Predict with Data: Identify Patient Voice

Binary Text Classification of Unlabeled Data

Prasanthi Desiraju

Prasanthi.Desiraju@gmail.com

Motivation

Leverage ML to extract patterns within patient data and convert this information to actionable intelligence

- Analytics to improve quality of care : What treatments are preferred, symptoms exhibited
- Influence Product Development
- Enable Competitive Advantage

- Collect data from Patient's experience - Identify "Right" Data
- Leverage Social Media – Twitter, PatientsLikeMe, SocialGest
- Avoid Garbage In, Garbage Out – Build a classification model to extract required data

- Lack of Pre-Trained Data – Key Differentiator for any ML model
- How to pick the right training data? **OUR FOCUS TOPIC**

Background

- Advent of Transformers revolutionized the NLP domain because of better word representations that are Context-based
- Easy to use Pipelines – sentiment analysis, text classification, translations with high accuracies based on BERT or advanced models.
- Access the pre-trained models and fine-tune them as required.

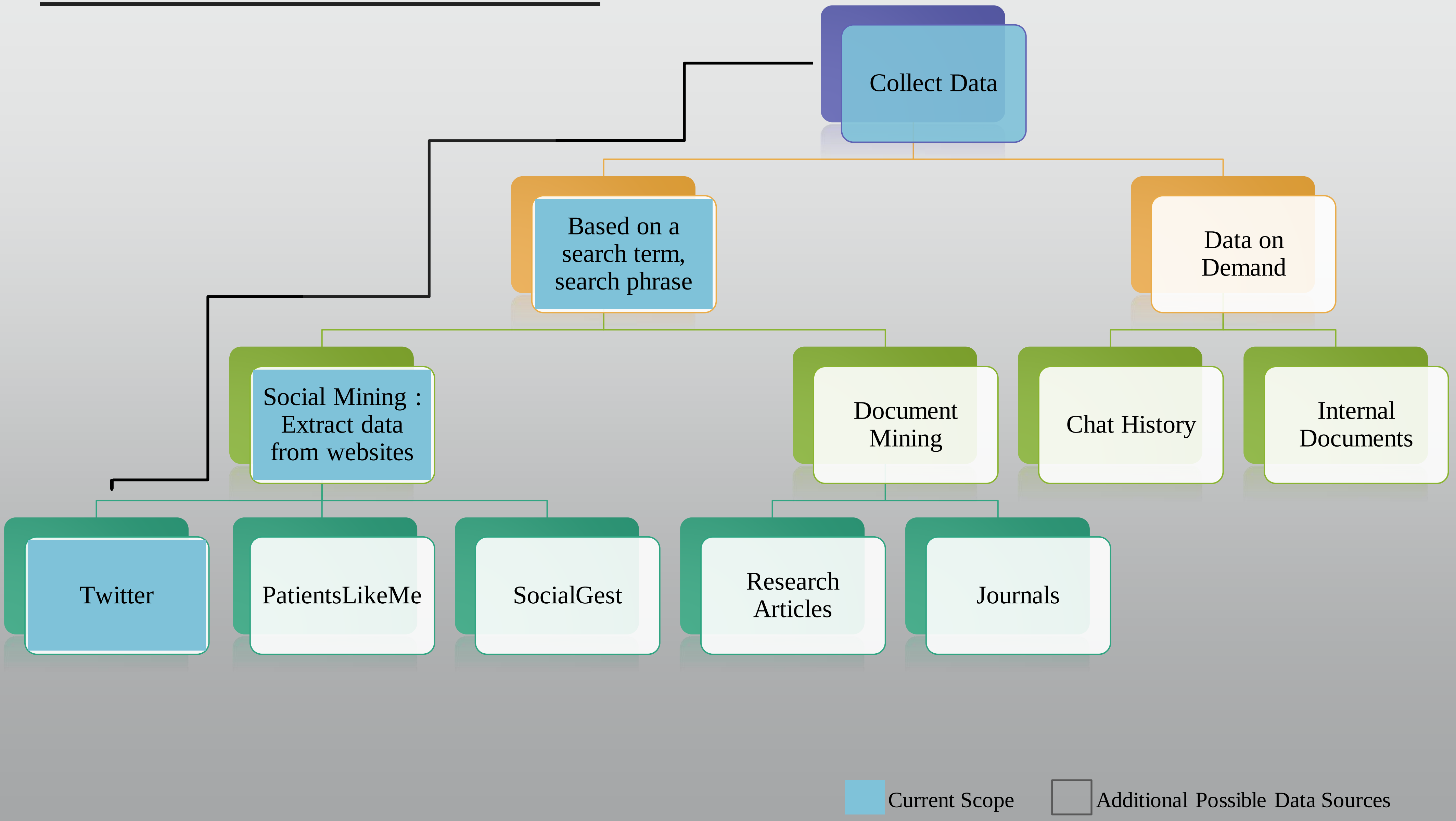
BUT!! We need a lot of labeled data to quantify DNN models
Also, most of these pipelines do not cover medical domain-specific data

Not a New Problem! We have active research addressing the issue and providing a framework to generate labeled data.

- Weak Supervision (Snorkel (2017), skweak (2021) etc.)
- Active Learning (ModAl (2018), small-text(2021) etc.)

This work focuses on leveraging and combining these methods to try and see how efficiently we can classify the patient-specific information with minimal annotation efforts

Where to Find Data



Data Preparation

Data Governance

- Use specific search terms to extract only relevant data (drug names)
- Validate the Search Term, Source Credibility, Profanity Check

Multi-Lingual Support

- We focus only on English Tweets for this study

Eliminate Noise

- Remove URL's, Mentions, Unwanted Tokens
- Remove Duplicate Information
- Convert HTML Text to General Text
- Remove short sentences (less than 6 words)

Final Metrics

- Initial Data : 120,833 Tweets
- Final Data :   63,321 Tweets

Given the sheer volume of tweets, we end up with tweets that do not relate to patient experience.

Example:
What to Know About the New MS Drug XXX .. (Other)

I am on Drug XXX and it seems to be working good so far (Patient)

Problem Statement : Build this binary classification model on unstructured data (text) without any labeling data efficiently.

Libraries Used : Snscrape, Profanity-check, Tweet-preprocessor, Beautiful Soup

We have Data, What Next?

Manual Annotation

- Manually labeled 3200 records
  (Patient – 1270, Other – 1930)
- Acts as our own test data set for binary classification model
- To be used as initial seed dataset for the data labeling techniques

Explore data labeling techniques

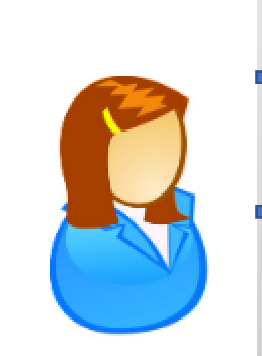
- Weak supervision with Snorkel AI
- Active Learning with small-text

Quantify the results with SOTA

- BERT: State-of-art-algorithm trained on Wikipedia and Brown Corpus
- BioBERT: Variant of BERT trained additionally on PubMed Articles, PMC articles

Weak Supervision

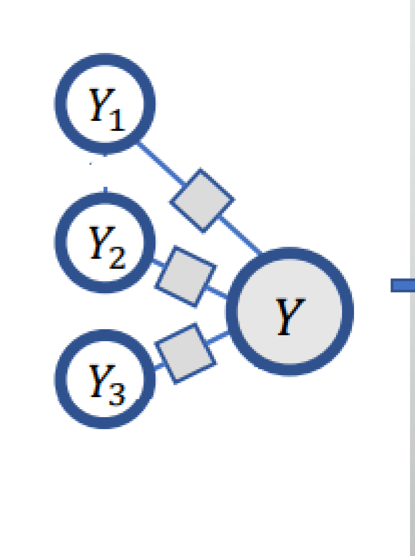
Pattern Matching LF



LF1: Contains Personal Pronoun – 'Patient'

LF2: Contains Drug Manufacturer – 'Other'

LF3: Contains Symptoms – 'Patient'



The resulting model helps predict labels associated with a probability Interval.

Labelled Dataset

User creates labeling functions(LFs) to assign a class for a given datapoint.

LFs can conflict with one another

Snorkel Cleans and Combines LFs to Create one final output class based on few true labels

User has no control over this Model

Ex: I am doing great after taking..

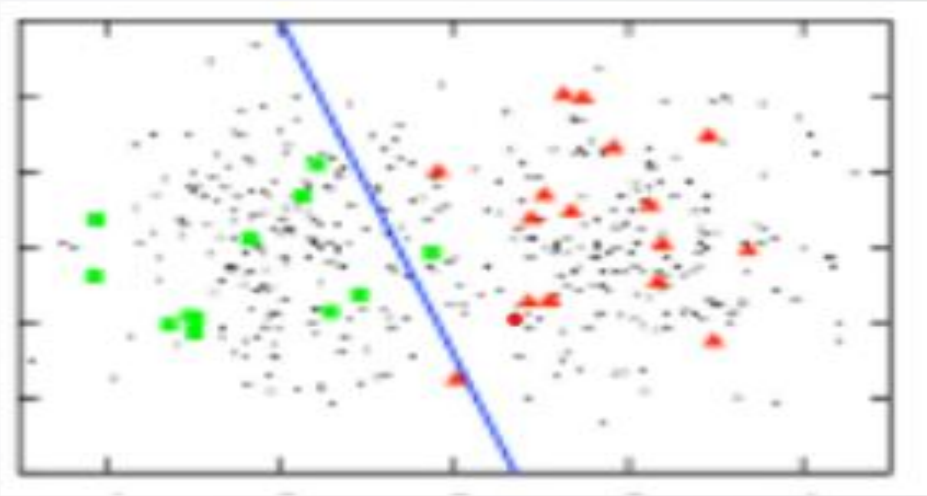
Probability Interval
0.98 -> Patient
0.02 -> Other

- Quality depends on the labeling functions
- Oracle to pick predictions with low confidence intervals and label the data

Active Learning

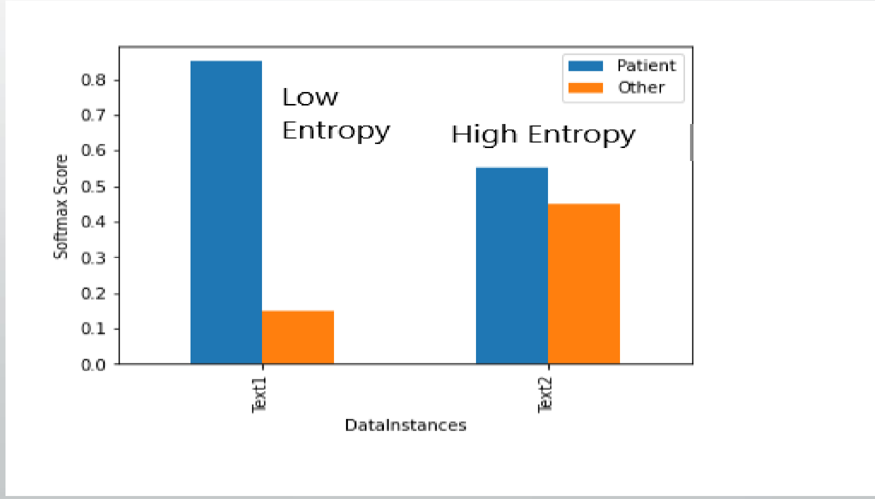
What is AL?

- Semi-supervised learning algorithm that can query a user interactively for labels and adjust the performance over each iteration.
- Select instances from a large pool of unlabeled data based on some informative measure.



Terminologies:

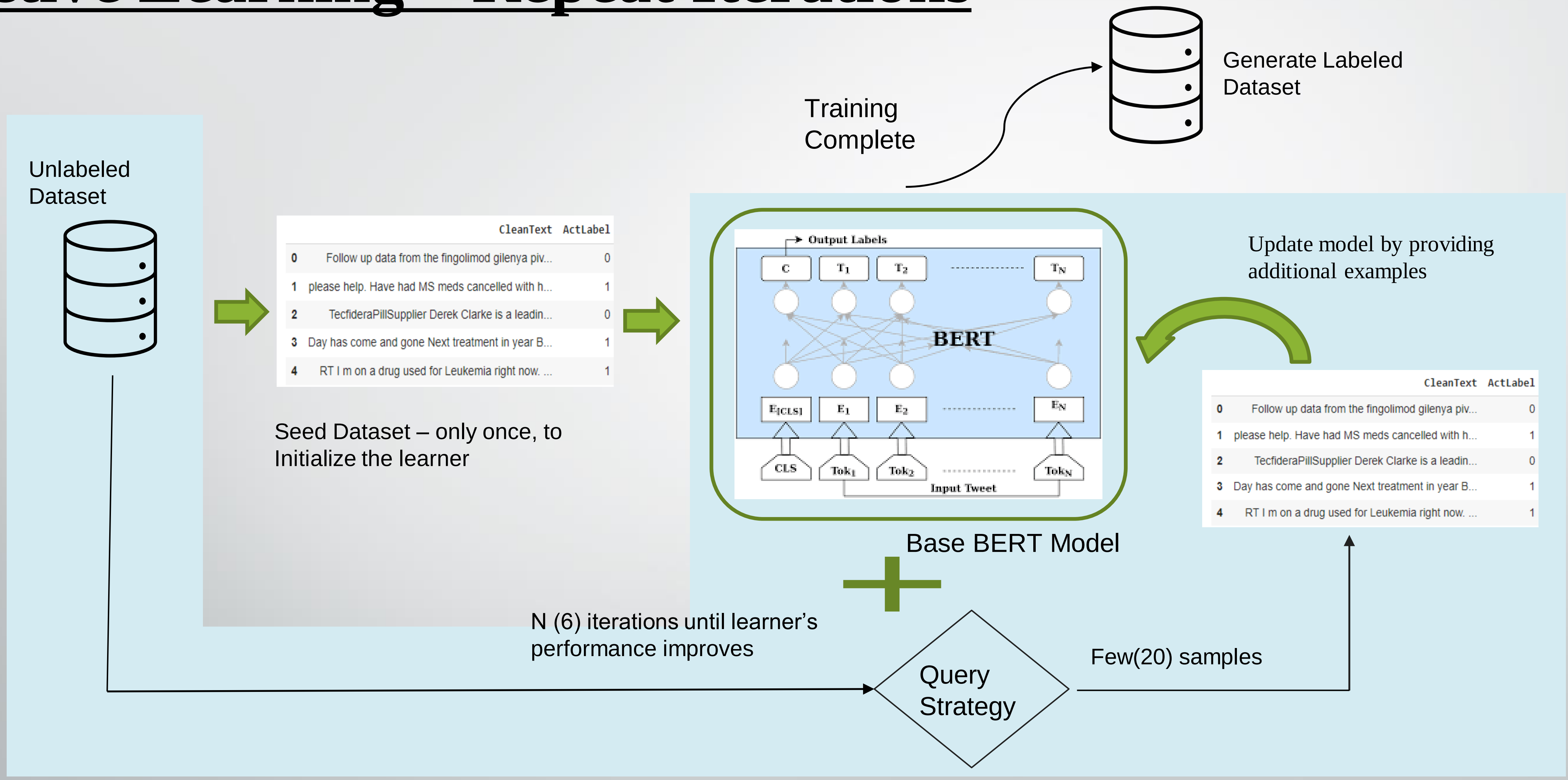
- Seed Dataset – Initial labeled data provided by annotator (We provided 1000 Tweets)
- Query Strategy – Sampling data using specified criteria, Prediction Entropy, Random Sampling



Prediction Entropy:

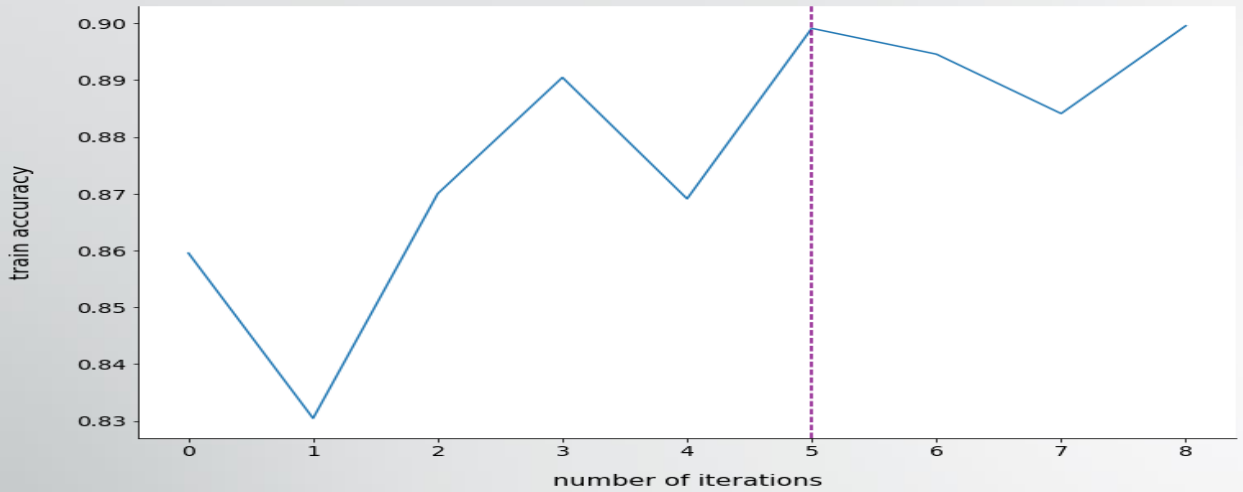
- Learner assigns probabilities to each data point based on the current model.
- Entropy formula is applied on each data instance and the instance with largest entropy value is queried.

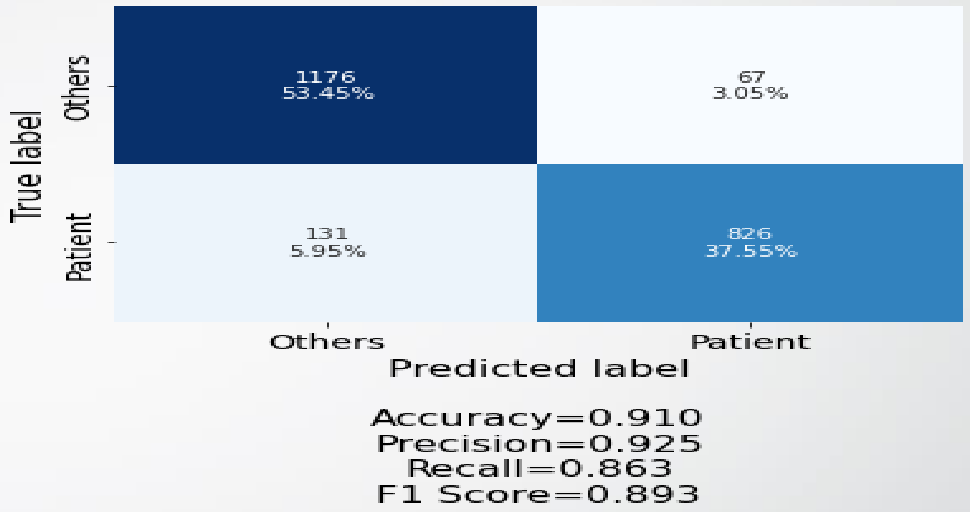
Active Learning – Repeat Iterations



Results (Data Labeling)

Active Learning

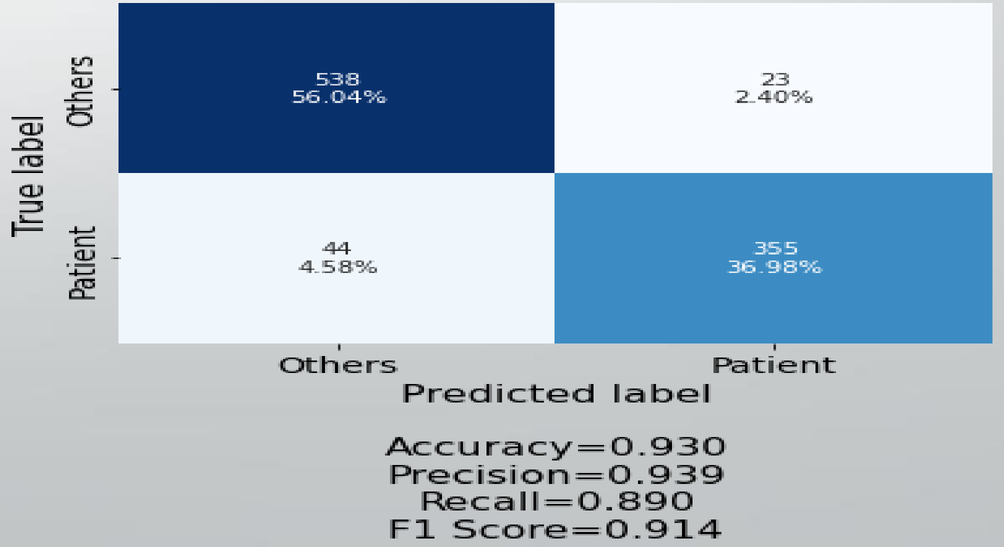




Improvement of Active Learning Accuracy over multiple iterations

Weak Supervision

CleanText	Pred1	Pred2	LabelPredText
When your body is still throwing a hissy fit and vomiting every day week post lemtrada you know just because glitterbrainproblems On the upside being housebound has meant more booking adventures for later in the year	0.71	0.29	Other
day to lemtrada feeling nervous now	0.71	0.29	Other

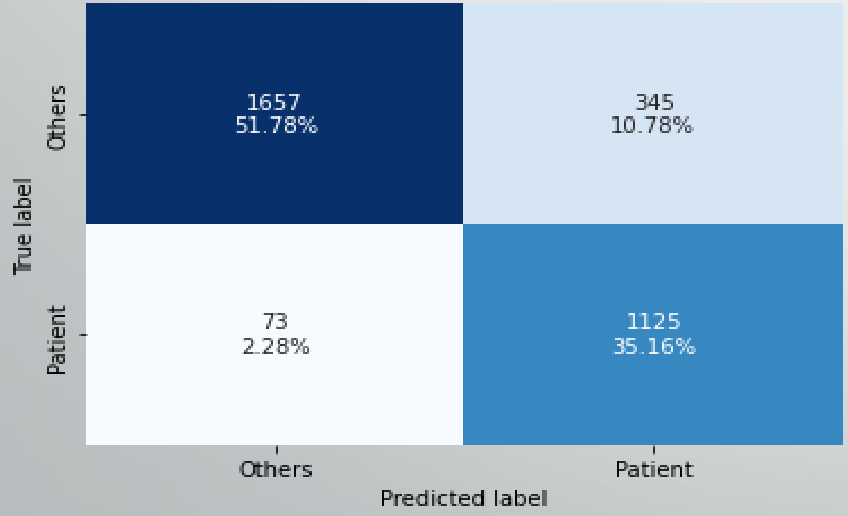


Couple of Examples with Incorrect Label and Low Prediction

Quantify Results With BERT/BioBERT

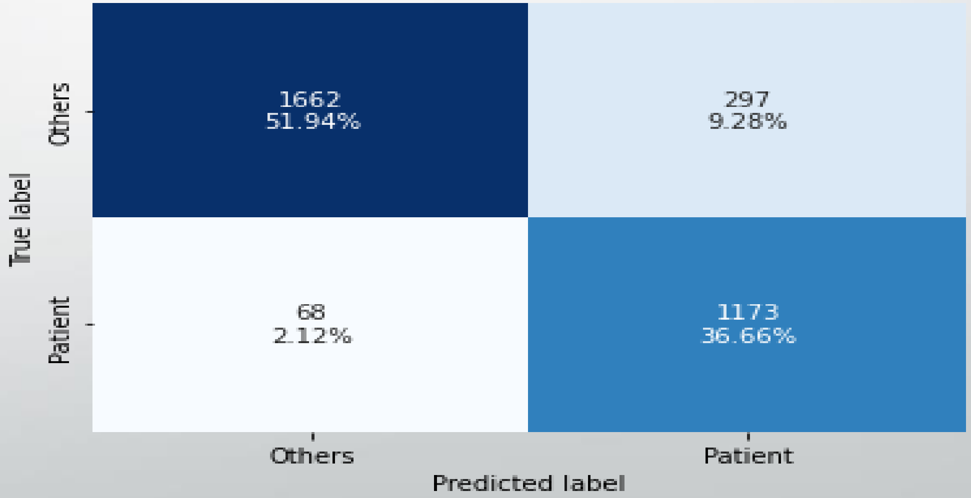
We achieved good accuracy with both models, and we chose Active Learning for this use case given:

- Weak supervision needs more initial training data compared to Active learning (More manual annotation - additional expense)
- The robustness of the model depends on the quality of rules and domain expertise is required to frame better rules



Accuracy=0.869
Precision=0.765
Recall=0.939
F1 Score=0.843

Metrics on Test Dataset (BERT Model)



Accuracy=0.886
Precision=0.798
Recall=0.945
F1 Score=0.865

Metrics on Test Dataset (BioBert Model)

Final Results

Labeled Corpus (61,021 tweets) generated via Active learning is used to fine-tune the base models of BERT and BioBERT and the test results are validated on the initial manually annotated data.

Our model achieved an accuracy of 87% and 89% respectively on base models. To perform additional hyperparameter tuning before concluding the final model.

Experiment	Accuracy	Precision	Recall	F1 Score
BERT	86.9	76.5	93.9	84.3
BioBERT	88.6	79.8	94.5	86.5

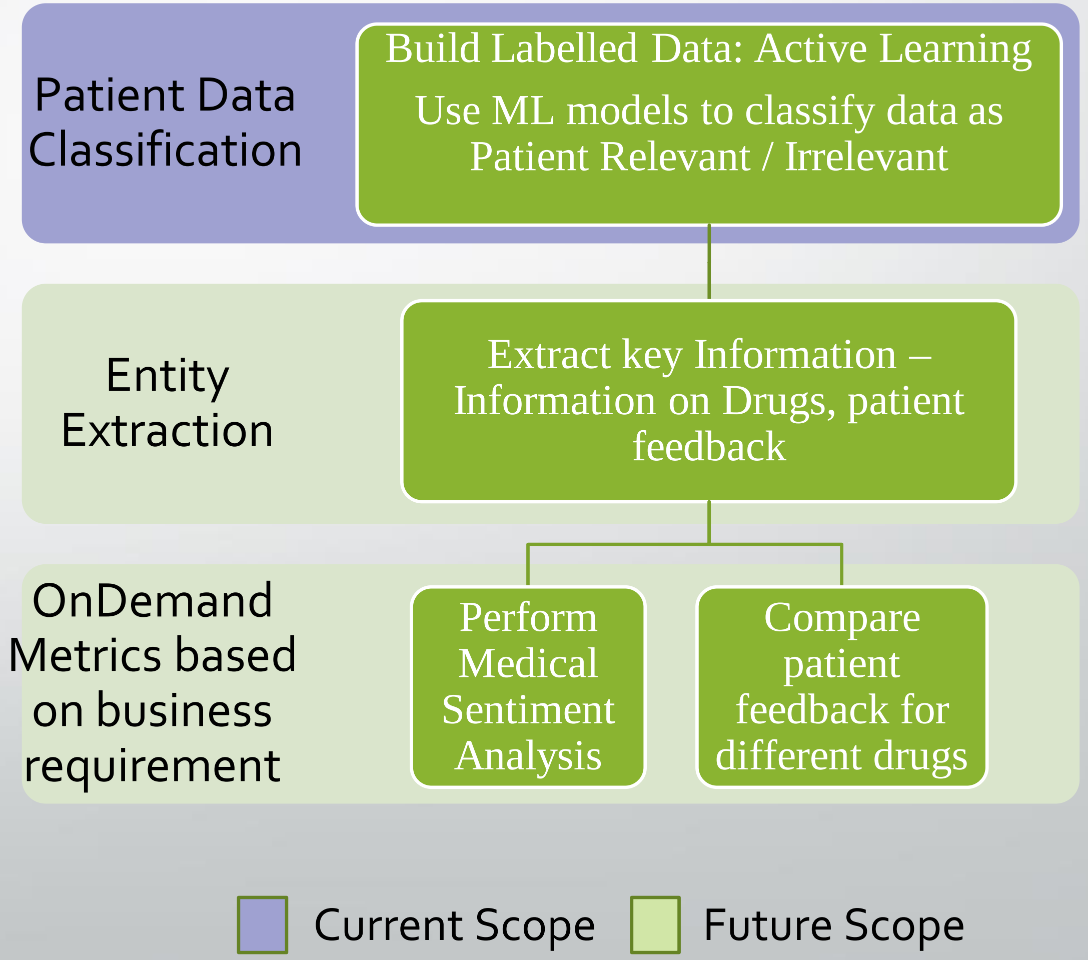
Additional Metrics for Binary Classification

Both models exhibit high recall, which indicates model can filter most of the patient-relevant experiences – the key functionality of our project !!

Summary

We address a common Industry use case -- There is no lack of models, but lack of trained/labeled examples to leverage AI

- Our study shows we can utilize SOTA algorithms without requiring spending hours manually annotating the data. – Promoting AI
- Invest time in more meaningful tasks
- The results have a high recall, indicating we can extract most of the patient-relevant experiences from the past two years just by annotating 3200 records.
- This unveiled data opens up a wide variety of possibilities that can benefit pharma industries and provide better patient care



Many thanks to Prathamesh Karmalkar for introducing me to this problem.

Thank You for listening!

Detailed Report and relevant code can be found at

https://github.com/PrasanthiDesiraju/TextClassification-of-Unlabeled-Data

Please reach out to me at prasanthi.desiraju@gmail.com for any questions, suggestions, or improvements.