TOLSA-M Codebook

Detravious Jamari Brinkley, Kingdom Man

PhD Student in Computer Science
University of Florida (UF)
UF Data Studio Lab
dj.brinkley@ufl.edu

UF Herbert Wertheim College of Engineering
Department of Computer & Information Science & Engineering
UNIVERSITY *of* FLORIDA

Outline

1. Motivation: What are we examining?
2. Motivation: What are current limitations?
3. Motivation: What are our solutions—TOLSA-M?
4. Motivation: Why is identifying TOLSA-Ms a hard problem?
5. Methods: Collecting Data
6. Methods: Models
7. Results

Outline

1. **Motivation: What are we examining?**
2. Motivation: What are current limitations?
3. Motivation: What are our solutions—TOLSA-M?
4. Motivation: Why is identifying TOLSA-Ms a hard problem?
5. Methods: Collecting Data
6. Methods: Models
7. Results

Motivation

What are we examining?

Winter Forecast 2025-26: Snowy Season For Parts of the US
191K views • 2 weeks ago

AccuWeather

Daily Mail
https://www.dailymail.co.uk › health › pandemic-expert-big-one.html

Haunting prediction of where next pandemic will strike... and how ...

Oct 12, 2025 · The next pandemic will begin with the death of a single baby in Africa – and go on to kill more than seven million Americans, with 20 times more fatalities worldwide. That's the ...

NFL Predictions and Picks For EVERY Week 7 Game [Steelers at Bengals] | Best Bets ✅
223K views • 4 days ago

CBS Sports

Race to the WH
https://www.racetothewh.com

Predicting the 2026 Elections | Polls and Political News

New Feature: Explore all the latest polls in one place, from 2026 midterm races to presidential approval rating polls. Includes links to the original polls, and to our in depth-polling averages.

7 stock market predictions for 2022 from Yahoo Finance's Brian Sozzi
7K views • 3 years ago

Yahoo Finance

Never doubt Shaq #nba #finals #predictions ...
195K views

Why examine Predictions?

1. To inform the public

   Why? Because decisions are based on predictions

2. To know which voice to listen to

   Why? At worse, can result in emotional distress and a loss of time, energy, and money.

...

Outline

1. Motivation: What are we examining?
2. **Motivation: What are current limitations?**
3. Motivation: What are our solutions—TOLSA-M?
4. Motivation: Why is identifying TOLSA-Ms a hard problem?
5. Methods: Collecting Data
6. Methods: Models
7. Results

Motivation

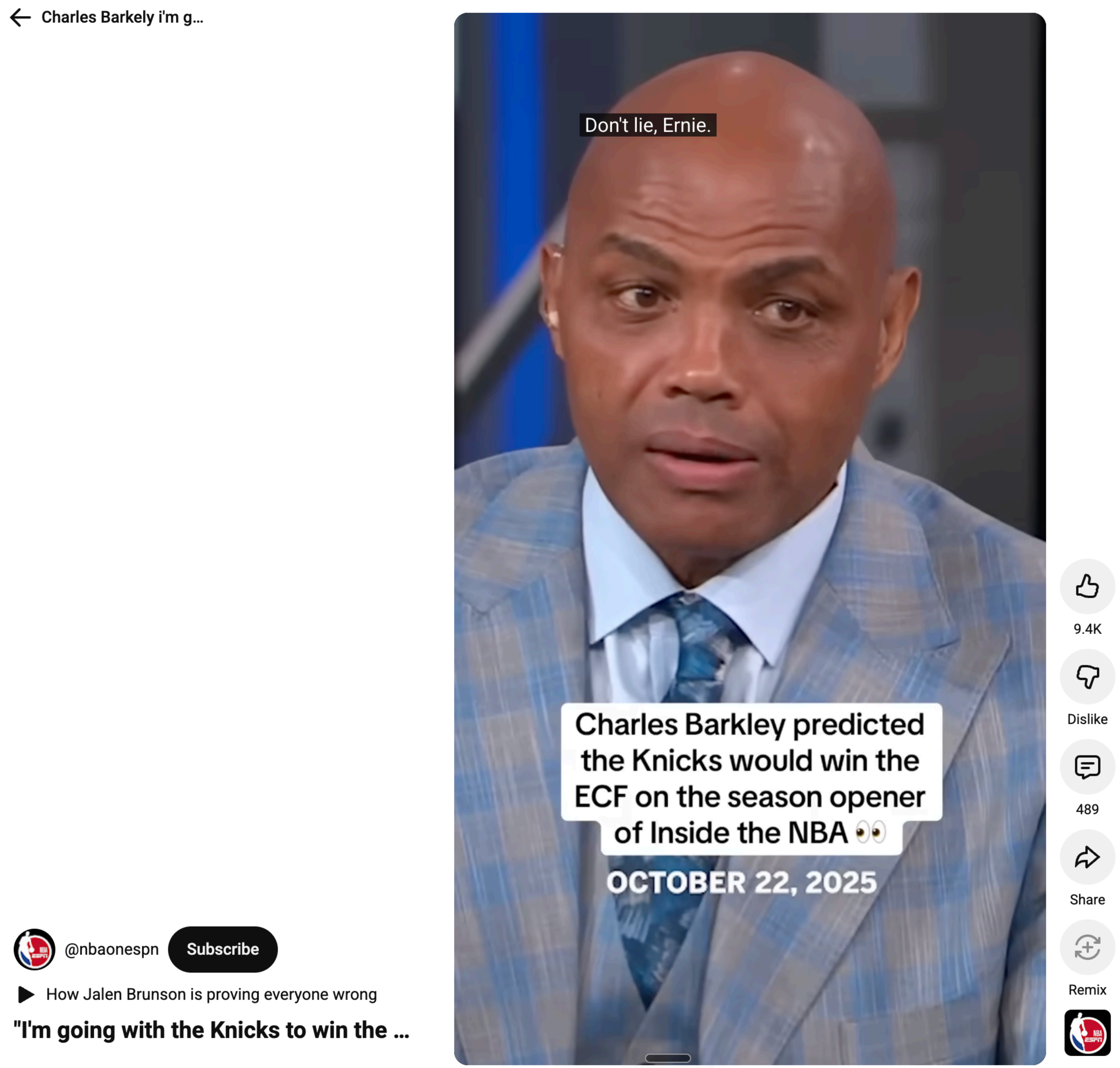
What are current limitations?

1. Not many works classify text documents as predictive/non-predictive
2. Previous works only capture the future tense for predictions
3. Previous works look at keywords such as predictions and synonyms (modal verbs/hedge?)
4. Previous works do not extract common properties
5. Rarity of metadata with original data
6. Lack of connection to discourse genres (beliefs, claims, etc)

Motivation

What are c

1. **Not many works classify text documents as predictive/ non-predictive**
2. Previous works only capture the future tense for predictions
3. Previous works look at keywords such as predictions and synonyms
4. Previous works do not extract common properties
5. Rarity of metadata with original data
6. Lack of connection to discourse genres (beliefs, claims, etc)



Motivation

What are current limitations?

1. Not many works classify text documents as predictive/non-predictive

2. **Previous works only capture the future tense for predictions**

3. Previous works look at keywords such as predictions and synonyms

4. Previous works do not extract common properties

5. Rarity of metadata with original data

6. Lack of connection to discourse genres (beliefs, claims, etc)

Why? Because by definition of predictions is future

Motivation

<u>What are current limitations?</u>

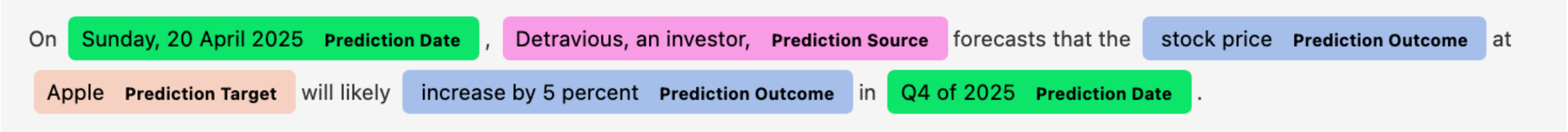
1. Not many works classify text documents as predictive/non-predictive
2. Previous works only capture the future tense for predictions
3. **Previous works look at keywords such as predictions and synonyms**
4. Previous works do not extract common properties
5. Rarity of metadata with original data
6. Lack of connection to discourse genres (beliefs, claims, etc)

Why? Because these words are more gauranteed vs other words/phrases

Motivation

What are current limitations?

1. Not many works classify text documents as predictive/non-predictive
2. Previous works only capture the future tense for predictions
3. Previous works look at keywords such as predictions and synonyms
4. **Previous works do not extract common properties**
5. Rarity of metadata with original data
6. Lack of connection to discourse genres (beliefs, claims, etc)



Count	Variable	Definition
1	p_s	**Any source entity (person, organization, etc)**
2	p_t	**Any target entity (person, organization, etc)**
3	p_d	**Date/Time range when p is expected to come to fruition and when p was made**
4	p_o	**Outcome**

Table 1: At minimum, a prediction must contain these properties (variables and definitions)

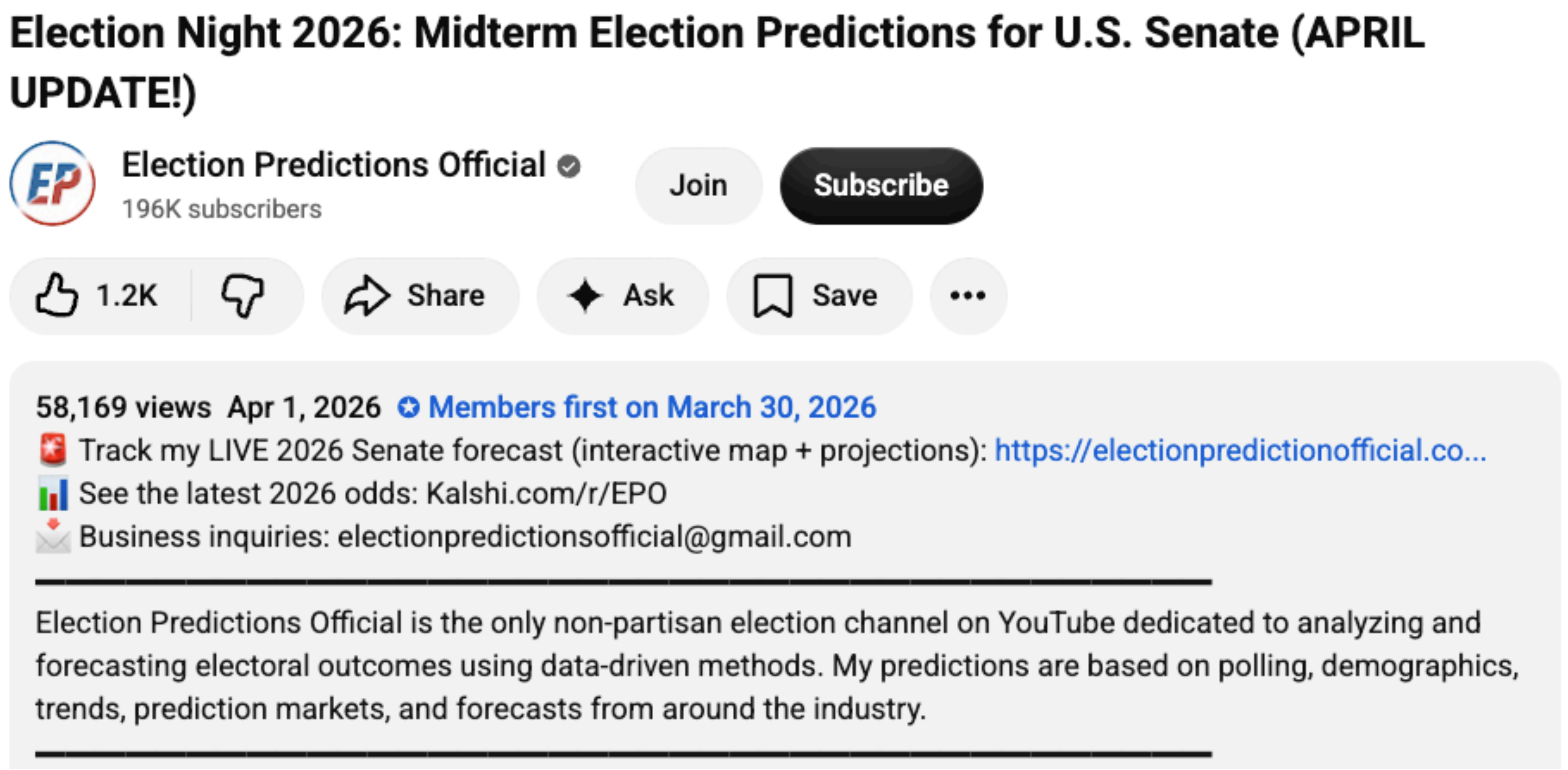
Why should we?

dj.brinkley@ufl.edu | www.ufdatastudio.com/people/

Motivation

What are current limitations?

1. Not many works classify text documents as predictive/non-predictive
2. Previous works only capture the future tense for predictions
3. Previous works look at keywords such as predictions and synonyms
4. Previous works do not extract common properties
5. **Rarity of metadata with original data**
6. Lack of connection to discourse genres (beliefs, claims, etc)



Why lack of metadata and granularity/taxonomy in current datasets?

Motivation

What are current limitations?

1. Not many works classify text documents as predictive/non-predictive
2. Previous works only capture the future tense for predictions
3. Previous works look at keywords such as predictions and synonyms
4. Previous works do not extract common properties
5. Rarity of metadata with original data
6. **Lack of connection to discourse genres (beliefs, claims, etc)**

Why haven't the connections been explored?

Outline

1. Motivation: What are we examining?
2. Motivation: What are current limitations?
3. **Motivation: What are our solutions—TOLSA-M?**
4. Motivation: Why is identifying TOLSA-Ms a hard problem?
5. Methods: Collecting Data
6. Methods: Models
7. Results

Motivation

What are our solutions?

TOLSA-M = **T**arget **O**utcome with optiona**L** **S**ource, d**A**te, and **M**etadata

dj.brinkley@ufl.edu | www.ufdatastudio.com/people/

Motivation

What are our solutions?

TOLSA-M = **T**arget **O**utcome with optiona**L** **S**ource, d**A**te, and **M**etadata

is defined as a proposal of any tense (past, present, future) document with properties—must contain some target (the entities of interests within the document) and measurable outcomes (either being an attribute, metric, or slope) with optional source (the entity declaring the document) and date (of declaration and of fruition)—while possibly exhibiting some uncertainty about the outcome.

dj.brinkley@ufl.edu | www.ufdatastudio.com/people/

Motivation

What are our solutions?

TOLSA-M = **T**arget **O**utcome with optiona**L** **S**ource, d**A**te, and **M**etadata

TOLSA-M Taxonomy

Granularity Class	Granularity Subclass
TOLSA-M Modality	Text, Audio, Visual, Text-Audio, Text-Visual, Audio-Visual, Text-Audio-Visual
TOLSA-M Structure / Match	Full Match (Structured), Partial Match (Semi-Structured), Unstructured, Semi-Unstructured
TOLSA-M-Observation Association	Full Match, Partial Match, No Match
TOLSA-M Temporal Horizon	Short-Term, Long-Term, Unverifiable, Updated
TOLSA-M Quantification Type	Instantaneous Value / Value at an Instant, Statistical Function, Interval Change / Changes over Interval, Second-Order Effect, Comparison, Recurrent Pattern
TOLSA-M Tense Validity	Past TOLSA-M, Present/Future TOLSA-M
TOLSA-M Meta-data/Meta-source	Human, Non-Human
TOLSA-M Check-Worthiness	Check-Worthy, Not Check-Worthy
TOLSA-M Domain	Financial, Policy, Sports, Weather, Health, Biblical, Other
TOLSA-M Dialogue Act Type	Question, Assertion, Commitment
TOLSA-M Epistemic	Speculative, Probabilistic, Assertive, Certain
TOLSA-M Outcome Type	Binary / Boolean, Categorical, Ordinal, Numeric / Value (Point), Numeric / Value (Interval), Trend / Persistence
TOLSA-M Outcome Reaction	Motivational, Disappointed, Behavioral, Skeptical
TOLSA-M Source "Knowledge"	Human Individual (Expert, Novice, Tipster), Human Collaborative (TOLSA-M Market, Polls, Wisdom of Crowd, Community/Advice), Models (External/Internal Ratings)
TOLSA-M Motivation	Hedging, Adversarial, Entertain, Impress, Comfort, Manipulate, Bias
TOLSA-M Application	Everyday / Routine, Certain Moments
TOLSA-M Document/Representation	Phrase, Sentence, Span, Paragraph
TOLSA-M Theory	Ideological Ties, Humanitarian Concerns, Resource Theories/Competition, Great Power Rivalry, Leadership Traits, Hierarchical Intervention Theory
TOLSA-M Timing	Intuitive or Automated, Thoughtful or Deliberate
TOLSA-M Context	Without Context, With Context

Table 2: TOLSA-M Taxonomy: Granularity classes and subclasses.

Motivation

What are our solutions?

TOLSA-M = **T**arget **O**utcome with optiona**L** **S**ource, d**A**te, and **M**etadata

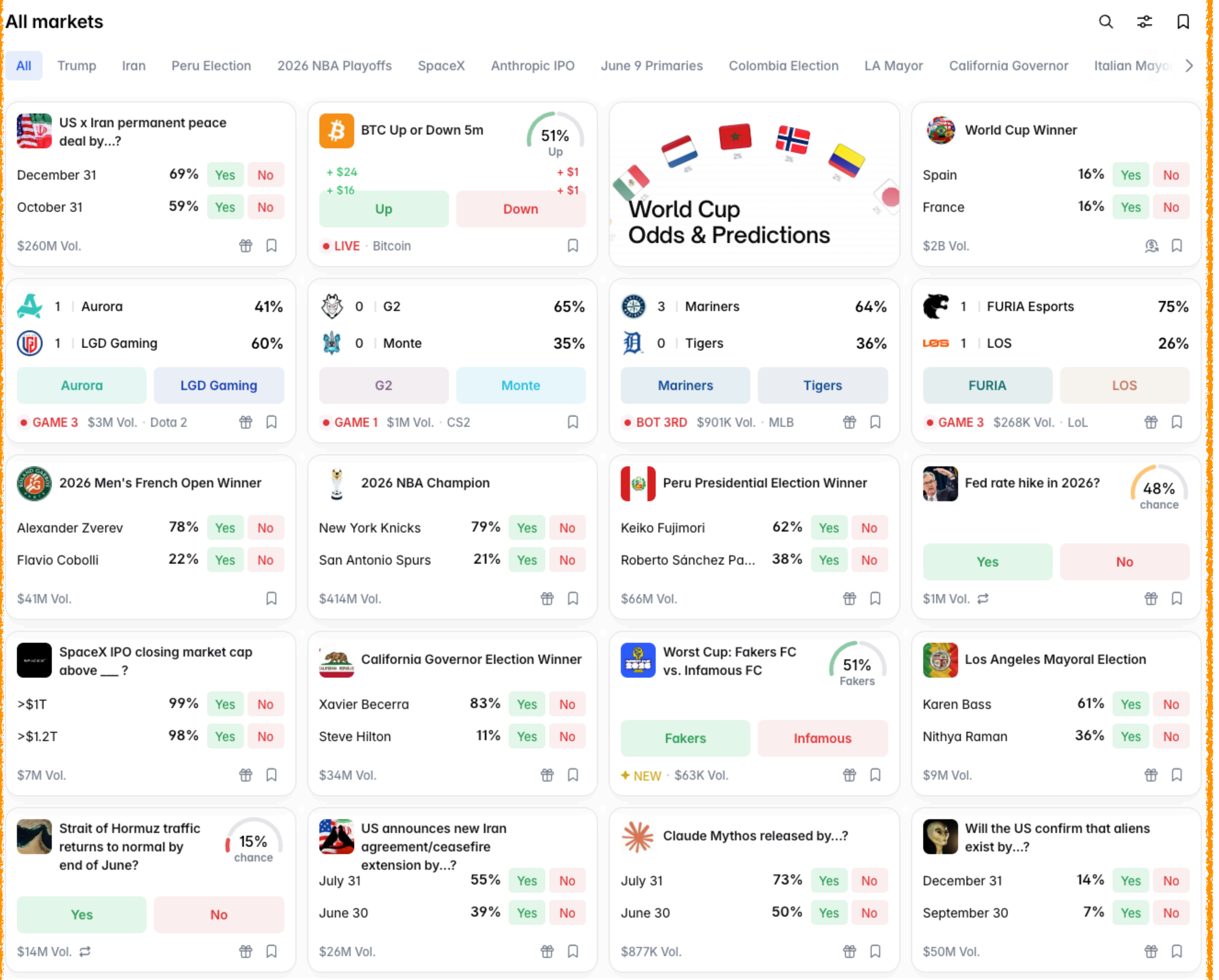
TOLSA-M Taxonomy

TOLSA-M Class	Subclass	Example
Dialogue	Question	So, who makes it to the Final Four this time around? placeholder2024
Epistemic	Assertive	The Mavericks will upset the Thunder in the Western Conference Finals. placeholder2024
Outcome Type	Ranking	2013 NBA Playoff rankings: 1. Oklahoma City Thunder, 2. San Antonio Spurs, 3. Denver Nuggets, 4. Los Angeles Clippers, 5. Memphis Grizzlies, 6. Golden State Warriors, 7. Los Angeles Lakers, 8. Houston Rockets. placeholder2024
Document	Textual Clause	Government promised to extend the maternity leave... placeholder2024
Application	Routine	I predict the bus will arrive on time today. placeholder2024

Table 3: TOLSA-M Taxonomy examples mapping the main class and subclass to document text.

Motivation

What are our solutions?



Source Knowledge = Markets

Outcome Type = Binary

Outcome Reaction = Behavioral

Observation-Association and Correctness = Full Match

Outline

1. Motivation: What are we examining?
2. Motivation: What are current limitations?
3. Motivation: What are our solutions—TOLSA-M?
4. **Motivation: Why is identifying TOLSA-Ms a hard problem?**
5. Methods: Collecting Data
6. Methods: Models
7. Results

Motivation

Why is identifying TOLSA-Ms a hard problem?

1. Not many works classify text documents as predictive/non-predictive

2. Tense as we not only include future, but past and present

3. Relationship to potential synonyms like projection, prediction, forecast, etc + keywords (modal verbs/hedge?)

4. Extraction: Common properties using rule-based or not at all

5. Vague TOLSA-Ms, so required to have metadata, which may/may not be available and if available, the structure may not be easily parsible

6. Relationship to discourse genres (ie: beliefs, claims, events, facts, etc)

Why is identifying TOLSA-Ms a hard problem?

1. **Not many works classify text documents as predictive/non-predictive**
2. Tense as we not only include future, but past and present
3. Relationship to potential synonyms like projection, prediction, forecast, etc + keywords (modal verbs/hedge?)
4. Extraction: Common properties using rule-based or not at all
5. Vague TOLSA-Ms, so required to have metadata, which may/may not be available and if available, the structure may not be easily parsible
6. Relationship to discourse genres (ie: beliefs, claims, events, facts, etc)

Why is identifying TOLSA-Ms a hard problem?

1. Not many works classify text documents as prediction/non-predictive

- Papers that specifially classify predictive/non-predictive, even if they encode predictive as future and non-predictive as non-future. If they don't map future —> predictive, don't add here.

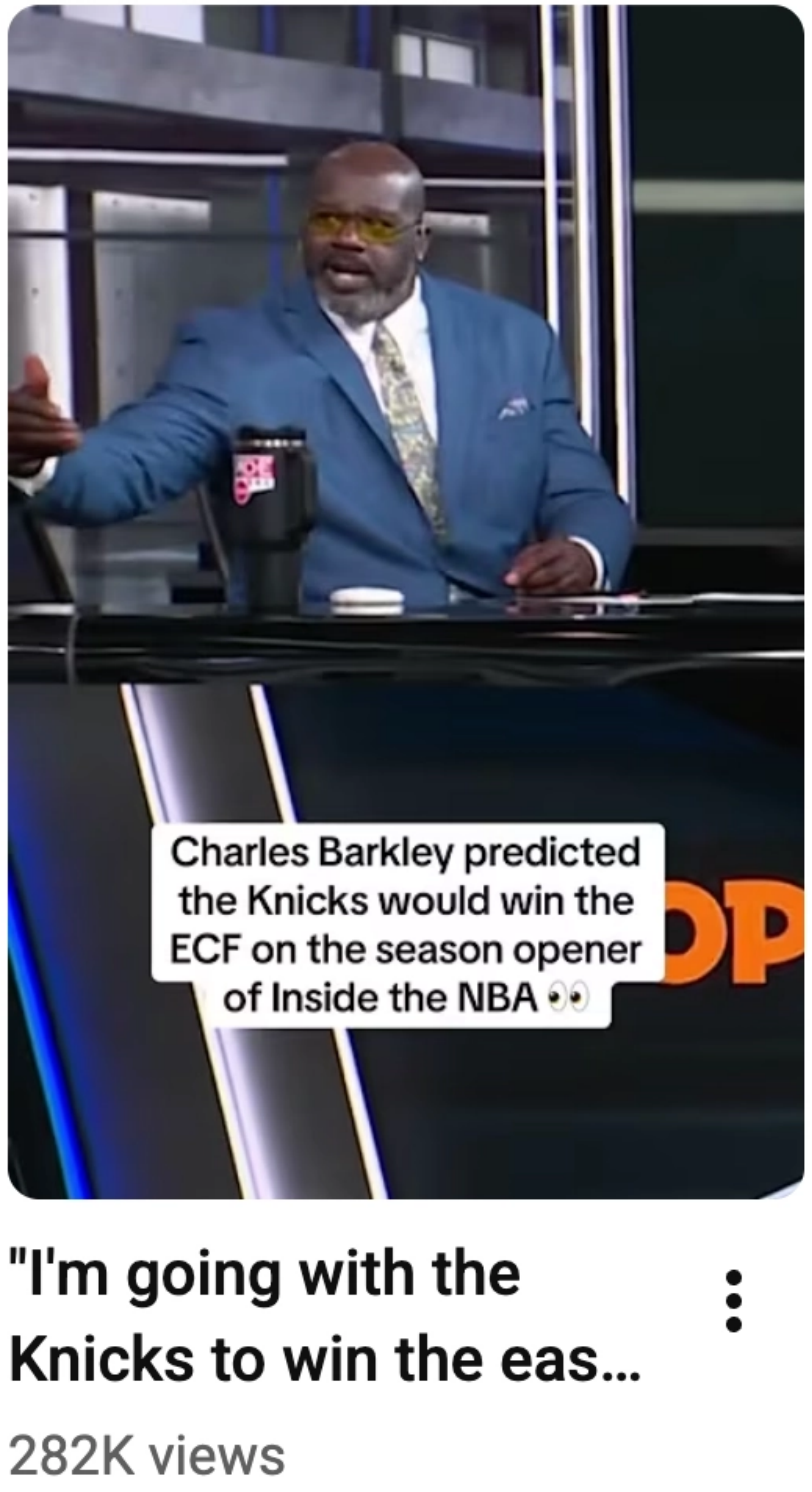
- Current systems will miss this
  - PAPER: Extracting Predictive Statements with Their Scope from News Articles
  - PAPER: Detection of separate reality at discourse level on financial news by combining natural language processing and what machine learning by Garci'a Me'ndez it all (2022)

Why is identifying TOLSA-Ms a hard problem?

1. Not many works classify text documents as predictive/non-predictive
2. **Tense as we not only include future, but past and present**
3. Relationship to potential synonyms like projection, prediction, forecast, etc + keywords (modal verbs/hedge?)
4. Extraction: Common properties using rule-based or not at all
5. Vague TOLSA-Ms, so required to have metadata, which may/may not be available and if available, the structure may not be easily parsible
6. Relationship to discourse genres (ie: beliefs, claims, events, facts, etc)

Why is identifying TOLSA-Ms a hard problem?

1. Tense as we not only include future, but past and present



- Current systems will miss this
  - PAPER: Extracting Predictive Statements with Their Scope from News Articles
  - PAPER: Detection of separate reality at discourse level on financial news by combining natural language processing and what machine learning by Garci'a Me'ndez it all (2022)
  - PAPER: chronicle2050
  - PAPER: Automatic detection of relevant information, predictions and forecasts in financial news through topic modelling with Latent Dirichlet Allocation
  - PAPER: Past + On-going + Future
  - PAPER: Investigation of Future Reference Expressions in Trend Information

- TOLSA-M encorporate because
  - We can track narritive (ie: We now know Charles' prediction at the season opener. How about throughout season and before ECF? Did he change his prediction any? What's his latest prediction?
    - TOLSA-M Updated = tolsa-m_1 @date (previous), tolsa-m_2 @date (previous), tolsa-m_3 @date (latest)

Why is identifying TOLSA-Ms a hard problem?

1. Tense as we not only include future, but past and present

- Current systems will miss this
  - PAPER: chronicle2050 by Katerenchuk and Levitan

Category	Count (%)
Future-Related and Prediction	1106 (17.3%)
Future-Related and NOT Prediction	2089 (32.7%)
NOT Future-Related but Prediction	33 (0.5%)
NOT Future-Related and NOT Prediction	3163 (49.5%)

Table 4: Distribution of chronicle2050 vs. Our Projection labels.

Why is identifying TOLSA-Ms a hard problem?

1. Not many works classify text documents as predictive/non-predictive
2. Tense as we not only include future, but past and present
3. **Relationship to potential synonyms like projection, prediction, forecast, etc + keywords (modal verbs/hedge?)**
4. Extraction: Common properties using rule-based or not at all
5. Vague TOLSA-Ms, so required to have metadata, which may/may not be available and if available, the structure may not be easily parsible
6. Relationship to discourse genres (ie: beliefs, claims, events, facts, etc)

- Work that define predictions, synonymns

- PAPER: Extracting Predictive Statements with Their Scope from News Articles
- PAPER: The CoNLL-2010 Shared Task: Learning to Detect Hedges and their Scope in Natural Language Tex by Farkas
- PAPER: ‘'You should probably read this'': Hedge Detection in Text by Katerenchuk and Levitan
- PAPER: Past + On-going + Future
- PAPER: Investigation of Future Reference Expressions in Trend Information

dj.brinkley@ufl.edu | www.ufdatastudio.com/people/

Why is identifying TOLSA-Ms a hard problem?

1. Not many works classify text documents as predictive/non-predictive
2. Tense as we not only include future, but past and present
3. Relationship to potential synonyms like projection, prediction, forecast, etc + keywords (modal verbs/hedge?)
4. **Extraction: Common properties using rule-based or not at all**
5. Vague TOLSA-Ms, so required to have metadata, which may/may not be available and if available, the structure may not be easily parsible
6. Relationship to discourse genres (ie: beliefs, claims, events, facts, etc)

- PAPER: Extracting Predictive Statements with Their Scope from News Articles

- PAPER: Detection of separate reality at discourse level on financial news by combining natural language processing and what machine learning by Garci'a Me'ndez it all (2022)
- PAPER: Investigation of Future Reference Expressions in Trend Information

dj.brinkley@ufl.edu | www.ufdatastudio.com/people/

Why is identifying TOLSA-Ms a hard problem?

1. Not many works classify text documents as predictive/non-predictive
2. Tense as we not only include future, but past and present
3. Relationship to potential synonyms like projection, prediction, forecast, etc + keywords (modal verbs/hedge?)
4. Extraction: Common properties using rule-based or not at all
5. **Vague TOLSA-Ms, so required to have metadata, which may/may not be available and if available, the structure may not be easily parsible**
6. Relationship to discourse genres (ie: beliefs, claims, events, facts, etc)

Why is identifying TOLSA-Ms a hard problem?

1. Not many works classify text documents as predictive/non-predictive
2. Tense as we not only include future, but past and present
3. Relationship to potential synonyms like projection, prediction, forecast, etc + keywords (modal verbs/hedge?)
4. Extraction: Common properties using rule-based or not at all
5. Vague TOLSA-Ms, so required to have metadata, which may/may not be available and if available, the structure may not be easily parsible
6. **Relationship to discourse genres (ie: beliefs, claims, events, facts, etc)**

Why is identifying TOLSA-Ms a hard problem?

5. Relationship to discourse genres (ie: beliefs, claims, events, facts, etc)

1. Sparse previous works to identify similarities and differences
2. Want to know how each overlap and not overlap linguistically for better identification and extraction
3. Reduce the #FP (model predicts 1 and ground truth is 0), so model hallucinates —> inflating dataset, wasting compute on downstream tasks
4. Reduce the #FN (model predicts 0 and ground truth is 1), so miss out on a TOLSA-M —> not continuing with downstream tasks
    1. Further labelling within our taxonomy
    2. Finding matching observations
    3. Indicating if correct or not
    4. Updating correctness score of source

Why is identifying TOLSA-Ms a hard problem?

- PAPER: Extracting Predictive Statements with Their Scope from News Articles
- PAPER: Detection of separate reality at discourse level on financial news by combining natural language processing and what machine learning by Garci’a Me’ndez it all (2022)
- PAPER: Past + On-going + Future
- PAPER: Investigation of Future Reference Expressions in Trend Information

dj.brinkley@ufl.edu | www.ufdatastudio.com/people/

Why is identifying TOLSA-Ms a hard problem?

5. Relationship to discourse genres (ie: beliefs, claims, events, facts, etc)

dj.brinkley@ufl.edu | www.ufdatastudio.com/people/

Outline

1. Motivation: What are we examining?
2. Motivation: What are current limitations?
3. Motivation: What are our solutions—TOLSA-M?
4. Motivation: Why is identifying TOLSA-Ms a hard problem?
5. **Methods: Collecting Data**
6. Methods: Models
7. Results

dj.brinkley@ufl.edu | www.ufdatastudio.com/people/

Methods

Collecting Data

Table 1: Dataset name versus ground truth breakdown, including percentages of the total for each dataset.

Dataset	NON-TOLSA-M (0)	TOLSA-M (1)	Total
Synthetic	1112 (46.08%)	1301 (53.92%)	2413
Financial PhraseBank	4388 (90.55%)	458 (9.45%)	4846
Chronicle2050	5259 (82.15%)	1143 (17.85%)	6402
TimeBank	2553 (84.54%)	467 (15.46%)	3020
YouTube	0 (0.00%)	1005 (100.00%)	1005
News API	0 (0.00%)	72 (100.00%)	72
MF Climate	0 (0.00%)	3622 (100.00%)	3622
Clients, Rivals, Rogues	0 (0.00%)	6 (100.00%)	6
ForecastBench	0 (0.00%)	968 (100.00%)	968
Smart Hospitals	0 (0.00%)	23 (100.00%)	23
Total	13312 (59.49%)	9065 (40.51%)	22377

Adjudication protocol (interviewing labelers, using shared guidelines, group discussions, and majority voting for difficult sentences).

Methods

Collecting Data

Category	Count (%)
Future-Related and Prediction	1106 (17.3%)
Future-Related and NOT Prediction	2089 (32.7%)
NOT Future-Related but Prediction	33 (0.5%)
NOT Future-Related and NOT Prediction	3163 (49.5%)

Table 4: Distribution of chronicle2050 vs. Our Projection labels.

Outline

1. Motivation: What are we examining?
2. Motivation: What are current limitations?
3. Motivation: What are our solutions—TOLSA-M?
4. Motivation: Why is identifying TOLSA-Ms a hard problem?
5. Methods: Collecting Data
6. **Methods: Models**
7. Results

Outline

1. Motivation: What are we examining?
2. Motivation: What are current limitations?
3. Motivation: What are our solutions—TOLSA-M?
4. Motivation: Why is identifying TOLSA-Ms a hard problem?
5. Methods: Collecting Data
6. Methods: Models
7. **Results**

Error Analysis

TOLSA-M vs Event