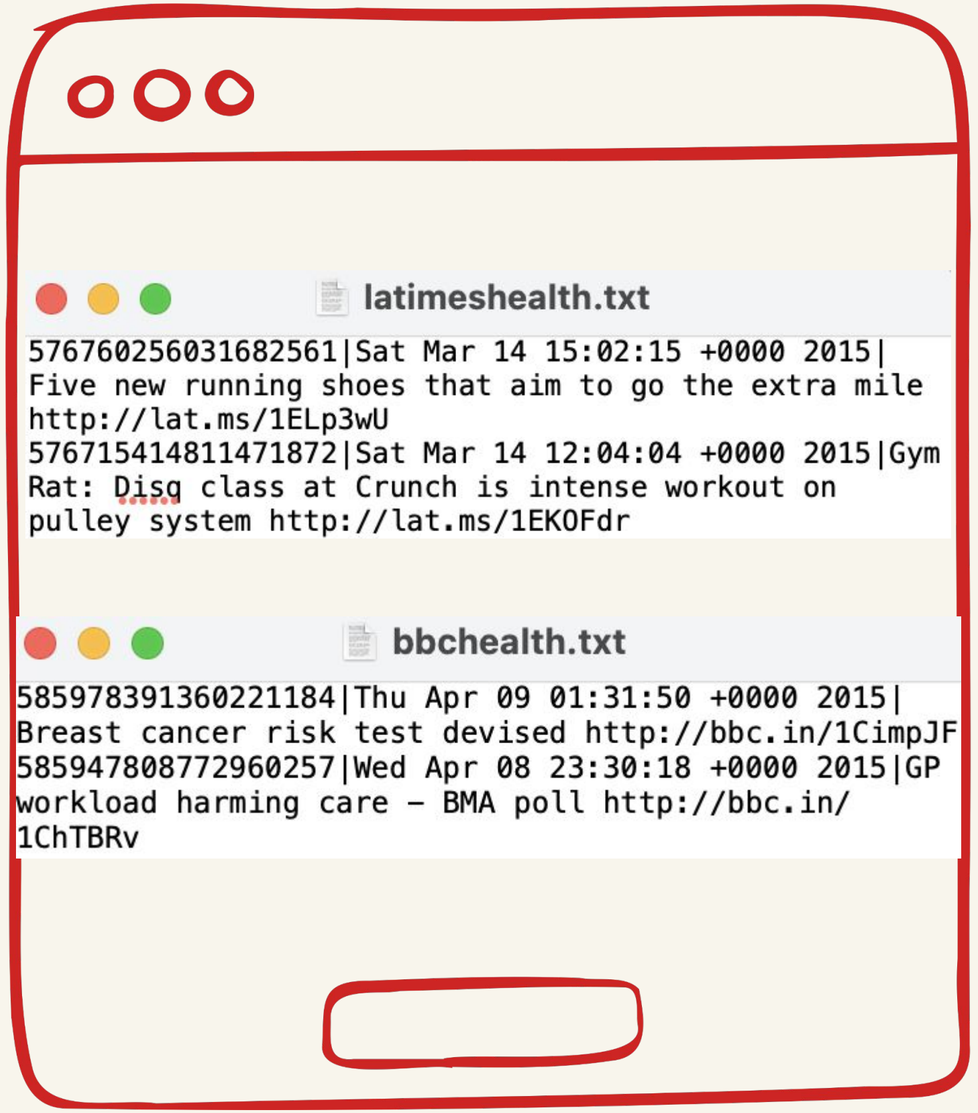
Health Tweets Classification & Recommendation

Yuki Yu & Naiqi Zhang

Data Overview

Health News in Twitter (UCI ML Repository)

- 27771 Tweets
  - tweet id|date & time|headline
- 16 News Sources (eg. LA Times, BBC, NPR, etc.)
- Up to 2015



1

Data Preprocessing

Processing the text data and embedding them

2

Labeling

Labeling the unclassified tweets

3

Classification Modeling

Modeling to classify unlabeled tweets

4

Recommendation

Recommend tweets based on user viewing history

Data Preprocessing

01 Standardization & Extraction

- Removed punctuation, symbols, etc.
- Extracting headlines, source, and date/time → Dataframe

02 BERT Tokenization

- Breaks down text into tokens → feed to models

03 BERT Embeddings

- Text → dense numerical vectors → semantic meaning and context

BERT is a Transformer-based NLP model that understands words in context, enabling tasks like text classification and question answering.

Labeling

01 ZeroShot Classification

- Using semantic understanding →Predicting category of headlines w/out prior training
  - Fitness, Viruses, Mental Health, Other, etc.

02 Keywords Classification

- Classifying based on common keywords in each category
  - Gym → Fitness, Ebola → Viruses

03 Iterate through 1 & 2 to refine the “Other” Category!

Categories

Compiled/regrouped all categories into 10 main ones:

1. **Health and Wellness**
2. **Viruses, Diseases, Pharmaceuticals**
3. **Public Health & Policy**
4. **Mental, Behavioral, & Neurological Health**
5. **Reproductive & Gender Health**
6. **Health Technology & Medical Research**
7. **Patient Care & Specialized Health**
8. **Cardiovascular & Organ Health**
9. **Disabilities & Diversities**
10. **Other**

Other models tried: Random Forest, XGBoost, OCT

Neural Networks

70-30 Train-Test-Split

Combating Class Imbalance

SMOTE: generates synthetic samples for minority classes based on nearest neighbors → creates balanced dataset

- Higher overall accuracy without SMOTE - likely due to inflated accuracy from majority classes
- Higher F1-scores with SMOTE

Use SMOTE because it classifies minority classes better, with only a 2-3% reduction in overall accuracy.

Hyperparameter Tuning

Bayesian Optimization: optimization process that searches for optimal hyperparameters by modeling the objective function and iteratively selecting the most promising values to evaluate

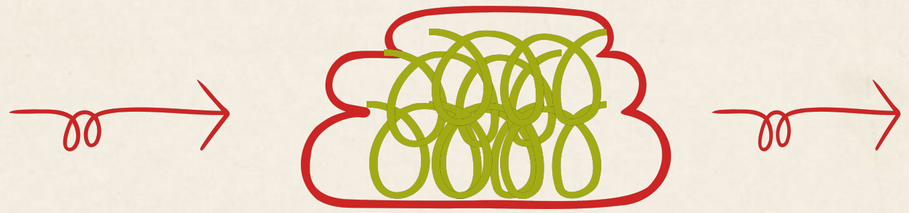
Baseline Majority Test Accuracy: 19.74%

Final Test Accuracy: 57.52%

User's Profile

Users' Profiles

- **Unlabeled tweeter news**
- **Recent 100 views**
- **Simulated from test data**



Labeling

- **Classification of Twitter news**
- **Use trained Neural Network model**

	Source	Text	Datetime
0	reuters	House Oversight Committee holds hearing on wha...	2015-03-13 17:43:06+00:00
1	cbchealth	Birth control pill recall expands to Esme-28 a...	2015-03-12 23:25:46+00:00
2	goodhealth	@jenniegarth is our guest blogger today! Get h...	2015-03-11 16:05:44+00:00
3	cbchealth	Urban flooding likely to worsen, say experts	2015-03-10 22:40:21+00:00
4	latimes	The American Dietetic Assn. gets a new name	2015-03-10 20:57:41+00:00
5	latimes	For Russian men, the more vodka you drink, the...	2014-01-04 19:39:17+00:00
6	goodhealth	No olive oil on hand? Try using one of these 3...	2013-10-06 8:30:26+00:00
7	msn	Hospital Programs to Reduce Antibiotic Resista...	2013-10-06 9:13:45+00:00
8	npr	Freaky Friday: Autonomous Tissue Grabbers Are ...	2011-12-08 19:30:16+00:00

Category
Public Health and Policy
Public Health and Policy
Health and Wellness
Public Health and Policy
Health and Wellness
Mental, Behavioral, and Neurological Health
Reproductive and Gender Health
Public Health and Policy
Other

Recommendation system

Approach

- Identify the top 3 categories based on frequency
- Apply text vectorization to transform the news content into numerical format
- Compute cosine similarity scores between the vectorized texts
- For each of the top categories, recommend 5 news articles with the highest averaged similarity scores

Scikit-learn TfidfVectorizer: Convert texts into a numerical matrix, capturing the importance of words across texts
Scikit-learn Cosine Similarity: Measure the similarity between two vectors

User's Profile:

	Source	Text	Datetime	Category
0	reuters	House Oversight Committee holds hearing on wha...	2015-03-13 17:43:06+00:00	Public Health and Policy
1	cbchealth	Birth control pill recall expands to Esme-28 a...	2015-03-12 23:25:46+00:00	Public Health and Policy
2	goodhealth	@jenniegarth is our guest blogger today! Get h...	2015-03-11 16:05:44+00:00	Health and Wellness
3	cbchealth	Urban flooding likely to worsen, say experts	2015-03-10 22:40:21+00:00	Public Health and Policy
4	latimes	The American Dietetic Assn. gets a new name	2015-03-10 20:57:41+00:00	Health and Wellness
5	latimes	For Russian men, the more vodka you drink, the...	2014-01-04 19:39:17+00:00	Mental, Behavioral, and Neurological Health
6	goodhealth	No olive oil on hand? Try using one of these 3...	2013-10-06 8:30:26+00:00	Reproductive and Gender Health
7	msn	Hospital Programs to Reduce Antibiotic Resista...	2013-10-06 9:13:45+00:00	Public Health and Policy
8	npr	Freaky Friday: Autonomous Tissue Grabbers Are ...	2011-12-08 19:30:16+00:00	Other

Recommendation List:

	Source	Text	Category
0	KaiserHealthNews	House Oversight Panel Calls On Gruber, Tavenne...	Public Health and Policy
1	KaiserHealthNews	RT @philgalewitz: Willl be live tweeting the S...	Public Health and Policy
2	nytimes	U.S. Finds Many Failures in Medicare Health Plans	Public Health and Policy
3	reuters	Oregon holds first public hearing on proposed ...	Public Health and Policy
4	npr	A Single Insurer Holds Obamacare's Fate In 2 S...	Public Health and Policy

- Use more accurate labels for categorization
- Refine the neural network model to achieve higher prediction accuracy
- Apply matrix completion techniques to construct archetypal user profiles with latent interaction vectors
- Update recommendations dynamically based on user engagement and interaction
- ...

Future Direction

Thank You!

— Questions, Comments, Concerns?