Automated construction of patient cohorts from social media, to support clinical knowledge discovery

Almog Mor[1] **and Dr. Tom Hope**[2]

[1]**The Hebrew University of Jerusalem - almog.mor@mail.huji.ac.il**
[2]**The Hebrew University of Jerusalem - tom.hope@mail.huji.ac.il**

ABSTRACT

In an area where annotated data is crucial for clinical-knowledge research on the one hand and the increase of social-media traffic with medical knowledge on the other, led us to develop a pipeline which extract textual data of posts and comment that had been shared anonymously on social media.

Keywords:

INTRODUCTION

This project's motivation came from the assumption that medical forums contain data about cancer patients which doesn't exist in established medical knowledge sources such as medical literature. The main goal was to create high quality annotated dataset of cancer patient information from posts and comments that were posted on social media.

This data can be later used for extracting information like alternative drugs experiences among cancer patients and to learn about treatment and side effects. Even sensitive and exposed data that has been shared among users about their emotions during the treatment process.

We found Reddit cancer related threads as a good source for this kind of data. Reddit define itself as a 'network of communities where people can dive into their interests, hobbies and passions'. More specifically on medical threads, we found a many posts and comments from cancer patients. We were surprised to see how people tend to share their situation and the way the sickness effect them and their loved ones.

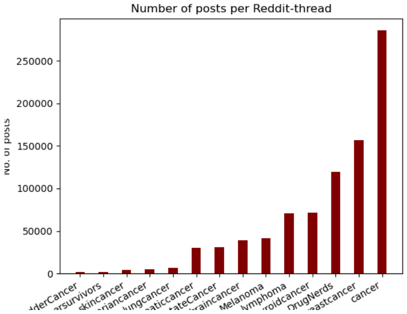
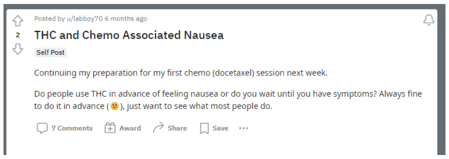


Figure 1. the distribution of the number of posts over reddit-threads

1 DATA EXTRACTION

Extraction the data from reddit includes 3 steps. (1) Downloading metadata from pushift.io in 'zft' format. Each file includes all the metadata about all the threads in Reddit per month. (2) The second stage gets as input the 'zft' files, filter the relevant threads, extract the relevant columns and convert them into 'csv' files. In this project we chose to filter a list of cancer related threads. (3) The last stage is a crawler which gets the ‘permalink‘ from the ‘csv‘ files of stage 2 and scraps the post, comment and title from the reddit website. The last stage outputs the raw data file per month in a ‘csv‘ format which is ready to later be annotated.



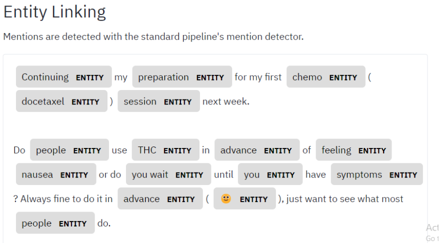
Posted by u/labboy70 6 months ago

THC and Chemo Associated Nausea

Continuing my preparation for my first chemo (docetaxel) session next week.

Do people use THC in advance of feeling nausea or do you wait until you have symptoms? Always fine to do it in advance (🙂), just want to see what most people do.

Figure 2. example for a post on Reddit



Entity Linking

Mentions are detected with the standard pipeline's mention detector.

Continuing ENTITY my preparation ENTITY for my first chemo ENTITY (docetaxel ENTITY) session ENTITY next week.

Do people ENTITY use THC ENTITY in advance ENTITY of feeling ENTITY nausea ENTITY or do you wait ENTITY until you ENTITY have symptoms ENTITY ? Always fine to do it in advance ENTITY (🙂 ENTITY), just want to see what most people ENTITY do.

Figure 3. annotation example

2 MODEL

2.1 Existing models - scispacy

As a way of getting a sense of the data, we tried annotation with existing **?** https://github.com/allenai/scispacy (**?**) models. Using UMLS as a large Knowledge-Base with different linkers like ‘umls‘ and ‘mesh‘.

We tried 3 different configurations: (1)‘en ner bionlp13cg md-0.5.1‘ with ‘mesh‘ linker, (2) ‘en ner bc5cdr md-0.5.1‘ with ‘mesh‘ linker. (3) en ner bc5cdr md-0.5.1 with ‘umls‘ linker.

In order to evaluate the annotation performance I made a small set for evaluations that contains 10 posts and 10 matching comments. I annotated the data myself analyzed using the following definitions (code appears here):

True-Positive: entities that appear exactly the same in both manual annotation and model annotations.
False-Positive: appears just in the model and not in the manual.
False-Negative: appears just in the manual and not in the model.

The rates are aggregated through the posts and comments where each entity is unique i.e. repeated tokens that were found and annotated counts as one.

COMMENTS	precision	recall
‘Model (1) ‘	0.275	**0.607**
‘Model (2)‘	**0.555**	0.126
‘Model (3)‘	0.268	0.278

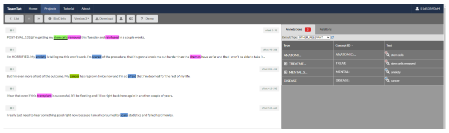
Table 1. Performance - Comments.

POSTS	precision	recall
‘Model (1)‘	0.341	**0.66**
‘Model (2)‘	**0.526**	0.097
‘Model (3)‘	0.307	0.495

Table 2. Performance - Posts.

2.2 New Model

While analyzing the existing models we thought about the possibility that creating a new model might lead to better results. Training a model on a specific Name-Entitity-Recognition as downstream task and annotation manually posts and comments from our data in order to increase the f1 score.For Training we used 300 posts and comments from the data and annotated them manually in teamtate framework. We used the following classes for annotation: ‘DISEASE‘, ‘OTHER RELEVANT‘, ‘SYMPTOM‘, ‘TREATMENT‘, ‘TEST‘, ‘DEMOGRAPHICS‘, ‘SEVERITY‘, ‘MENTAL STATE‘, ‘ANATOMICAL‘. We used the hugging-face framework and the following configuration: 0.8/0.2 percent of the annotated data for train/test, model=distilbert-base-uncased, optimizer=‘AdamWeightDecay‘, ‘learning rate‘=‘PolynomialDecay‘,‘initial learning rate‘=2e-05,‘decay steps‘=300, ‘end learning rate‘=0.0‘.This model recived these results:



POST-EVAL_132@I'm getting my stem cells removed this Tuesday and reinfused in a couple weeks.

I'm HORRIFIED. My anxiety is telling me this won't work. I'm scared of the procedure, that it's gonna knock me out harder than the chemos have so far and that I won't be able to take it...

But I'm even more afraid of the outcome. My cancer has regrown twice now and I'm so afraid that I'm doomed for the rest of my life.

I fear that even if this transplant is successful, it'll be fleeting and I'll be right back here again in another couple of years.

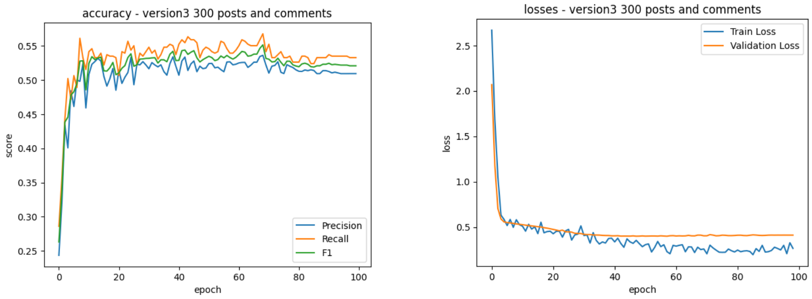
I really just need to hear something good right now because I am all consumed by scary statistics and failed testimonies.

Type	Concept ID	Text
ANATOMI...	ANATOMIC:...	stem cells
TREATME...	TREAT:	stem cells removed
MENTAL_S...	MENTAL:	anxiety
DISEASE	DISEASE:	cancer

Figure 4. teamtat annotation framework

Training Loss	0.2654
Validation Loss	0.2654
Precision	0.5093
Recall	0.5327
F1	0.5208
Accuracy	0.9330

Table 3. Model results



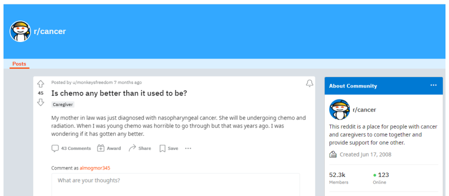
(a) precision/recall

(b) loss

Figure 5. Model performance

Inference examples

Let’s see an examples of the model’s inference: The first post was written in cancer-subreddit (r/cancer) by the author ’monkeysfreedom’ at 2022-10.



Posted by u/monkeysfreedom 7 months ago

Is chemo any better than it used to be?

My mother in law was just diagnosed with nasopharyngeal cancer. She will be undergoing chemo and radiation. When I was young chemo was horrible to go through but that was years ago. I was wondering if it has gotten any better.

Comment as almogmor345

What are your thoughts?

This reddit is a place for people with cancer and caregivers to come together and provide support for one other.

Created Jun 17, 2008

Figure 6. post from 2022-10

Example n.1:

”my mother in law was just diagnosed with **nasopharyngeal cancer** [DISEASE]. she will be undergoing **chemo**[TREATMENT] and **radiation**[TREATMENT] when I was young **chemo**[TREATMENT] was horrible to go through but that was years ago. I was wondering if it has gotten any better.” The model performed well in this example. Recognized all the interesting entities, classified them right and didn’t annotated any other word in the sentence.

Example n.2 (from ’lymphoma’ subreddit):

”My dad was diagnosed with **stage 3B**[SEVERITY] **NHL lymphoma**[DISEASE] back in April and today his doctor told him there was no evidence of **cancer**[DISEASE] on his most recent scan. He is almost **70**[DEMOGRAPHICS] years old (so no spring chicken) but he was able to mostly tolerate R-Chop well. I hope this gives hope to anyone going to through the same thing.” The model classified well almost all the tokens in the text and tagged right entities like ’SEVERITY’ and ’DEMOGRAPHICS’ which are more rare to see in the data. ‘R-Chop‘ is a recall loss. It’s a type of combination therapy which the model should have annotated.

Example n.3 (from ’breastcancer’ subreddit):

’Hi, Did you tell your employer of your **breast cancer**[DISEASE] diagnosis[TEST] or keep it confidential? I’ve read that it may be beneficial to tell them in order to be protected by ADA. I’ve already informed them of a “future need to use **FMLA**[TREATMENT] but didn’t exactly say why. How did you all handle your work situation and taking time off?’ The model miss-classified the token ‘FMLA‘. ‘FMLA‘ stands for Family and Medical Leave Act. This token is a false-positive because it shouldn’t get any special tagging.