#MyIBDHistory on Twitter
Predicting IBD Type From Personal Tweets

Maya Stemmer, Gilad Ravid, Yisrael Parmet
Department of Industrial Engineering and Management,
Ben-Gurion University of the Negev (BGU), Israel
mayast@post.bgu.ac.il

SOTICS 2022

Meet the presenter

Maya Stemmer

- PhD Student in the Department of Industrial Engineering and Management at Ben-Gurion University of the Negev, Israel.
- Research area: social media mining for health insights.

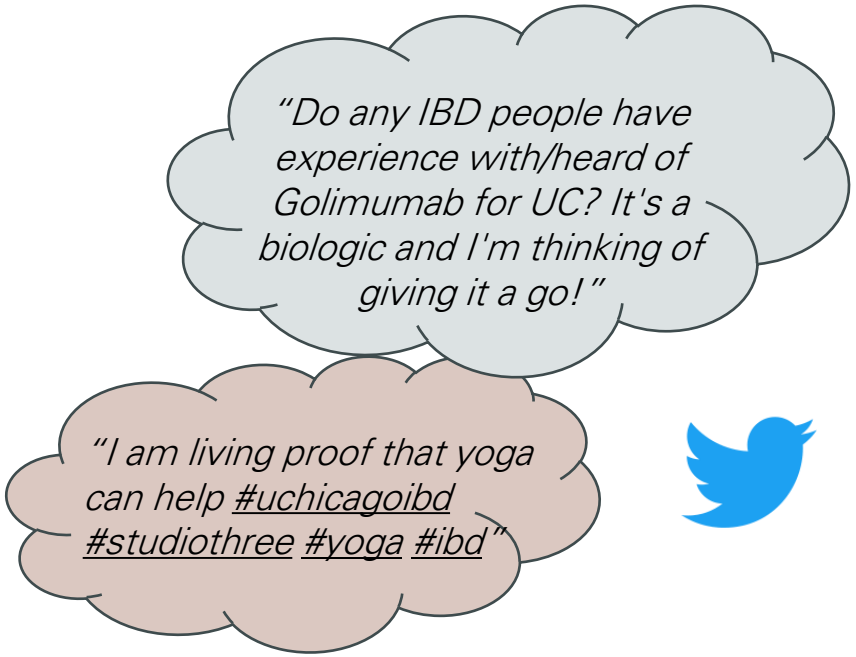
Recent publications

Stemmer, M., Parmet, Y., & Ravid, G. (2022). Identifying Patients With Inflammatory Bowel Disease on Twitter and Learning From Their Personal Experience: Retrospective Cohort Study. Journal of Medical Internet Research, 24(8), e29186. https://doi.org/10.2196/29186.

Stemmer, M., Parmet, Y., & Ravid, G. (2021, November). What Are IBD Patients Talking About on Twitter?. In International Conference on ICT for Health, Accessibility and Well-being (pp. 206-220). Springer, Cham. https://doi.org/10.1007/978-3-030-94209-0_18.

Introduction

Motivation



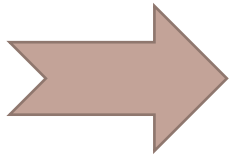
Learning from Patients' Personal Experience

- Social networks serve as alternative information sources for patients.
- Patients use social media to share **daily experiences** including health and treatment information.
- Chronically ill patients exchange experiential knowledge about their everyday life with the disease that extends far beyond medical care.
- Mining these informative conversations may shed some light on patients' ways of life and **support the research of chronic conditions**.

Inflammatory Bowel Disease (IBD)

- Chronic inflammation condition of the digestive system
- Flares and remissions
- Two main diseases: Crohn's Disease (CD), Ulcerative Colitis (UC)
- Inconclusive cases: IBD Unclassified (IBD-U)

Characteristics	Chronic and incurable
	Lifestyle changes can help with symptoms
	Diversity in treatments' effectiveness
Social Implications	Disruption of daily activities
	Embarrassment
	Lack of public awareness



Mining social media

#MyIBDHistory on Twitter

- The hashtag #MyIBDHistory was initiated in 2018 by a Twitter account promoting IBD-related discussion called @bottomlineibd.
- Patients with IBD wrote their own medical history in a single tweet and signed it with the hashtag #MyIBDHistory.
- In their tweets, patients mentioned their disease type (CD or UC), their age at diagnosis, the medications they have tried over the years, whether they underwent any surgeries, and more.

> *"UC, diagnosed 2011 (age 30). Colonoscopy showed severe ulceration. Asacol didn't work, pred did. Tried Pentasa, more pred, Balsalazide (still on 9/day). 3 flares in a year lead to Aza and Mercap but intolerant to both. On Golimumab for a year, still seems to work. #MyIBDHistory"*

Research questions

Is it possible to derive insights into the well-being of patients with IBD by mining Twitter data?

Can we use the information shared by patients with IBD on Twitter to determine whether a patient with IBD suffers from CD or UC?

What are the key features that help to distinguish between CD and UC?

Methodology

Overview

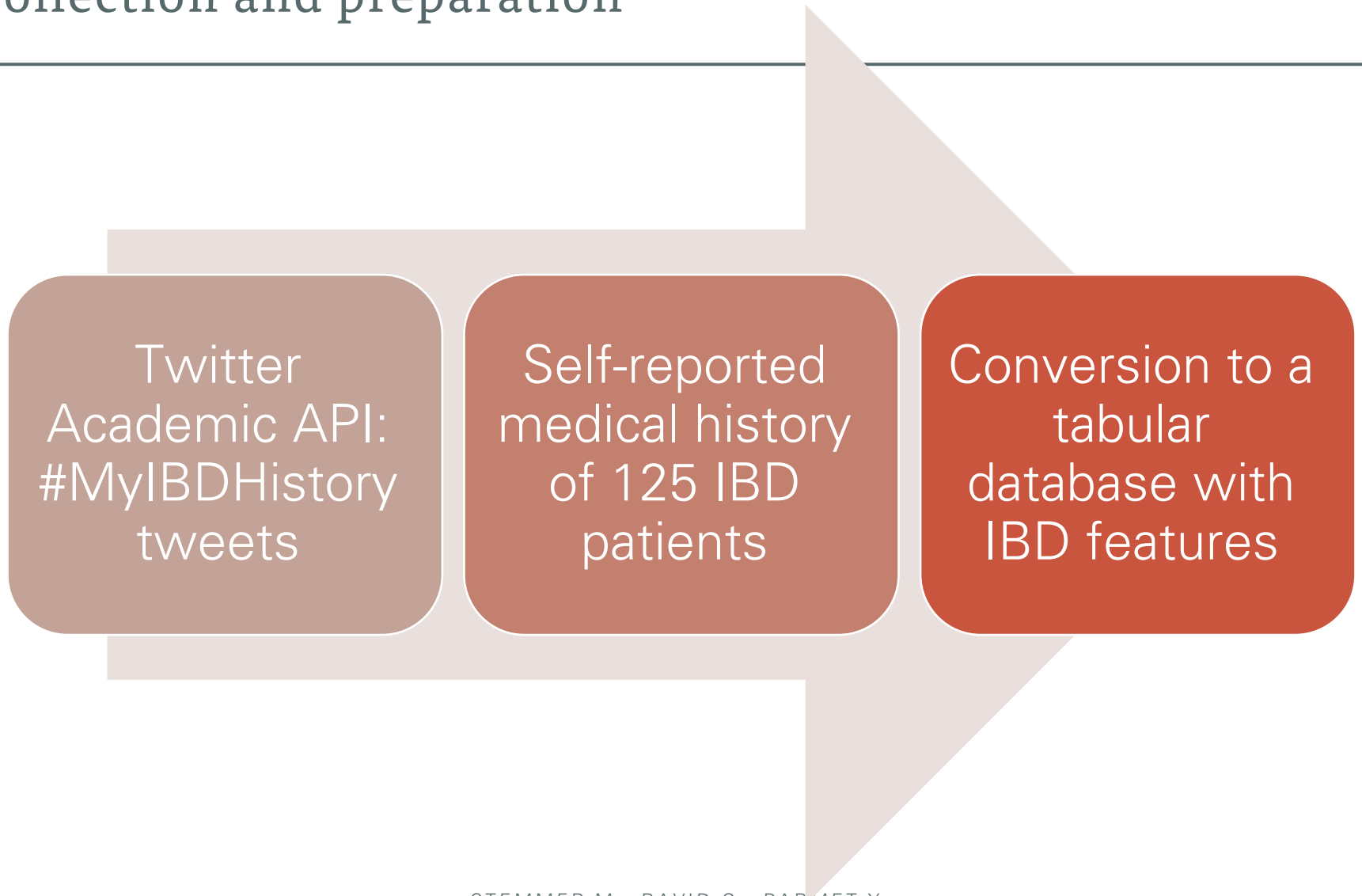
This research aimed to analyze patients' tweets containing the hashtag #MyIBDHistory and to determine the disease type of an IBD patient based on their symptoms and treatments.

We constructed a list of classification features and used LASSO logistic regression to predict whether a patient suffered from CD or UC.

We identified key features and our results correlated with IBD-related research.

To adhere to ethical norms and maintain user privacy, we publish aggregated results that do not reveal the specific users. Directed quotes are presented here after we obtained informed consent from their authors.

Data collection and preparation



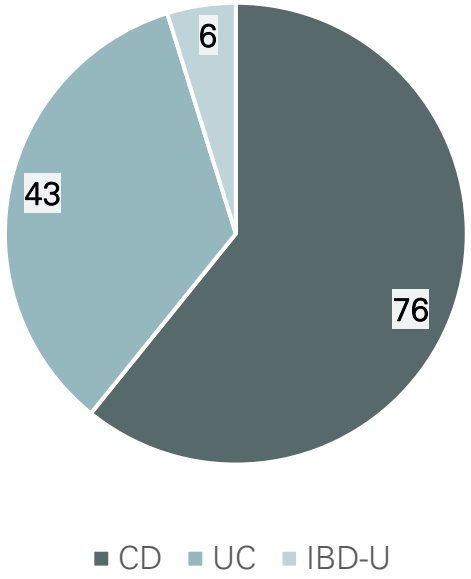
Data collection and preparation – example

"UC, diagnosed 2011 (age 30). Colonoscopy showed severe ulceration. Asacol didn't work, pred did. Tried Pentasa, more pred, Balsalazide (still on 9/day). 3 flares in a year lead to Aza and Mercap but intolerant to both. On Golimumab for a year, still seems to work. #MyIBDHistory"

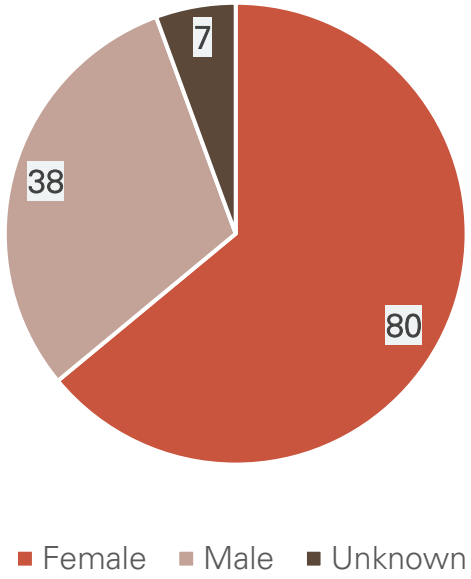
Disease type	Gender	Age group	Meds: anti-inflammatory	Meds: steroids	Meds: antibiotics	Meds: biologics	Meds: Immune suppressors	Meds: others	Wrong diagnosis	Fistula	Weight	Diet	Pre-diagnosis	Hospital
UC	M	15-35	1	1	0	1	1	0	0	0	0	0	0	0

A bit of descriptive statistic...

Disease type

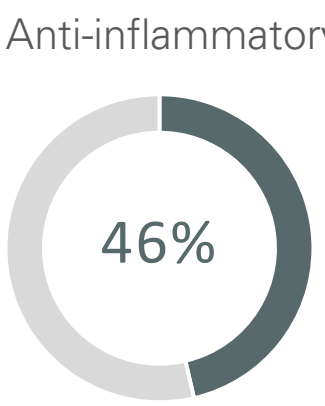


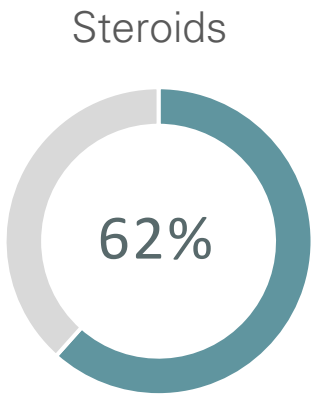
Gender

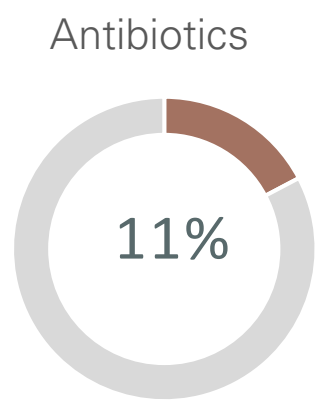


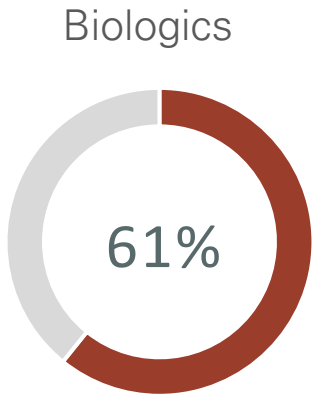
A bit of descriptive statistic...

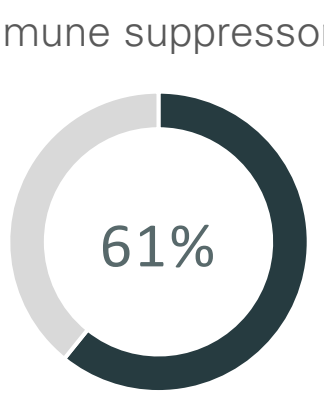
Medications

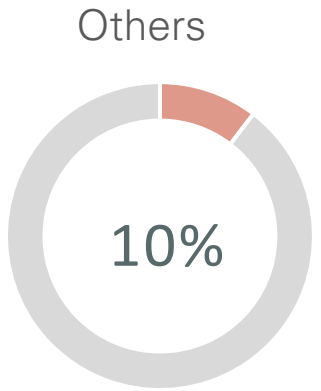




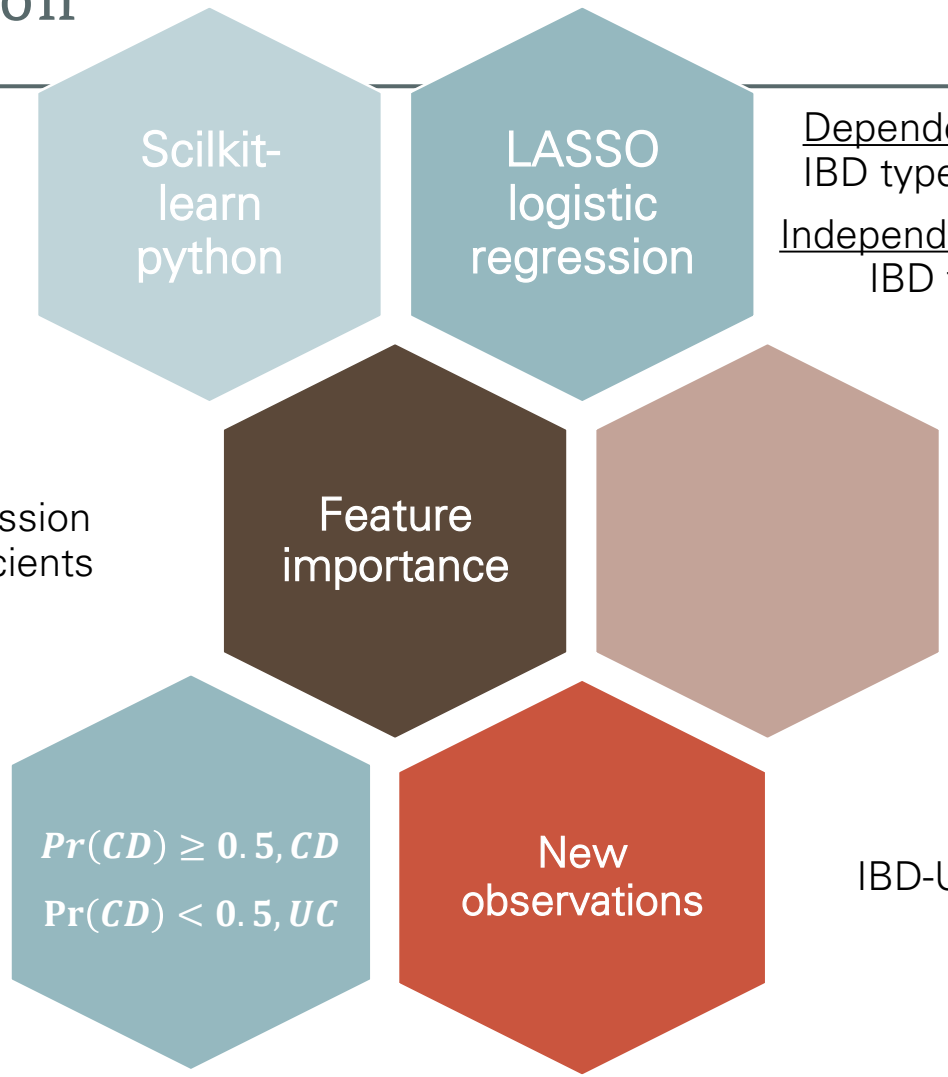








Disease type prediction



Dependent variable:
IBD type (CD or UC)

Independent variables:
IBD features

Regression coefficients

IBD-U patients

Results

Disease type prediction

Regression model evaluation results

Evaluation Measure	Evaluation Data	
	Test Set	Entire Dataset
Accuracy	0.75	0.7563
Precision	0.7273	0.7527
Recall	1.0	0.9211
F1	0.8421	0.8284
Area Under the Receiver Operating Characteristic Curve (AUC ROC)	0.625	0.6931

Confusion matrices for the regression model

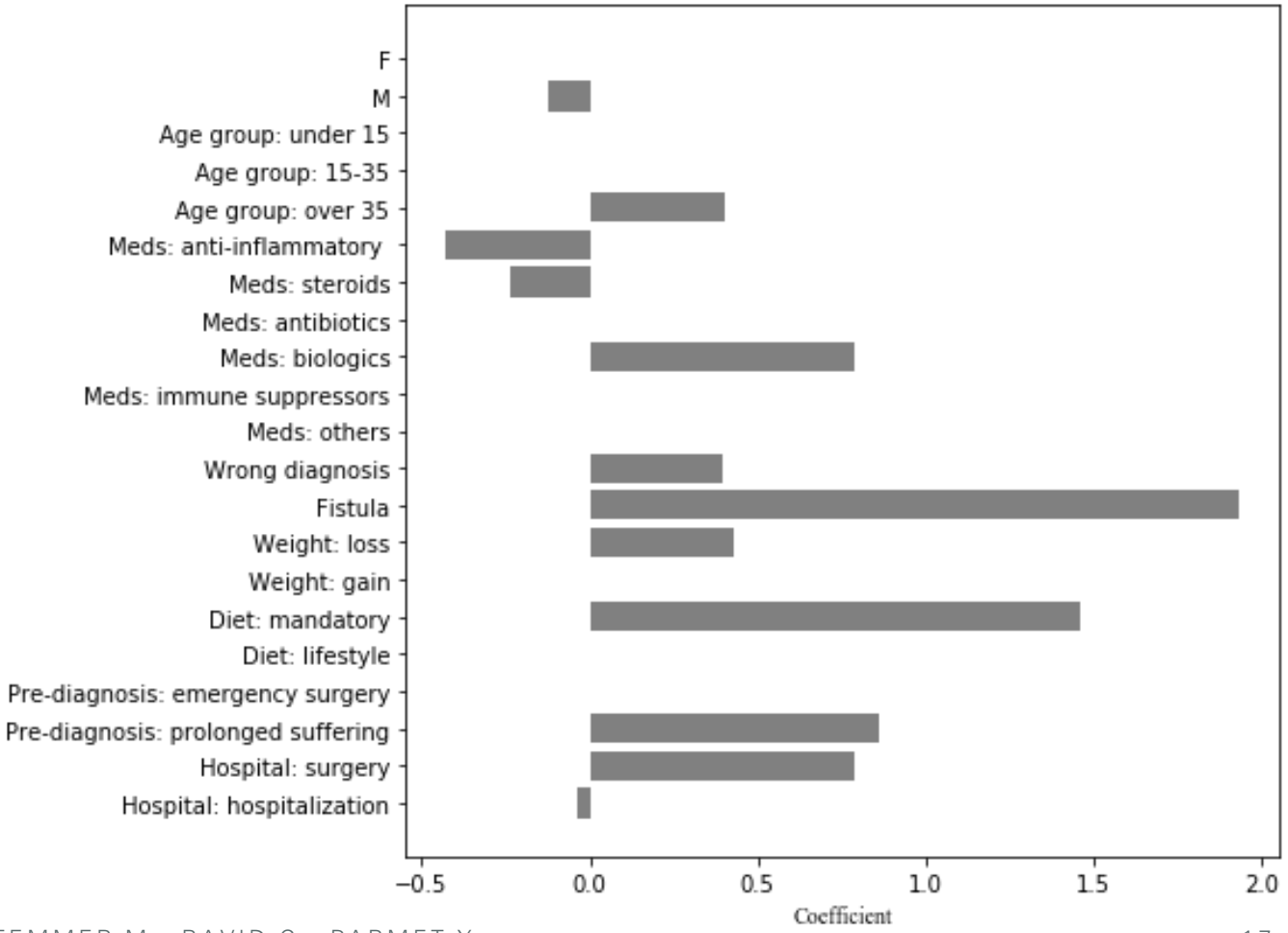
Evaluation Data		Predictions	
		Predicted UC	Predicted CD
Test Set	True UC	2	6
	True CD	0	16
Entire Dataset	Ture UC	20	23
	True CD	6	70

Feature importance

Strong (CD) features:

1. Fistula
2. Diet: mandatory
3. Pre-diagnosis: prolonged suffering
4. Meds: biologics
5. Hospital: surgery

Bar chart of feature importance based on regression coefficients



IBD-U prediction

Features and predictions for IBD-U patients

Highly confident classification

Confident classification

Inconclusive classification

Feature/ Prediction	Patient IBD1	IBD2	IBD3	IBD4	IBD5	IBD6
Gender: female	0	1	1	0	1	1
Gender: male	1	0	0	1	0	0
Age group: under 15	0	0	1	0	0	0
Age group: 15-35	0	0	0	1	0	0
Age group: over 35	1	0	0	0	1	0
Meds: anti-inflammatory	1	0	0	0	1	0
Meds: steroids	1	0	1	1	1	0
Meds: antibiotics	0	0	1	0	0	0
Meds: biologics	**1**	0	**1**	**1**	1	1
Meds: immune suppressors	1	0	1	0	0	0
Meds: others	0	0	0	0	0	0
Wrong diagnosis	0	0	0	0	0	0
Fistula	0	0	**1**	0	0	0
Weight: loss	0	0	1	0	0	0
Weight: gain	0	0	0	0	0	0
Diet: mandatory	0	0	**1**	0	0	0
Diet: lifestyle	0	0	0	0	0	0
Pre-diagnosis: emergency surgery	0	1	0	0	0	0
Pre-diagnosis: prolonged suffering	**1**	0	0	0	0	0
Hospital: surgery	0	**1**	**1**	**1**	0	0
Hospital: hospitalization	1	0	0	0	0	0
Probability	0.622	0.567	0.994	0.603	0.428	0.589
Class	1	1	1	1	0	1

IBD-U prediction

- **Patient IBD3**: fistula + mandatory diet change, the two strongest features => unambiguous CD classification.
- **Patients IBD1 and IBD4**: biologic meds + prolonged suffering/ surgery => CD classification.
- **Patients IBD2, IBD5, and IBD6**: only one medium level feature => indecisive classification with probabilities ~0.5.
- **Patient IBD5**: anti-inflammatory meds + steroids, the strongest UC features => UC classification.

	Patient					
Feature/ Prediction	IBD1	IBD2	IBD3	IBD4	IBD5	IBD6
Gender: female	0	1	1	0	1	1
Gender: male	1	0	0	1	0	0
Age group: under 15	0	0	1	0	0	0
Age group: 15-35	0	0	0	1	0	0
Age group: over 35	1	0	0	0	1	0
Meds: anti-inflammatory	1	0	0	0	1	0
Meds: steroids	1	0	1	1	1	0
Meds: antibiotics	0	0	1	0	0	0
Meds: biologics	**1**	0	**1**	**1**	**1**	**1**
Meds: immune suppressors	1	0	1	0	0	0
Meds: others	0	0	0	0	0	0
Wrong diagnosis	0	0	0	0	0	0
Fistula	0	0	**1**	0	0	0
Weight: loss	0	0	1	0	0	0
Weight: gain	0	0	0	0	0	0
Diet: mandatory	0	0	**1**	0	0	0
Diet: lifestyle	0	0	0	0	0	0
Pre-diagnosis: emergency surgery	0	1	0	0	0	0
Pre-diagnosis: prolonged suffering	**1**	0	0	0	0	0
Hospital: surgery	0	**1**	**1**	**1**	0	0
Hospital: hospitalization	1	0	0	0	0	0
Probability	0.622	0.567	0.994	0.603	0.428	0.589
Class	1	1	1	1	0	1

Discussion

Principal findings

- Analysis of #MyIBDHistory tweets: extraction of categorical features from natural language text describing the medical history of IBD patients.
- LASSO logistic regression: prediction of IBD type (CD or UC) and identification of key features that help distinguish CD from UC.
- The two strongest CD features – having a fistula and resorting to mandatory diet changes – two common CD phenomena that seldom occurs with UC.
- The LASSO regulation helped to eliminate unnecessary independent classification variables.
- Differentiation complexity: with CD distinctive characteristics – confident CD classification. Otherwise, equivocal classification.

Limitations and future research

Relatively small patient dataset

Enriching the dataset by identifying more patients on Twitter or expanding the search to other social media.

Imbalanced data with 0.5 classification threshold

Balancing the groups by adding more UC patients or changing the classification threshold in favor of UC.

Contribution and Conclusion

- The first study to use social media data for differentiating between CD and UC.
- The first study to use the #MyIBDHistory hashtag to scientifically draw conclusions for IBD.
- The methods can also help to explore other medical conditions.
- Findings from Twitter research on IBD patients correlate with existing medical knowledge regarding the disease => Room for collaboration between physicians and engineers.
- Further mining Twitter for health-related data may complement and enhance healthcare research.

Thank you

Maya Stemmer

Ben-Gurion University of the Negev

mayast@post.bgu.ac.il