Lilly Project 3: Regulatory Natural Language Processing

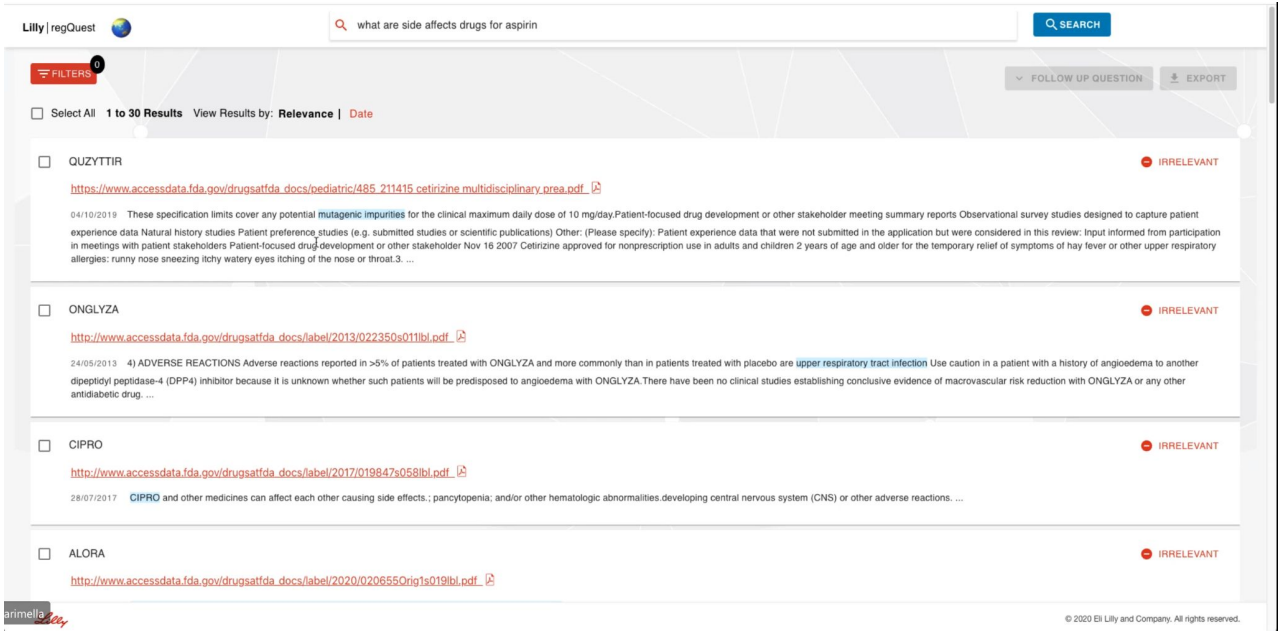
--- **By Team Members:**
Prateesh Reddy
Himani Anil Deshpande
Krishna Vamsi Guntupalli
Venkata Kartheek Janapati
Victor Zitao Zhang
Chaitanya Deshpande

LUDDY
SCHOOL OF INFORMATICS,
COMPUTING, AND ENGINEERING

Overview of the Project

Lilly RegQuest :

To Provide Cognitive search capability to search against database like FDA and EMA (European medical agency) via a natural language question and return relevant results in order to help with accelerating regulatory submissions (which is important for Lilly).



Project Objectives

- To understand the data from dailymed.
- To Scrape relevant data to the NLP question asked.
- To understand and analyse state of the art text summarization algorithms.
- To compare performance of Advanced NLP techniques on medical data and weigh out limitations.
- To generalise and automate the summarization for searched NLP question.

Text preprocessing

- Need to filter out keywords/patterns which can obviously identify the sentences
- Finding paragraphs by keywords will include a lot of “red herring” informations which will subsequently confuse the abstraction algorithm

Correct:	NSAIDs cause an increased risk of serious gastrointestinal GI adverse events including bleeding ulceration and perforation of the stomach or intestines…
Incorrect:	Table 1a depicts adverse events that occurred in ≥2% of the MOBIC treatment groups in a 12-week placebo- and active-controlled osteoarthritis trial…
Incorrect:	See ADVERSE REACTIONS and DOSAGE AND ADMINISTRATION for monitoring recommendations

Pre-processing of question 1

- What are the most common adverse reaction?
  - Find block text that is a <paragraph> or <content>
  - Contain the word “adverse”
  - Exclude sentence like:
    - contact FDA …
    - adverse effects are identified

```
if repr(item.tag).find('paragraph') != -1 or repr(item.tag).find('content') != -1:
    # print(item.tag, item.attrib, item.text)
    txt = repr(item.text)
    excludes = ['contact', 'label', 'identify']  exclude keyword
    if txt.lower().find('adverse') != -1:
```

```
for exclude in excludes:
    if txt.lower().find(exclude) != -1:
```

Pre-processing for question 1

- What are the most common adverse reaction?
  - Find the Xpath of the Tag that we want
  - We focused on tag **ADVERSE REACTIONS SECTION** and all the text inside the tag

```
root = tree.getroot()
for parent in root.findall("./x:component/x:structuredBody/x:component/x:section/x:code/[@displayName='ADVERSE REACTIONS SECTION']/..",
                           namespaces={'x':"urn:hl7-org:v3"}):
    for child in parent.iter():
        if repr(child.tag).find('paragraph') != -1 or repr(child.tag).find('content') != -1 :
            data = data + child.text.rstrip()
```

Pre-processing of question 2

The MOBIC Phase 2/3 clinical trial database includes 10,122 OA patients and 1012 RA patients treated with MOBIC 7.5 mg/day, 3505 OA patients and 1351 RA patients treated with MOBIC 15 mg/day.

- How many patients were studied in total?
  - Find number followed by patient(s)/individual(s)/participant(s)
  - Accommodate some number written in xxx,xxx format (remove ‘,’ first)
  - Ignore fractions

```
found = re.findall('(?:(-?[0-9]+/0*[1-9][0-9]*)|(\d+))(\s+\w*\s?)(patient|individual|participant)' txt.lower())
if len(found) > 0:
    #print(found, end = '')
    pos = -1
    for item in found:
        itemstr = ''.join(''.join(elems) for elems in item)
        print(itemstr, end = '')

        pos += 1
        if pos != len(found):
            print(',', end = '')
        # not fraction
        if itemstr.find('/') == -1:
            sum += int(re.search('\d+', itemstr).group())
```

Issues on pre-processing

- Will always find new pattern that need excluded as language is complex
- Pattern matching can not exclude subset recounting for problems like “finding the total number of …”

  A total of 519 RA patients 65 years of age and older including 107 patients 75 years of age and older received HUMIRA

  The pattern is found twice but only one should be valid

- Searching will not include sentences that use an alternative expression to keywords, for further development, the system will need to have a dictionary of synonyms

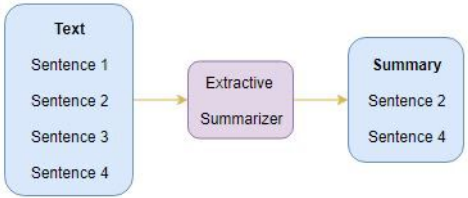
Learnings from data

- We have the clinical data from the DailyMed website. On that website we have an XML format for any particular drug.
- Each XML file has several sections, out of which we will be focusing on the Adverse Effects section.
- The Adverse effects section is the 6th section in the XML file, and may contain subsections like 6.1, 6.2, etc.
- We need to combine the data from the <paragraph> tags from each of the subsections and pass it to the summarization models.
- The data might also contain table entries, which can be ignored because it doesn't give us much information about adverse effects.

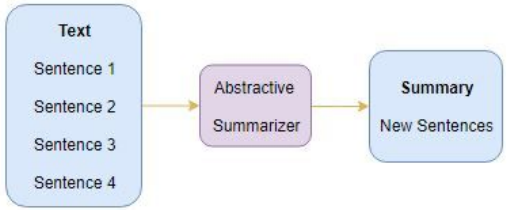
Types on Text Summarization

Two categories — **Extractive Summarization** and **Abstractive Summarization**.

1. **Extractive Summarization:** Rely on the existing text that has phrases, sentences to create a new summary. So, we need to identify key words or sentences of the existing text.



1. **Abstractive Summarization:** Uses advanced NLP techniques to generate an entirely new summary which will not contain phrases or sentences that exist in the original text. It is closer to what humans usually expect from text summarization. The process is to understand the original document and rephrase the document to a shorter text while capturing the key points.



Extractive Summarization

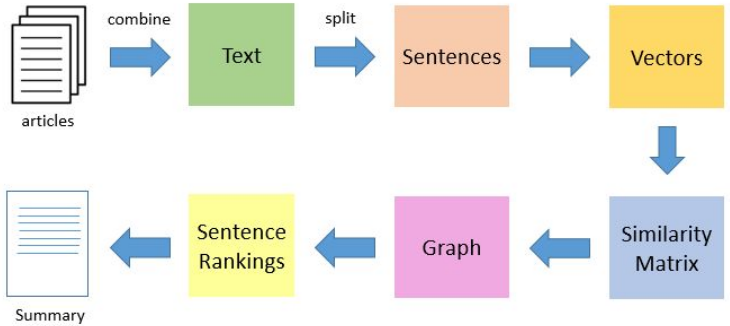
- A subset of words that represent the most important points is pulled from a piece of text and combined to make a summary.
- Unsupervised approach: *TextRank (Genism, TextRank)
- Heuristic approach: TextTeaser

TextRank Algorithm

Graph-based Ranking Algorithms

Unsupervised Algorithm for Keyword Extraction and Text Summarization

- Similarities between sentence vectors are then calculated and stored in a matrix
- The similarity matrix is then converted into a graph, with sentences as vertices and similarity scores as edges, for sentence rank calculation
- Finally, a certain number of top-ranked sentences form the final summary



Our Summary for TextRank

What are the most common adverse reaction?

Text Rank Algorithm Output :

Discontinue fenofibrate and treat patients appropriately if SCAR is suspected..Most common adverse reactions (> 2% and at least 1% greater than placebo) are abnormal liver tests, increased AST, increased ALT, increased CPK, and rhinitis (.Because clinical trials are conducted under widely varying conditions, adverse reaction rates observed in the clinical trials of a drug cannot be directly compared to rates in the clinical trials of another drug and may not reflect the rates observed in clinical practice..The following adverse reactions have been identified during post approval use of fenofibrate.

Photosensitivity reactions have occurred days to months after initiation; in some of these cases, patients reported a prior photosensitivity reaction to ketoprofen..Limited available data with fenofibrate use in pregnant women are insufficient to determine a drug associated risk of major birth defects, miscarriage or adverse maternal or fetal outcomes.

Limitations (Extractive)

- It summarizes most relevant sentences. So, It might not make a meaningful sentences everytime.
- Ignores important word that appeared in low frequency.
- Not feasible to implement summarization on medical data

Other Extractive Text Summarization Algorithm

Apart from TextRank algorithm, we have a few more extractive algorithms under “sumy” library:

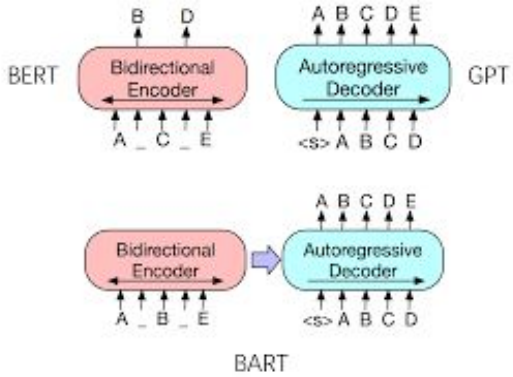
1. **LexRank**- This algorithm uses the concept wherein a sentence which is similar to many other sentences of the text has a high probability of being important. Based on these probabilities, summary is decided.
2. **Luhn**- Luhn’s algorithm is a naive approach based on TF-IDF and looking at the “window size” of non-important words between words of high importance.
3. **Latent Semantic Analysis**- This algorithm combines Term frequency with Singular Value Decomposition.
4. **Kullback-Lieber-Sum**- This algorithm is a greedy method which adds sentences to the summary as long as the KL Divergence (a measure of entropy) is decreasing.

Abstractive Summarization

- *A neural network approach
  - Recurrent neural network
    - LSTM/BRU
    - BERT
    - BART
    - T5
    - GPT2/3 (Open AI)
- Tree-based approach
- Template-based approach
- Rule-based approach

BART Transformers

Bidirectional and Auto-Regressive Transformers (Lewis et al., 2019)



BERT is a Bidirectional Transformer with a Masked Language Modelling uses seq2seq/machine translation architecture.
GPT is a autoregressive model which uses left to right decoder.

Allows Token Masking, Token Deletion, Text Infilling

from_pretrained() -> tokenizer.encode() -> model.generate() -> tokenizer.decode

Our Summary for BART

What are the most common adverse reaction?

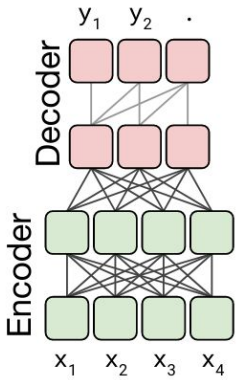
Most common adverse reactions are: Rheumatoid and Psoriatic Arthritis, upper respiratory tract infection, nasopharyngitis, diarrhea, and headache. The proportion of patients who discontinued treatment due to any adverse reaction during the 0 to 3 months exposure was 4% for patients taking XELJANZ and 3% for placebo-treated patients. The most common types of malignancy were lung and breast cancer.

Limitations

- The maximum number of tokens which can be generated is 1024.
- LM and pre-training techniques do not provide the full picture.
- It can only deal with fixed-length text strings. The text has to be split into a certain number of segments or chunks before being fed into the system as input
- This chunking of text causes **context fragmentation**. For example, if a sentence is split from the middle, then a significant amount of context is lost. In other words, the text is split without respecting the sentence or any other semantic boundary

T5 Algorithm

- The T5 Algorithm is a pre-trained abstractive summarization algorithm introduced by Google, which also uses Transformers along with the encode-decode approach. It is pre-trained on the Colossal Clean Crawled Corpus.
- We pass the string generated from the “Adverse Effects” section of the XML file along with the tokens from T5Tokenizer and generate a tokenized text.
- The maximum number of tokens which can be generated by the tokenizer is 512.
- Using the model generated from T5ForConditionalGeneration, we create the summary IDs, which are further decoded to generate the summary of a particular text.



Output for T5 Algorithm

- We ran the T5 Algorithm on Adverse Effects on the drug AVEED, and we got a summary as follows:

AVEED was evaluated in an 84-week clinical study using a dose regimen of 750 mg (3 mL) at initiation, at 4 weeks, and every 10 weeks thereafter in 153 hypogonadal men. the most commonly reported adverse reactions (>2%) were acne (5.2%), injection site pain (4.6%), prostate specific antigen increased (4.6%), hypogonadism (2.6%) and estradiol increased (2.6%).

Example of T5-large vs Wikihow-T5-Small

- T5-large

  "Most common adverse reactions to COUMADIN are fatal and nonfatal hemorrhage from any tissue or organ" warfarin crosses the placenta, and concentrations in fetal plasma approach the maternal values. exposure to warfarin during the first trimester of pregnancy caused a pattern of congenital malformations in about 5% of exposed offspring. warfarin embryopathy is characterized by nasal hypoplasia with or without stippled epiphyses...
- wikihow-T5-small

  'Other adverse reactions to COUMADIN are fatal and nonfatal hemorrhage from any tissue or organ.'Consider fetal plasma exposure to warfarin during the first trimester of pregnancy.Consider nasal hypoplasia with or without stippled epiphyses (chondrodysis).Consider stipled epiphyses.Consider a stipled stippled stipple stipples.Consider the effects of warfarin in humans.

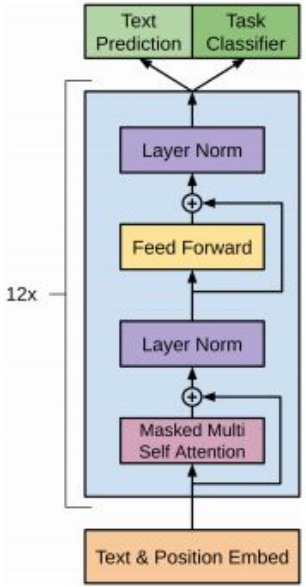
Limitations of T5 Algorithm

The Limitations of the T5 Algorithm are:

- Despite changing the length_penalty parameter for our text, the output size does not have a significant difference in the summary.
- Since the maximum number of tokens which can be generated by the T5 algorithm is 512, we only get a small sized summary for our text. If we want a larger size summary, we will not get desirable results.
- This algorithm also requires good processing speed and power like GPUs, so it isn't very convenient to use it on local machines if we wish the scan all the XML files.
- Algorithm is not context aware, can not effectively exclude sentences that does not fit the theme

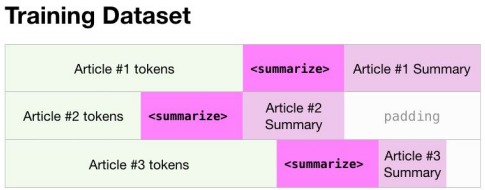
GPT-2 Algorithm

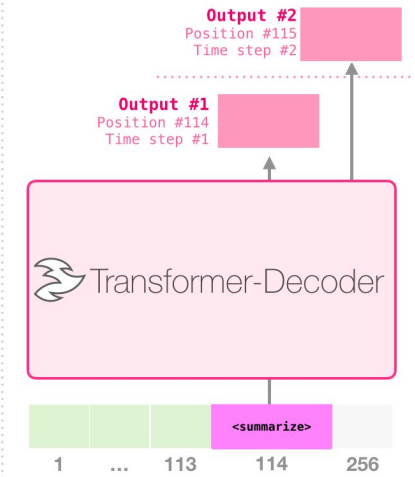
- Is a seq2seq algorithm developed by OpenAI which stands for Generative Pre-trained T
- Uses masked self-attention.
- Uses byte pair encoding(BPE)
- “text = summary”
- Uses larger context and vocabulary size



GPT-2 Coding Format

- Import GPT2Tokenize and GPT2LMHeadModel
- Encode the text
- Pass them to model.generate()
- Decode to generate the text summary.





Text Summary Output for GPT-2

For the question **What are the most common adverse reactions?**

The following clinically significant adverse reactions are described elsewhere in the labeling: Serious Infections. Most common adverse reactions are: Rheumatoid and Psoriatic Arthritis: Reported during the first 3 months in rheumatoid arthritis controlled clinical trials and occurring in ≥2% of patients treated with XELJANZ monotherapy or in combination with DMARDs: upper respiratory tract infection, nasopharyngitis, diarrhea, and headache. Because clinical studies are conducted under widely varying conditions, adverse reaction rates observed in the clinical studies of a drug cannot be directly compared to rates in the clinical studies of another drug and may not predict the rates observed in a broader patient population in clinical practice. All seven protocols included provisions for patients taking placebo to receive treatment with XELJANZ at Month 3 or Month 6 either by patient response (based on uncontrolled disease activity) or by design, so that adverse events cannot always be unambiguously attributed to a given treatment.

OUTPUT:
The most common Adverse Reactions are Rheumatoid and Psoriatic Arthritis: Reported during the first 3 months in rheumatoid arthritis controlled clinical trials and occurring in ≥2% of patients treated with XELJANZ monotherapy or in combination with DMARDs: upper respiratory tract infection, nasopharyngitis, diarrhea, and headache.

GPT-2 Limitations

- Output size does not have a significant difference in the summary
- The maximum number of tokens GPT-2 can be generated is 512
- Were at most only 70% of the time correct, independent of the model used
- Not equipped to handle complex and long language formations

Generalise & Automate Text Summarization

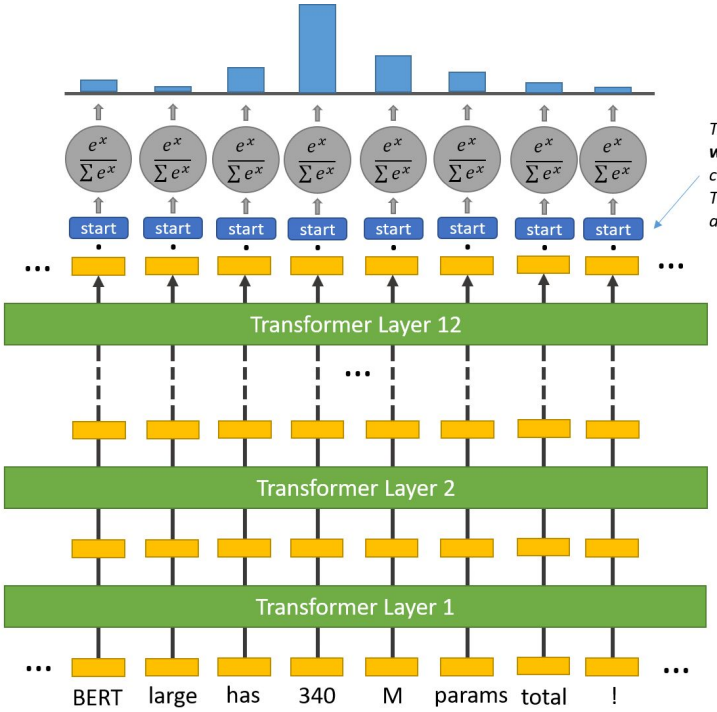
Inspiration :

- Until now summarization by keywords from the NLP question
- Need for Question Answering type approach
- Inspired by BERT for Question Answering on SQuAD 2.0 by Stanford
- Summarize the most appropriate answers instead of keywords using BART

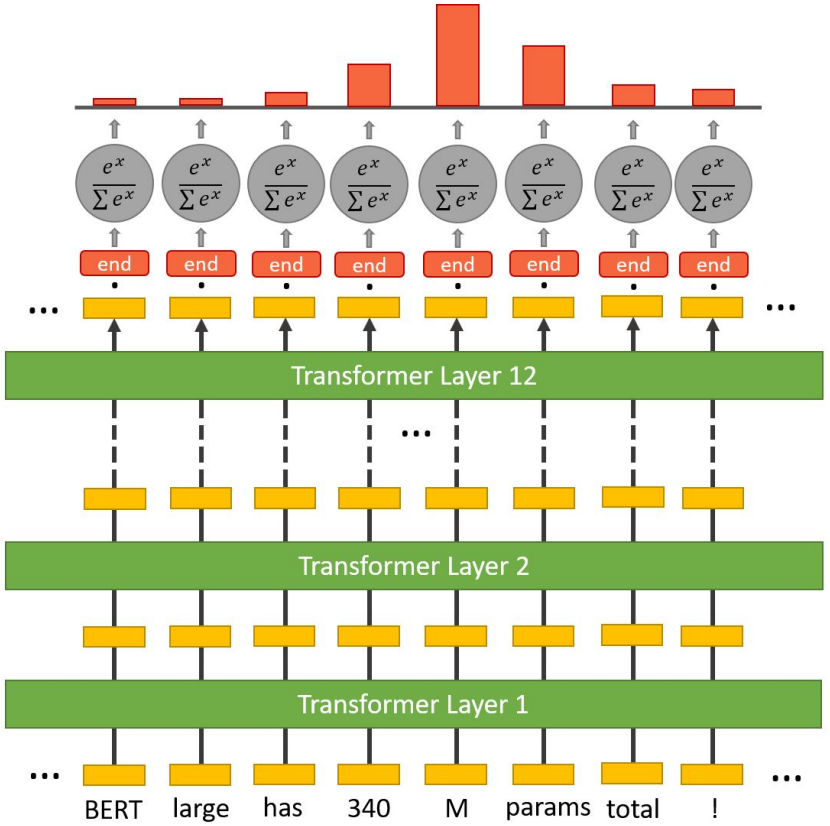
BART Question and Answer

- Pre-trained Using SQuAD 2.0
- Search for best token based on question token using BartForQuestionAnswering
- Append these into answer_array
- Input this array to our previously implemented BART to generate summary

BART Question Answering



*This length 768 vector is the **weights** for the start token classifier.*
*The **same weights** are applied to **every position**.*



Limitation

- It tries to find words to get a best answer
- If the best answer is wrong, it will still produce it
- Summary Concentrates more on correctness than human readable.
- Cutting Edge Research.Thus, No concrete evidence or research on this approach yet.

Comparing Models

What are the most common adverse reaction?

BART: Most common adverse reactions are: Rheumatoid and Psoriatic Arthritis, upper respiratory tract infection, nasopharyngitis, diarrhea, and headache. The proportion of patients who discontinued treatment due to any adverse reaction during the 0 to 3 months exposure was 4% for patients taking XELJANZ and 3% for placebo-treated patients. The most common types of malignancy were lung and breast cancer.

BART QnA: NUZYRA was evaluated in three Phase 3 clinical trials (Trial 1, Trial 2 and Trial 3). The most common adverse reactions (incidence ≥2%) are nausea, vomiting, infusion site reactions, alanine aminotransferase increased, and hypertension.

GPT-2: The most common Adverse Reactions are Rheumatoid and Psoriatic Arthritis: Reported during the first 3 months in rheumatoid arthritis controlled clinical trials and occurring in ≥2% of patients treated with XELJANZ monotherapy or in combination with DMARDs: upper respiratory tract infection, nasopharyngitis, diarrhea, and headache.

T5: 'Most common adverse reactions to COUMADIN are fatal and nonfatal hemorrhage from any tissue or organ" warfarin crosses the placenta, and concentrations in fetal plasma approach the maternal values. exposure to warfarin during the first trimester of pregnancy caused a pattern of congenital malformations in about 5% of exposed offspring. warfarin embryopathy is characterized by nasal hypoplasia with or without stippled epiphyses...

Challenges we faced

- Original problem of patient narrative didn’t have processable data, so we got a new topic at around week 3/4
- Semi-structured Data
- Not all the XML had ADVERSE EFFECTS Tag
- The maximum tokens we could work with were 512/1024, which was less as we had a large input sequence to work with.

Future work

- How to build context awareness for text preprocessor
- Proper evaluation metric (ROUGE/BLEU only deal with grammar)
  - Find discriminator algorithm for recognizing obviously wrong answers
- Optimize the question-answering model into the text of second question

References

BART: Denoising Sequence-to-Sequence Pre-training for NLG (Research Paper Walkthrough) - Youtube

Transformers In NLP | State-Of-The-Art-Models - Analytics Vidhya

BERT for Question Answering on SQuAD 2.0 - Stanford

TextRank: Bringing Order into Texts - Rada Mihalcea and Paul Tarau (UMitch)

Thank You !
&
Questions ?