CS 1671 / CS 2071 / ISSP 2071
Human Language Technologies

Session 15: Project proposal presentations

March 4, 2026

Schedule

1. Raina, Brett, Hannah, Aidan, Kee
2. Fatimah, Heather, Vivien, Ifemi, Ryder
3. Daley, Cole, Griffin, Jeana, David
4. Joshua, Forest, Rose, Chris, Raymond
5. Jack, Hongyu, Patrick, Aaron, Marcus
6. Laxmi, Ciara, Sanjana, Yifei, Irisin
7. Matthew, Enzo, Nate, Owen, Kevin, Nihal
8. Grace, Michelle, Kiana, Sarina, Amyia
9. Ryan, Pier, Justin, Wyatt, Praz
10. Jonah, Amanda, Lyndsey, Saung, Jay

Instructions

- Plan for **5 min presentations max** not including Q&A
- Cover at least these key points
  - Project motivation (what is the value of this work?)
  - What data you are planning to use
  - What approach/methods you plan to take
  - How you will evaluate your approach
- Put your slides in this presentation after your project name slide by **class session, 1pm on Wed Mar 4**

1. Raina, Brett, Hannah, Aidan, Kee

Movie Summarization

By: Raina, Brett, Hannah, Aidan, Kee

Motivation (value of the work)

Can check to see if summaries posted on sites like Wikipedia & IMDB are accurate

Generate summaries for movies that lack them

Remember key plot points of a movie you just watched

Data

- Custom dataset of movie subtitles and gold summaries
- Subtitles come from Open Subtitles
- Summaries come from Wikipedia
- 1322 unique datapoints
- Average of 1255 lines of dialog per movie
- Average of 7942 words per movie

Approach/methods

Feed to LLMs in zero-shot approach

Given only the subtitles and prompt to summarize

Compare results from Deepseek-R1 and Llama 3.1

Evaluation

2 key metrics

- ROUGE score measures accuracy
- SummaC measures faithfulness

Human evaluation too (on small subset of data)

2. Fatimah, Heather, Vivien, Ifemi, Ryder

Classifying Restaurants by Yelp Reviews

Fatimah Alfaraj, Heather Guzik, Vivien Lim,
Ifemi Olojo-Kosoko, Ryder Pham

Data

Dataset Source:

1. Yelp Review Data Set (has over 6 million reviews)
2. We will filter it down locations in Pittsburgh only
3. Focused on Restaurants only

Target Labels:

- Cuisine Types (derived from categories)
- Formality level (derived from business attributes):
  - Causal
  - Mid-range
  - Fine dining

Data

https://soyummy.com/entertaining-so-yummy/17-funniest-reviews-yelp/

Review dataset

Columns in dataset: ['review_id', 'user_id', 'business_id', 'stars', 'useful', 'funny', 'cool', 'text', 'date']
Loaded 25 rows

First 25 rows of the dataset:

	review_id	user_id	business_id	stars	useful	funny	cool	text	date
0	KU_O5udG6zpxOg-VcAEodg	mh_-eMZ6K5RLWhZyISBhwA	XQfwVwDr-v0ZS3_CbbE5Xw	3	0	0	0	If you decide to eat here, just be aware it is...	2018-07-07 22:09:11
1	BiTunyQ73aT9WBnpR9DZGw	OyoGAe7OKpv6SyGZT5g77Q	7ATYjTIgM3jUlt4UM3IyqQ	5	1	0	1	I've taken a lot of spin classes over the year...	2012-01-03 15:28:18
2	saUsX_uimxRlCVr67Z4Jig	8g_iMtfSiwikVnbP2etR0A	YjUWPpI6HXG530lwP-fb2A	3	0	0	0	Family diner. Had the buffet. Eclectic assortm...	2014-02-05 20:30:30
3	AqPFMleE6RsU23_auESxiA	_7bHUi9Uuf5__HHc_Q8guQ	kxX2SOes4o-D3ZQBkiMRfA	5	1	0	1	Wow! Yummy, different, delicious. Our favo...	2015-01-04 00:01:03
4	Sx8TMOWLNuJBWer-0pcmoA	bcjbaE6dDog4jkNY91ncLQ	e4Vwtrqf-wpJfwesgvdgxQ	4	1	0	1	Cute interior and owner (?) gave us tour of up...	2017-01-14 20:54:15
5	JrlxlS1TzJ-iCu79ul40cQ	eUta8W_HdHMXPzLBBZhL1A	04UD14gamNjLY0IDYVhHJg	1	1	2	1	I am a long term frequent customer of this est...	2015-09-23 23:10:31
6	6AxgBCNX_PNTOxmbRSwcKQ	r3zeYsv1XFBRA4dJpL78cw	gmjsEdUsKpj9Xxu6pdjH0g	5	0	2	0	Loved this tour! I grabbed a groupon and the p...	2015-01-03 23:21:18
7	_ZeMknuYdlQcUqng_Im3yg	yfFzsLmaWF2d4Sr0UNbBgg	LHSTtnW3YHCeUkRDGyJOyw	5	2	0	0	Amazingly amazing wings and homemade bleu chee...	2015-08-07 02:29:16
8	ZKvDG2sBvHVdF5oBNUOpAQ	wSTuiTk-sKNdcFyprzZAjg	B5XSoSG3SfvQGtKEGQ1tSQ	3	1	1	0	This easter instead of going to Lopez Lake we ...	2016-03-30 22:46:33
9	pUycOfUwM8vqX7KjRRhUEA	59MxRhNVhU9MYndMkz0wtw	gebiRewfieSdtt17PTW6Zg	3	0	0	0	Had a party of 6 here for hibachi. Our waitres...	2016-07-25 07:31:06
10	rGQRf8UafX7OTlMNN19I8A	1WHRWwQmZOZDAhp2Qyny4g	uMvVYRgGNXf5boolA9HXTw	5	2	0	0	My experience with Shalimar was nothing but wo...	2015-06-21 14:48:06
11	I3Wk_mvAog6XANluGQ9C7Q	ZbqSHbgCjzVAqaa7NKWn5A	EQ-TZ2eeD_E0BHuvoaeG5Q	4	0	0	0	Locals recommended Milktooth, and it's an amaz...	2015-08-19 14:31:45
12	XW_LfMv0fV2119c6xQd_Iw	9OAtfnWag-ajVxRbUTGlyg	lj-E32x9_FA7GmUrBGBEWg	4	0	0	0	Love going here for happy hour or dinner! Gre...	2014-06-27 22:44:01
13	8JFGBuHMoiNDyfcxuWNtrA	smOvOajNG0lS4Pq7d8g4JQ	RZtGWDLCAtuipwaZ-UfjmQ	4	0	0	0	Good food--loved the gnocchi with marinara\nth...	2009-10-14 19:57:14
14	UBp0zWyH60Hmw6Fsasei7w	4Uh27DgGzsp6PqrH913giQ	otQS34_MymijPTdNBoBdCw	4	0	2	0	The bun makes the Sonoran Dog. It's like a snu...	2011-10-27 17:12:05
15	OAhBYw8IQ6wlfw1owXWRWw	1C2lxzUo1Hyye4RFIXly3g	BVndHaLihEYbr76Z0CMEGw	5	0	0	0	Great place for breakfast! I had the waffle, w...	2014-10-11 16:22:06
16	oyaMhzBSwfGgemSGuZCdwQ	Dd1jQj7S-BFGqRbApFzCFw	YtSqYv1Q_pOltsVPSx54SA	5	0	0	0	Tremendous service (Big shout out to Douglas) ...	2013-06-24 11:21:25
17	LnGZB0fjfgeVDVz5IHuEVA	j2wlzrntrbKwyOcOiB3l3w	rBdG_23USc7DletfZ11xGA	4	1	0	0	The hubby and I have been here on multiple occ...	2014-08-10 19:41:43
18	u2vzZaOqJ2feRshaaF1doQ	NDZvyYHTUWWu-kqgQzzDGQ	CLEWowfkj-wKYJlQDqT1aw	5	2	0	1	I go to blow bar to get my brows done by natal...	2016-03-07 00:02:18
19	Xs8Z8lmKkosqW5mw_sVAoA	lQsF3Rc6lgCzjVV9DE8KXg	eFvzHawVJofxSnD7TgbZtg	5	0	0	0	My absolute favorite cafe in the city. Their b...	2014-11-12 15:30:27

Business dataset

Columns in dataset: ['business_id', 'name', 'address', 'city', 'state', 'postal_code', 'latitude', 'longitude', 'stars', 'review_count', 'is_open', 'attributes', 'categories', 'hours']
Loaded 25 rows

First 25 rows of the dataset:

	business_id	name	address	city	state	postal_code	latitude	longitude	stars	review_count	is_open	attributes	categories	hours
0	Pns2l4eNsfO8kk83dixA6A	Abby Rappoport, LAC, CMQ	1616 Chapala St, Ste 2	Santa Barbara	CA	93101	34.426679	-119.711197	5.0	7	0	{'ByAppointmentOnly': 'True'}	Doctors, Traditional Chinese Medicine, Naturop...	None
1	mpf3x-BjTdTEA3yCZrAYPw	The UPS Store	87 Grasso Plaza Shopping Center	Affton	MO	63123	38.551126	-90.335695	3.0	15	1	{'BusinessAcceptsCreditCards': 'True'}	Shipping Centers, Local Services, Notaries, Ma...	{'Monday': '0:0-0:0', 'Tuesday': '8:0-18:30', ...
2	tUFrWirKiKi_TAnsVWiNQQ	Target	5255 E Broadway Blvd	Tucson	AZ	85711	32.223236	-110.880452	3.5	22	0	{'BikeParking': 'True', 'BusinessAcceptsCredit...	Department Stores, Shopping, Fashion, Home & G...	{'Monday': '8:0-22:0', 'Tuesday': '8:0-22:0', ...
3	MTSW4McQd7CbVtyjqoe9mw	St Honore Pastries	935 Race St	Philadelphia	PA	19107	39.955505	-75.155564	4.0	80	1	{'RestaurantsDelivery': 'False', 'OutdoorSeati...	Restaurants, Food, Bubble Tea, Coffee & Tea, B...	{'Monday': '7:0-20:0', 'Tuesday': '7:0-20:0', ...
4	mWMc6_wTdE0EUBKIGXDVfA	Perkiomen Valley Brewery	101 Walnut St	Green Lane	PA	18054	40.338183	-75.471659	4.5	13	1	{'BusinessAcceptsCreditCards': 'True', 'Wheelc...	Brewpubs, Breweries, Food	{'Wednesday': '14:0-22:0', 'Thursday': '16:0-2...
5	CF33F8-E6oudUQ46HnavjQ	Sonic Drive-In	615 S Main St	Ashland City	TN	37015	36.269593	-87.058943	2.0	6	1	{'BusinessParking': 'None', 'BusinessAcceptsCr...	Burgers, Fast Food, Sandwiches, Food, Ice Crea...	{'Monday': '0:0-0:0', 'Tuesday': '6:0-22:0', '...
6	n_0UpQx1hsNbnPUSlodU8w	Famous Footwear	8522 Eager Road, Dierbergs Brentwood Point	Brentwood	MO	63144	38.627695	-90.340465	2.5	13	1	{'BusinessAcceptsCreditCards': 'True', 'Restau...	Sporting Goods, Fashion, Shoe Stores, Shopping...	{'Monday': '0:0-0:0', 'Tuesday': '10:0-18:0', ...
7	qkRM_2X51Yqxk3btlwAQlg	Temple Beth-El	400 Pasadena Ave S	St. Petersburg	FL	33707	27.766590	-82.732983	3.5	5	1	None	Synagogues, Religious Organizations	{'Monday': '9:0-17:0', 'Tuesday': '9:0-17:0', ...

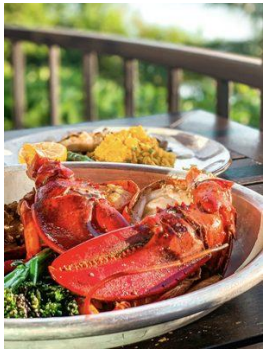
Approach & Method

Task:

- Multi-class text classification

Models:

- n-gram
- Logistic regression



Evaluation

$$F_1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

Dataset Split:

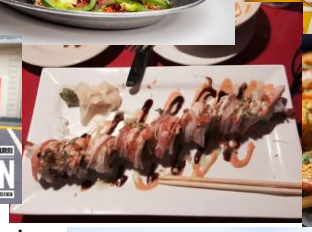
- Subset of dataset for testing

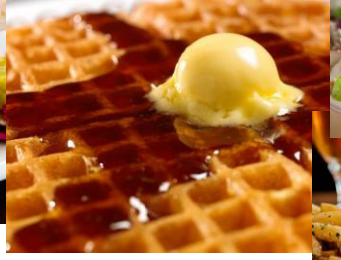
Performance metrics:

- Precision
- Recall
- F1-Score
- Accuracy?

Goal:

- Assess how well the model predicts
  - Cuisine Type
  - Formality level



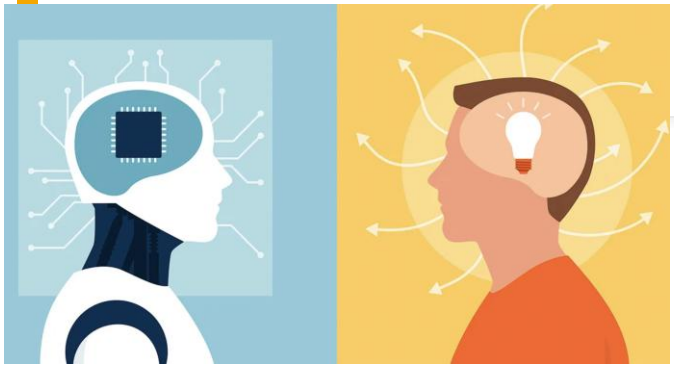


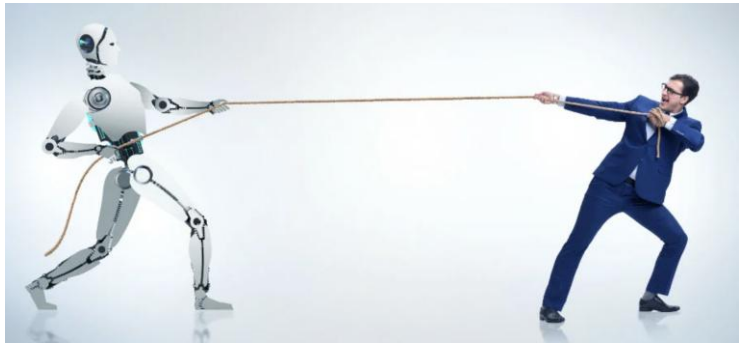
Q&A

3. Daley, Cole, Griffin, Jeana, David

Detecting Human vs. AI Generated Text

Project Motivation: Why Detect AI-Generated Text?





- AI-generated text is rapidly increasing across the internet
- Academic institutions face challenges verifying authorship
- False accusations of AI use can harm students
- AI detection remains an open research problem
- Even major AI companies have struggled with reliable detection

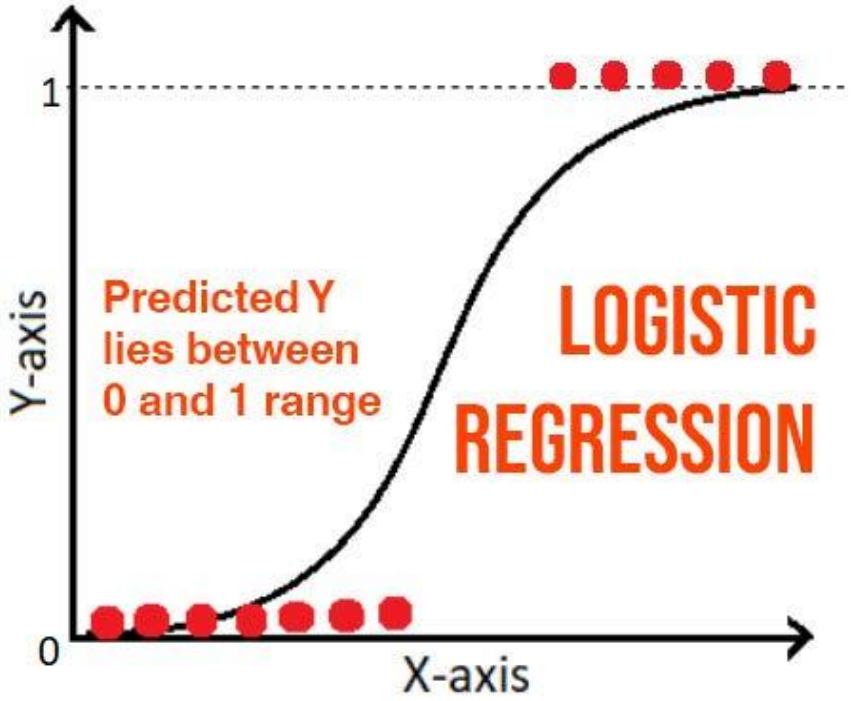
Dataset

label int64	text string
0	If you know the company you would like to work for, you can go to the "Careers" section of the company's website to apply for a position. If you…
0	As a deaf person , I 'd like to thank you for giving me way more credit than I probably deserve .
0	If you're aware of the weather patterns, you can prepare yourself properly before starting your hike, then properly evaluate when or whether to turn…
1	Charlie was having a lot of trouble getting a good night's sleep. His daytime productivity was starting to suffer severely. His doctor…
0	The Town of Ellenboro is a town located in Grant County, Wisconsin, United States. The population was 608 at the 2000 census. Transportation The town…

- A combination of human-written and machine-generated English text.
- 872,525 total lines
- Split into train, dev, and test
- Datapoints are labeled either human (0) or MGT (1) and range from 1 – 17,400 characters

Methods

- N-grams to extract features from the text
- Feature selection
- Stratify the data (80/10/10)
- Train a binary logistic regression model using cross-entropy loss
- Tune hyperparameters on development set



Evaluating Our Approach

- Balanced dataset but still focusing on F1-Score
- Emphasis on **precision** metric
  - Accusing one person of using AI is worse than failing to catch one AI-generated text
  - Ethical issue especially with students whose first language isn't English

Translating Customer Service Chats

Joshua Frank, Forest Maguire, Rose Resnick, Christopher White, Raymond Zong

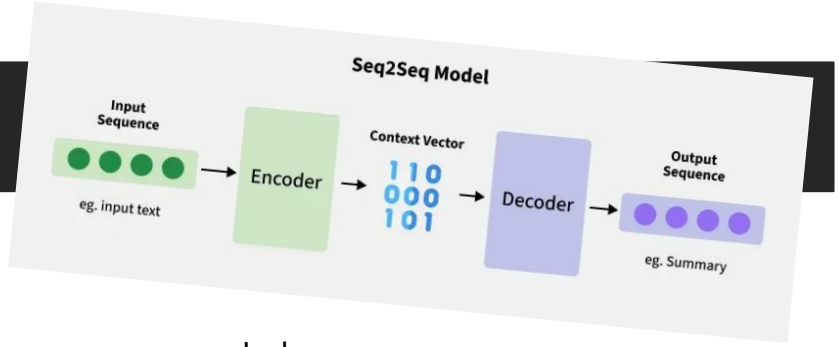
Motivations

- Reduce language barriers in customer service
- Test the limits of live translation
- Challenge ourselves

Data

- Data is sourced from the 2024 WMT chat task, which compiles real customer service messages
- We have 15,000 entries of translated messages between English and French
- The data entries also have context for whether the sender was a customer or service representative
- We also possess a separate set of data that is untranslated for testing purposes

Method



Method 1

- We will train a custom sequence-to-sequence model
  - Essentially 2 neural networks: encoder and decoder
- We will then use the provided development data set to tweak hyperparameters

Method 2

- Zero-shot prompting
  - Seeing how well an existing AI model can translate the chats
- There are also ways to change hyperparameters using this method

Evaluation

- We’re given the correct response to our training data.
- Therefore,
  - Black box evaluation approach will be used,
  - called METEOR.
- Final result will be an adjusted harmonic mean.

Questions?

5. Jack, Hongyu, Patrick, Aaron, Marcus

Annotating Online Gaming Voice Chat

Project Motivation

- One of the most active communication spaces, however also notoriously toxic
- Well-documented and widespread problems, yet gaming voice chat remain dramatically understudied
- Existing hate speech datasets and detection models often fail to capture unique dynamics

Resource Creation

- Very few labeled gaming voice chat datasets
- Producing an annotated dataset with a reproducible codebook creates a resource for the research community
- Codebook tailored to gaming language advances annotation practices in an understudied domain

Practical Impact

- A working hate speech detection model for gaming voice chat could help platforms with moderation
- Better moderation and hate speech detection fosters safer online communities, especially in high-risk environments

Generalizability

- Methods and codebooks developed could extend to other informal spoken domains beyond gaming voice chat
  - e.g., Twitch, Discord servers, communication tools

Data

- **Video Recordings of Gameplay/Livestream:** Games included are the 13 most popular games in June, 2024, and recordings range from 30 minutes to 5 hours long
  - Rust, Valorant, Counter Strike 2, League of Legends, DOTA 2, Rainbow Six Siege
- **Software-Generated Transcripts:** There are 480 software-generated, anonymized, English transcripts

Statistic	Sum	Avg. per transcript	Avg. per line
Transcripts	480	N/A	N/A
Lines	1,410,146	2,937	1
Words (Token)	15,000,000-17,000,000	31,250-35,416	10-12
Speakers	20,385	42	1

Methods

- **Literature Review:** Review existing hate speech definitions, annotation frameworks, and best practices to guide the structure of the annotation codebook and methodology.
- **Initial Codebook Development:** Create a preliminary codebook with operational definitions, annotation categories (e.g., hate speech, offensive speech, harassment, neutral speech), decision rules, and examples.
- **Pilot Annotation & Training:** All five annotators will label a small pilot set (~10 transcripts) to identify ambiguities and refine annotation guidelines.
- **Reliability Testing & Codebook Refinement:** Annotators independently label a shared subset (30–50 transcripts). Inter-annotator reliability will be assessed and used to refine categories and definitions.
- **Full Annotation & NLP Modeling:** Apply the finalized codebook to the remaining ~420–460 transcripts and use the annotated dataset to train and evaluate an NLP hate speech detection model.

Evaluation

- **Codebook Development:** The initial codebook will be evaluated by identifying and discussing inconsistencies in the annotations and refining definitions and guidelines accordingly.
- **Inter-Annotator Reliability:** For a shared subset of approximately 30–50 transcripts, inter-annotator reliability will be assessed using statistical measures for categorical data such as Cohen's kappa, Fleiss' kappa, or Gwet's AC1. Agreement levels will be interpreted using established benchmarks to determine the consistency of the annotations.
- **NLP Model Performance:** The resulting NLP model will be evaluated using standard classification metrics, such as the F1 score. Additional metrics (e.g., precision and recall) may also be used to better understand model performance.

6. Laxmi, Ciara, Sanjana, Yifei, Irisin

Predicting Daily S&P 500 Direction from Financial News

Using NLP to Capture Market Sentiment (2008–2024)

Presenter: Group 6: Market prediction from financial news - group1

Date: 03-04-2026

Team: Yifei, Asta, Ciara, Sanjana, Irisin

Agenda

1. Project Motivation & Task
2. Data & Label Engineering
3. Methodology: From Counting to Representation
4. Evaluation & Metrics
5. Ethics & Future Work
6. Roles & Responsibilities

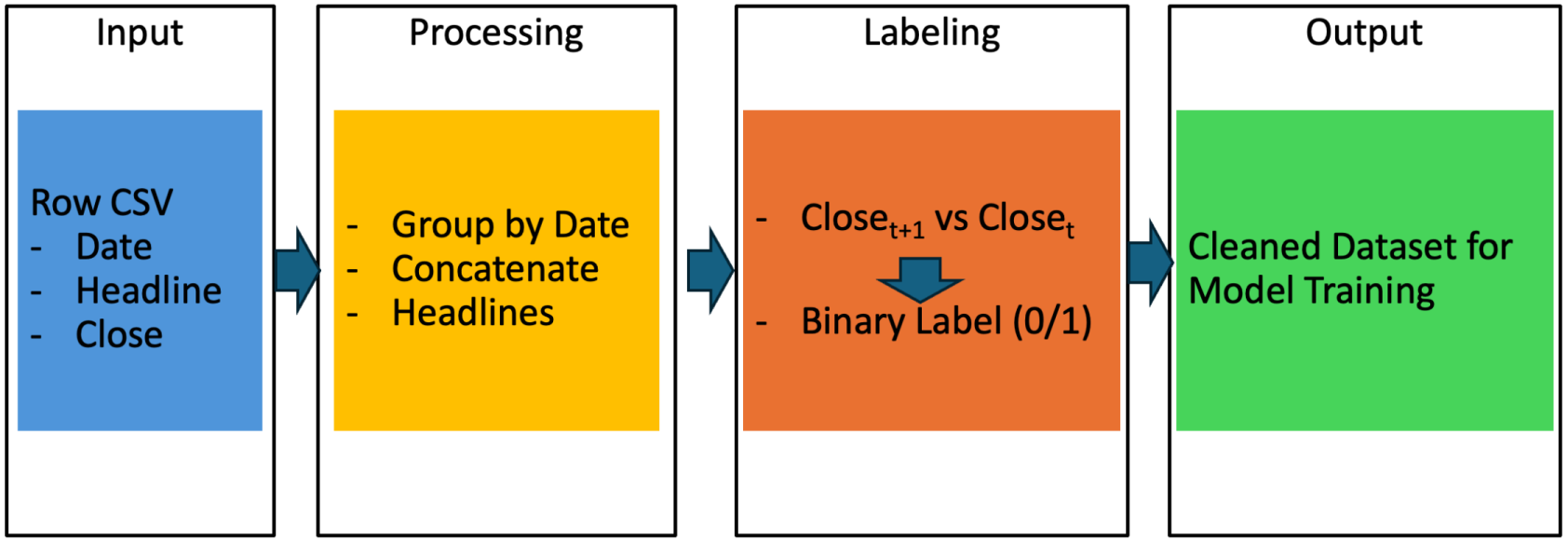
1. Project Motivation & Task

- **Task**: Binary classification of next-day S&P 500 movement (Up vs. Down/Flat).
- **Problem**: How does short-term public sentiment in news affect market behavior?.
- **Framework**: Supervised **Discriminative Classification** (Mapping text features to labels).
- **Goal**: Transition from "word counting" to "understanding financial semantics."

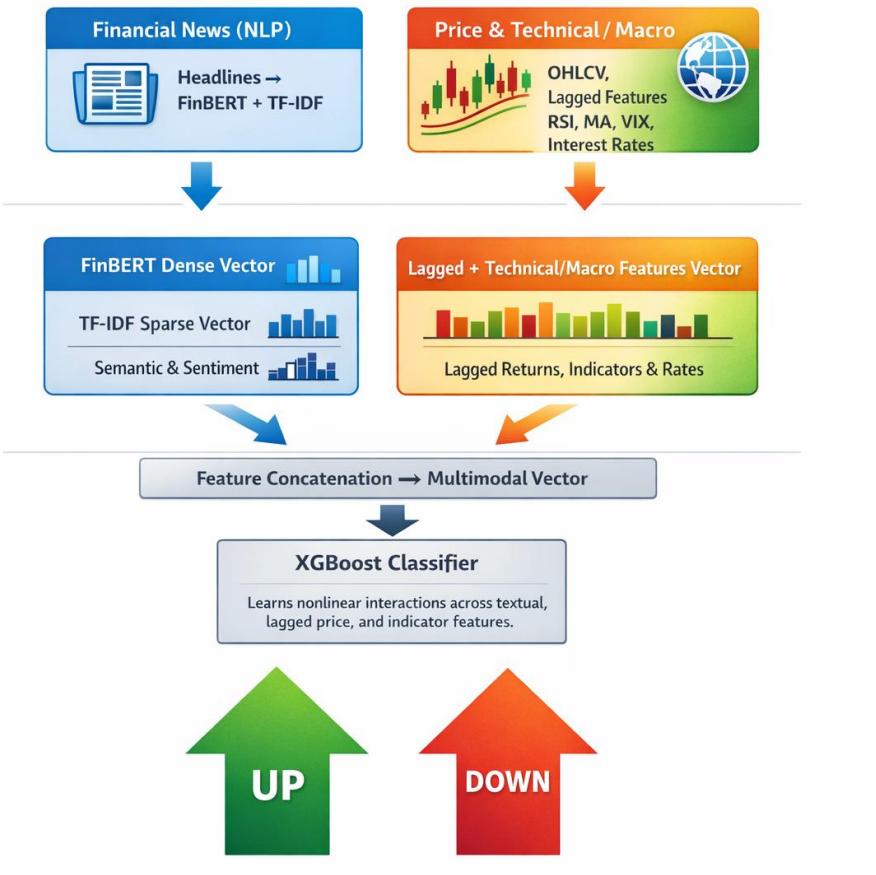
2. Data & Label Engineering

- **Source**: S&P 500 with Financial News Headlines (Kaggle, 2008–2024).
- **Dataset Size**: ~19,000+ daily headline-price pairs.
- **Data Cleaning**: Concatenating multiple headlines from the same trading day into a single document.
- **Labels**: Self-engineered binary targets (1 if $Close_{t+1} > Close_t$, else 0).

Data Pipeline & Label Engineering



Multimodal Binary Prediction:
S&P 500 Opening Direction



3. Methodology: From Counting to Representation

NLP Pipeline: From Text to Prediction

Traditional Baseline: N-gram

1. Text Processing
Text -> TF-IDF + n-gram

2. Feature Engineering
- Sentiment
- Numeric/Keyword Weighting

3. Model
-> Linear SVM(sklearn)

Financial News: S&P 500 Prediction ->

Advanced Approach: LLM

1. Text Processing
Text -> FinBERT Embeddings

2. Representation Learning
- Contextual Understanding
- Non-linear Patterns

3. Model
-> Direct Prediction or Sentiment Features

Evaluation (for both) -> Accuracy/F1/Precision

4 Model Performance & Reliability

Time-Series Split (Avoid Lookahead Bias)

- Train: 2008 – 2018
- Dev: 2019-2021
- Test: 2022-2024

Key Metrics

- Precision (Avoid false “UP” Prediction)
- Accuracy
- F-measure (eg, Macro F1)
- Confusion Matrix

Baseline for Comparison

Random Baseline. –> 50%

Majority Class –> >50%

Analysis & Insights

- Feature Impact Analysis
- Error Analysis
- Precision-Recall

5. Ethics & Future Work

- **Ethics**: News bias (sensationalism), risk of market manipulation, and LLM hallucinations.
- **Limitations**: News headlines are only one factor; ignores technical indicators (volume, etc.).
- **Next Steps**:
    i. Baseline SVM implementation.
    ii. FinBERT fine-tuning and prompt engineering.
    iii. Error analysis on "Dev set" to avoid overfitting Test set.

6. Roles & Responsibilities

- **Irisin**: Modeling Lead & Presentation.
- **Yifei**: Data Lead (Cleaning & Labeling).
- **Asta**: Baseline & Feature Engineering.
- **Ciara**: LLM Prompting & Ethics.
- **Sanjana**: Visualizations & Report Coordination.

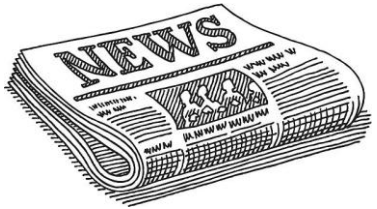
Q&A

Thank you! Any feedback is appreciated.

7. Matthew, Enzo, Nate, Owen, Kevin, Nihal

Project Motivation

- Financial markets react quickly to news and social media sentiment.
- Investors increasingly rely on automated tools to interpret large volumes of financial text.
- Our project investigates whether financial news, Twitter sentiment, or a combination of both provides the strongest signal for predicting short-term stock price direction.
- Goal: Build an NLP system that predicts whether a stock will rise within one week based on text data.



Datasets

Twitter Dataset

- 3.7M tweets mentioning major companies (2015–2020)
  - tweet text
  - timestamp
  - company ticker

Financial News Dataset

- ~5,000 financial articles
- Fields used:
  - title
  - description
  - publication date
  - ticker symbols

Stock Price Data

Retrieved using **yfinance API**

- Used to compute **price movement one week later**

Approach & Method

- Problem Framing
  - Supervised Binary Classification
- Feature Engineering
  - N-gram
  - Credibility Weighting
- Models
  - Logistic Regression

Evaluation Strategy

- Chronological Validation
  - Ensure no look-ahead bias
    - Ex: Train: 2015–2018 → Test: 2019
- Ablation Analysis (Comparison)
  - Text-only Model
  - Text + Publisher
  - Text + Publisher + Market features
- Performance Metrics
  - Accuracy, precision, recall, F1 score, ROC-AUC

8. Grace, Michelle, Kiana, Sarina, Amyia

Who's Who in *Modern Family*:
Character Attribution in a TV Sitcom

CS 1671 Project Proposal Presentation

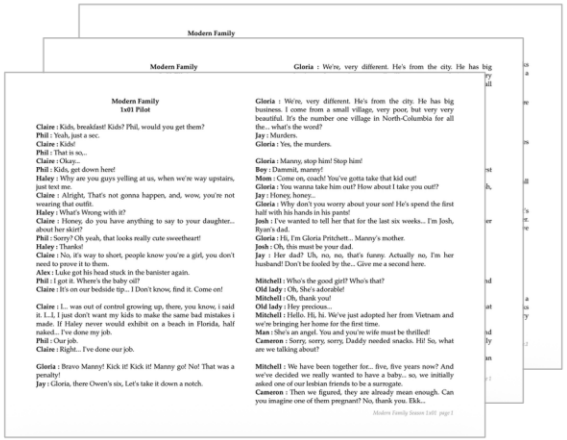
Michelle Hong, Amyia Singh, Grace Hines, Sarina Saran, Kiana Kazemi

Project Motivation

- Distinguish between character set of diverse personalities
- Test ability of model to identify characters based on script alone
- Provide insight on spoken language representation in American media

Datasets

SCRIBD Script PDFs



Modern Family

Modern Family

Modern Family
1x01 Pilot

Claire : Kids, breakfast! Kids? Phil, would you get them?
Phil : Yeah, just a sec.
Claire : Kids!
Phil : That is so...
Claire : Okay...
Phil : Kids, get down here!
Haley : Why are you guys yelling at us, when we're way upstairs, just text me.
Claire : Alright, That's not gonna happen, and, wow, you're not wearing that outfit.
Haley : What's Wrong with it?
Claire : Honey, do you have anything to say to your daughter... about her skirt?
Phil : Sorry? Oh yeah, that looks really cute sweetheart!
Haley : Thanks!
Claire : No, it's way to short, people know you're a girl, you don't need to prove it to them.
Alex : Luke got his head stuck in the banister again.
Phil : I got it. Where's the baby cat?
Claire : It's on our bedside tip... I Don't know, find it. Come on!

Claire : I... was out of control growing up, there, you know, i said it. I...I, I just don't want my kids to make the same bad mistakes i made. If Haley never would exhibit on a beach in Florida, half naked... I've done my job.
Phil : Our job.
Claire : Right... I've done our job.

Gloria : Bravo Manny! Kick it! Kick it! Manny go! No! That was a penalty!
Jay : Gloria, there Owen's six, Let's take it down a notch.

Gloria : We're, very different. He's from the city. He has big business. I come from a small village, very poor, but very very beautiful. It's the number one village in North-Columbia for all the... what's the word?
Jay : Murders.
Gloria : Yes, the murders.

Gloria : Manny, stop him! Stop him!
Boy : Dammit, manny!
Mom : Come on, coach! You've gotta take that kid out!
Gloria : You wanna take him out? How about I take you out!?
Jay : Honey, honey...
Gloria : Why don't you worry about your son! He's spend the first half with his hands in his pants!
Josh : I've wanted to tell her that for the last six weeks... I'm Josh, Ryan's dad.
Gloria : Hi, I'm Gloria Pritchett... Manny's mother.
Josh : Oh, this must be your dad.
Jay : Her dad? Uh, no, no, that's funny. Actually no, I'm her husband! Don't be fooled by the... Give me a second here.

Mitchell : Who's the good girl? Who's that?
Old lady : Oh, She's adorable!
Mitchell : Oh, thank you!
Old lady : Hey precious...
Mitchell : Hello. Hi, hi. We've just adopted her from Vietnam and we're bringing her home for the first time.
Man : She's an angel. You and you're wife must be thrilled!
Cameron : Sorry, sorry, sorry, Daddy needed snacks. Hi! So, what are we talking about?

Mitchell : We have been together for... five, five years now? And we've decided we really wanted to have a baby... so, we initially asked one of our lesbian friends to be a surrogate.
Cameron : Then we figured, they are already mean enough. Can you imagine one of them pregnant? No, thank you. Ekk...

Modern Family Season 1x01 page 1

Seasons 1-3
10 characters

CSV Dataset

Character	Line
Claire	Kids? Phil, would you get them?
Phil	Yeah, just a sec.

20,323 lines
214,066 preprocessed tokens
9,564 vocabulary size

Methods & Approach

Goal: Our goal is to train a model that learns patterns in dialogue and predicts which character is speaking given a line from the sitcom Modern Family.

Text Representation: each line of dialogue will be represented as a Bag-of-Words n-gram model, extracting both unigrams and bigrams (converting each into a vector with `CountVectorizor`)
↳ stop words will not be removed as they may provide stylistic information about a speaker

Feature Selection: we will use classification-based scoring functions from scikit-learn such as chi-square (chi^2), f_classif, or mutual information
↳ these methods will allow us to determine which words are strongly associated with which particular speaker

Model: we will be using multinomial linear regression as it provides interpretable coefficients

Additional Linguistic Features: stylistic features may also be investigated
↳ hedges: "maybe," "kind of," and "I think"
↳ punctuation usage

Packages & Libraries: model will be implemented with pandas, numpy, scikit-learn, and matplotlib

Testing & Evaluation

- We will split the data as so:
  1. Train on season 1 + ½ of season 2
  2. Validate on the other ½ of season 2 (for tuning parameters)
  3. Test on season 3
- This ensures that the model is evaluated on unseen episodes
- Evaluation metrics:
  1. Accuracy of predictions
  2. Precision, Recall, and F1 Score
  3. Confusion Matrix: which characters are frequently confused with each other

Questions?

9. Ryan, Pier, Justin, Wyatt, Praz

Training a Small Language Model for Code Generation

Goal: Build a small language model that generates Python code from natural language prompts

Train from scratch on code data, then teach it to follow prompts

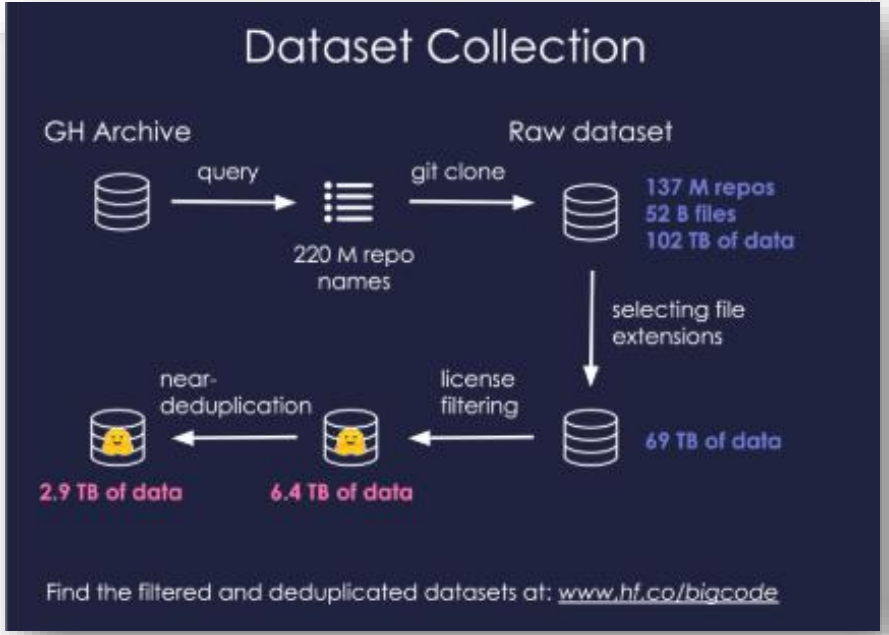
Focus areas:

- Pretraining on large code dataset
- Instruction tuning for prompts
- Evaluation of generated code quality

Datasets

Pretraining Dataset

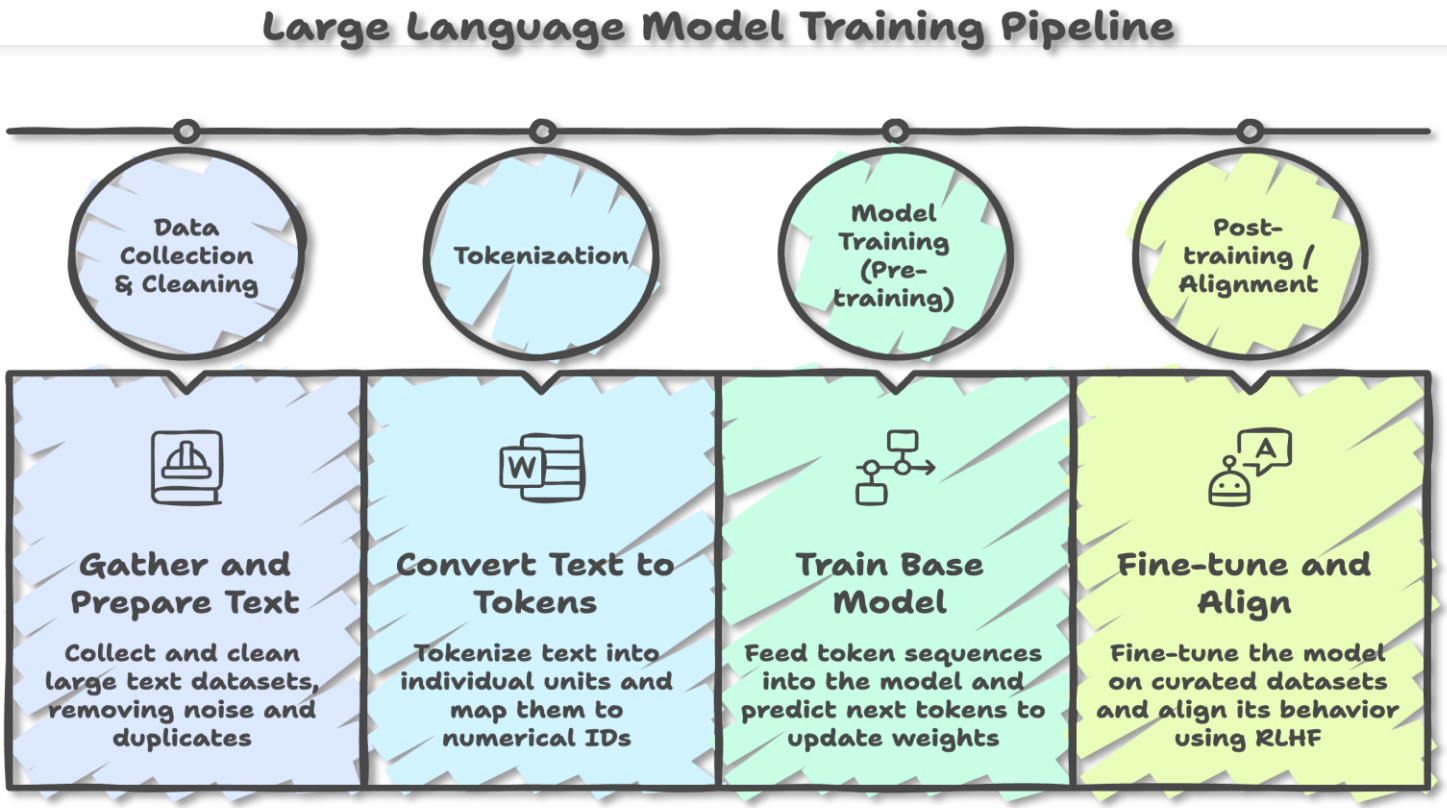
- "The Stack" from HuggingFace
- ~3TB deduplicated source code dataset
- Python subset



Instruction Dataset

- Code Alpaca 20k
- Used for post-training/instruction tuning

Pre-training/Post-training Pipeline



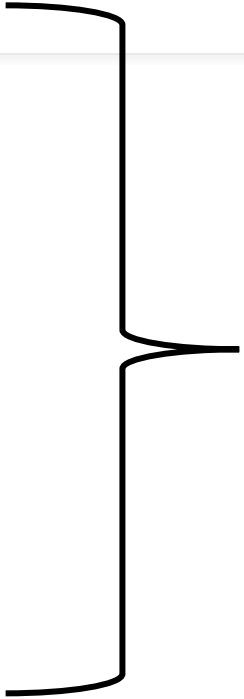
Model Evaluation

Quantitative

- LiveBench benchmark
- Validation perplexity

Qualitative

- Test prompts
- Run generated code in Python
- Measure:
  - Syntax correctness
  - Number of fixes needed



Goal

- Generate runnable Python code from prompts

10. Jonah, Amanda, Lyndsey, Saung, Jay

PREDICTING ARTIST FROM SONG LYRICS

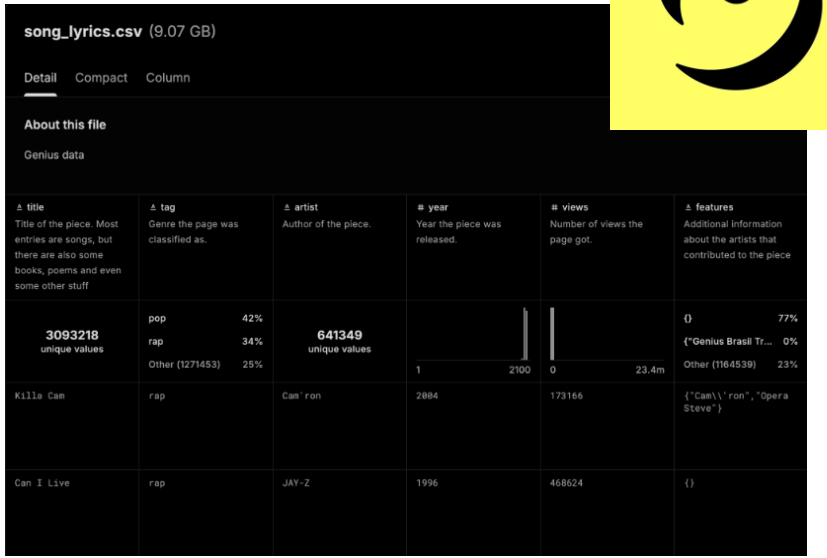
Project Motivation

- Artists often have distinct lyrical style and vocabulary
- Shows how stylistic patterns can be captured statistically
- Demonstrates real-world multi-class NLP classification
- Useful for:
  - authorship attribution
  - plagiarism detection
  - stylistic analysis in music/literature

- Our project explores whether simple statistical models and logistic regression can effectively distinguish artist style at scale.

Data

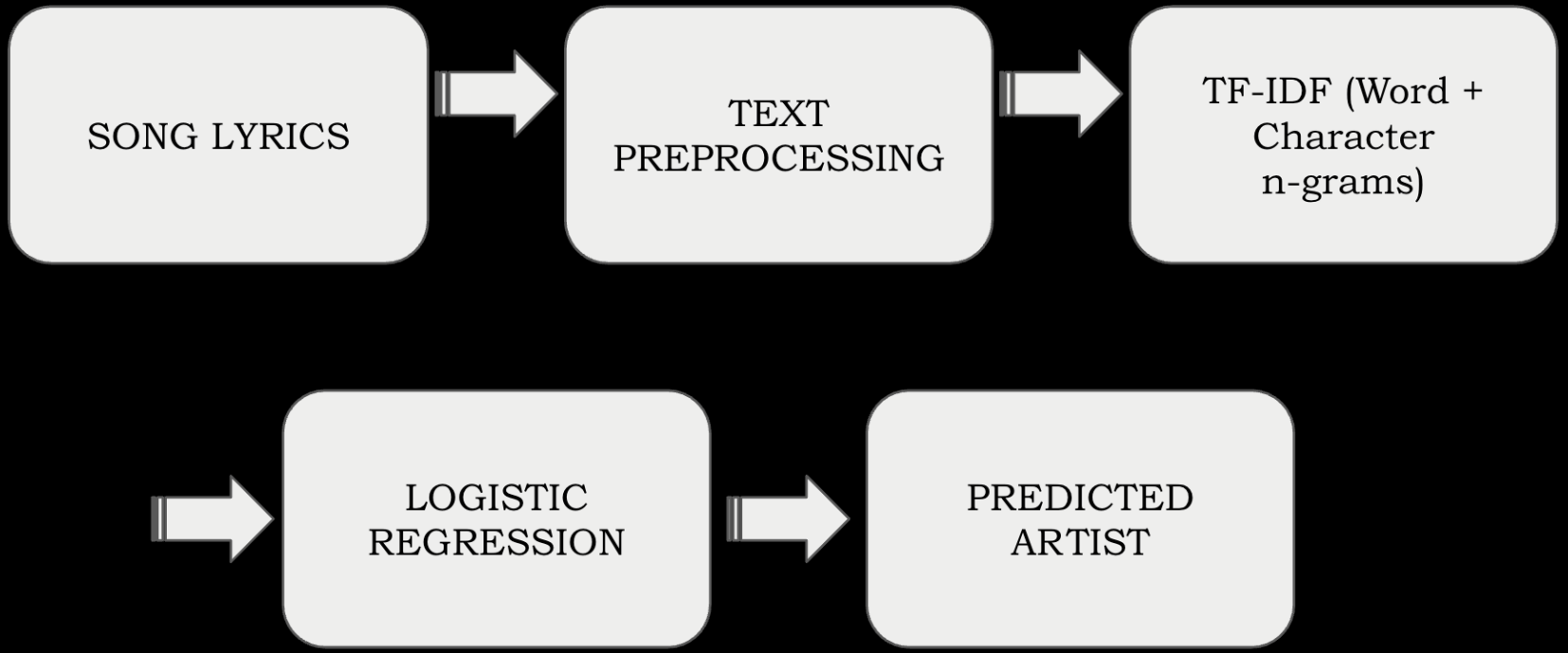
- **Dataset:**
  - Genius Song Lyrics (Kaggle)
  - ~5 million lyric entries
  - ~641k artists
- **Important note:**
  - Highly imbalanced
  - We restrict to top-K artists (20–50)
- **Input:** Full song lyrics
- **Output:** Predicted artist label



Approach / Methods

- **Text Processing:**
  - Normalize text
  - Filter by language
  - Remove duplicates
- **Feature Extraction:**
  - Character n-grams (capture stylistic spelling/patterns)
  - Character n-grams help capture stylistic patterns beyond just word choice.
- **Modeling:**
  - Multinomial Logistic Regression
  - Softmax output over K artists

MODEL PIPELINE



Evaluation

- **Data Split:**
  - 70% train
  - 15% validation
  - 15% test
  - Stratified sampling so each artist shows up in all data splits
- **Metrics:**
  - Accuracy
  - Precision / Recall
  - F1 Score
  - Macro-F1 (primary metric)
- **Error:**
  - Confusion Matrix: Artists model confuses most