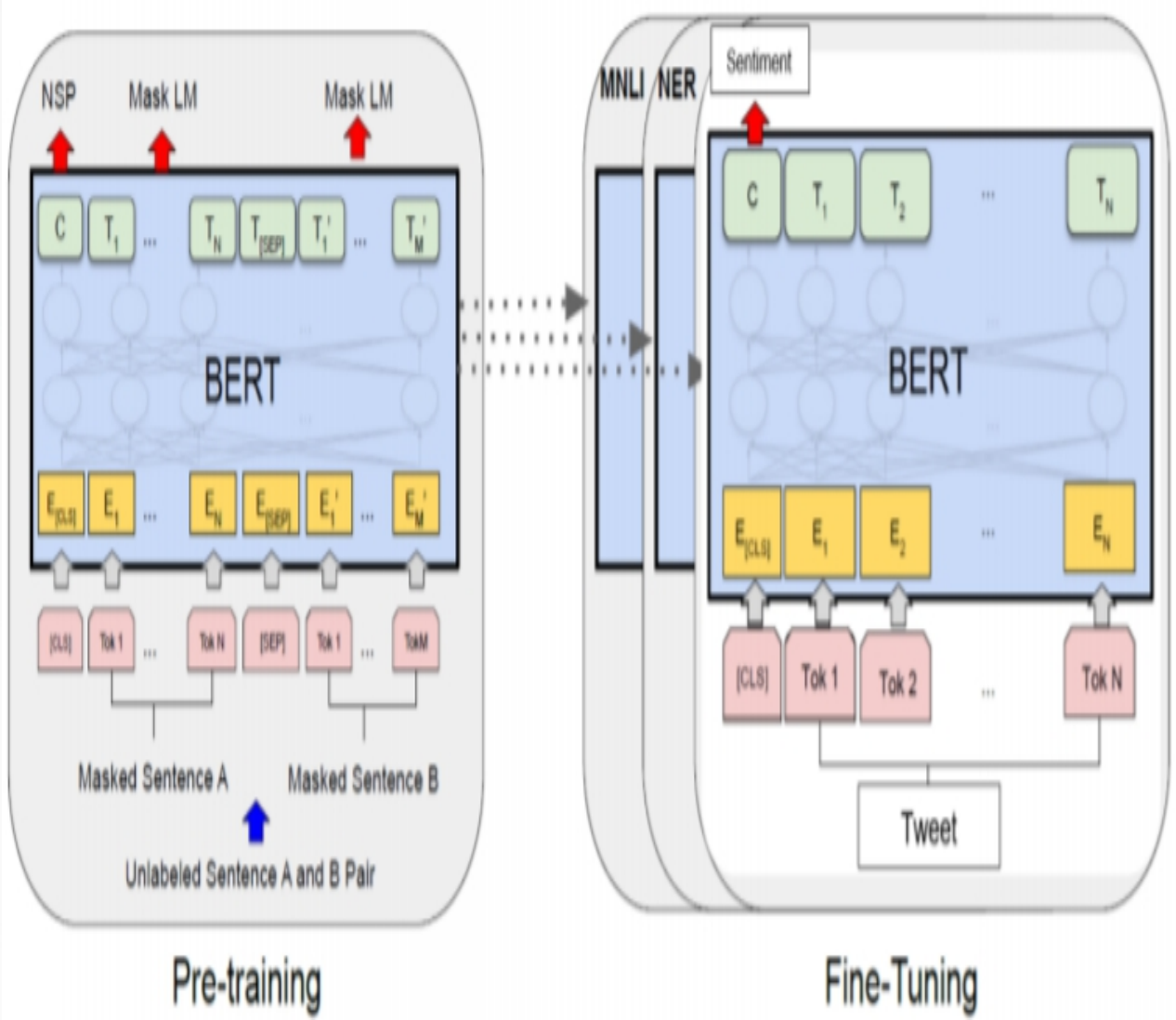
- This has an encoder-decoder architecture. It uses attention and feed-forward neural networks on top of encoder and decoder to improve the efficiency and calculation time for sequential problems



Other approaches in word dependency

- **BERT:(Bidirectional Encoder Representation for Transformers)**:

- There are two tasks in the BERT model training pre-training and fine-tuning and they use Transformers Neural Networks.



Challenges:

Although Textual data is hugely available it also corruptable and gathering this data from all possible resources is a challenge

Conclusion:

- 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.

Referenced Research paper:

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

THANK YOU !