Classifying Misinformation:

A Machine Learning Approach

Presented by:
Kasper Melberg Hansen, Patrick Bates, Hugo René Olsen and David Just Højgaard Jensen

Applied Machine Learning

UNIVERSITY OF COPENHAGEN

Overview of this presentation

- Problem & Motivation
- Project Overview
- Our Data
- Data Processing and EDA
- Models
  - XGBoost, DistilBERT
- Results & Comparisons
- Conclusion

Source: https://www.humanrightsresearch.org/post/misinformation-and-disinformation-what-s-the-difference

Problem & Motivation

- The Problem:
  - Tons of misinformation on the internet, reaches large populations before fact-checkers have the opportunity to respond
  - According to a research paper in 2018, false stories were 70% more likely to be reposted than true stories, and that true stories take roughly 6 times longer to reach an audience. (Vosoughi, Roy & Aral, Science, 2018)
  - Manual fact-checking is slow, expensive and can't scale with the amount of content currently being published

Problem & Motivation

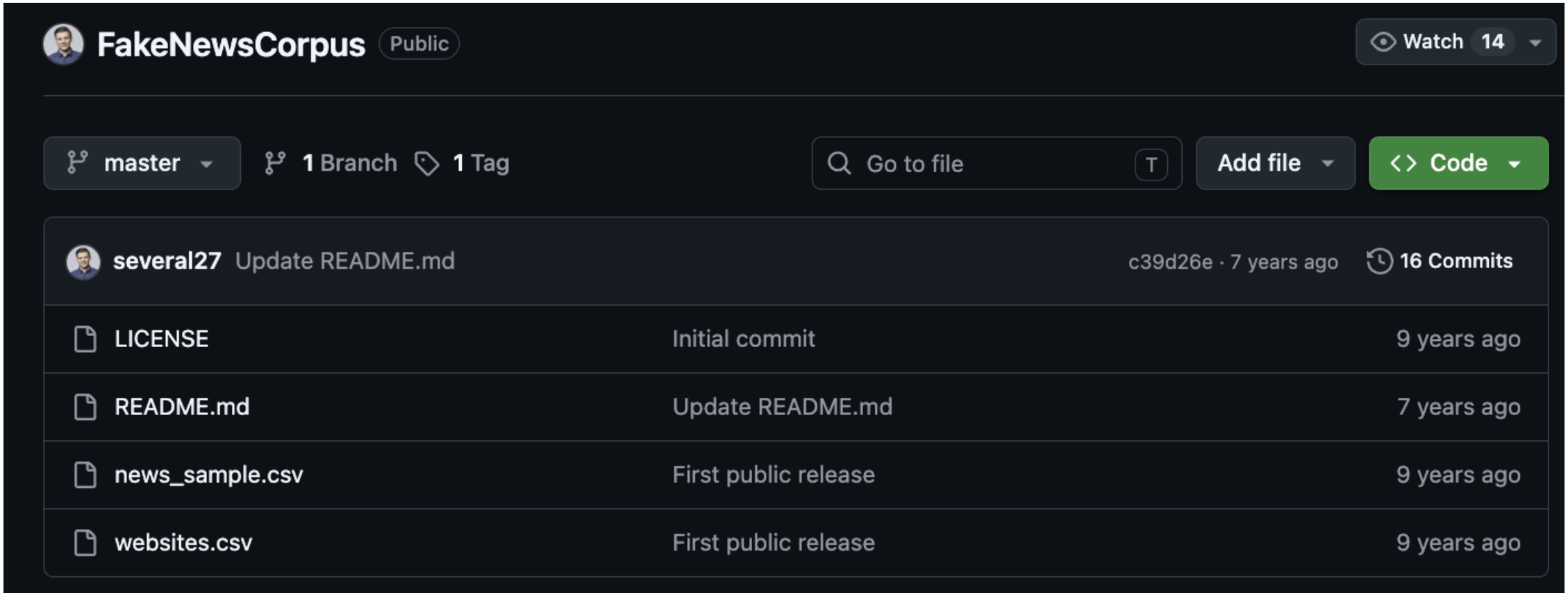
- The Task:
  - Build one or more classifiers which can distinguish between 'reliable' or 'unreliable' news articles based on their text content
  - Dataset: 9.4 million articles scraped from 745 domains, labelled by source (entire domains or websites instead of individual articles)

Our Data

- 50/50 ratio for reliable/unreliable news articles for training/validation
  - Tried to prevent models from learning to just predict the majority class
  - Political content dominates the raw data
    - Our models could have decent accuracy just by predicting 'reliable' for everything
  - Trade-off: Real world distribution isn't 50/50, models are trained on artificial distribution

Our Data

- Source: Github repo, FakeNewsCorpus



Our Data

- Source: Github repo, FakeNewsCorpus
    - CSV containing the following fields:

- id
- domain
- type
- url
- content
- scraped_at
- inserted_at
- updated_at
- title
- authors
- keywords
- meta_keywords
- meta_description
- tags
- summary
- source (opensources, nytimes, or webhose)

Our Data

- Source: Github repo, FakeNewsCorpus
  - CSV containing the following fields:
  - 11 different types of news articles:

Type	Tag	Count (so far)	Description
Fake News	fake	928,083	Sources that entirely fabricate information, disseminate deceptive content, or grossly distort actual news reports
Satire	satire	146,080	Sources that use humor, irony, exaggeration, ridicule, and false information to comment on current events.
Extreme Bias	bias	1,300,444	Sources that come from a particular point of view and may rely on propaganda, decontextualized information, and opinions distorted as facts.
Conspiracy Theory	conspiracy	905,981	Sources that are well-known promoters of kooky conspiracy theories.
State News	state	0	Sources in repressive states operating under government sanction.
Junk Science	junksci	144,939	Sources that promote pseudoscience, metaphysics, naturalistic fallacies, and other scientifically dubious claims.
Hate News	hate	117,374	Sources that actively promote racism, misogyny, homophobia, and other forms of discrimination.
Clickbait	clickbait	292,201	Sources that provide generally credible content, but use exaggerated, misleading, or questionable headlines, social media descriptions, and/or images.
Proceed With Caution	unreliable	319,830	Sources that may be reliable but whose contents require further verification.
Political	political	2,435,471	Sources that provide generally verifiable information in support of certain points of view or political orientations.
Credible	reliable	1,920,139	Sources that circulate news and information in a manner consistent with traditional and ethical practices in journalism (Remember: even credible sources sometimes rely on clickbait-style headlines or occasionally make mistakes. No news organization is perfect, which is why a healthy news diet consists of multiple sources of information).

Data Processing

- As previously mentioned, the dataset from FakeNewsCorpus contains ≈ 9.4 million articles
    - For our project, we sampled sets of 60,000 articles for computational feasibility
        - To train models, a sparse matrix was made over entire vocabulary of the sample.
        - We decided 60,000 articles was large enough to train our models but small enough to where we could process, vectorization and training times manageable.

Data Processing and Exploratory Data Analysis (EDA)

- Why process data?

Data Processing and Exploratory Data Analysis (EDA)

- Why process data?
  - Raw, scraped text data from websites is often noisy, inconsistent and not directly usable by ML models

	index_col	id	domain	type	url	content	scraped_at	inserted_at	updated_at	title	authors	keywords	meta_keywords	meta_description	tags	summary	source
1112	3389	2819528	veteranstoday.com	bias	https://www.veteranstoday.com/tag/x-rays/page/4/	Email * Yes, I would like to receive emails from VT. (You can unsubscribe anytime) Constant Contact Use.	2017-11-18T20:01:27.400599	2018-02-07 23:39:33.852671	2018-02-07 23:39:33.852696	Veterans Today	nan	nan	['']	nan	2, 3, 1	nan	nan

Data Processing and Exploratory Data Analysis (EDA)

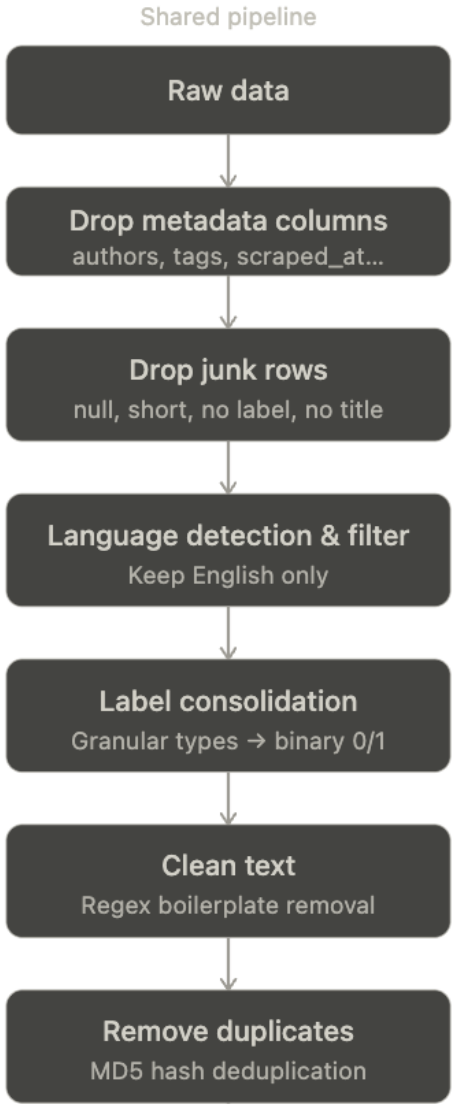
- Why process data?
  - Raw, scraped text data from websites is often noisy, inconsistent and not directly usable by ML models

	index_col	id	domain	type	url	content	scraped_at	inserted_at	updated_at	title	authors	keywords	meta_keywords	meta_description	tags	summary	source
1112	3389	2819528	veteranstoday.com	bias	https://www.veteranstoday.com/tag/x-rays/page/4/	Email * Yes, I would like to receive emails from VT. (You can unsubscribe anytime) Constant Contact Use.	2017-11-18T20:01:27.400599	2018-02-07 23:39:33.852671	2018-02-07 23:39:33.852696	Veterans Today	nan	nan	['']	nan	2, 3, 1	nan	nan

  - Unnecessary columns
  - Not actual article content (scraping artifact?), ML models trained on this would associate 'Constant Contact Use' and 'unsubscribe anytime' as signals of bias or unreliable news

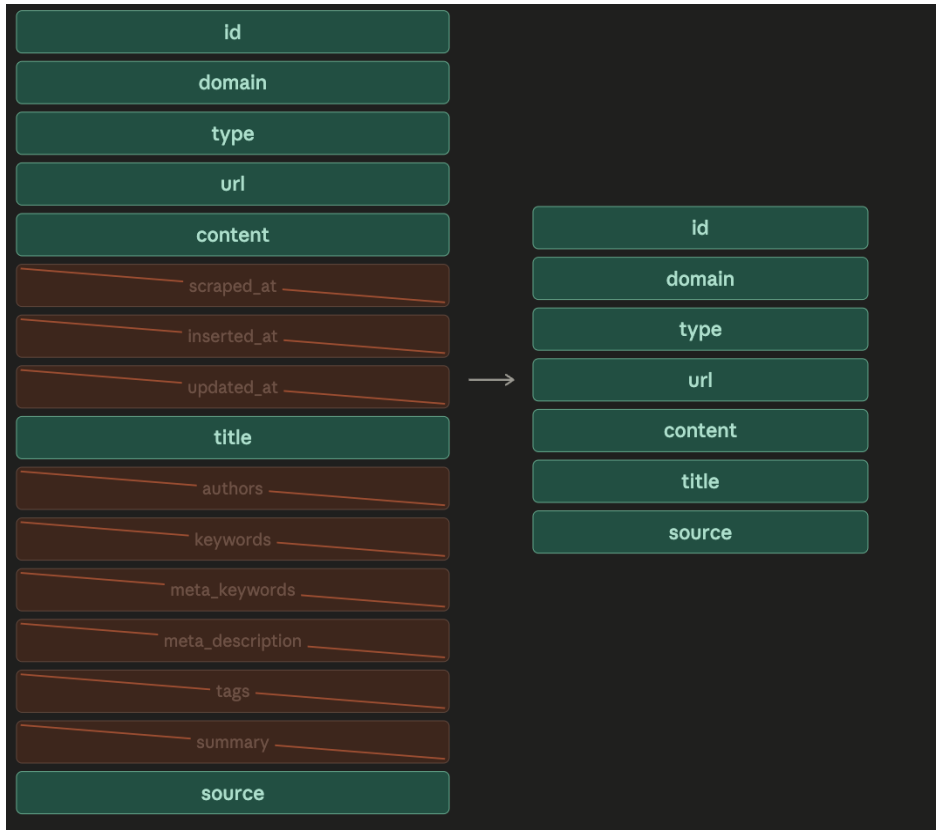
Data Processing and Exploratory Data Analysis (EDA)

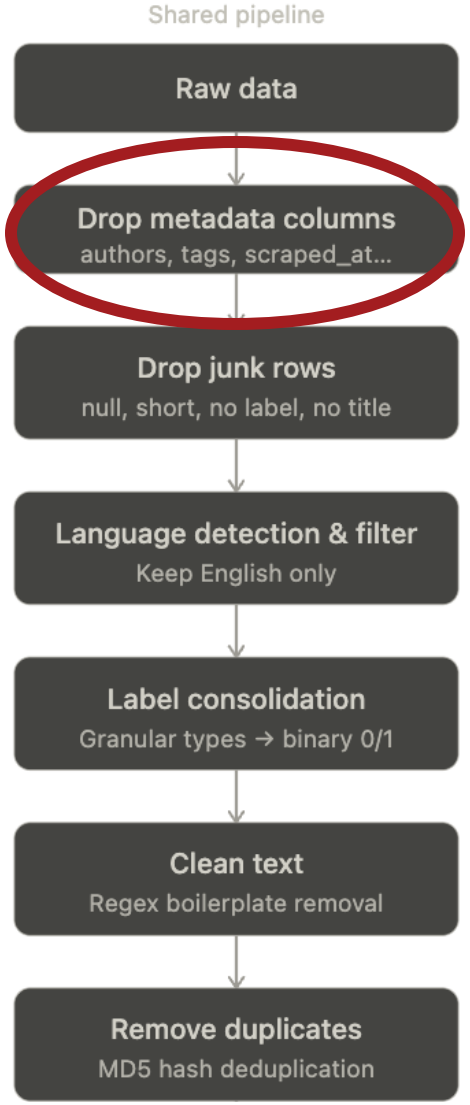
- Our data processing pipeline:



Data Processing and Exploratory Data Analysis (EDA)

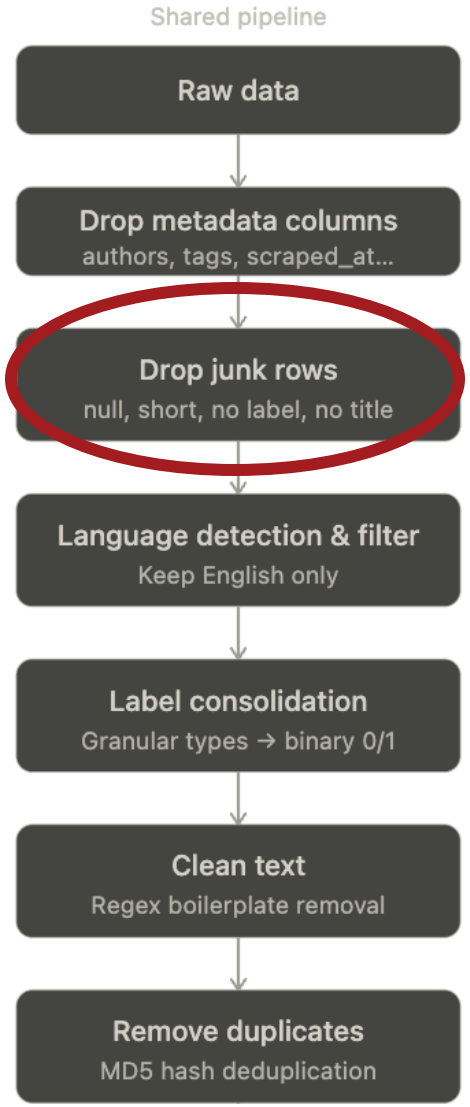
- Drop metadata columns
  - URL and id are used to find duplicates and are later dropped as well





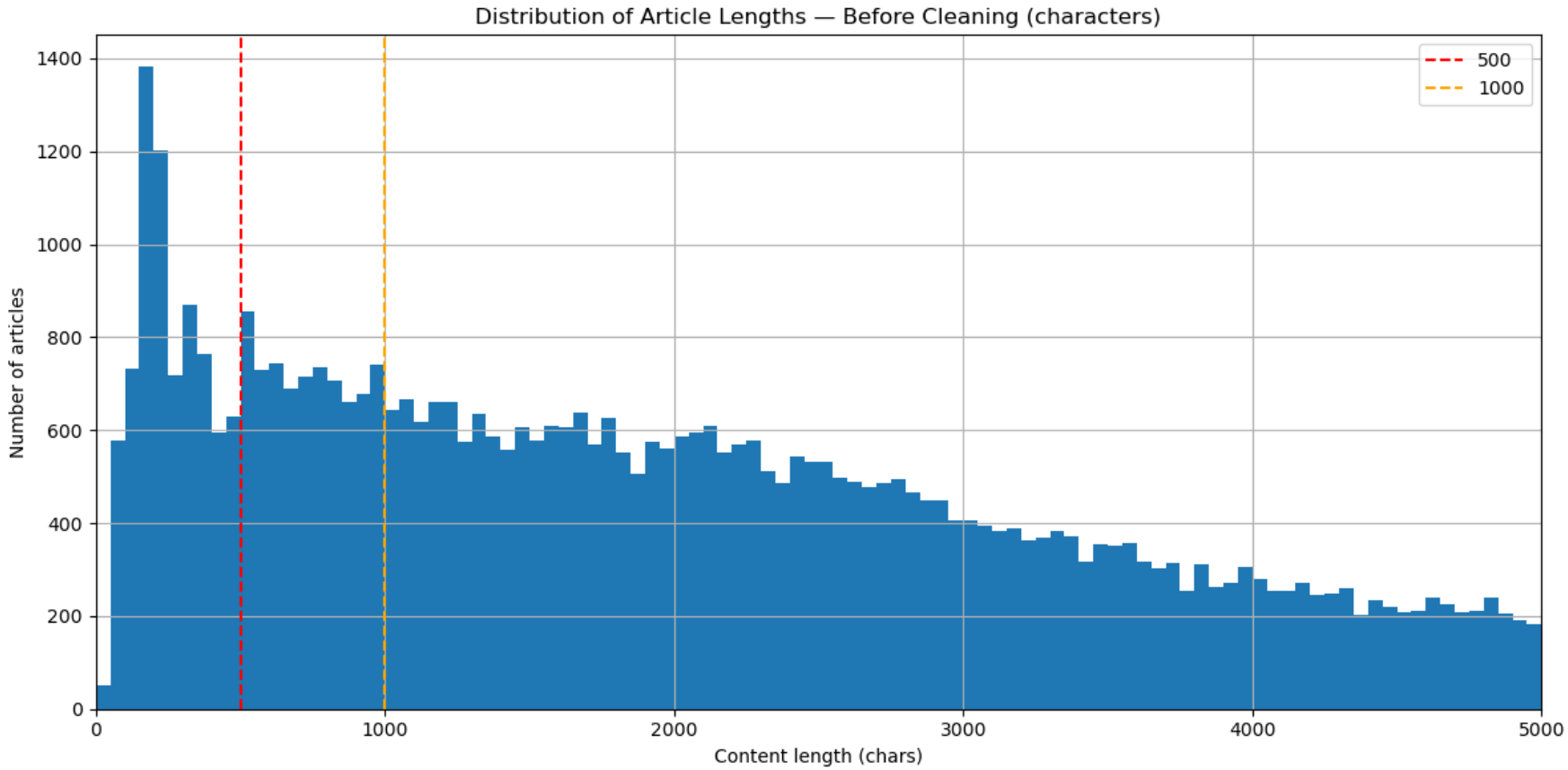
Data Processing and Exploratory Data Analysis (EDA)

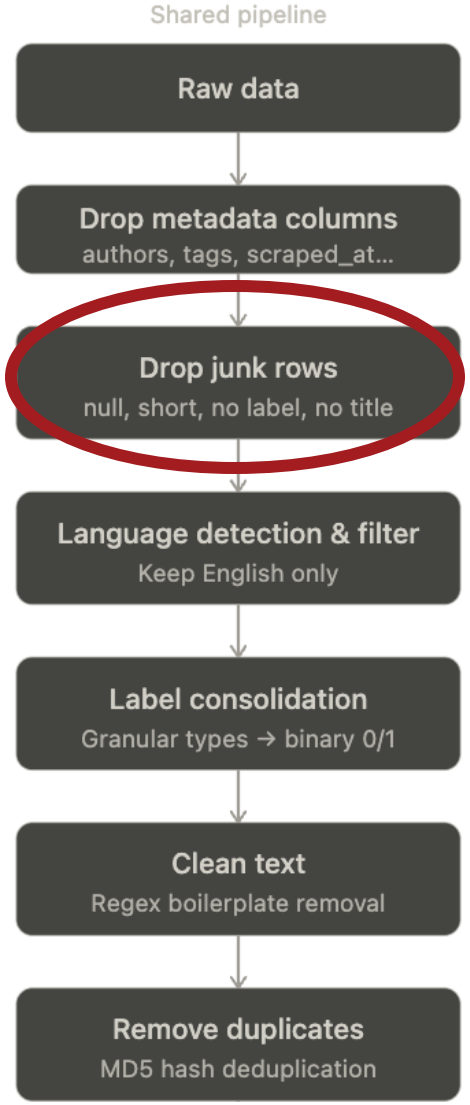
- Drop junk rows
  - Drops rows where content is NaN
  - After stripping whitespace, content less than 500 characters is dropped.
    - Could be bad scrapes or artifacts which introduce extra noise.



Data Processing and Exploratory Data Analysis (EDA)

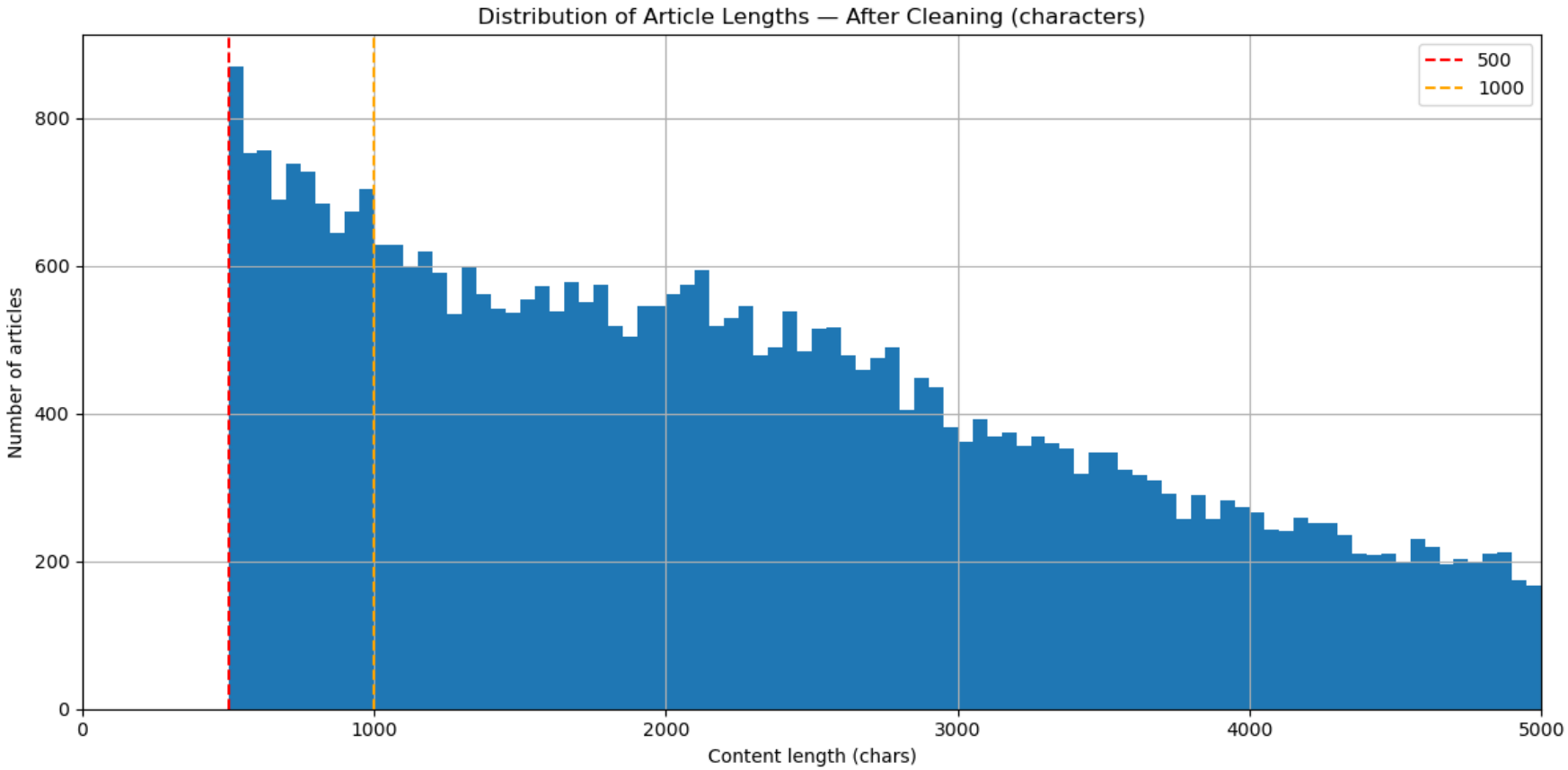
- Before dropping:

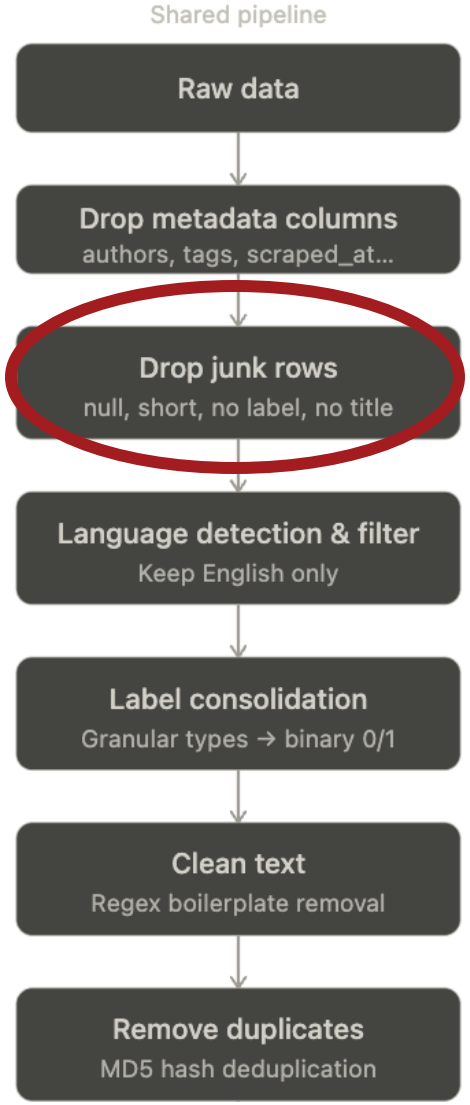




Data Processing and Exploratory Data Analysis (EDA)

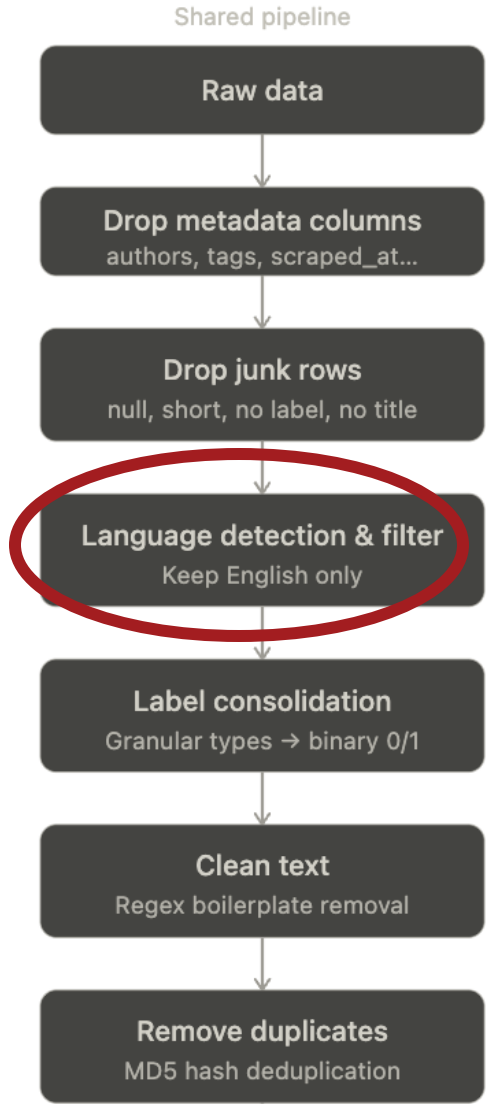
- After dropping:





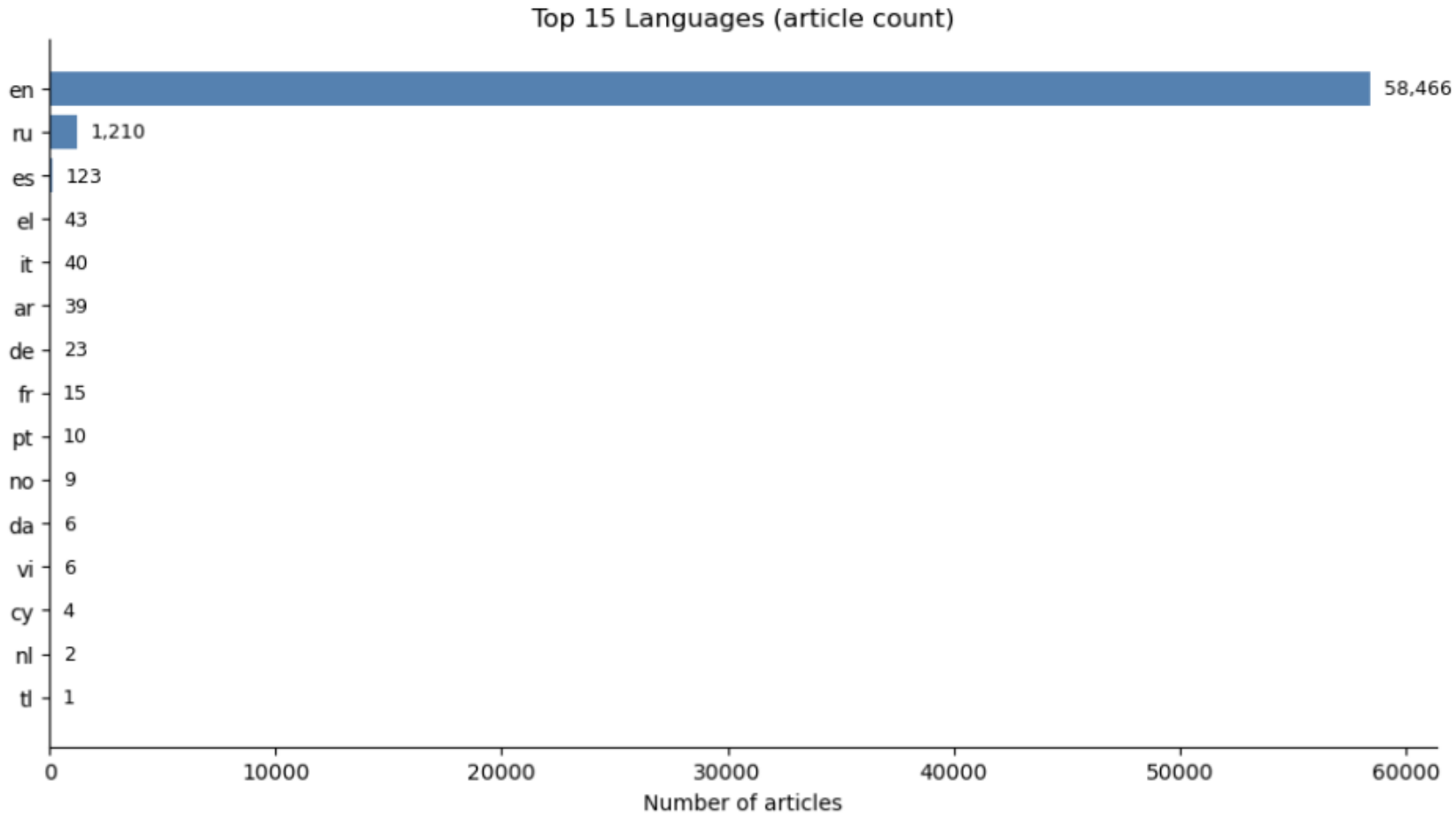
Data Processing and Exploratory Data Analysis (EDA)

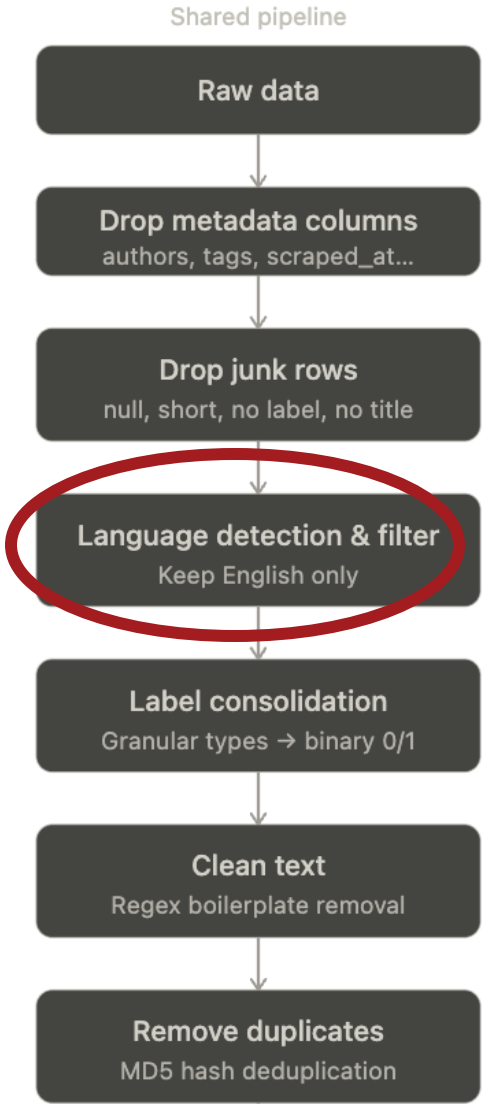
- Language detection & Filter:



Data Processing and Exploratory Data Analysis (EDA)

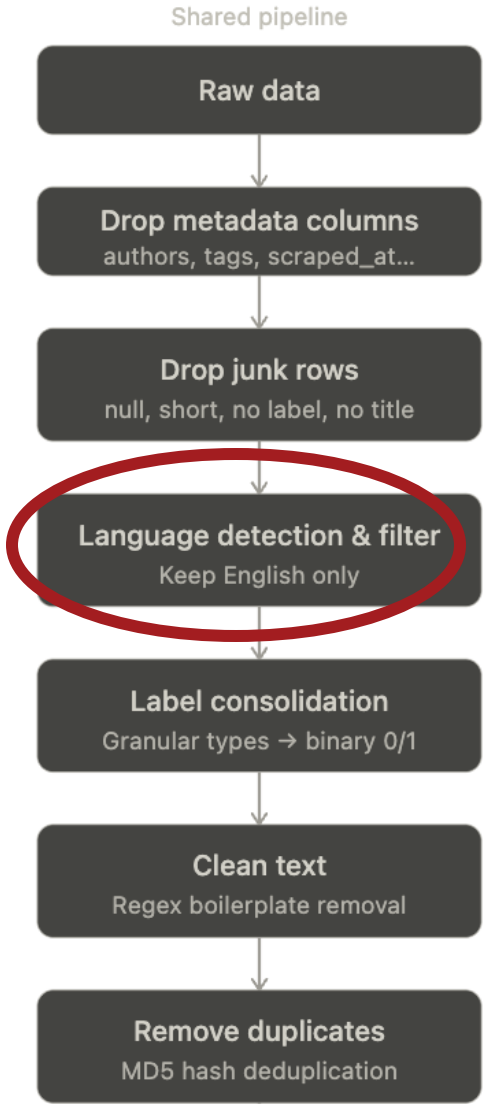
- Language detection & Filter:





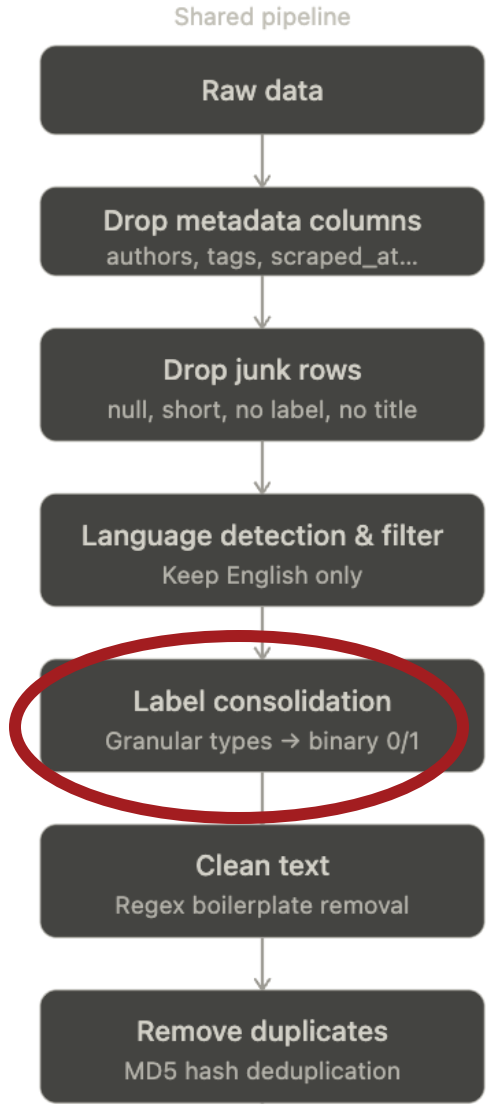
Data Processing and Exploratory Data Analysis (EDA)

- Language detection & Filter:
  - Retained only English articles for consistency in the model training and removed everything else



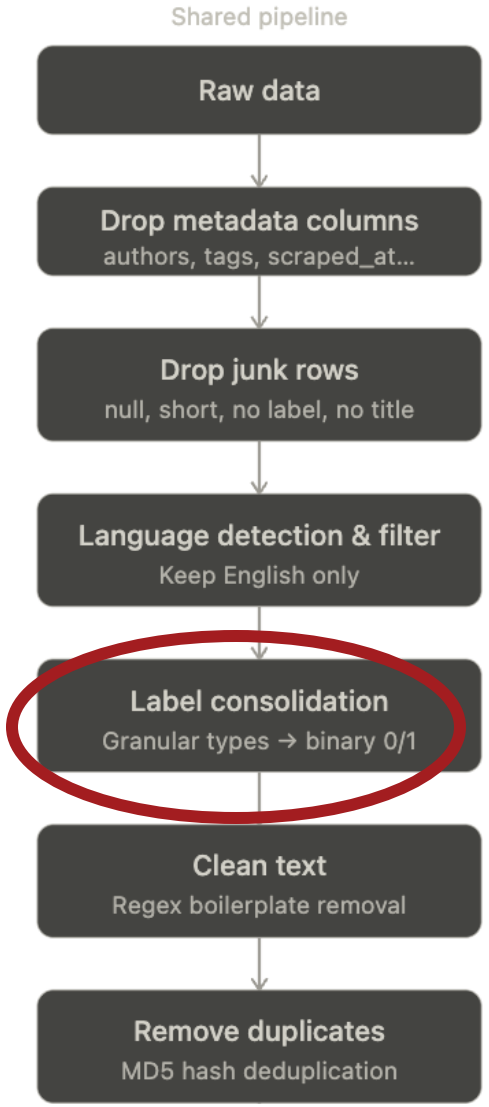
Data Processing and Exploratory Data Analysis (EDA)

- Label Consolidation:
  - We had to decide how we wanted to define the classification.



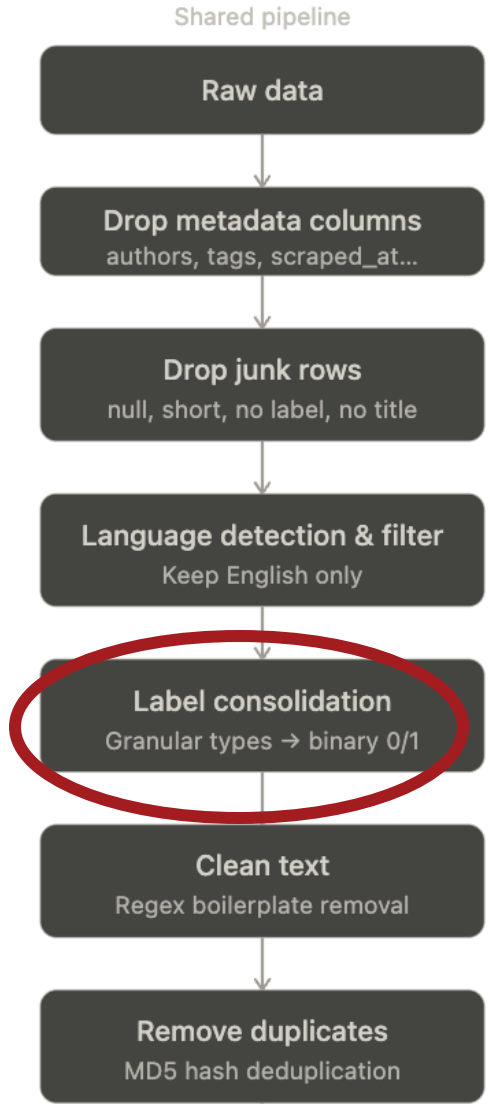
Data Processing and Exploratory Data Analysis (EDA)

- Label Consolidation:
  - We had to decide how we wanted to define the classification.
- Two approaches were considered:
  - Multiclass classification, where the ML models were trained to predict the specific categories
  - Binary Classification, where the ML models predicted whether an article is **reliable** or **unreliable**



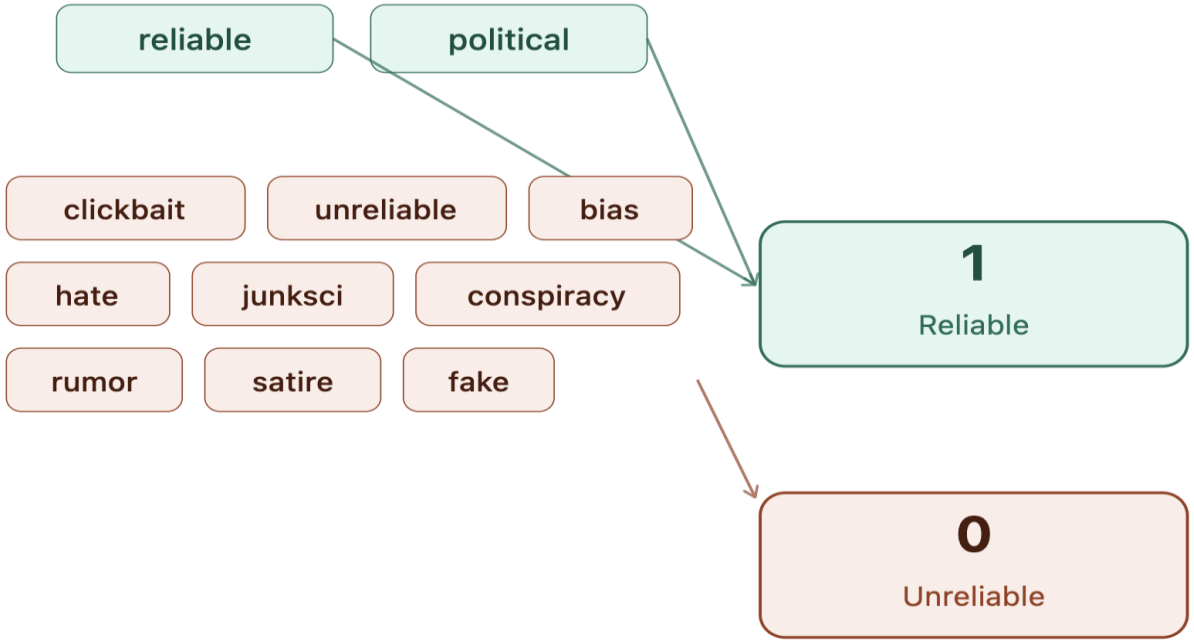
Data Processing and Exploratory Data Analysis (EDA)

- Label Consolidation:
  - We had to decide how we wanted to define the classification.
- Two approaches were considered:
  - Multiclass classification, where the ML models were trained to predict the specific categories
  - Binary Classification, where the ML models predicted whether an article is **reliable** or **unreliable**
- To create a simpler and more efficient classifier, we chose Binary Classification.



Data Processing and Exploratory Data Analysis (EDA)

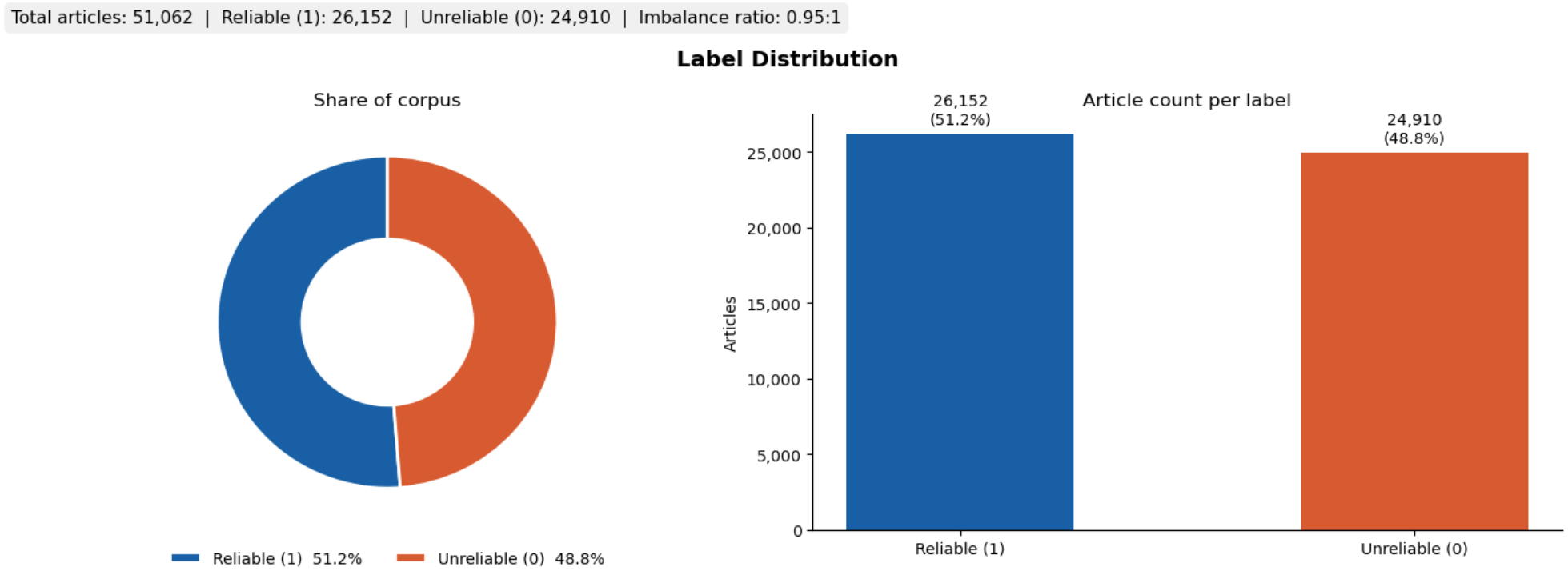
- Label Consolidation:
- Our binary label mapping involved a collection of subjective decisions
  - Clickbait articles could originate from credible news sources but use exaggerated or misleading headlines.
  - Political articles can vary widely in quality and objectivity.

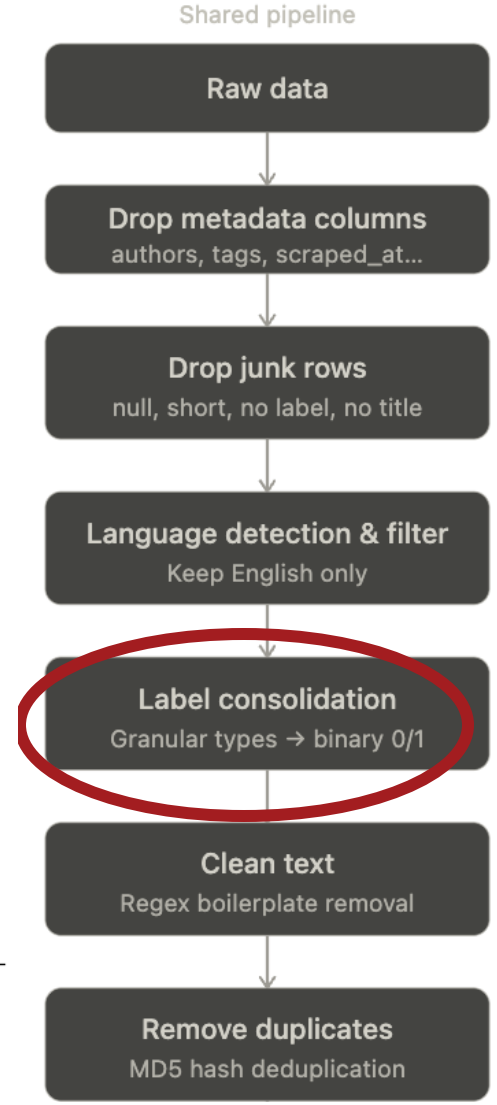


Type
Fake News
Satire
Extreme Bias
Conspiracy Theory
State News
Junk Science
Hate News
Clickbait
Proceed With Caution
Political
Credible

Data Processing and Exploratory Data Analysis (EDA)

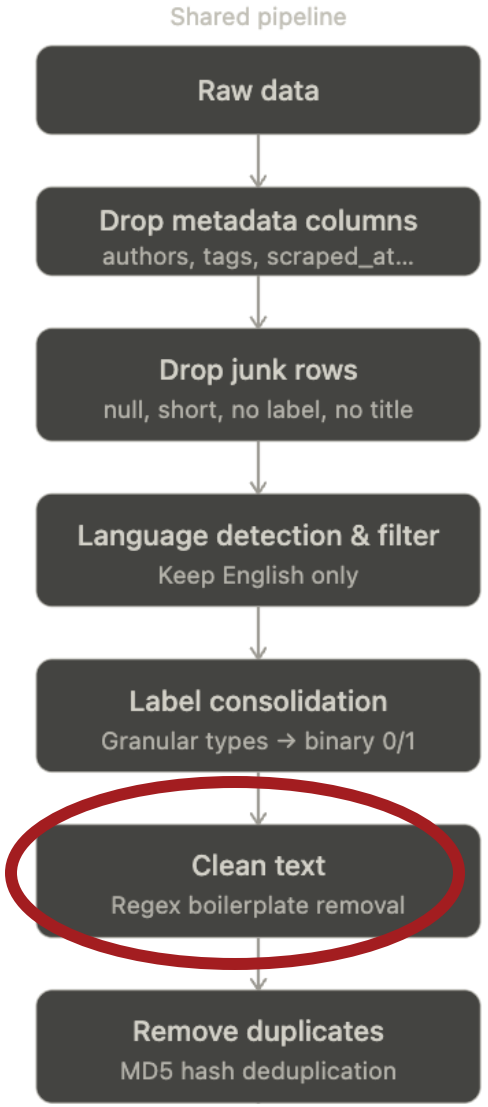
- Label Consolidation:





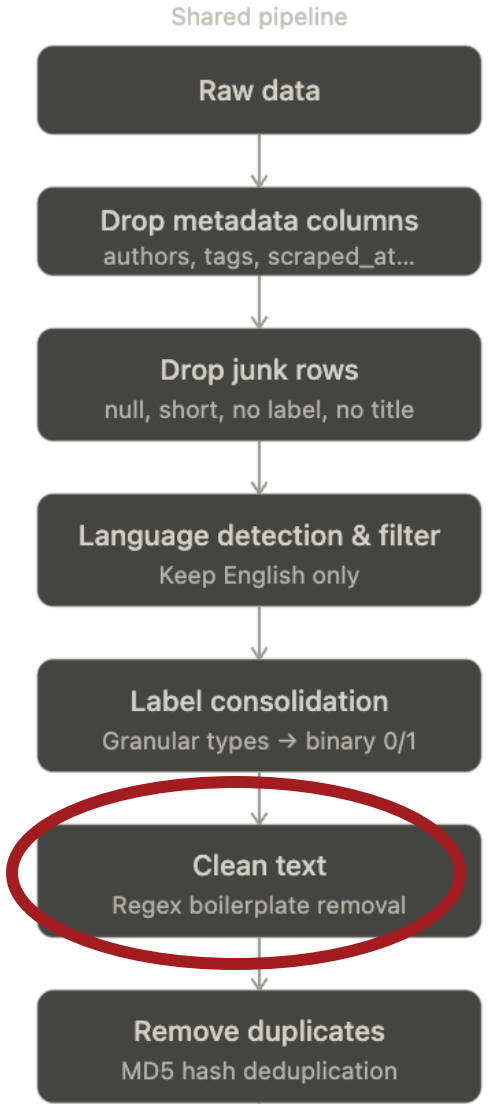
Data Processing and Exploratory Data Analysis (EDA)

- Text cleaning (Boilerplate removal):
  - Boilerplate refers to (in the context of data cleaning) irrelevant, repetitive textual noise.



Data Processing and Exploratory Data Analysis (EDA)

- Text cleaning (Boilerplate removal):
  - Boilerplate refers to (in the context of data cleaning) irrelevant, repetitive textual noise.
    - Disclaimers, website headers/footers for example.



Data Processing and Exploratory Data Analysis (EDA)

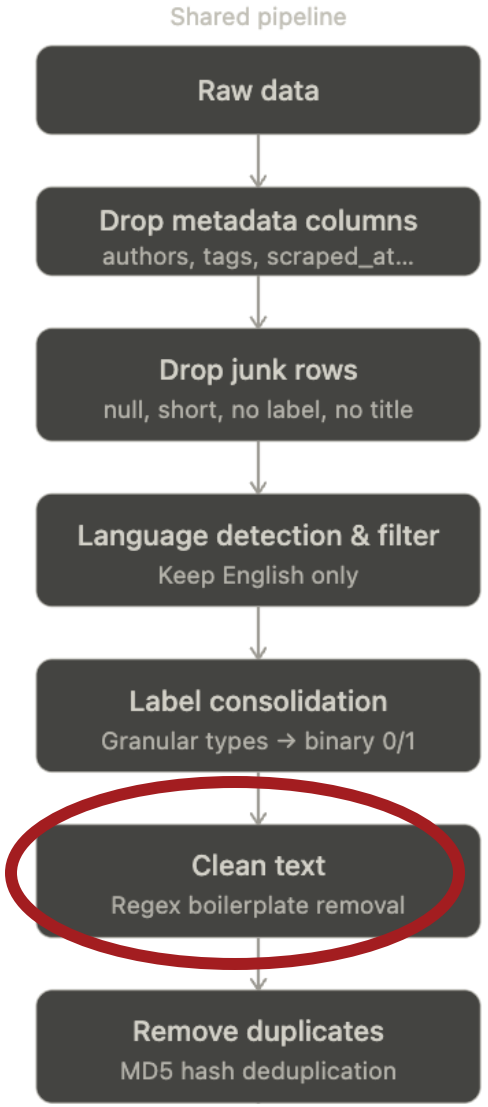
- Text cleaning (Boilerplate removal):
  - Boilerplate refers to (in the context of data cleaning) irrelevant, repetitive textual noise.
    - Disclaimers, website headers/footers for example.

Sputnik News | World

The Kremlin said on Friday that US sanctions imposed on Russian energy firms would have "serious consequences" for global oil markets. Read more at https://sputniknews.com/world/sanctions-full

Spokesman Dmitry Peskov told reporters the move was "counterproductive" and risked destabilising supply chains ahead of winter.

© 2023 Sputnik. All rights reserved.
Subscribe to Sputnik's newsletter for daily updates.
Get the latest news in our Telegram channel.



Data Processing and Exploratory Data Analysis (EDA)

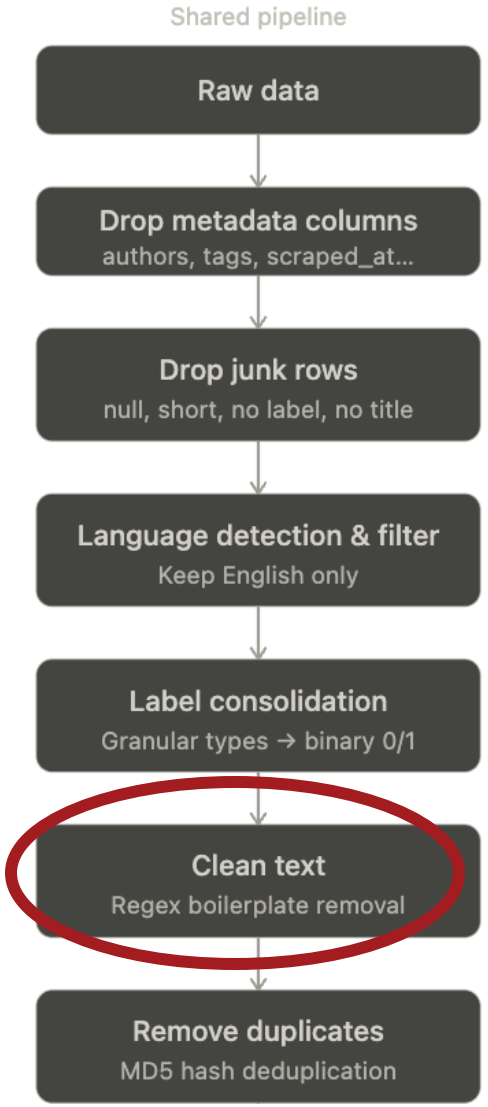
- Text cleaning (Boilerplate removal):
  - Boilerplate refers to (in the context of data cleaning) irrelevant, repetitive textual noise.
    - Disclaimers, website headers/footers for example.

Sputnik News | World

The Kremlin said on Friday that US sanctions imposed on Russian energy firms would have "serious consequences" for global oil markets. Read more at https://sputniknews.com/world/sanctions-full

Spokesman Dmitry Peskov told reporters the move was "counterproductive" and risked destabilising supply chains ahead of winter.

© 2023 Sputnik. All rights reserved.
Subscribe to Sputnik's newsletter for daily updates.
Get the latest news in our Telegram channel.



Data Processing and Exploratory Data Analysis (EDA)

- Text cleaning (Boilerplate removal):
  - Boilerplate refers to (in the context of data cleaning) irrelevant, repetitive textual noise.
    - Disclaimers, website headers/footers for example.

Sputnik News | World

The Kremlin said on Friday that US sanctions imposed on Russian energy firms would have "serious consequences" for global oil markets. Read more at https://sputniknews.com/world/sanctions-full

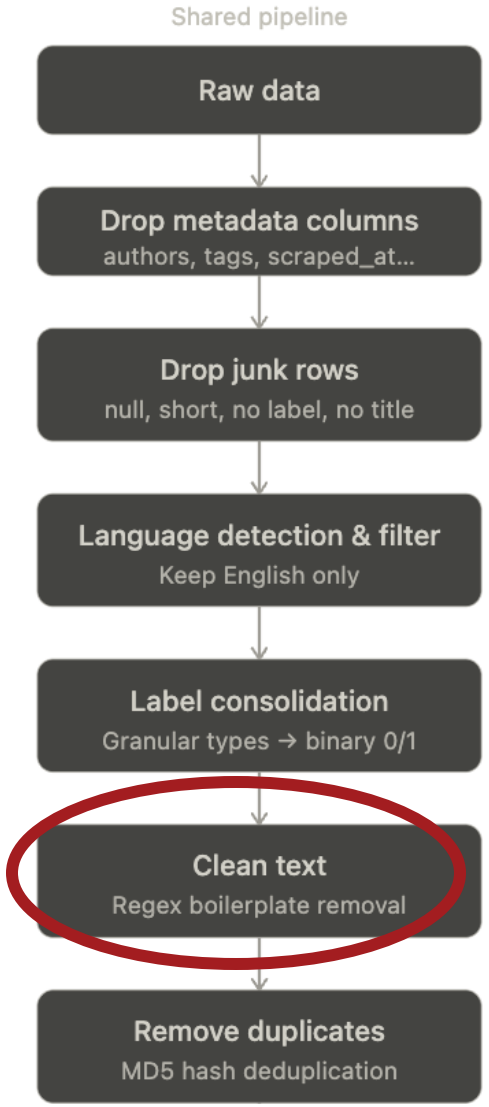
Spokesman Dmitry Peskov told reporters the move was "counterproductive" and risked destabilising supply chains ahead of winter.

© 2023 Sputnik. All rights reserved.
Subscribe to Sputnik's newsletter for daily updates.
Get the latest news in our Telegram channel.

→

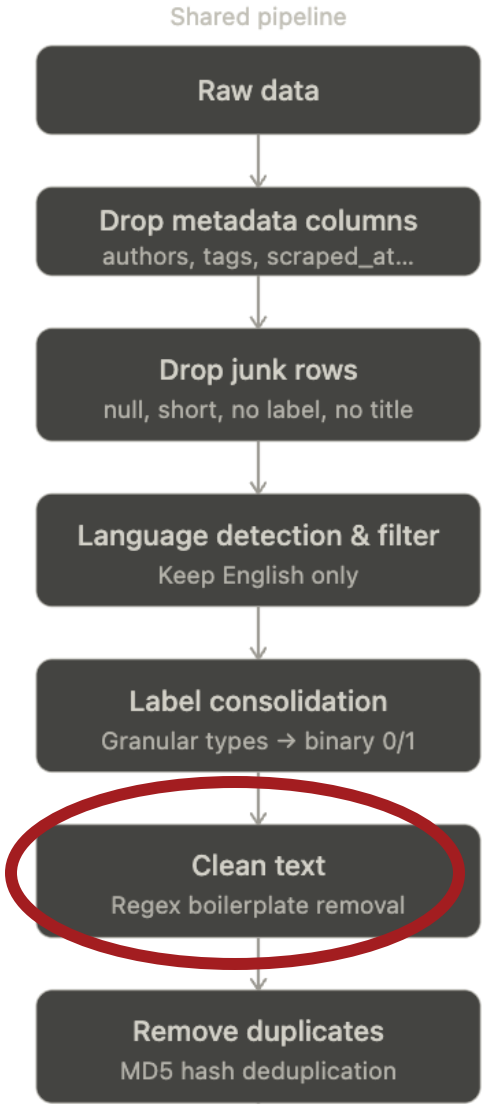
The Kremlin said on Friday that US sanctions imposed on Russian energy firms would have "serious consequences" for global oil markets.

Spokesman Dmitry Peskov told reporters the move was "counterproductive" and risked destabilising supply chains ahead of winter.



Data Processing and Exploratory Data Analysis (EDA)

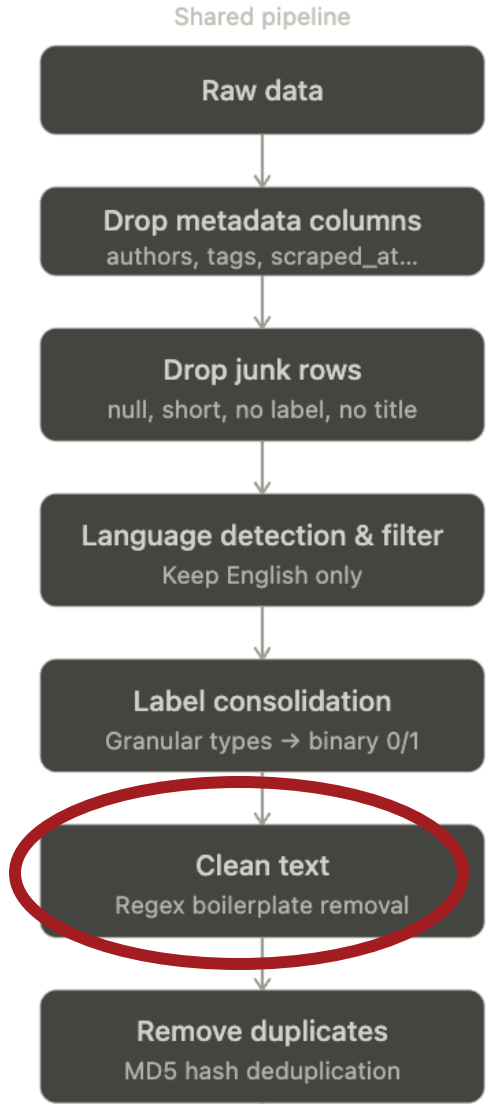
- Text cleaning (Boilerplate removal):
  - Using Regular Expressions (REGEX), Boilerplate patterns (12 in total) are defined and if found, the matching text is stripped from every article.



Data Processing and Exploratory Data Analysis (EDA)

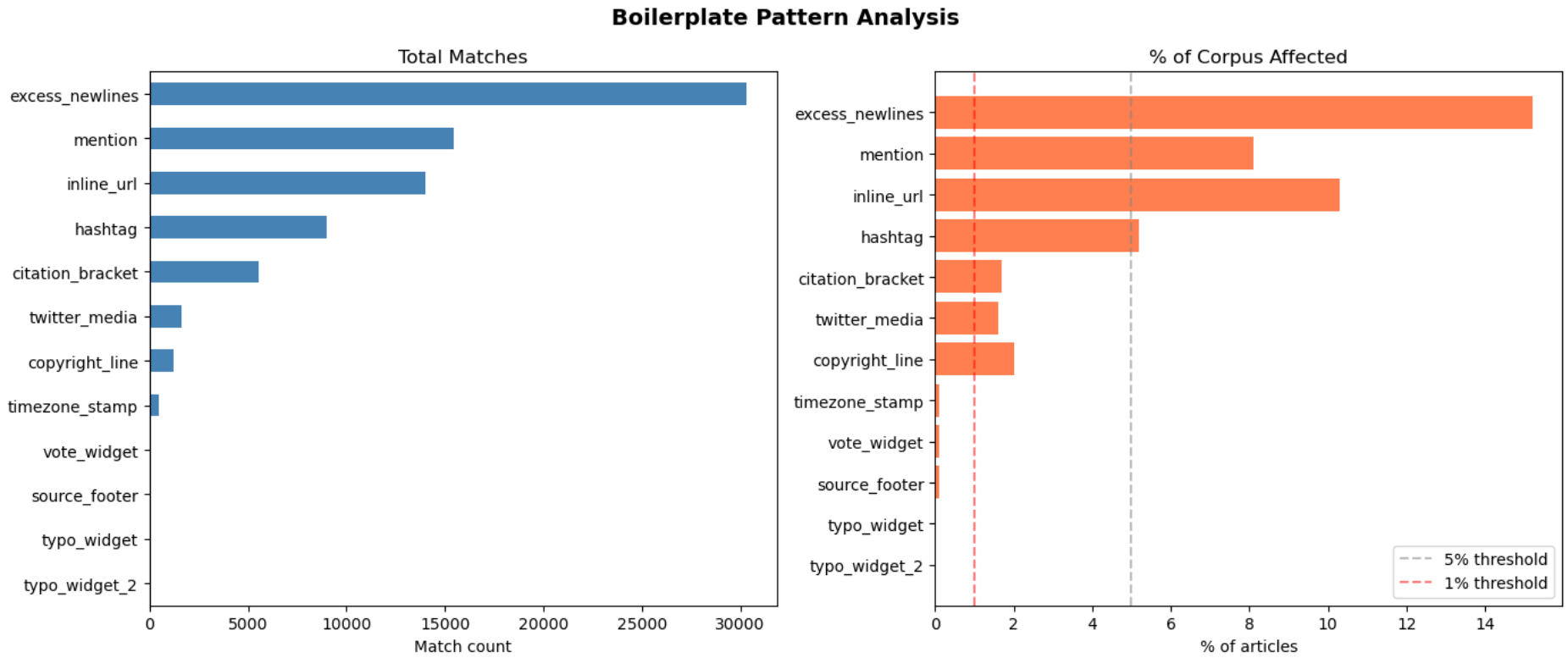
- Text cleaning (Boilerplate removal):
  - Using Regular Expressions (REGEX), Boilerplate patterns (12 in total) are defined and if found, the matching text is stripped from every article.

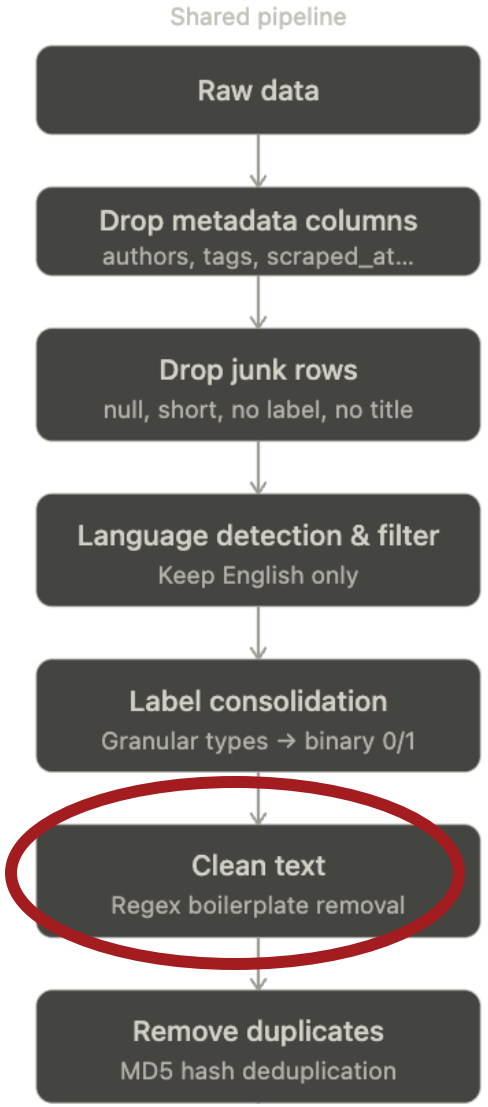
PATTERN NAME	REGEX	EXAMPLE MATCH
metadata **timezone_stamp** Metadata artifact — not linguistic content	`\(EDT\)|\(EST\)|\(UTC\)`	`3:45 PM (EDT)`
source identity **copyright_line** Publisher identity — model would cheat on label	`©\s.*`	`© 2023 Sputnik. All rights reserved.`



Data Processing and Exploratory Data Analysis (EDA)

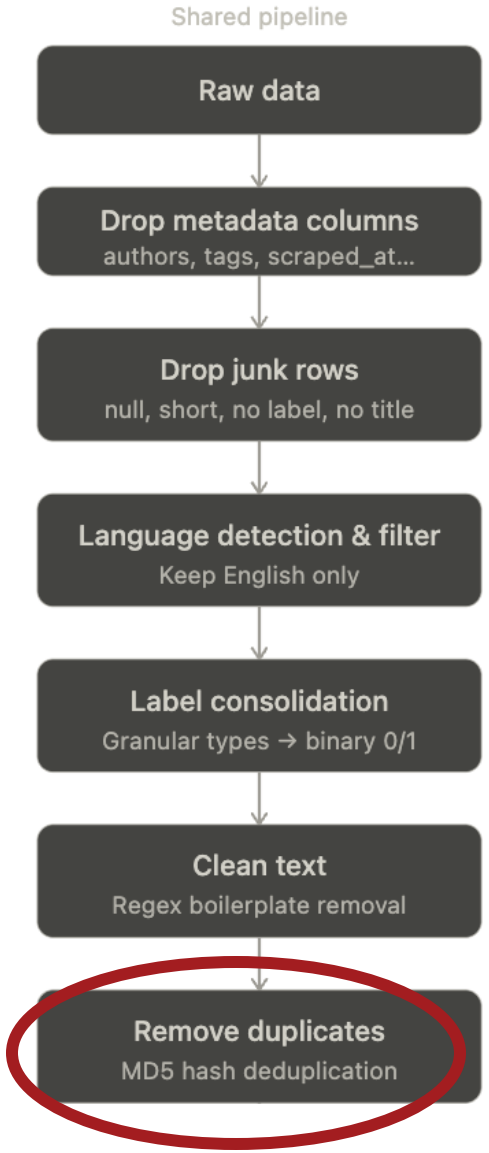
- Text cleaning (Boilerplate removal):
  - Threshold counts just show how prevalent the patterns are throughout the text and how well the REGEX definitions work.





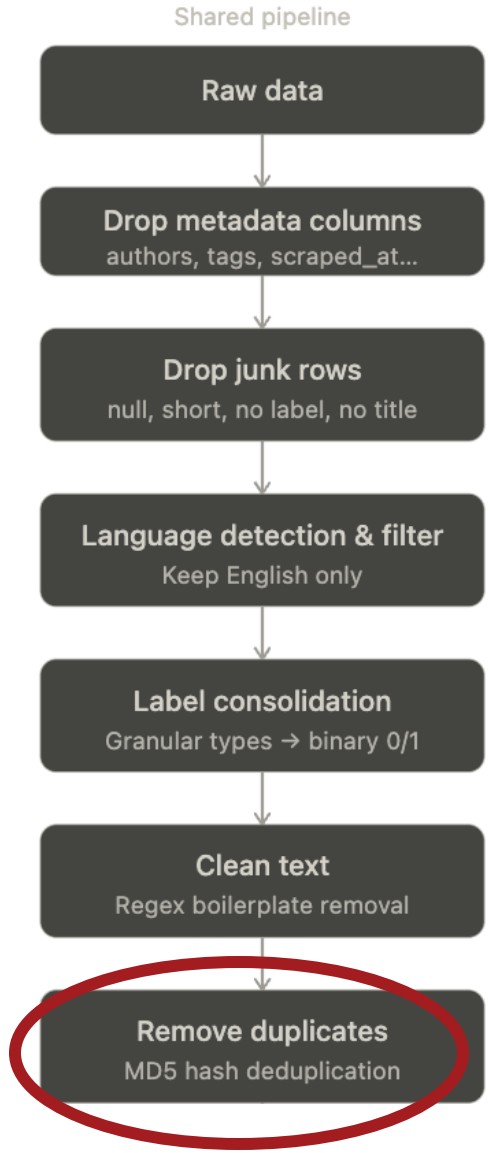
Data Processing and Exploratory Data Analysis (EDA)

- Removing Duplicates:
  - Can cause model to see identical content multiple times
    - Inflates confidence on those examples
    - Risks data leakage between training and evaluation sets



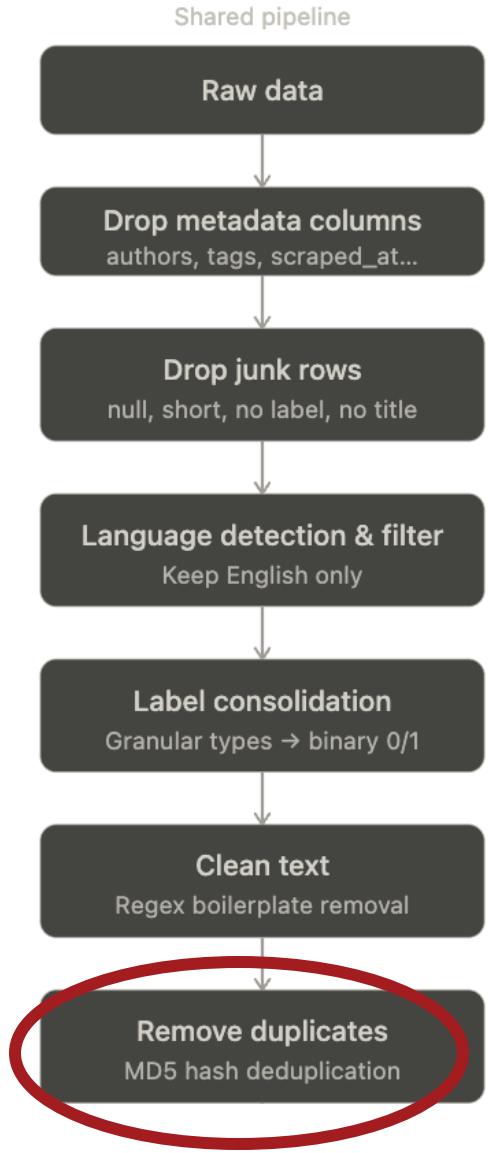
Data Processing and Exploratory Data Analysis (EDA)

- Removing Duplicates:
  - Two passes removing duplicates:
    - 1st pass: Deduplication occurs on the main Fake News Corpus before making our sample dataset.



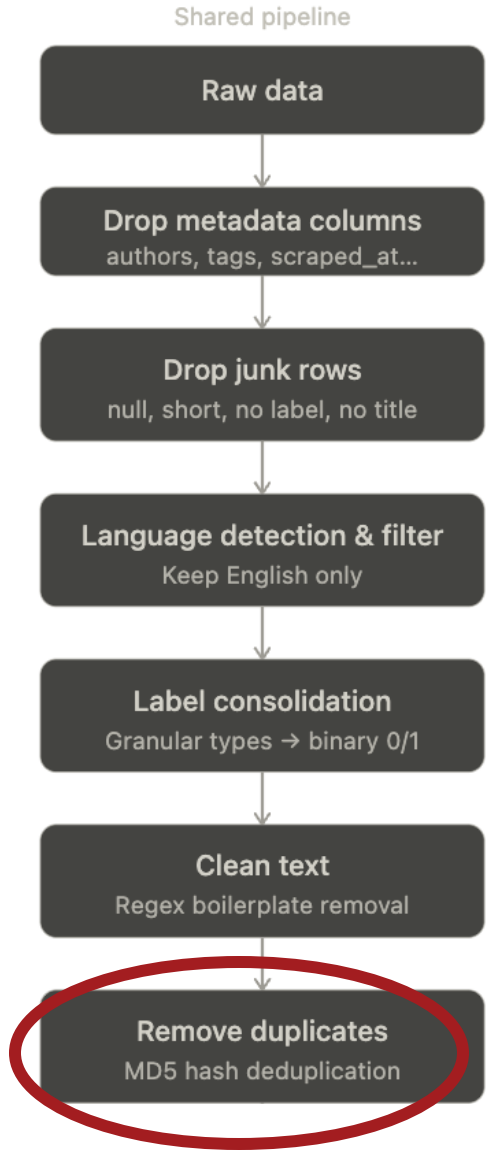
Data Processing and Exploratory Data Analysis (EDA)

- Removing Duplicates:
  - Two passes removing duplicates:
    - 1st pass: Deduplication occurs on the main Fake News Corpus before making our sample dataset.
    - Drops using exact ID and/or URL duplicates



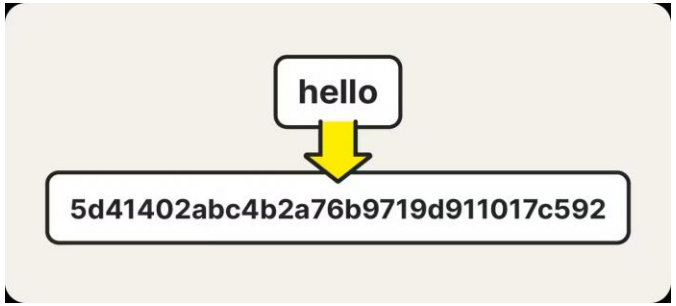
Data Processing and Exploratory Data Analysis (EDA)

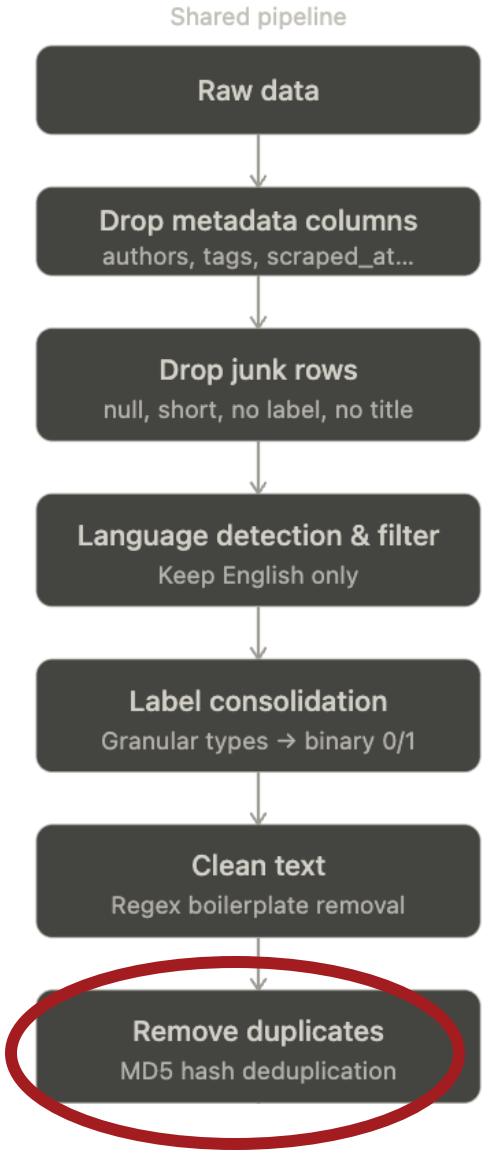
- Removing Duplicates:
  - Two passes removing duplicates:
    - 1st pass: Deduplication occurs on the main Fake News Corpus before making our sample dataset.
    - Drops using exact ID and/or URL duplicates
    - Near duplicate content also dropped:
      - First 500 characters hashed (fingerprints generated from text), articles with matching hashes are dropped



Data Processing and Exploratory Data Analysis (EDA)

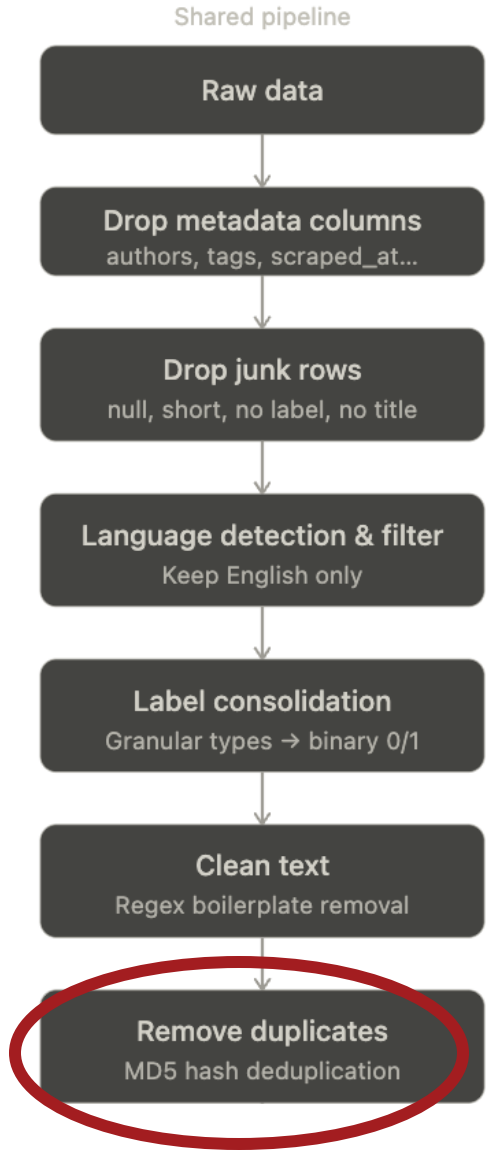
- Removing Duplicates:
  - Two passes removing duplicates:
    - 1st pass: Deduplication occurs on the main Fake News Corpus before making our sample dataset.
    - Drops using exact ID and/or URL duplicates
    - Near duplicate content also dropped:
      - First 500 characters hashed (fingerprints generated from text), articles with matching hashes are dropped





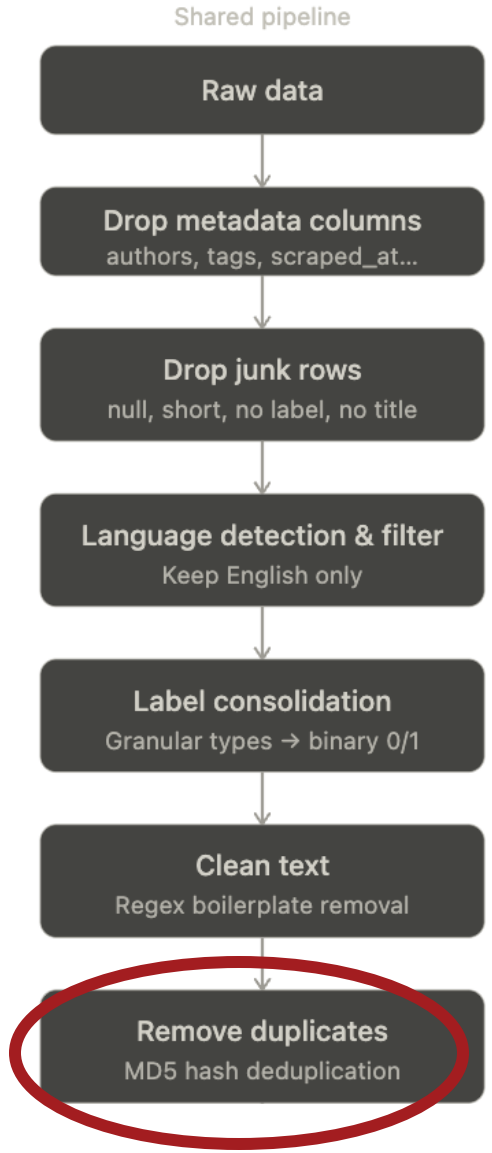
Data Processing and Exploratory Data Analysis (EDA)

- Removing Duplicates:
  - Two passes removing duplicates:
    - 2nd pass: After using REGEX boilerplate removal, the same process is repeated.



Data Processing and Exploratory Data Analysis (EDA)

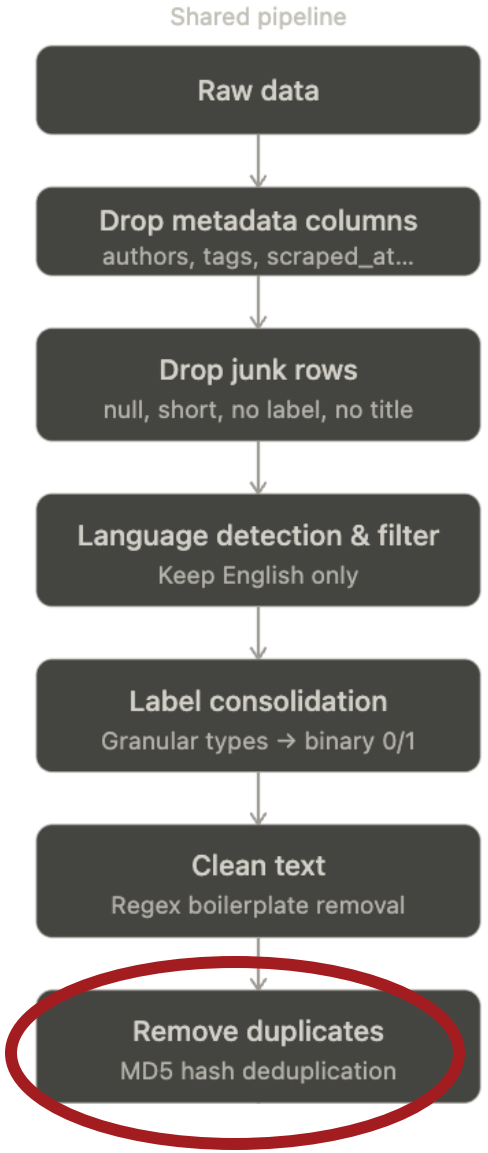
- Removing Duplicates:
  - Two passes removing duplicates:
    - 2nd pass: After using REGEX boilerplate removal, the same process is repeated.
    - Edge cases where two articles were different in raw text (e.g. one had an extra copyright line) become identical after boilerplate removal



Data Processing and Exploratory Data Analysis (EDA)

- Removing Duplicates:
  - Results from first pass:
    - ≈1.2 Million articles removed from the original dataset (≈ 9.4 Million articles)

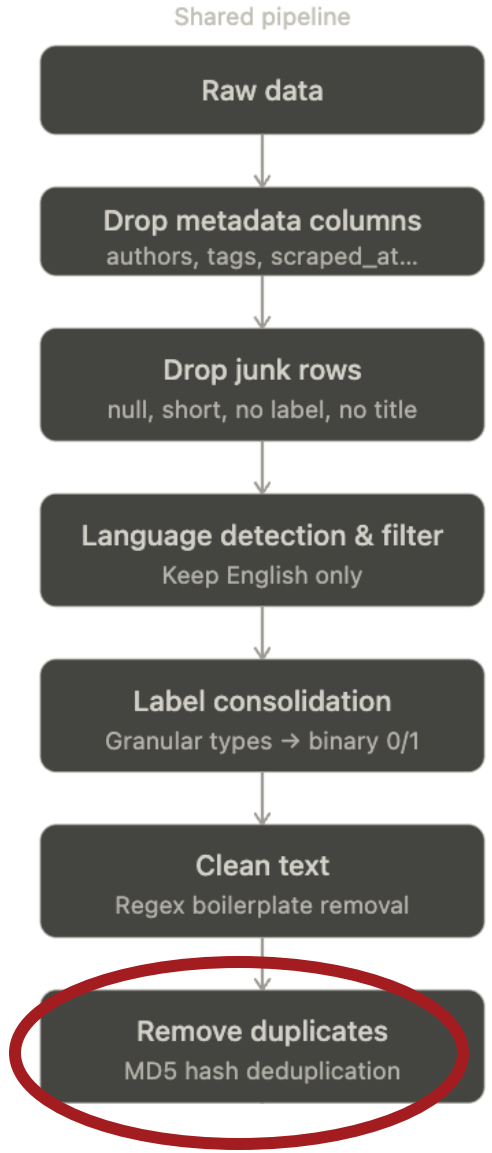
```
type
political     1013695
bias           822594
conspiracy     652989
fake           515256
rumor          328969
unreliable     281296
unknown        223367
clickbait      194759
junksci         95341
satire          89669
hate            68435
reliable        27516
Name: count, dtype: int64
INFO    Dedup summary:
    Exact ID duplicates:              0
    Exact URL duplicates:        20,426
    Near-duplicate content:   1,218,327
    ──────────────────────────────────
    Total removed:            1,238,753
```



Data Processing and Exploratory Data Analysis (EDA)

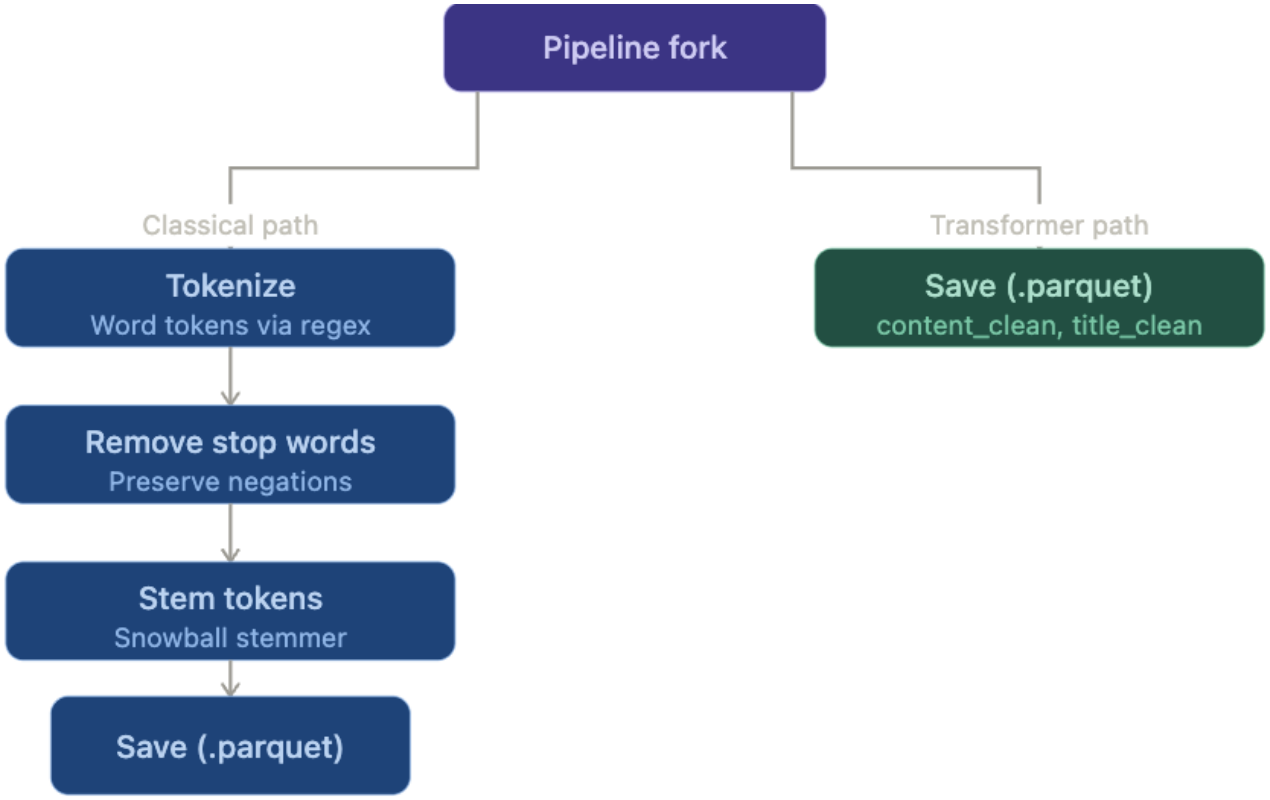
- Removing Duplicates:
  - Results from second pass:
    - 0 duplicates found after Boilerplate removal

```
INFO    Duplicate removal:
    Duplicates removed: 0
```



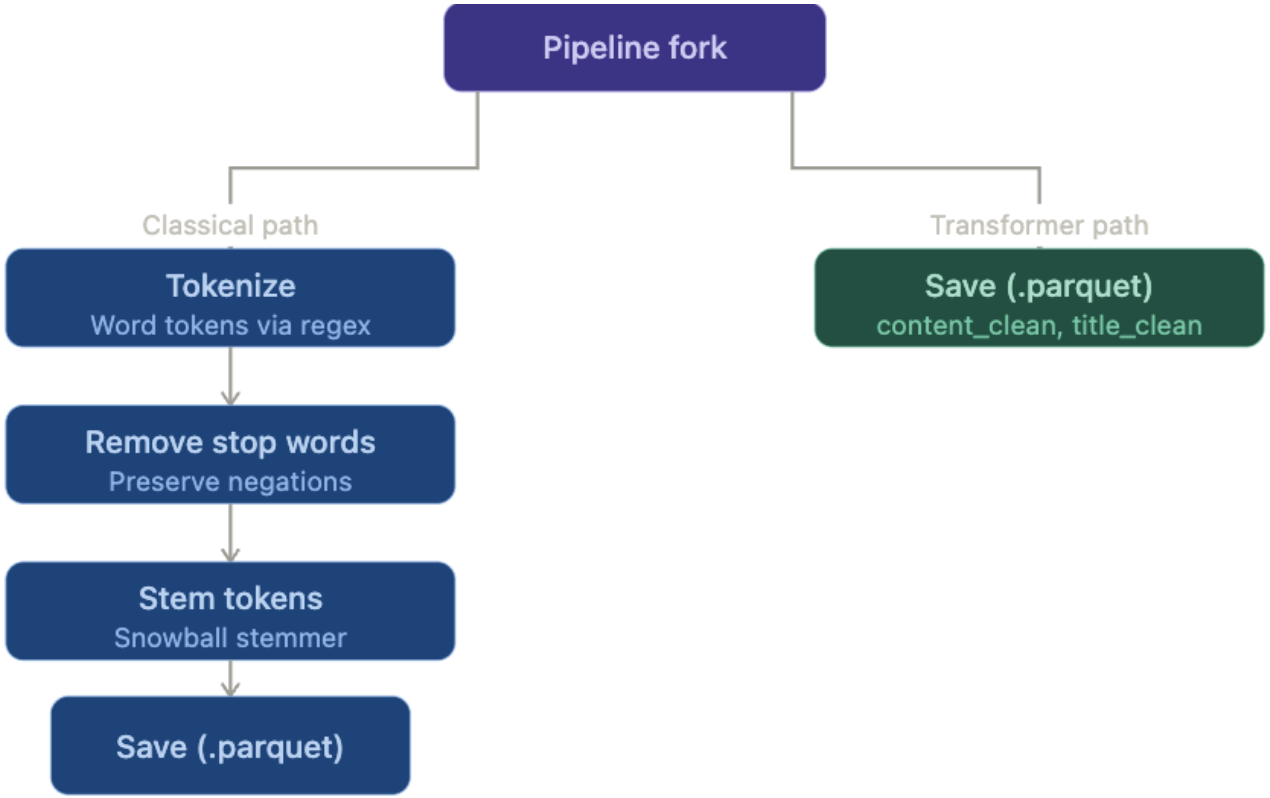
Data Processing and Exploratory Data Analysis (EDA)

- Further processing depended on which ML method we chose to use



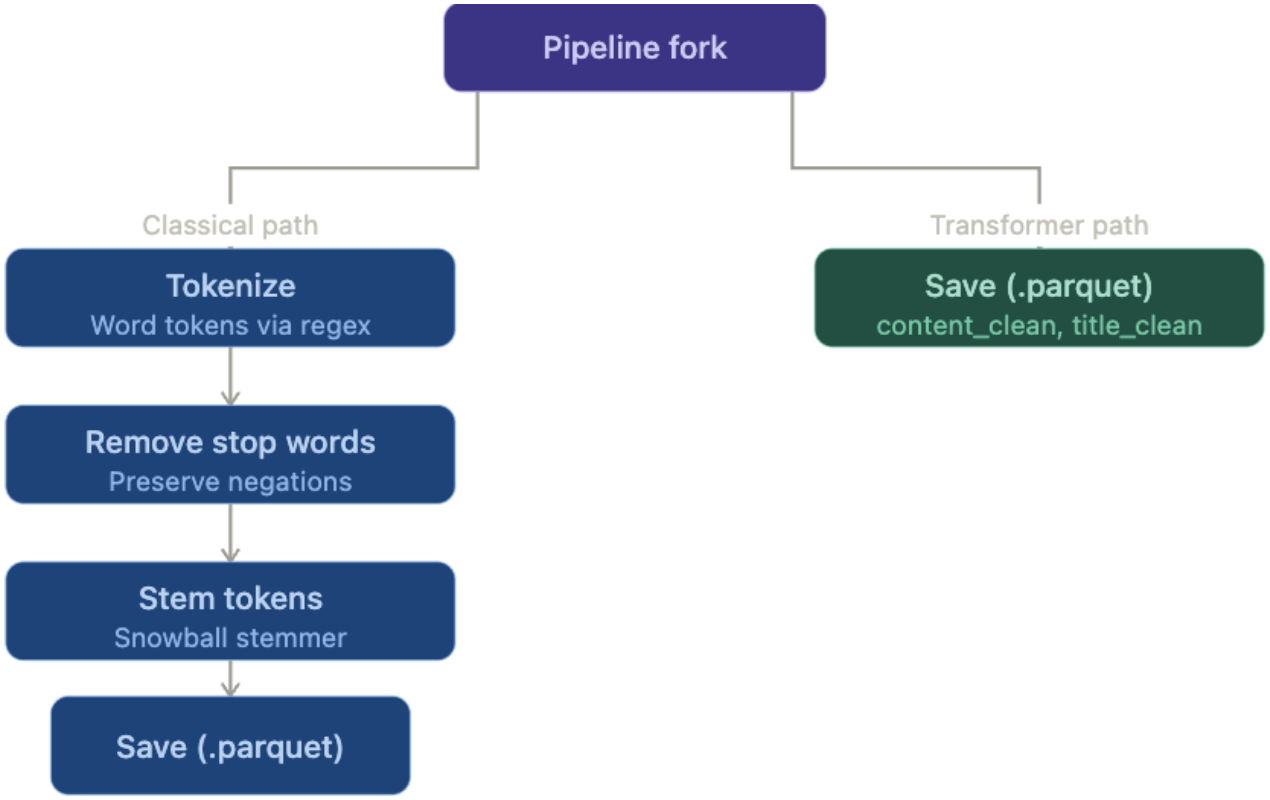
Data Processing and Exploratory Data Analysis (EDA)

- Further processing depended on which ML method we chose to use
- Two methods for two different models:



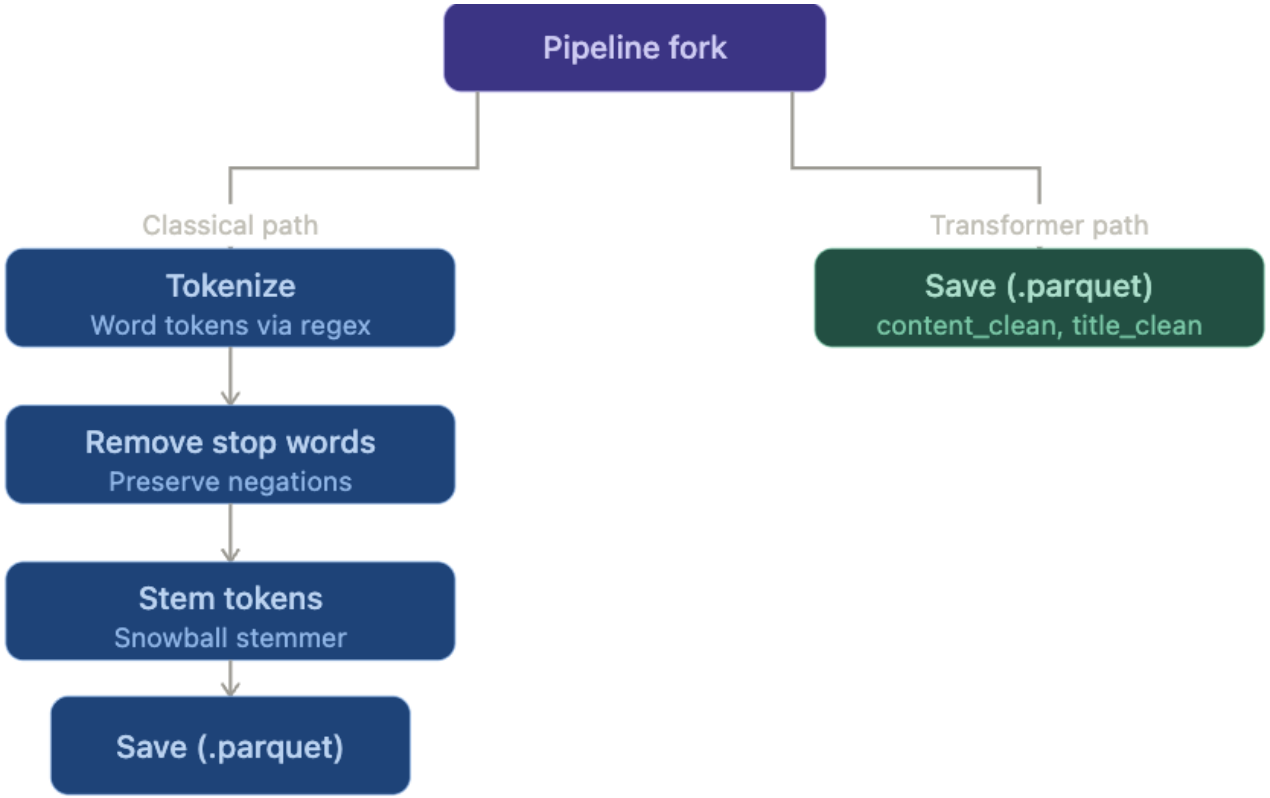
Data Processing and Exploratory Data Analysis (EDA)

- Further processing depended on which ML method we chose to use
- Two methods for two different models:
  - Classical models: Require tokenization, stop word removal and stemming



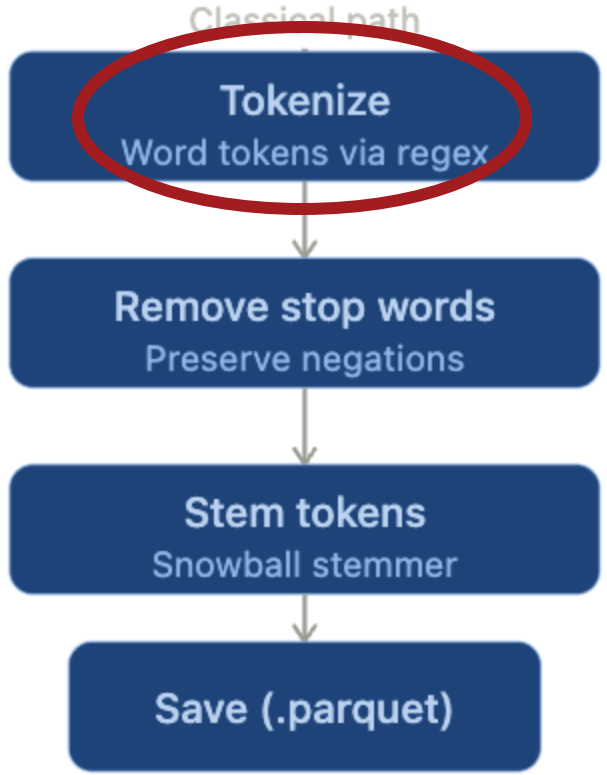
Data Processing and Exploratory Data Analysis (EDA)

- Further processing depended on which ML method we chose to use
- Two methods for two different models:
  - Classical models: Require tokenization, stop word removal and stemming
  - Transformer models: No stemming needed, stop words are kept, tokenization is completed internally



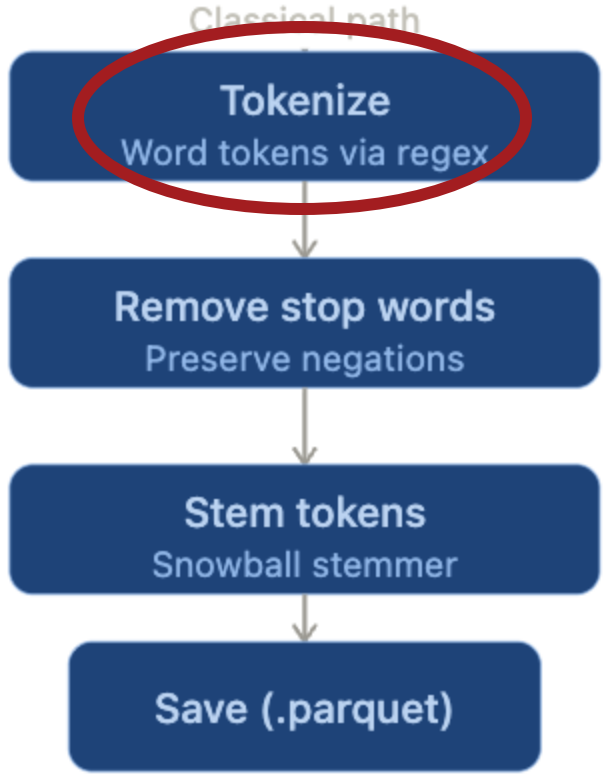
Data Processing and Exploratory Data Analysis (EDA)

- Tokenization:



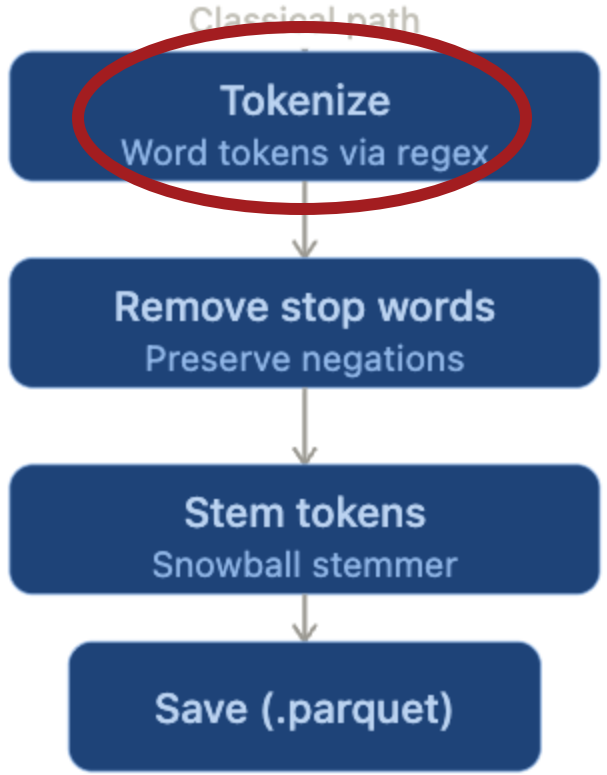
Data Processing and Exploratory Data Analysis (EDA)

- Tokenization:
  - Splits raw text into discrete units (or tokens)



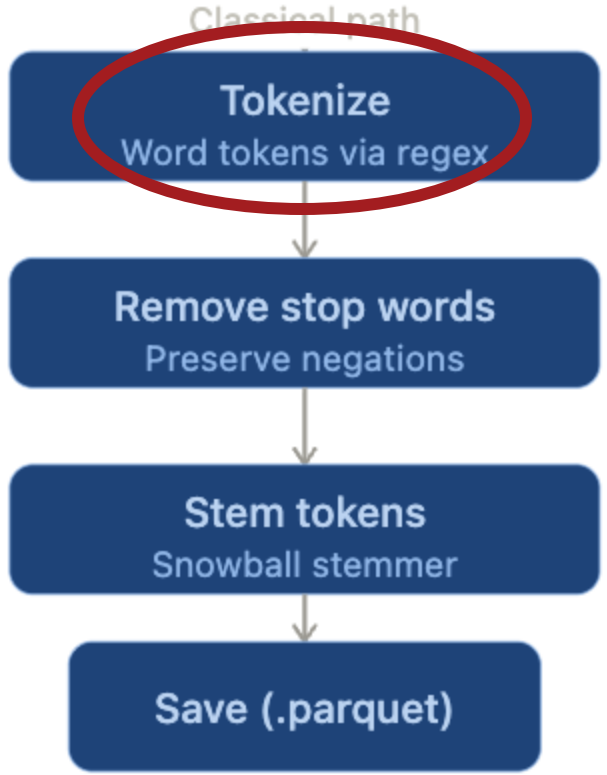
Data Processing and Exploratory Data Analysis (EDA)

- Tokenization:
  - Splits raw text into discrete units (or tokens)
  - Draws boundaries which determine what counts as a 'feature' in our article vocabulary



Data Processing and Exploratory Data Analysis (EDA)

- Tokenization:
  - Splits raw text into discrete units (or tokens)
  - Draws boundaries which determine what counts as a 'feature' in our article vocabulary
  - Prerequisite before removing stop words and stemming can happen

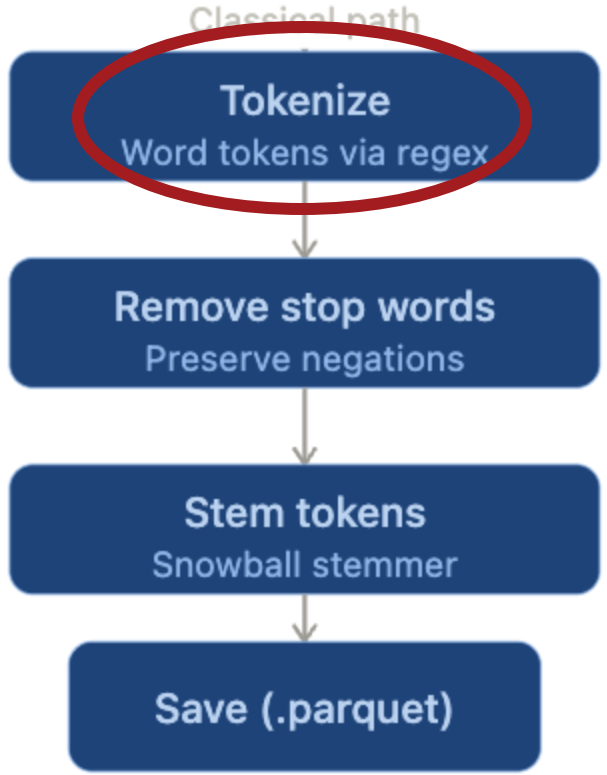


Data Processing and Exploratory Data Analysis (EDA)

- Tokenization:

The Kremlin said on Friday that US sanctions imposed on Russian energy firms would have "serious consequences" for global oil markets.

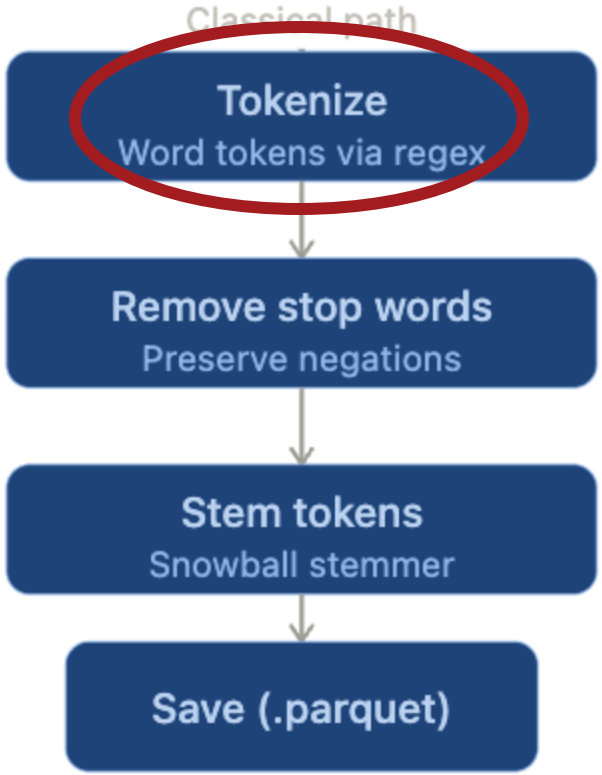
Spokesman Dmitry Peskov told reporters the move was "counterproductive" and risked destabilising supply chains ahead of winter.



Data Processing and Exploratory Data Analysis (EDA)

The Kremlin said on Friday that US sanctions imposed on Russian energy firms would have "serious consequences" for global oil markets.

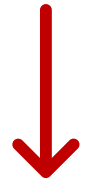
Spokesman Dmitry Peskov told reporters the move was "counterproductive" and risked destabilising supply chains ahead of winter.



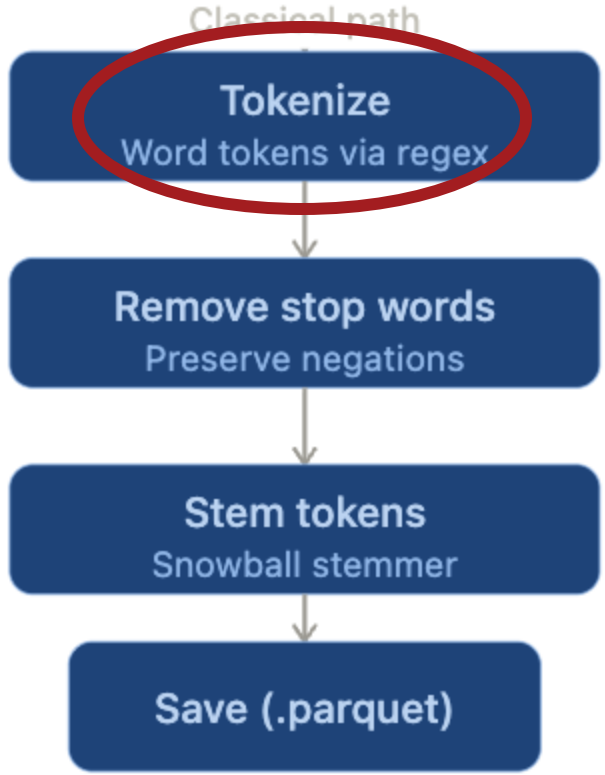
Data Processing and Exploratory Data Analysis (EDA)

The Kremlin said on Friday that US sanctions imposed on Russian energy firms would have "serious consequences" for global oil markets.

Spokesman Dmitry Peskov told reporters the move was "counterproductive" and risked destabilising supply chains ahead of winter.

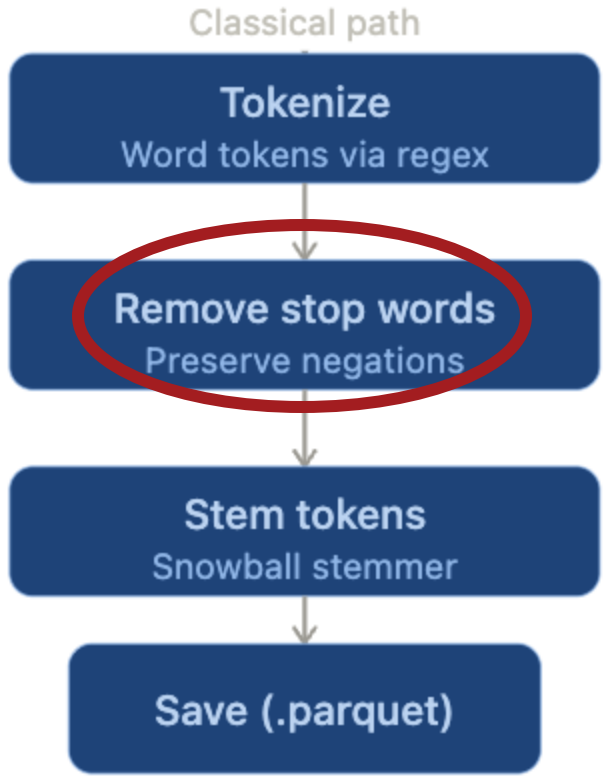


```
['The', 'Kremlin', 'said', 'on', 'Friday', 'that', 'US', 'sanctions',
 'imposed', 'on', 'Russian', 'energy', 'firms', 'would', 'have', '"',
 'serious', 'consequences', '"', 'for', 'global', 'oil', 'markets', '.',
 'Spokesman', 'Dmitry', 'Peskov', 'told', 'reporters', 'the', 'move',
 'was', '"', 'counterproductive', '"', 'and', 'risked', 'destabilising',
 'supply', 'chains', 'ahead', 'of', 'winter', '.']
```



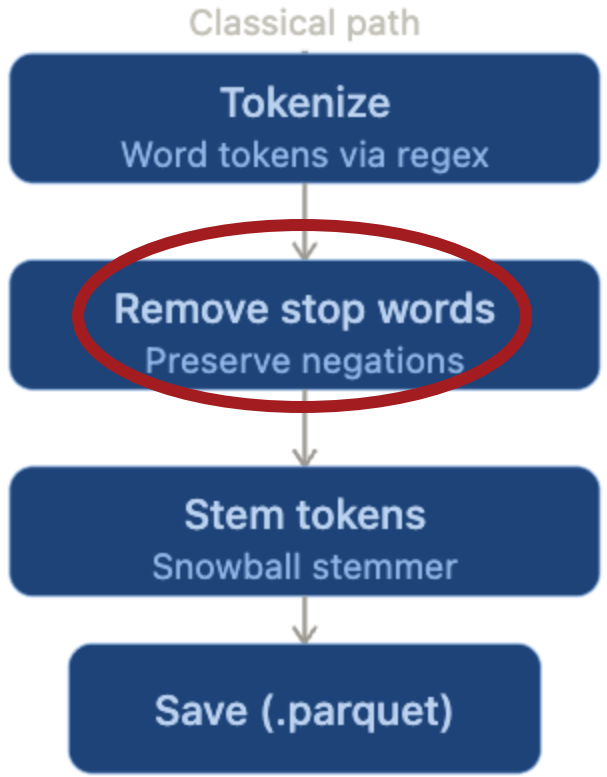
Data Processing and Exploratory Data Analysis (EDA)

- Stop words:



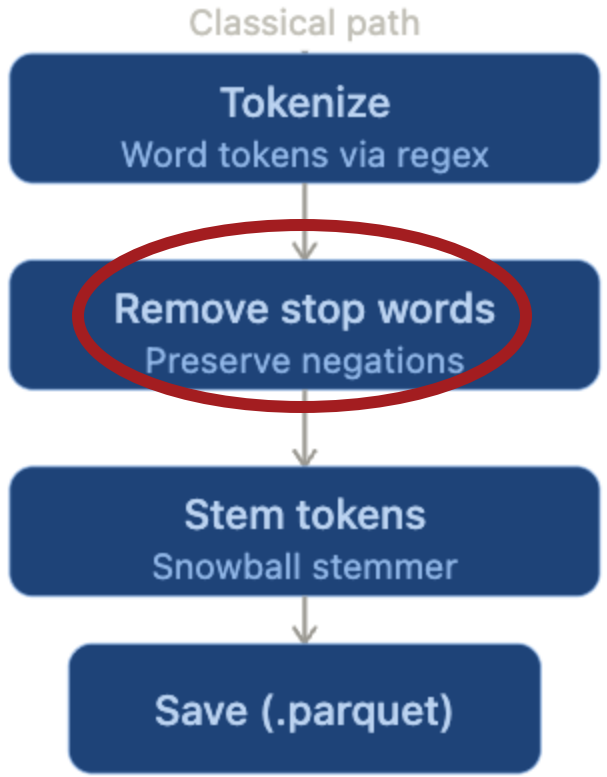
Data Processing and Exploratory Data Analysis (EDA)

- Stop words:
  - Common words that appear at high frequency across all texts, regardless of type or label



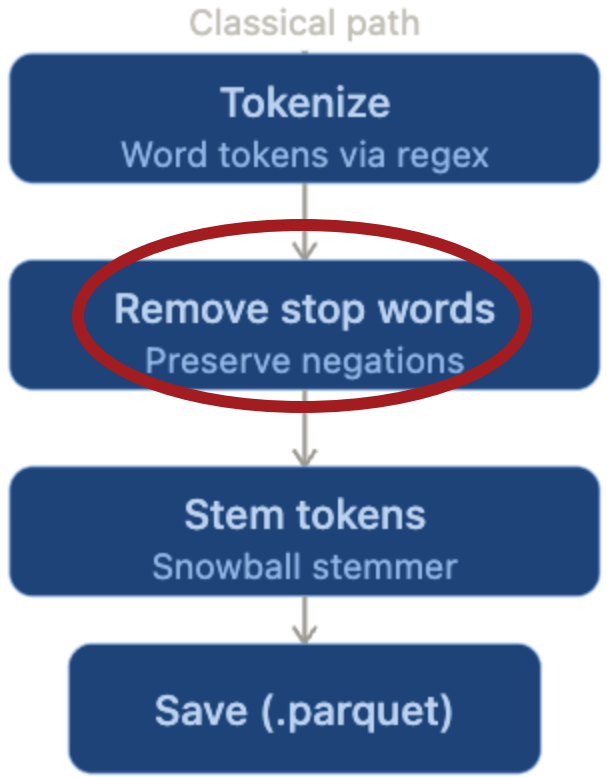
Data Processing and Exploratory Data Analysis (EDA)

- Stop words:
  - Common words that appear at high frequency across all texts, regardless of type or label
    - "the", "is", "and", "of", "in"



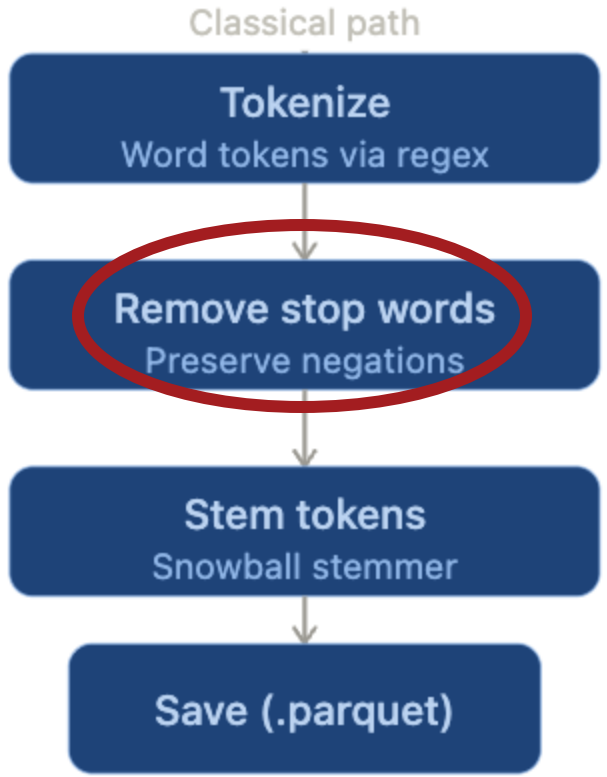
Data Processing and Exploratory Data Analysis (EDA)

- Stop words:
  - Common words that appear at high frequency across all texts, regardless of type or label
    - "the", "is", "and", "of", "in"
  - Carry little/no discriminative signal
    - Don' distinguish reliable from unreliable news



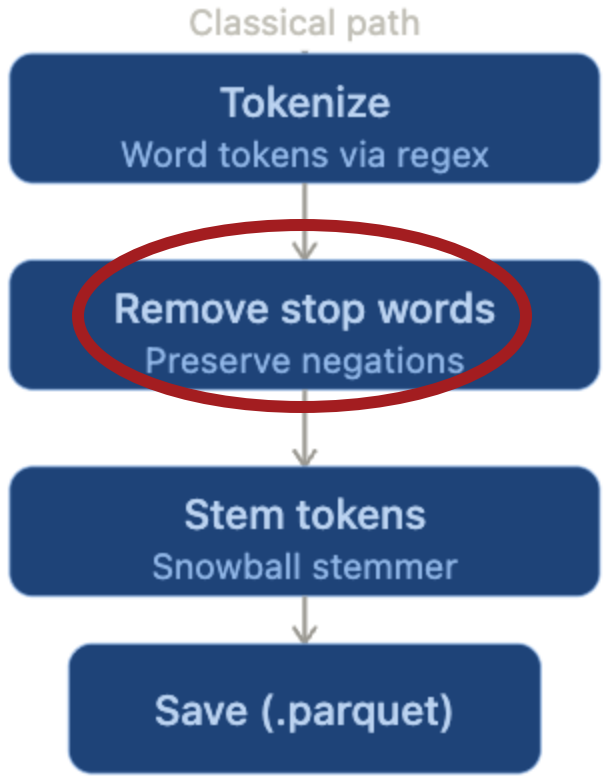
Data Processing and Exploratory Data Analysis (EDA)

```
['The', 'Kremlin', 'said', 'on', 'Friday', 'that', 'US', 'sanctions',
 'imposed', 'on', 'Russian', 'energy', 'firms', 'would', 'have', '"',
 'serious', 'consequences', '"', 'for', 'global', 'oil', 'markets', '.',
 'Spokesman', 'Dmitry', 'Peskov', 'told', 'reporters', 'the', 'move',
 'was', '"', 'counterproductive', '"', 'and', 'risked', 'destabilising',
 'supply', 'chains', 'ahead', 'of', 'winter', '.']
```



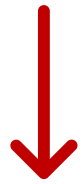
Data Processing and Exploratory Data Analysis (EDA)

```
['The', 'Kremlin', 'said', 'on', 'Friday', 'that', 'US', 'sanctions',
 'imposed', 'on', 'Russian', 'energy', 'firms', 'would', 'have', '"',
 'serious', 'consequences', '"', 'for', 'global', 'oil', 'markets', '.',
 'Spokesman', 'Dmitry', 'Peskov', 'told', 'reporters', 'the', 'move',
 'was', '"', 'counterproductive', '"', 'and', 'risked', 'destabilising',
 'supply', 'chains', 'ahead', 'of', 'winter', '.']
```

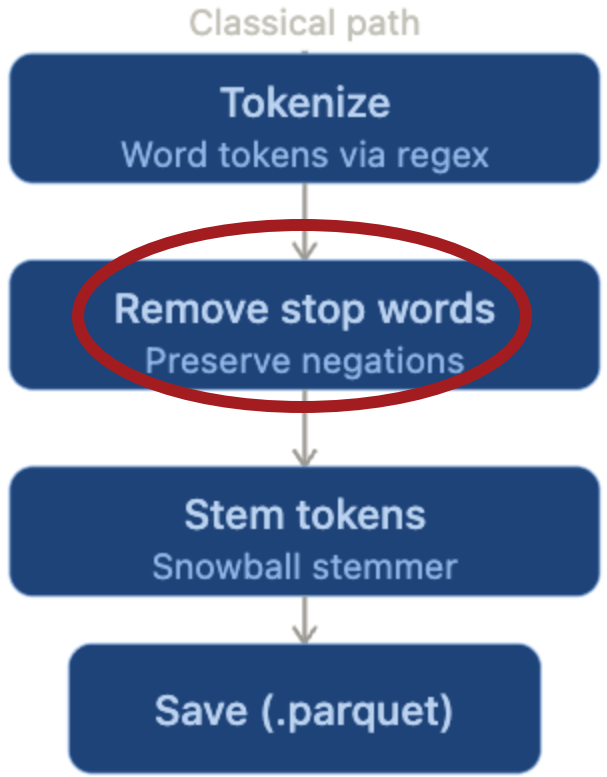


Data Processing and Exploratory Data Analysis (EDA)

```
['The', 'Kremlin', 'said', 'on', 'Friday', 'that', 'US', 'sanctions',
 'imposed', 'on', 'Russian', 'energy', 'firms', 'would', 'have', '"',
 'serious', 'consequences', '"', 'for', 'global', 'oil', 'markets', '.',
 'Spokesman', 'Dmitry', 'Peskov', 'told', 'reporters', 'the', 'move',
 'was', '"', 'counterproductive', '"', 'and', 'risked', 'destabilising',
 'supply', 'chains', 'ahead', 'of', 'winter', '.']
```

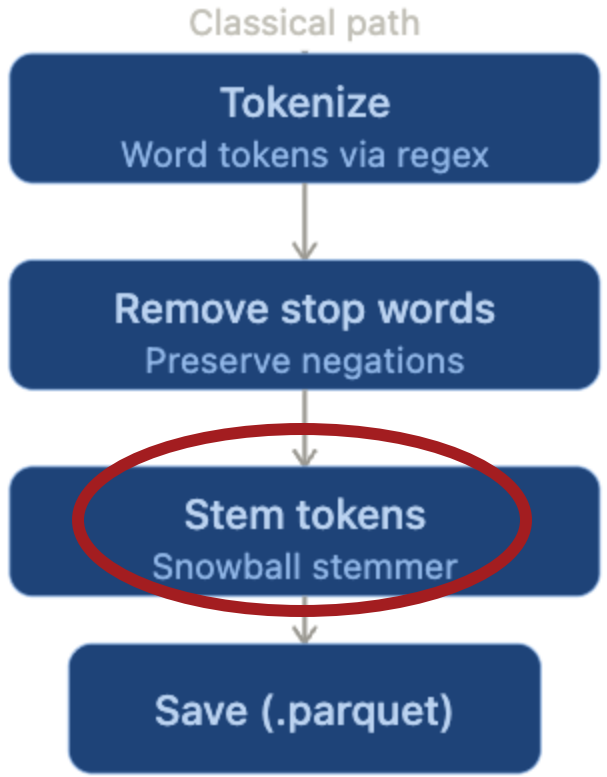


```
['Kremlin', 'said', 'Friday', 'US', 'sanctions', 'imposed', 'Russian',
 'energy', 'firms', 'would', '"', 'serious', 'consequences', '"', 'global',
 'oil', 'markets', '.', 'Spokesman', 'Dmitry', 'Peskov', 'told', 'reporters',
 'move', '"', 'counterproductive', '"', 'risked', 'destabilising', 'supply',
 'chains', 'ahead', 'winter', '.']
```



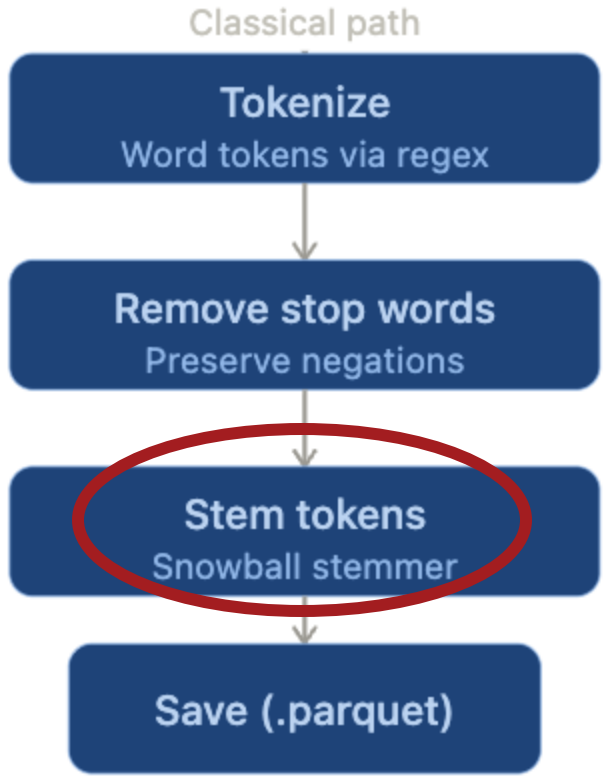
Data Processing and Exploratory Data Analysis (EDA)

- Stemming:



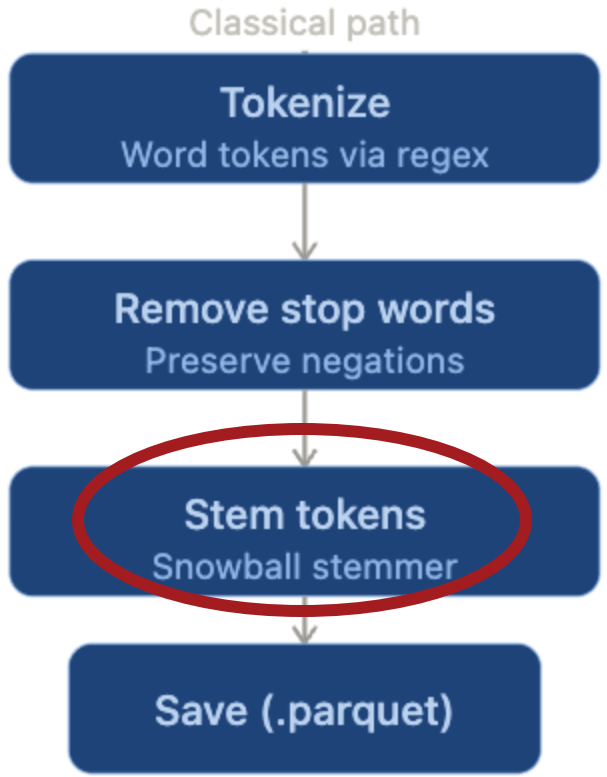
Data Processing and Exploratory Data Analysis (EDA)

- Stemming:
  - Collapses or reduces derived word forms to a common base form



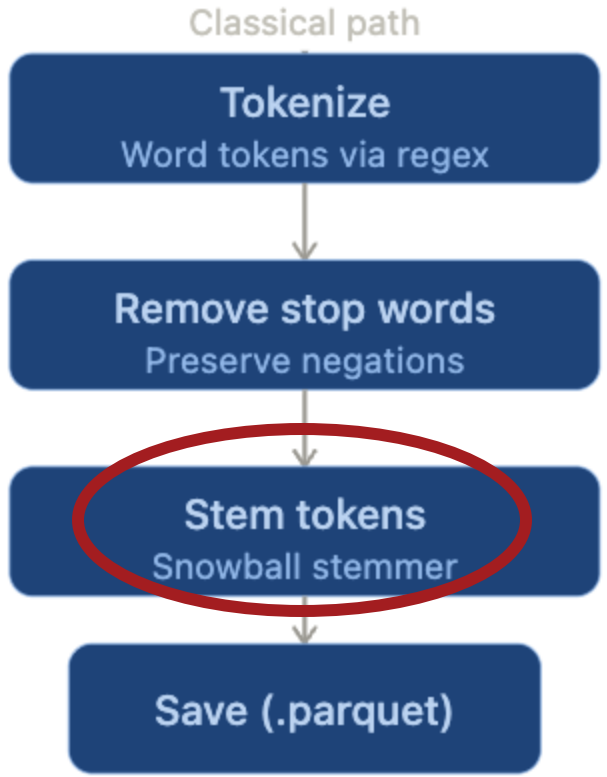
Data Processing and Exploratory Data Analysis (EDA)

- Stemming:
  - Collapses or reduces derived word forms to a common base form
    - Sanction, sanctions, sanctioned all become "sanction"



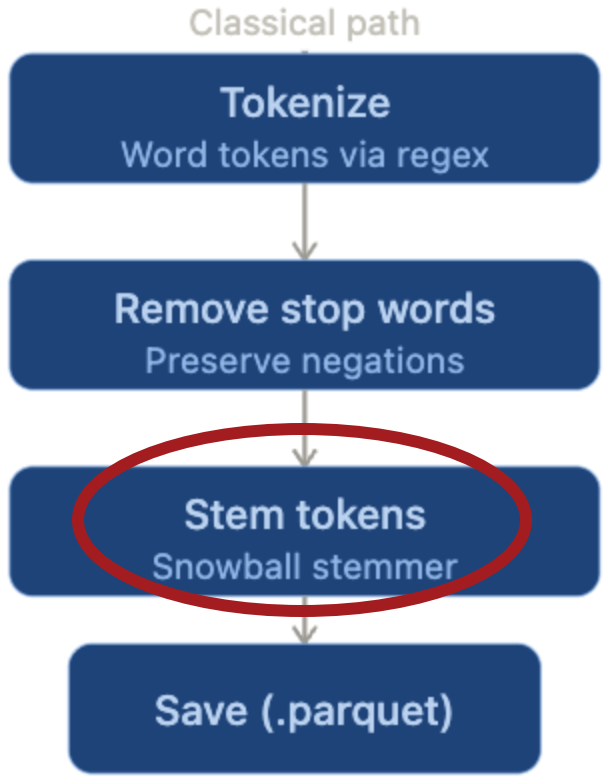
Data Processing and Exploratory Data Analysis (EDA)

- Stemming:
  - Collapses or reduces derived word forms to a common base form
    - Sanction, sanctions, sanctioned all become "sanction"
  - Helps ML models generalize across word forms



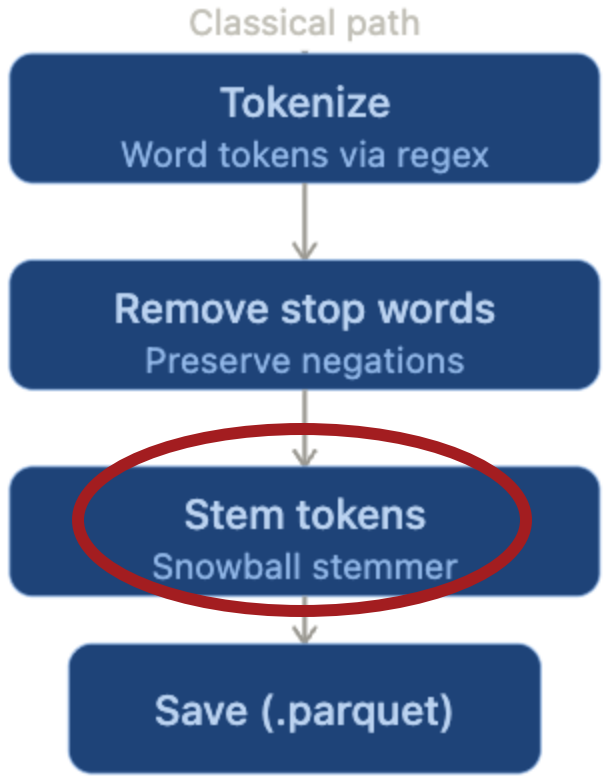
Data Processing and Exploratory Data Analysis (EDA)

- Stemming:
  - Collapses or reduces derived word forms to a common base form
    - Sanction, sanctions, sanctioned all become "sanction"
  - Helps ML models generalize across word forms
  - We want to extract as much signal as possible from each word



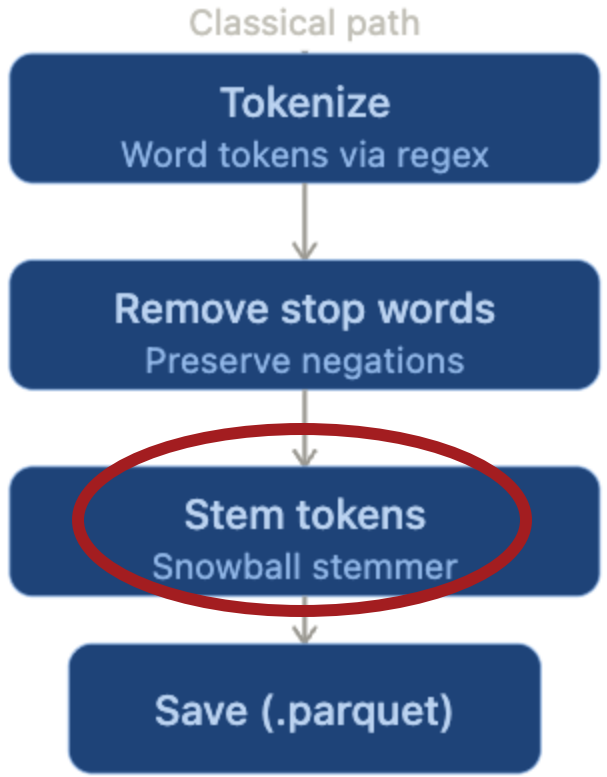
Data Processing and Exploratory Data Analysis (EDA)

- Stemming:
  - Collapses or reduces derived word forms to a common base form
    - Sanction, sanctions, sanctioned all become "sanction"
  - Helps ML models generalize across word forms
  - We want to extract as much signal as possible from each word
    - Without stemming, this signal is fragmented across variants of the same word



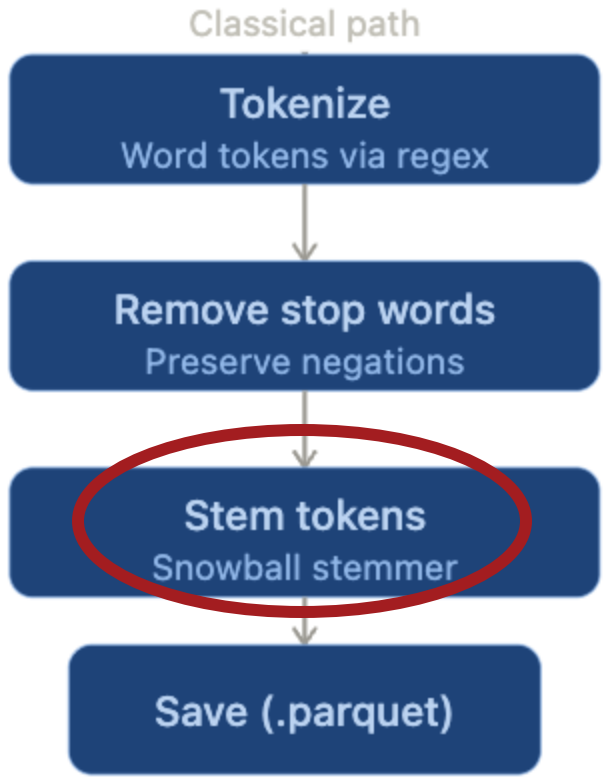
Data Processing and Exploratory Data Analysis (EDA)

- Stemming:
  - Collapses or reduces derived word forms to a common base form
    - Sanction, sanctions, sanctioned all become "sanction"
  - Helps ML models generalize across word forms
  - We want to extract as much signal as possible from each word
    - Without stemming, this signal is fragmented across variants of the same word



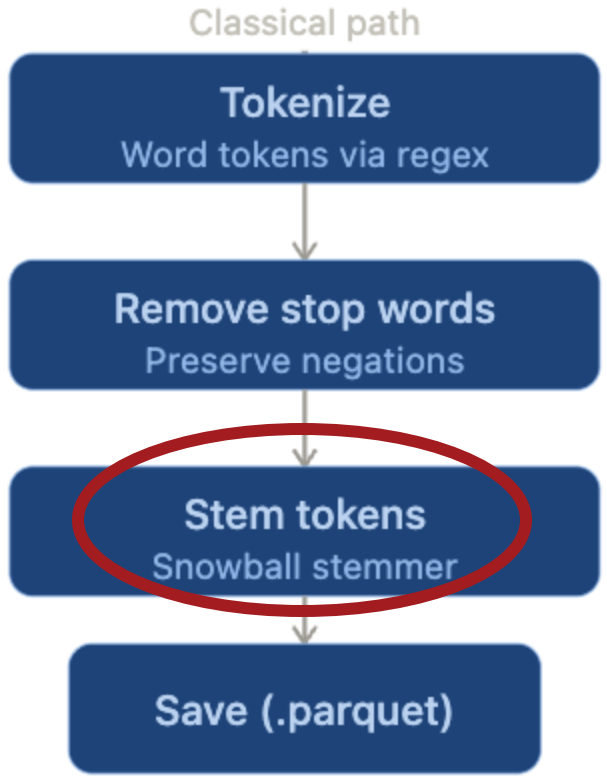
Data Processing and Exploratory Data Analysis (EDA)

```
['Kremlin', 'said', 'Friday', 'US', 'sanctions', 'imposed', 'Russian',
 'energy', 'firms', 'would', '"', 'serious', 'consequences', '"', 'global',
 'oil', 'markets', '.', 'Spokesman', 'Dmitry', 'Peskov', 'told', 'reporters',
 'move', '"', 'counterproductive', '"', 'risked', 'destabilising', 'supply',
 'chains', 'ahead', 'winter', '.']
```



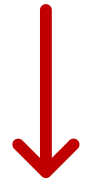
Data Processing and Exploratory Data Analysis (EDA)

```
['Kremlin', 'said', 'Friday', 'US', 'sanctions', 'imposed', 'Russian',
 'energy', 'firms', 'would', '"', 'serious', 'consequences', '"', 'global',
 'oil', 'markets', '.', 'Spokesman', 'Dmitry', 'Peskov', 'told', 'reporters',
 'move', '"', 'counterproductive', '"', 'risked', 'destabilising', 'supply',
 'chains', 'ahead', 'winter', '.']
```

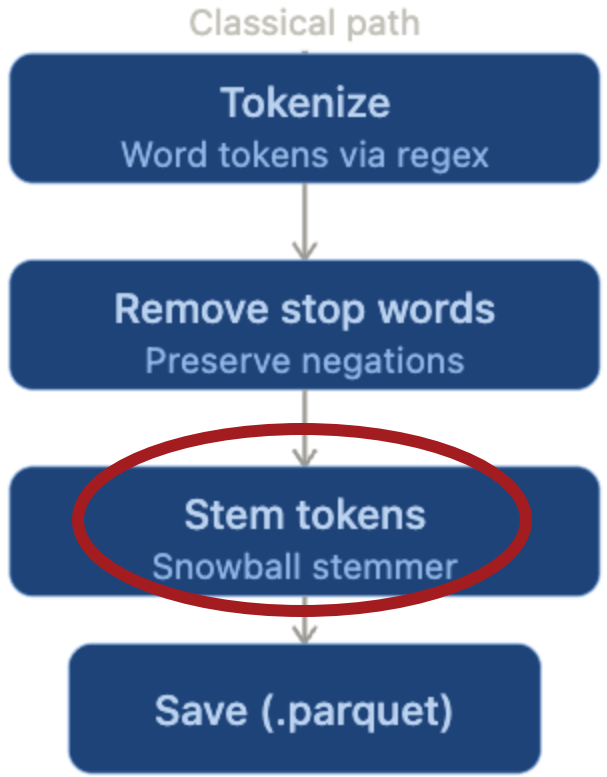


Data Processing and Exploratory Data Analysis (EDA)

```
['Kremlin', 'said', 'Friday', 'US', 'sanctions', 'imposed', 'Russian',
 'energy', 'firms', 'would', '"', 'serious', 'consequences', '"', 'global',
 'oil', 'markets', '.', 'Spokesman', 'Dmitry', 'Peskov', 'told', 'reporters',
 'move', '"', 'counterproductive', '"', 'risked', 'destabilising', 'supply',
 'chains', 'ahead', 'winter', '.']
```

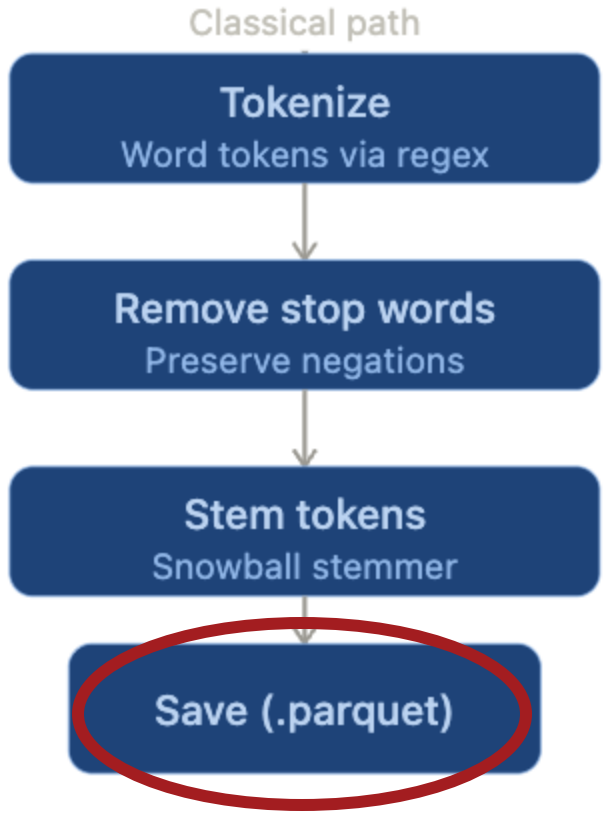


```
['kremlin', 'said', 'friday', 'us', 'sanction', 'impos', 'russian',
 'energi', 'firm', 'would', 'seriou', 'consequ', 'global', 'oil',
 'market', 'spokesman', 'dmitri', 'peskov', 'told', 'report', 'move',
 'counterproduc', 'risk', 'destabilis', 'suppli', 'chain', 'ahead', 'winter']
```



Data Processing and Exploratory Data Analysis (EDA)

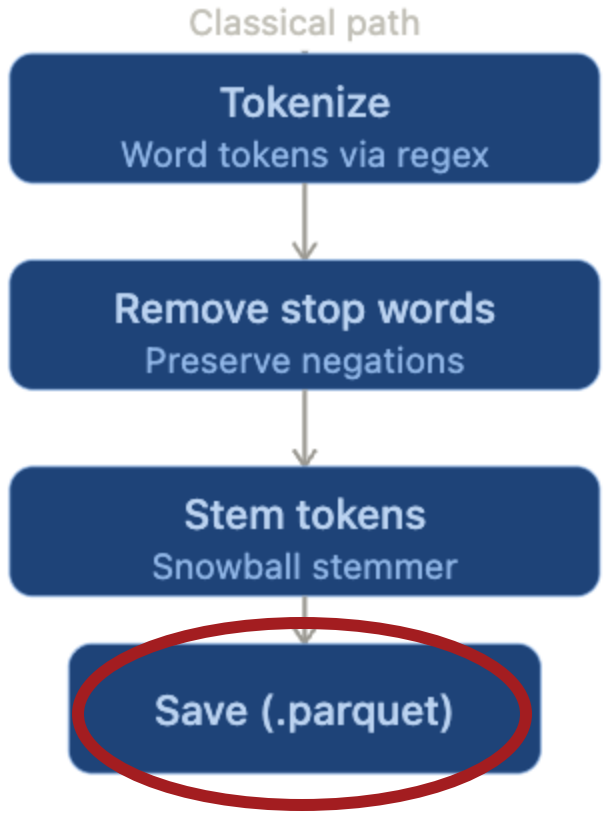
- Processed data is then saved and ready to be split for training, testing and evaluation



Data Processing and Exploratory Data Analysis (EDA)

- Processed data is then saved and ready to be split for training, testing and evaluation

- Before / After processing example:



Data Processing and Exploratory Data Analysis (EDA)

BEFORE

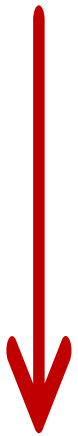
Sputnik News I World

The Kremlin said on Friday that US sanctions imposed on Russian energy firms would have "serious consequences" for global oil markets. Read more at https://sputniknews.com/world/sanctions-full

Spokesman Dmitry Peskov told reporters the move was "counterproductive" and risked destabilising supply chains ahead of winter.

Data Processing and Exploratory Data Analysis (EDA)

BEFORE



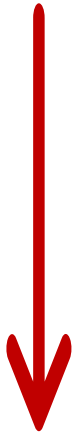
Sputnik News I World

The Kremlin said on Friday that US sanctions imposed on Russian energy firms would have "serious consequences" for global oil markets. Read more at https://sputniknews.com/world/sanctions-full

Spokesman Dmitry Peskov told reporters the move was "counterproductive" and risked destabilising supply chains ahead of winter.

Data Processing and Exploratory Data Analysis (EDA)

BEFORE



Sputnik News | World

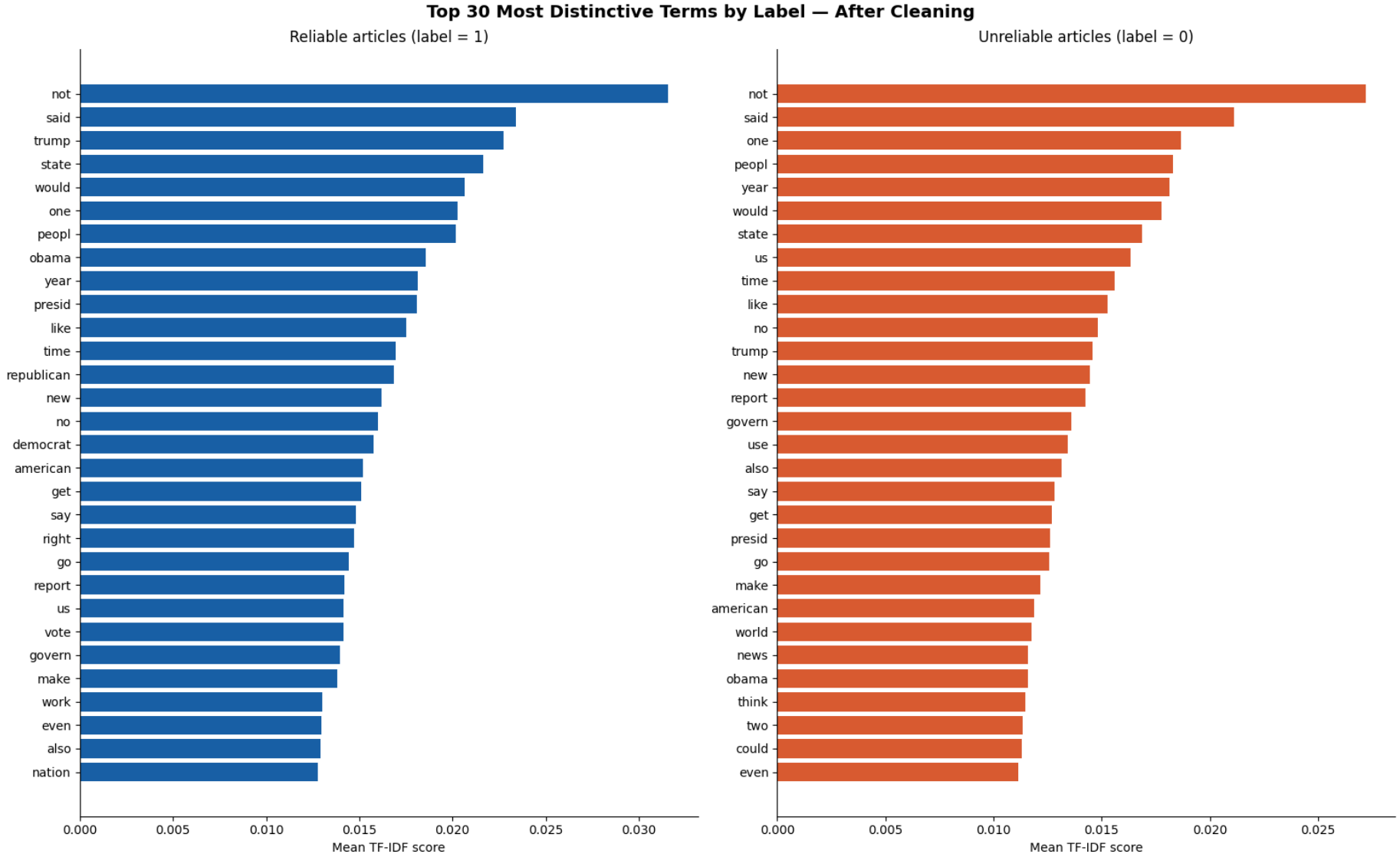
The Kremlin said on Friday that US sanctions imposed on Russian energy firms would have "serious consequences" for global oil markets. Read more at https://sputniknews.com/world/sanctions-full

Spokesman Dmitry Peskov told reporters the move was "counterproductive" and risked destabilising supply chains ahead of winter.

After

```
['kremlin', 'said', 'friday', 'us', 'sanction', 'impos', 'russian',
 'energi', 'firm', 'would', 'seriou', 'consequ', 'global', 'oil',
 'market', 'spokesman', 'dmitri', 'peskov', 'told', 'report', 'move',
 'counterproduc', 'risk', 'destabilis', 'suppli', 'chain', 'ahead', 'winter']
```

Data Processing and Exploratory Data Analysis (EDA)



Data Processing and Exploratory Data Analysis (EDA)

Data Processing and Exploratory Data Analysis (EDA)

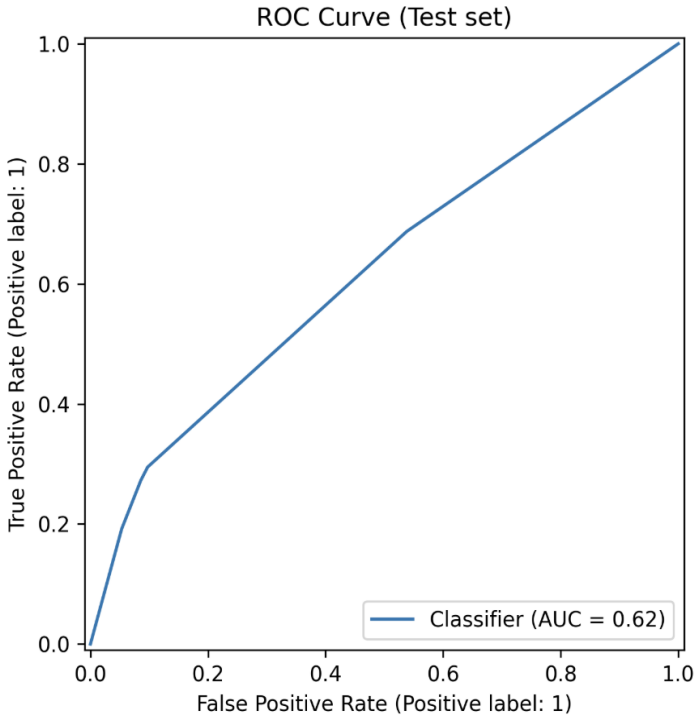
- Why process data?
  - TF-IDF vectorization
    - Term Frequency-Inverse Document Frequency
    - $\mathrm{TF}(t, d) = \frac{\text{N times term t appears in document d}}{\text{total numer of terms in document d}}$
    - $\mathrm{IDF}(t, D) = \log \frac{\text{total number of documents in corpus D}}{\text{number of documents containing term t}}$
    - Effective term ranking

XGBoost Model

- XGBoost:
  - Trained on balanced 50/50 data
  - Data set size of 60k articles
  - Train only on title + content
  - 70/15/15 split (Train, Validation, Test)
  - TF-IDF vectorizer on training data
  - Hyper Parameter Optimization

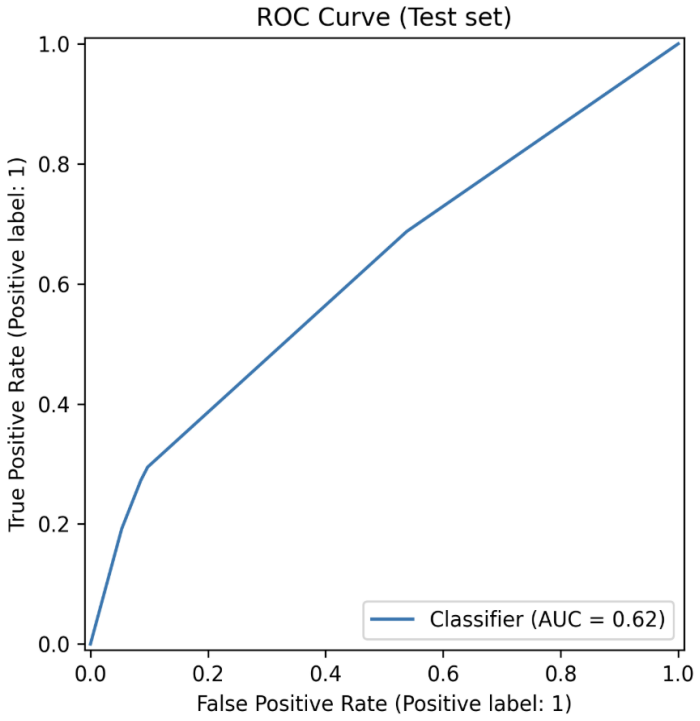
Results & Comparisons

- XGBoost
  - AUC development



Results & Comparisons

- XGBoost
  - AUC development

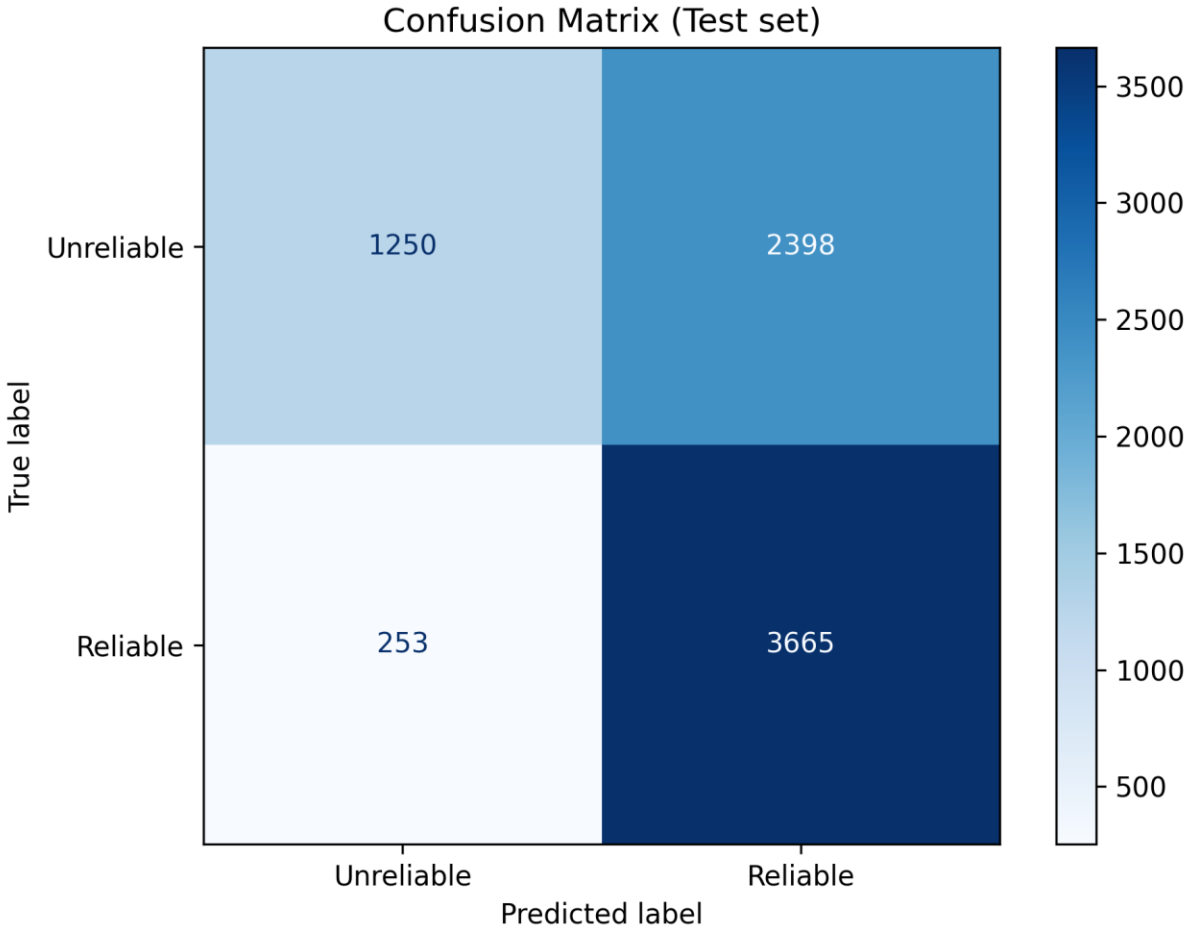


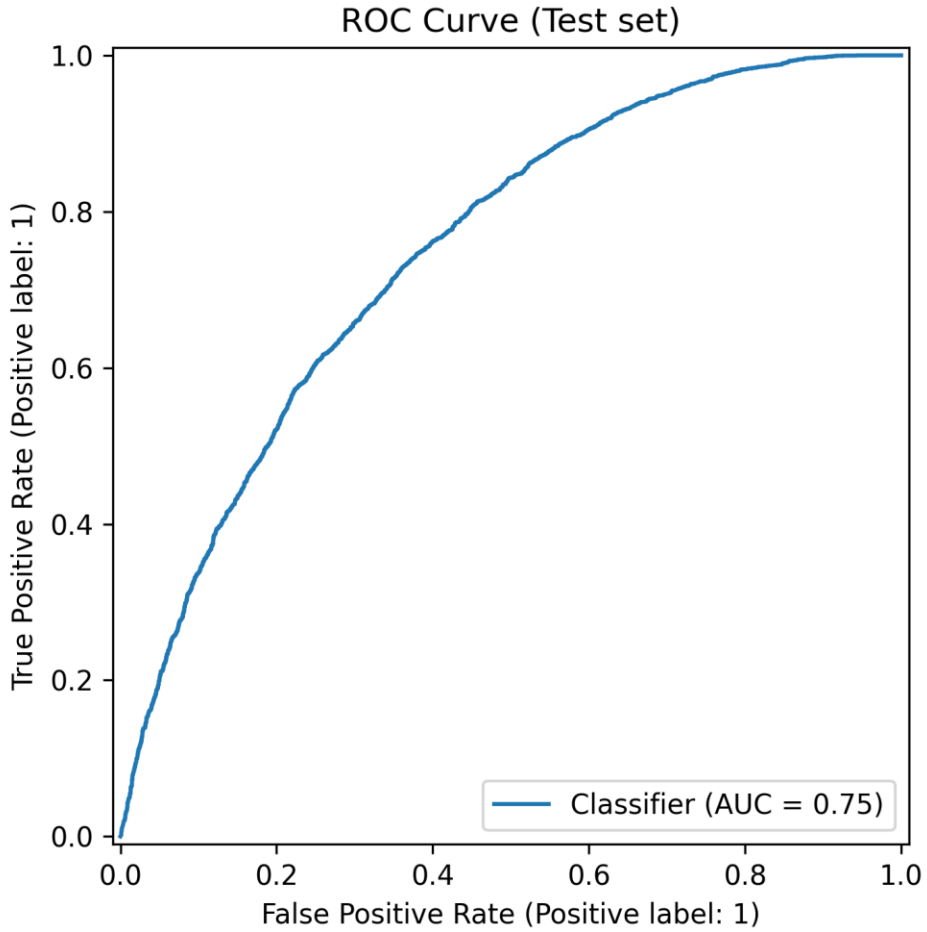
Trials	Baseline	10	50	100
Test set	0.6221	0.7484	0.7356	0.7502

AUC as function of Hyperparameter trials

Results & Comparisons

- XGBoost – Test set





Results & Comparisons

- XGBoost
- Challenges
  - Data set size limited by amount of RAM because of Vectorizer
  - Both Vectorizer and XGBoost model relying on the quality of data

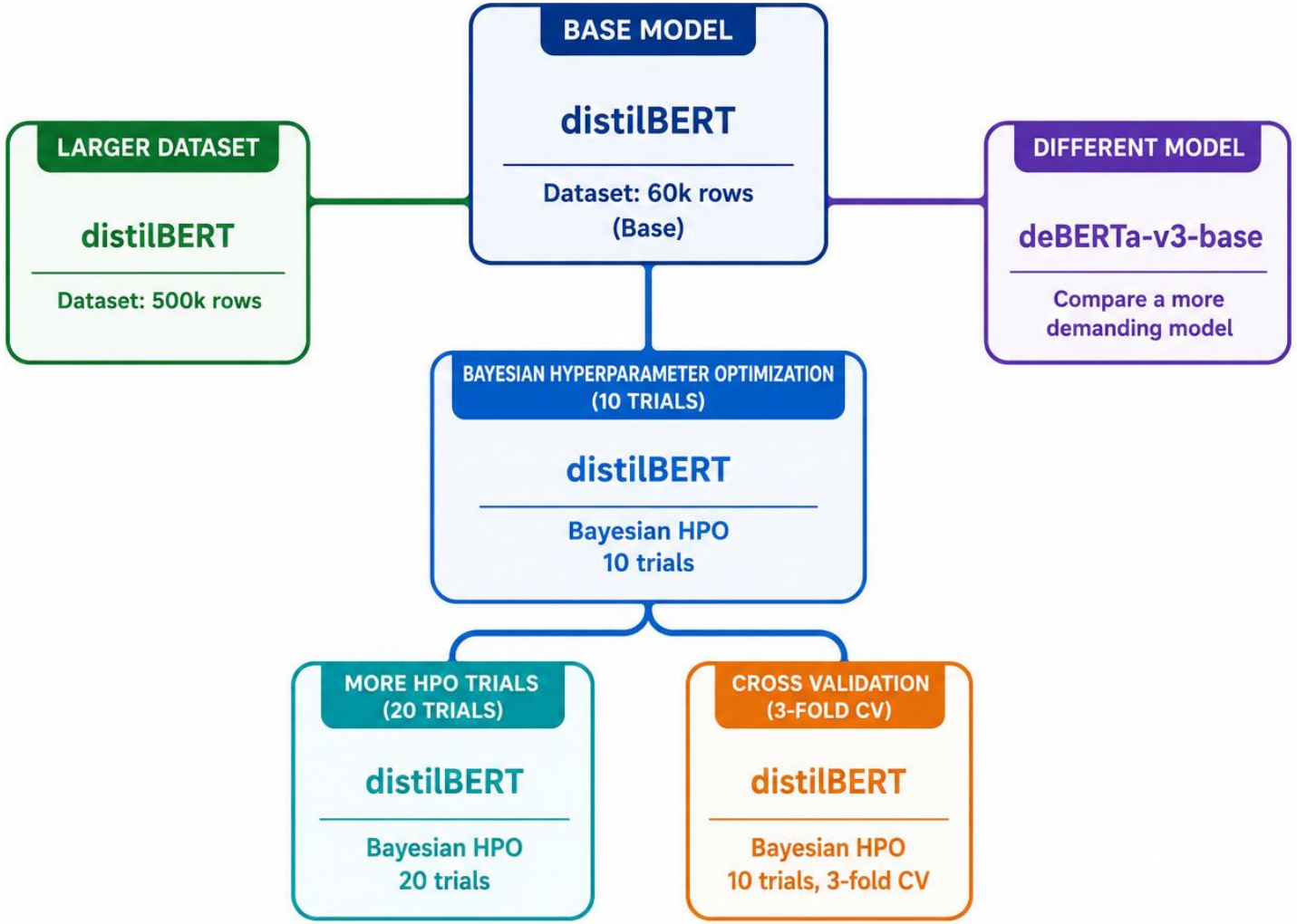
- What we would have done differently
  - Explore options for bigger data sets
  - Try with non-binary models

Results & Comparisons – DistilBERT and DeBERTa

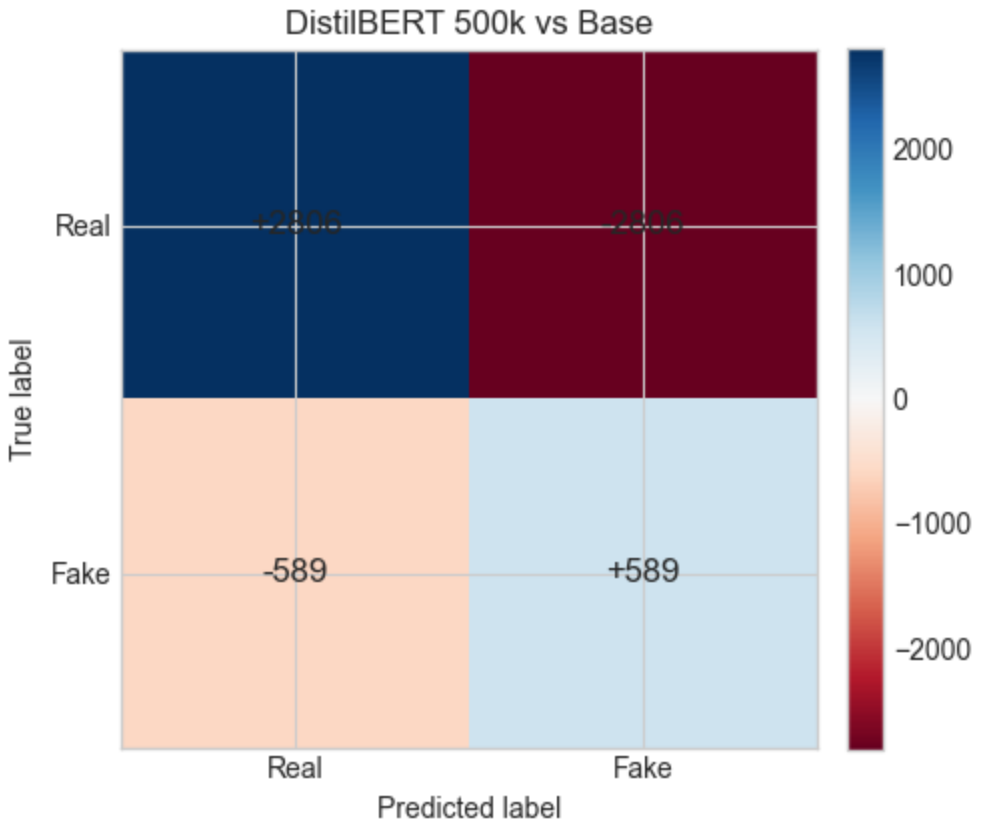
- Transformer NN pretrained
- Contextual meaning, not count words
- Tokenizer from models
- DestilBERT vs BERT
  - Less demanding still effective
  - student teacher knowledge distillation
    - Trained ground-truth labels and BERT guesses
- Tests
  - More data
  - Better Optimization
  - Stronger architecture

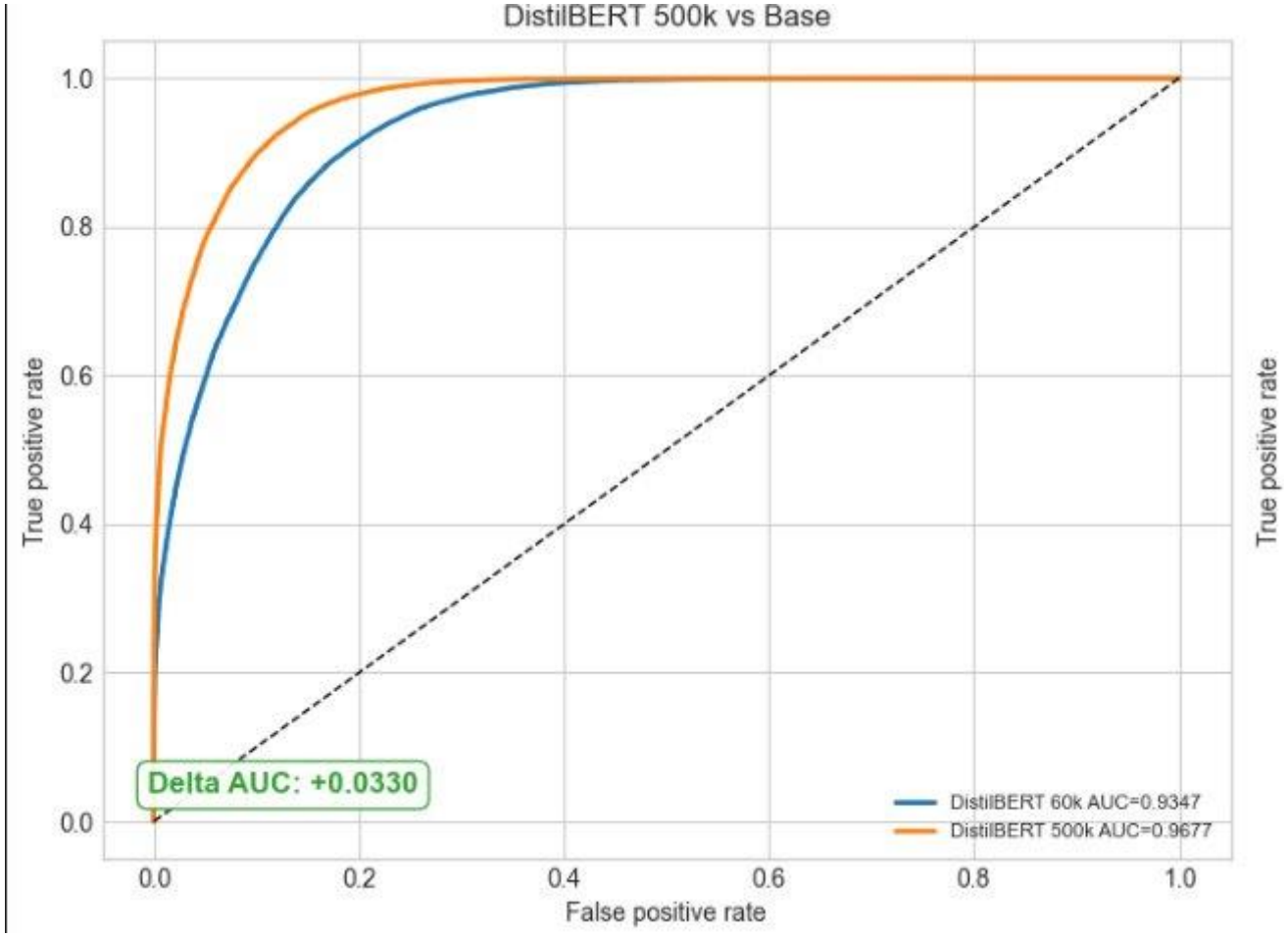
Results & Comparisons – Base

- 60k base
- 60/20/20 split
- Leak, 500k split test
- 60k still has "The"
  - 500k as well
  - DeBERTa doesn't

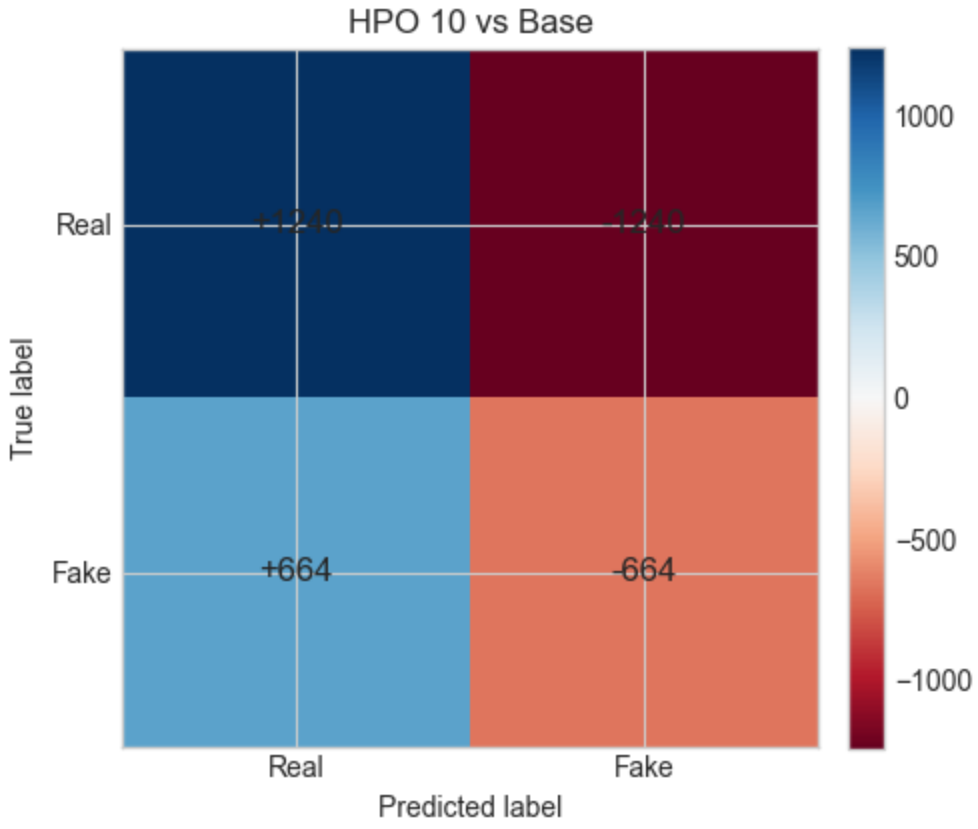


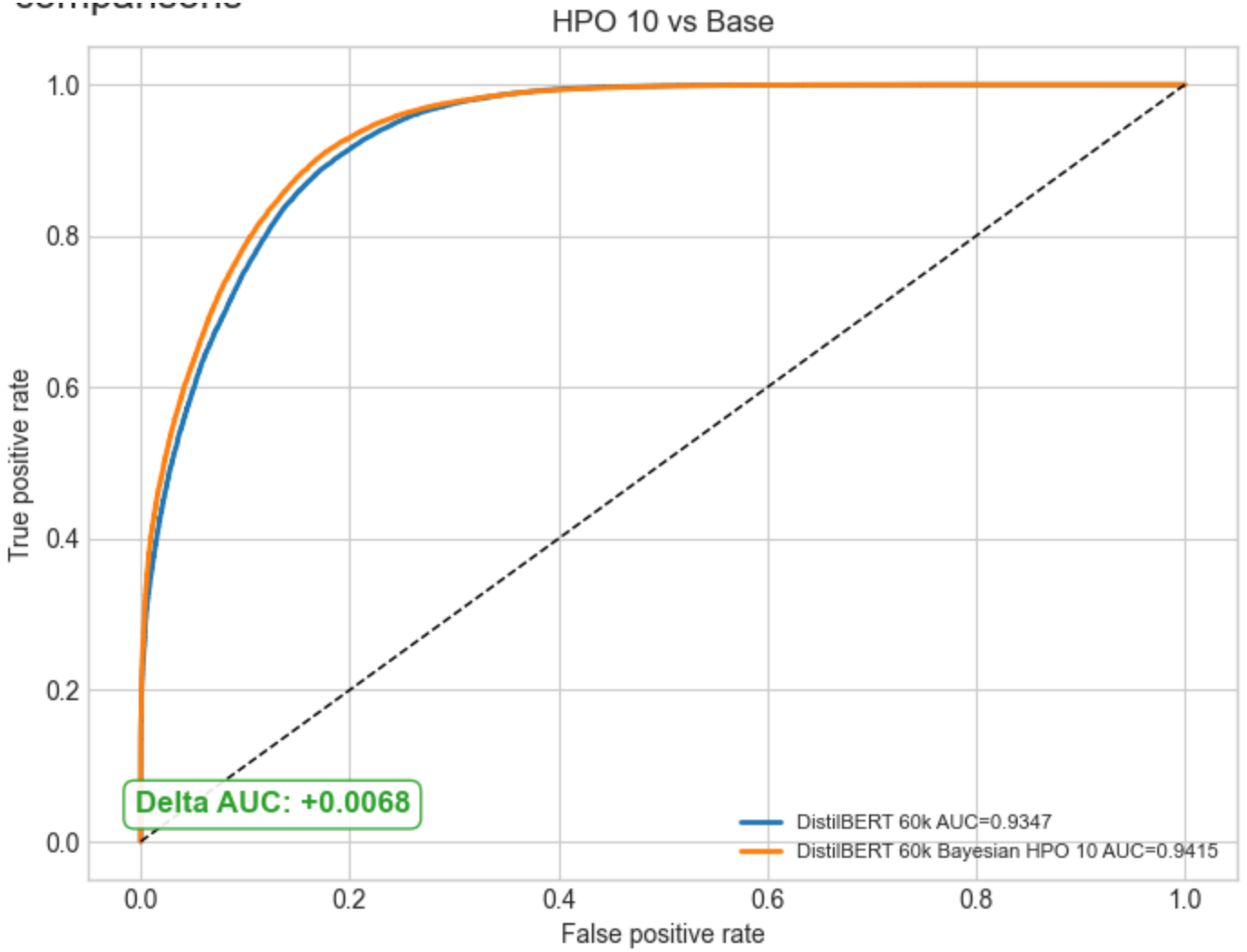
Results & Comparisons – 500k vs base



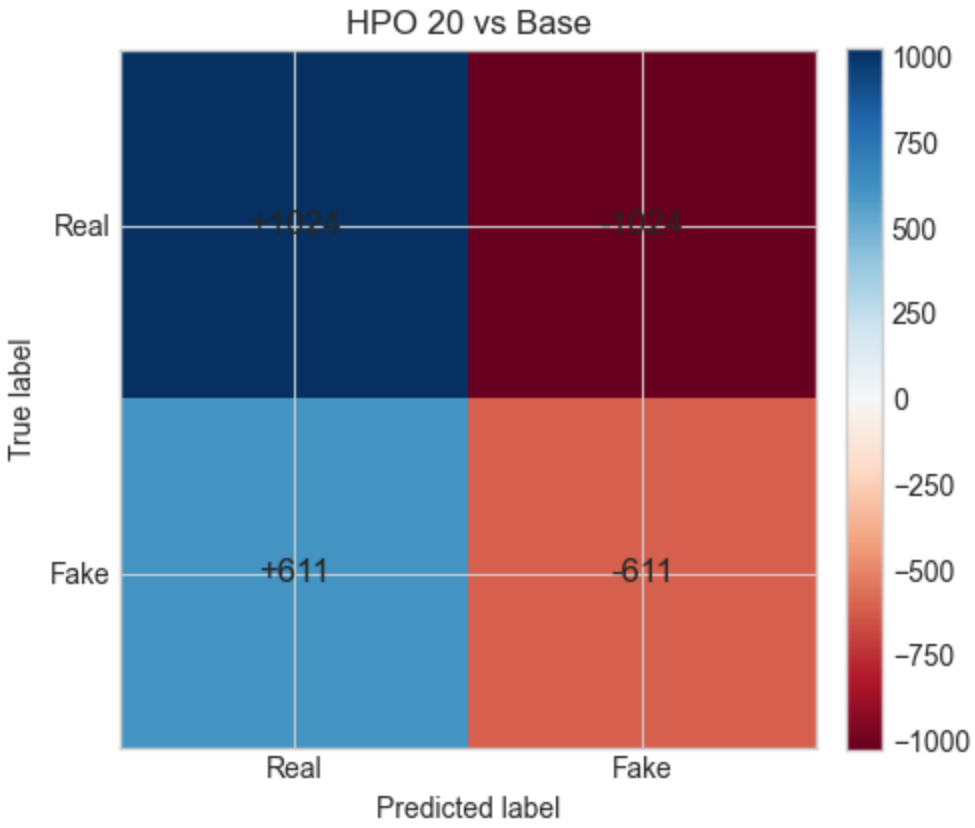


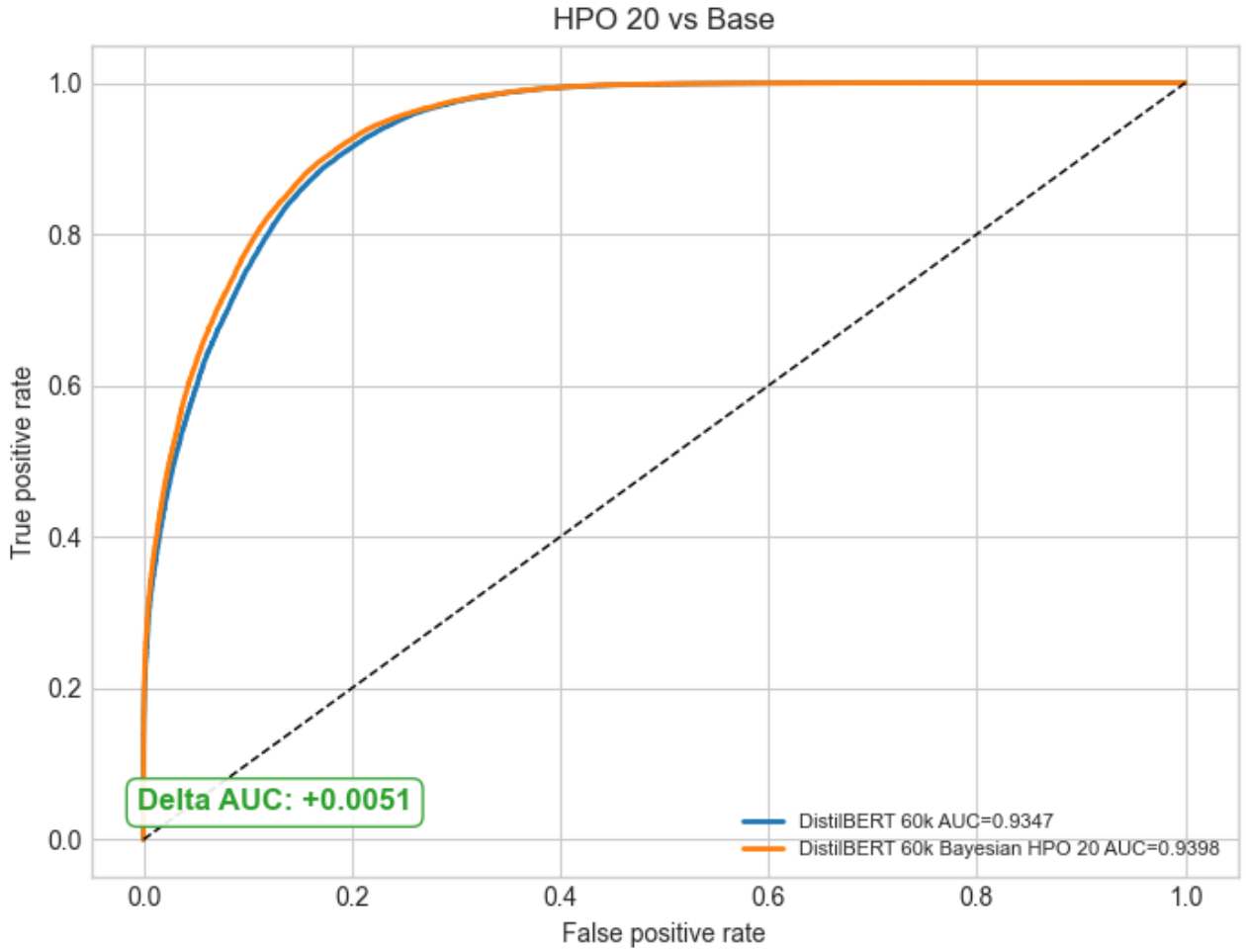
Results & Comparisons – 10 HPO vs base



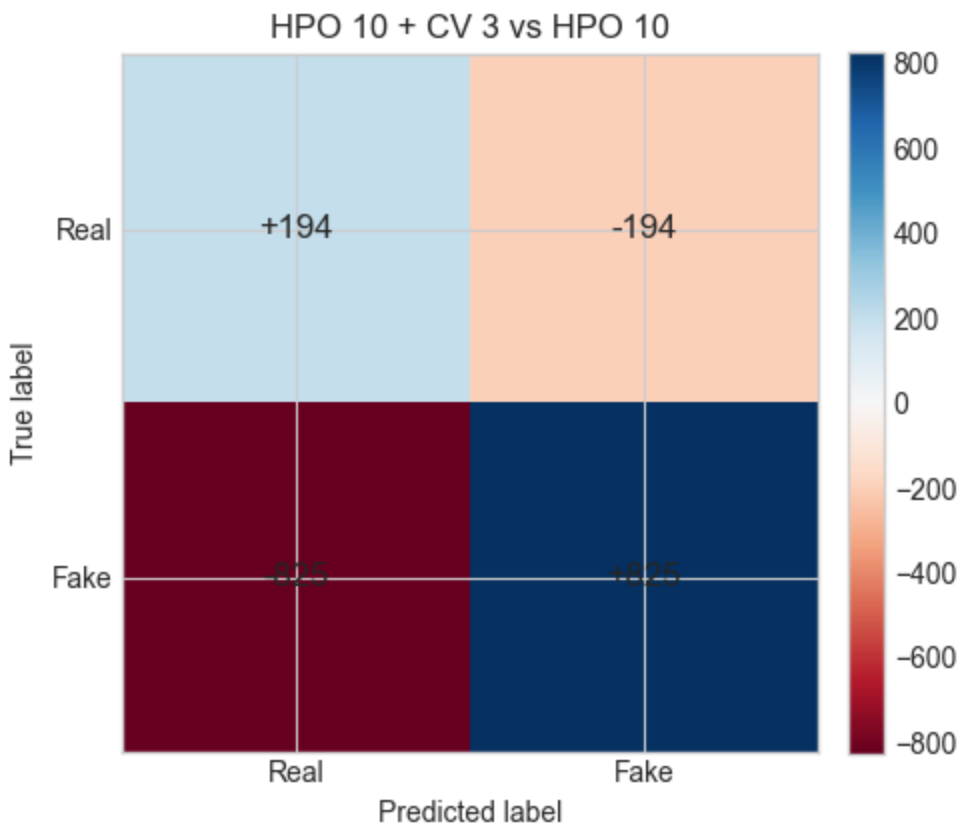


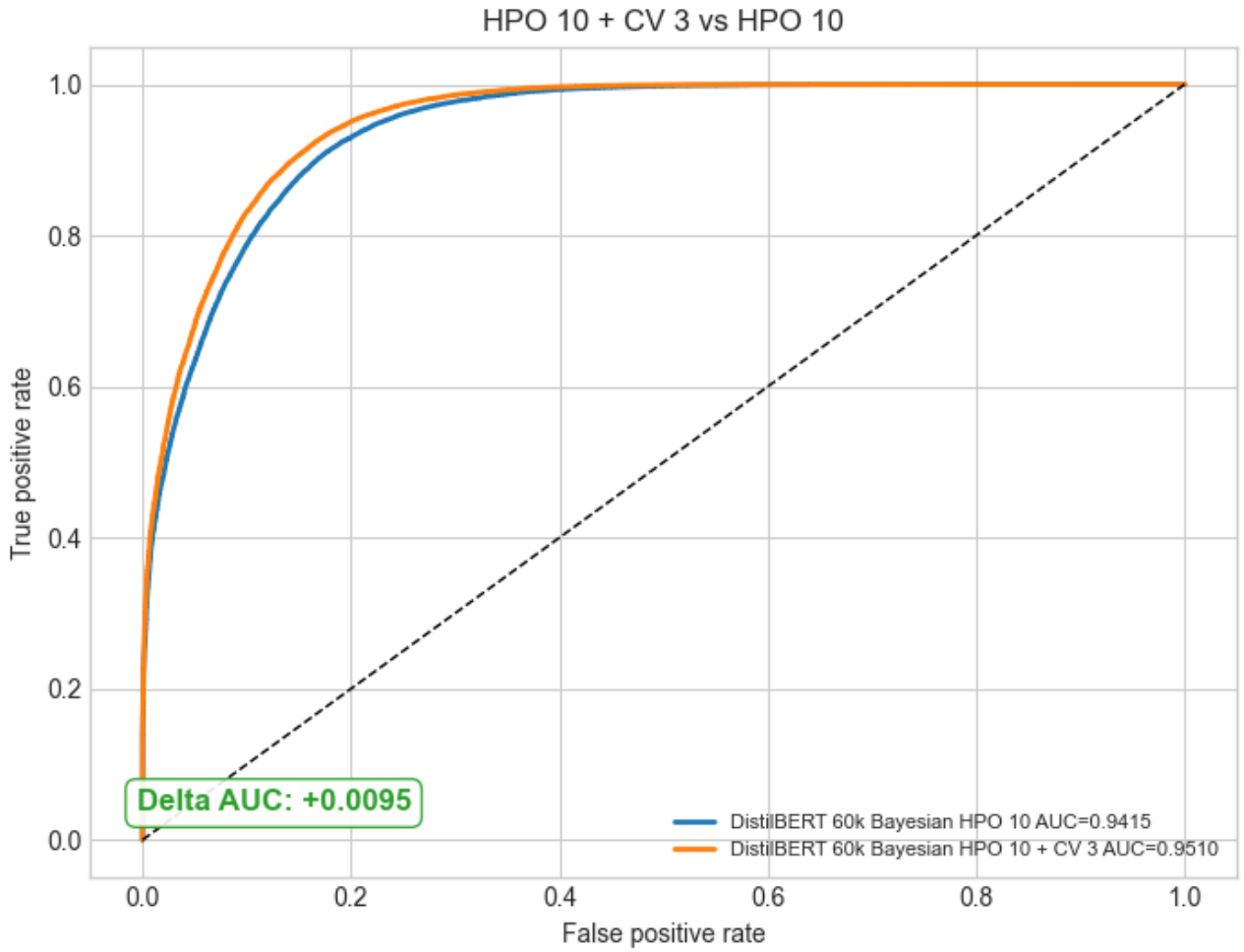
Results & Comparisons – 20 HPO vs base



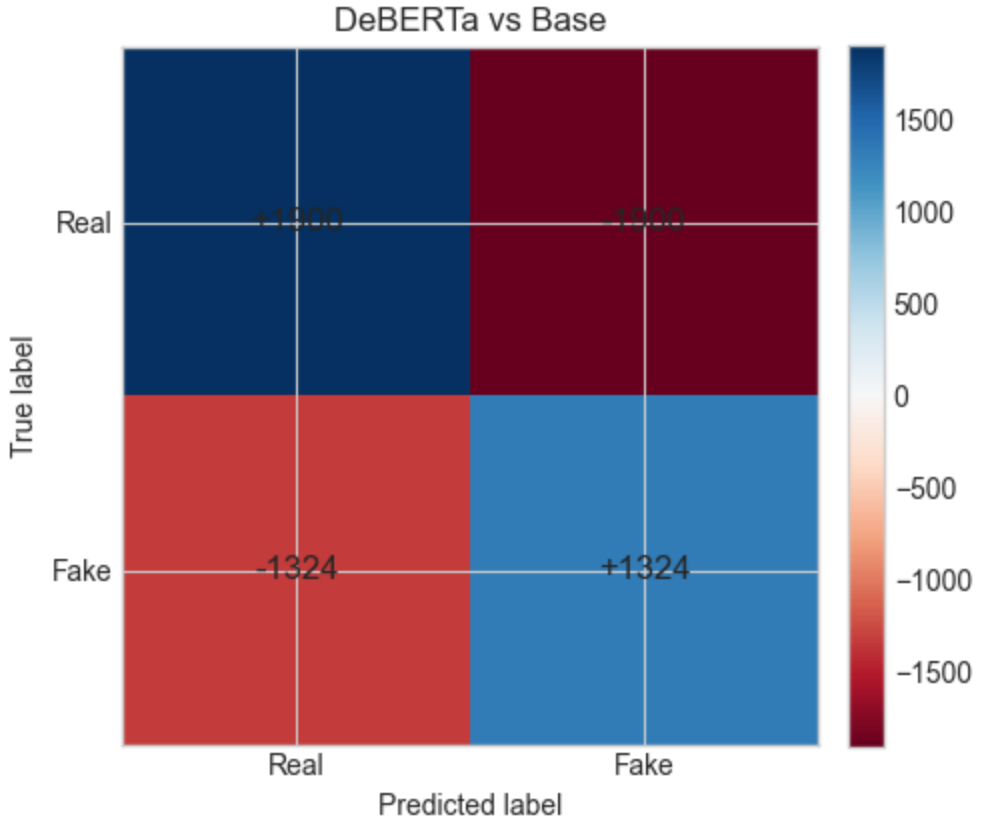


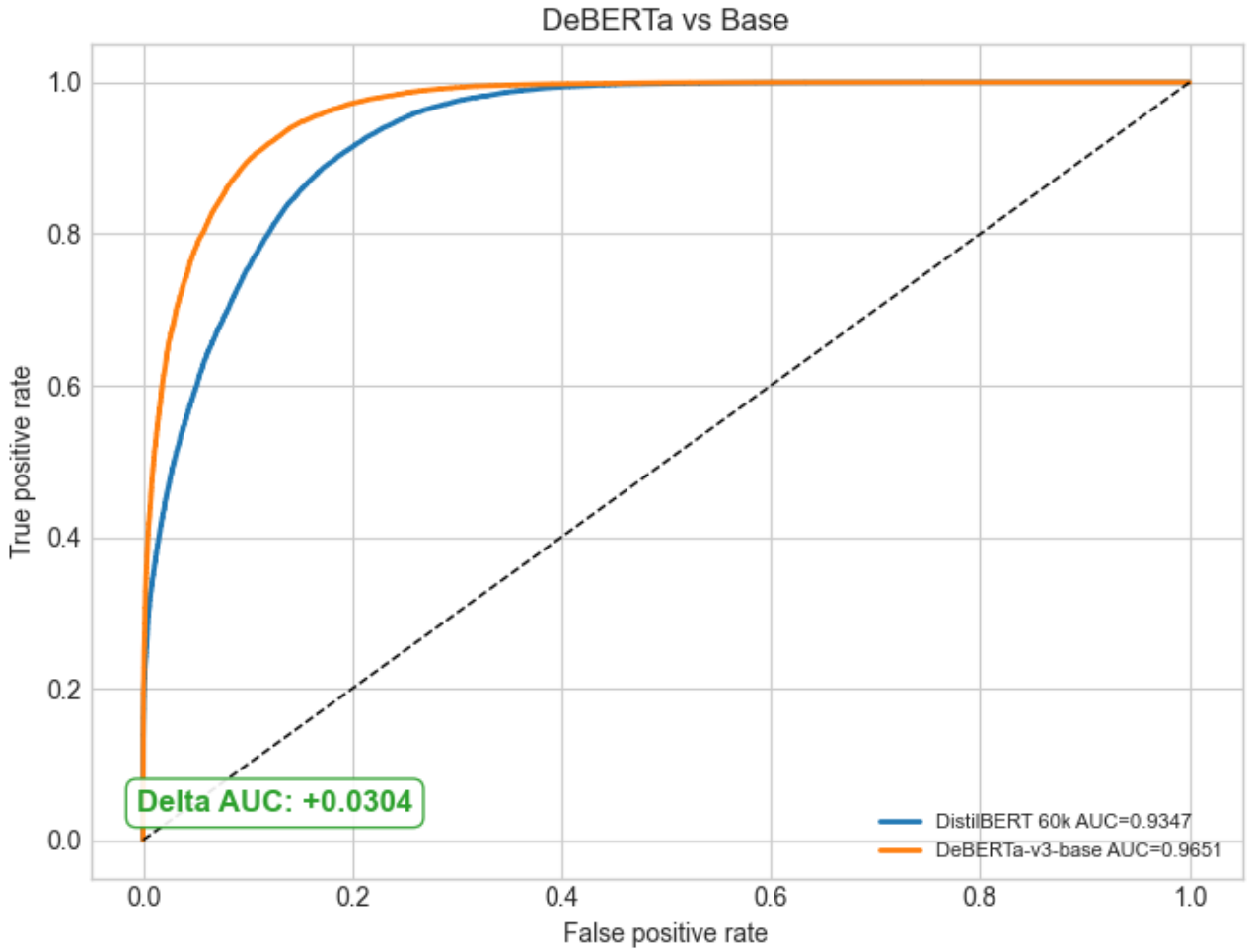
Results & Comparisons – 10 HPO w/ CV vs 10 HPO



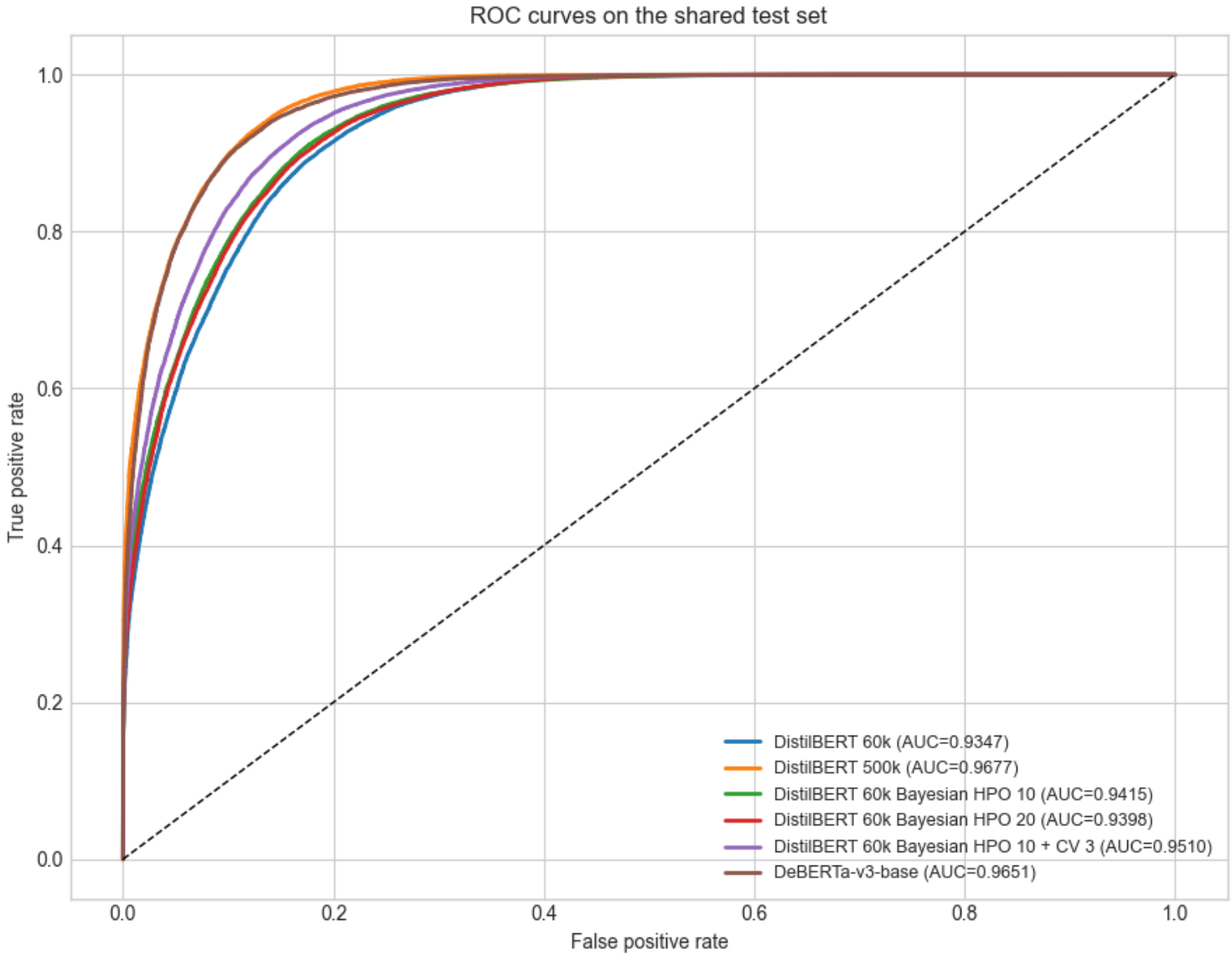


Results & Comparisons – DeBERTa vs base

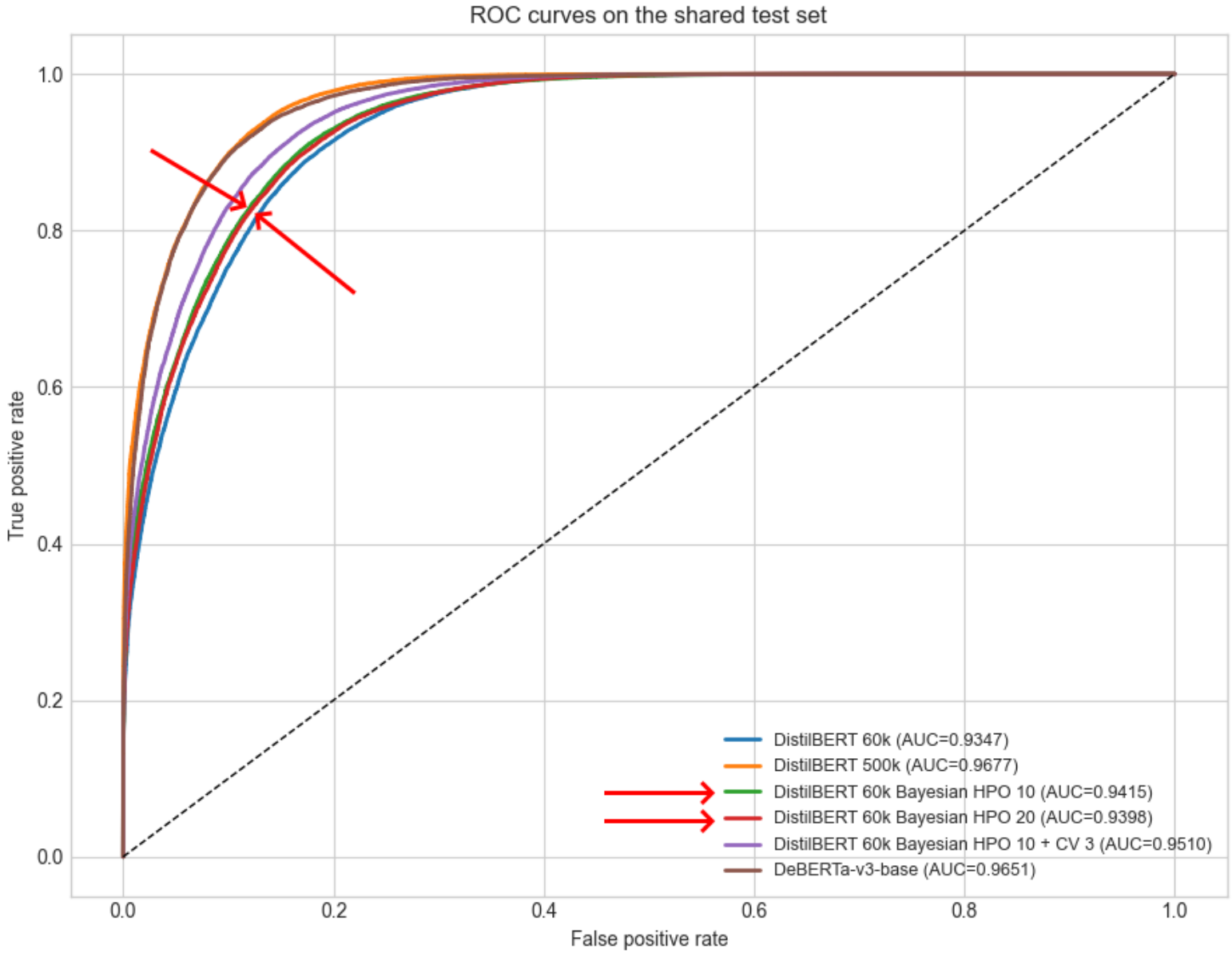




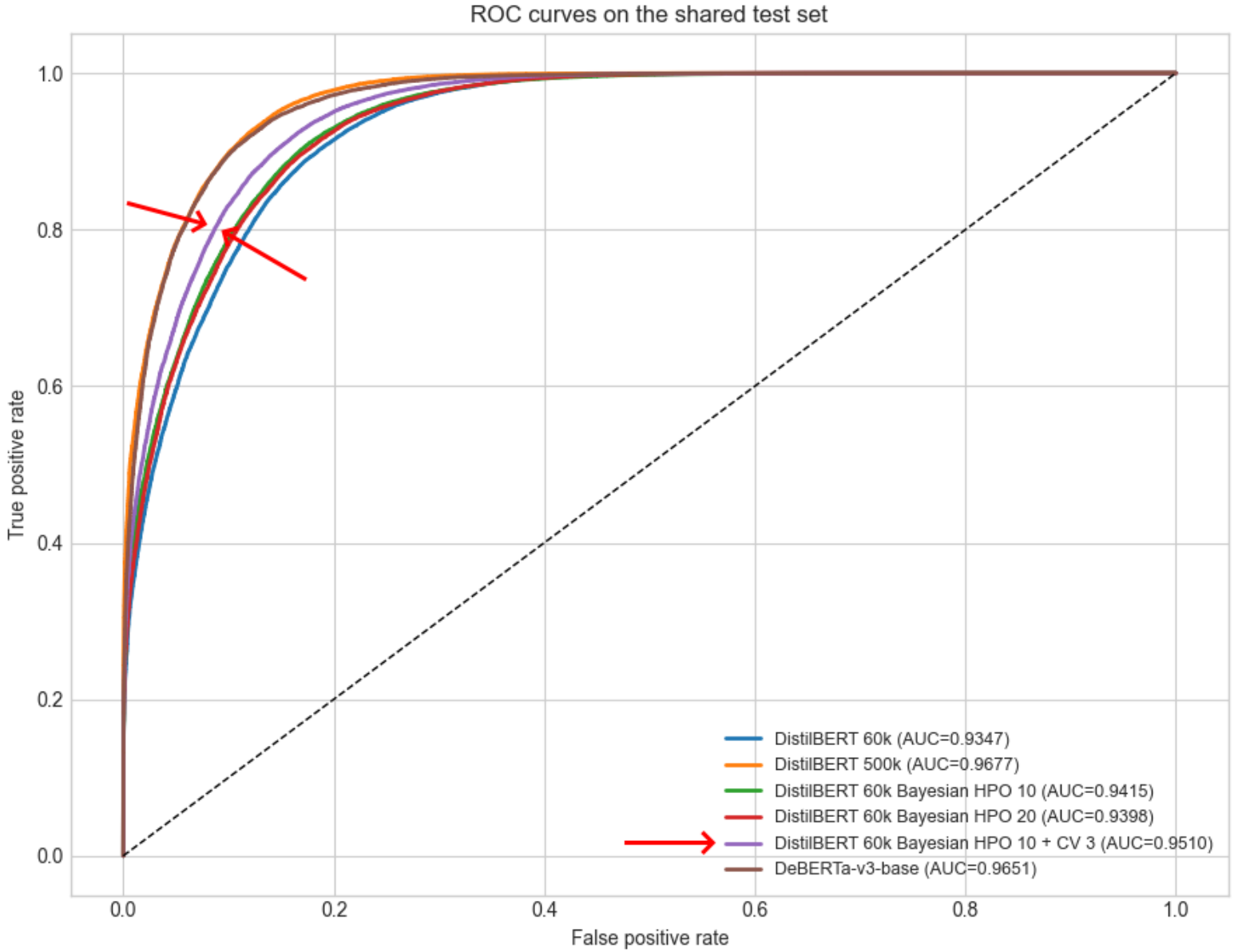
Results & Comparisons – Total



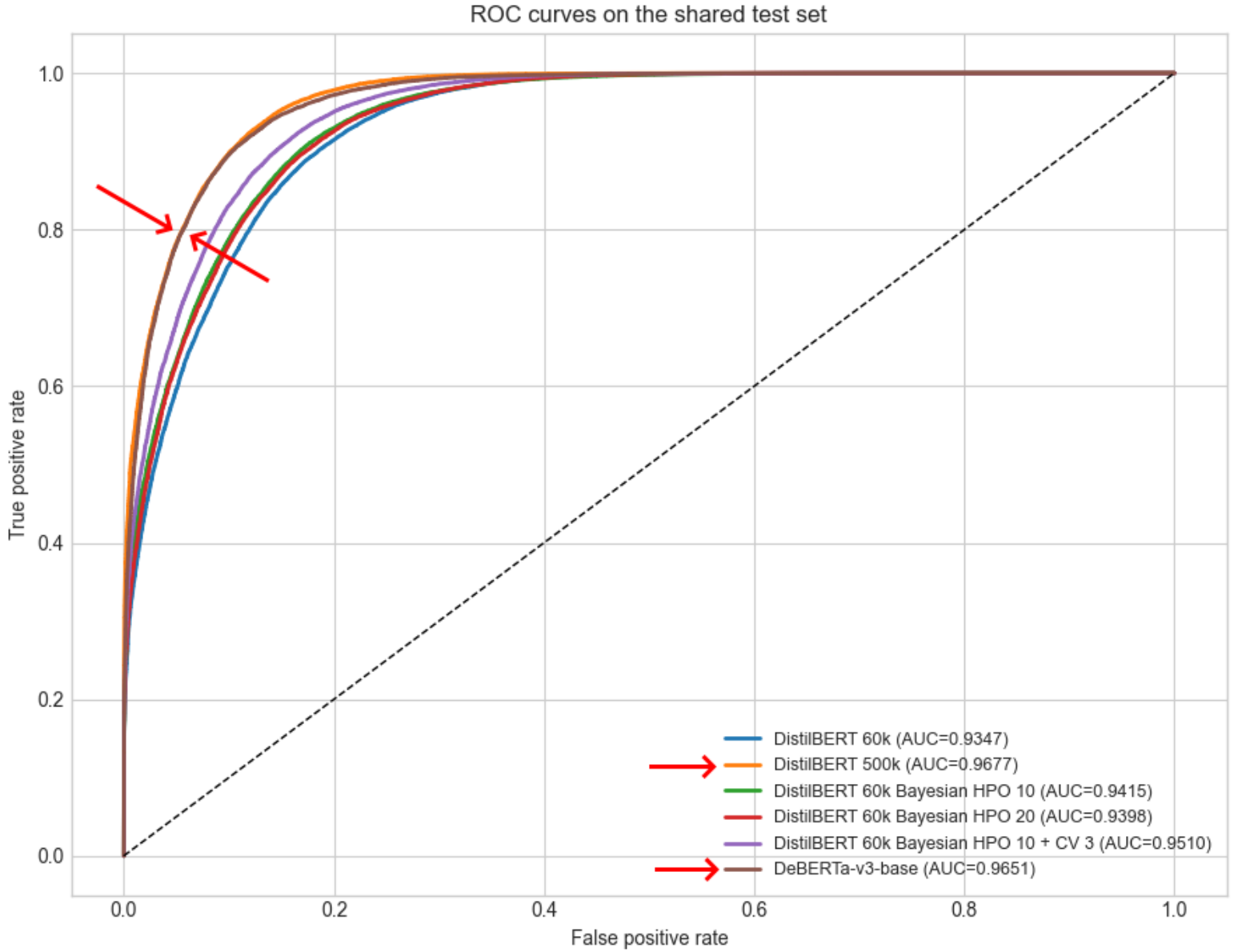
Results & Comparisons – Total



Results & Comparisons – Total



Results & Comparisons – Total

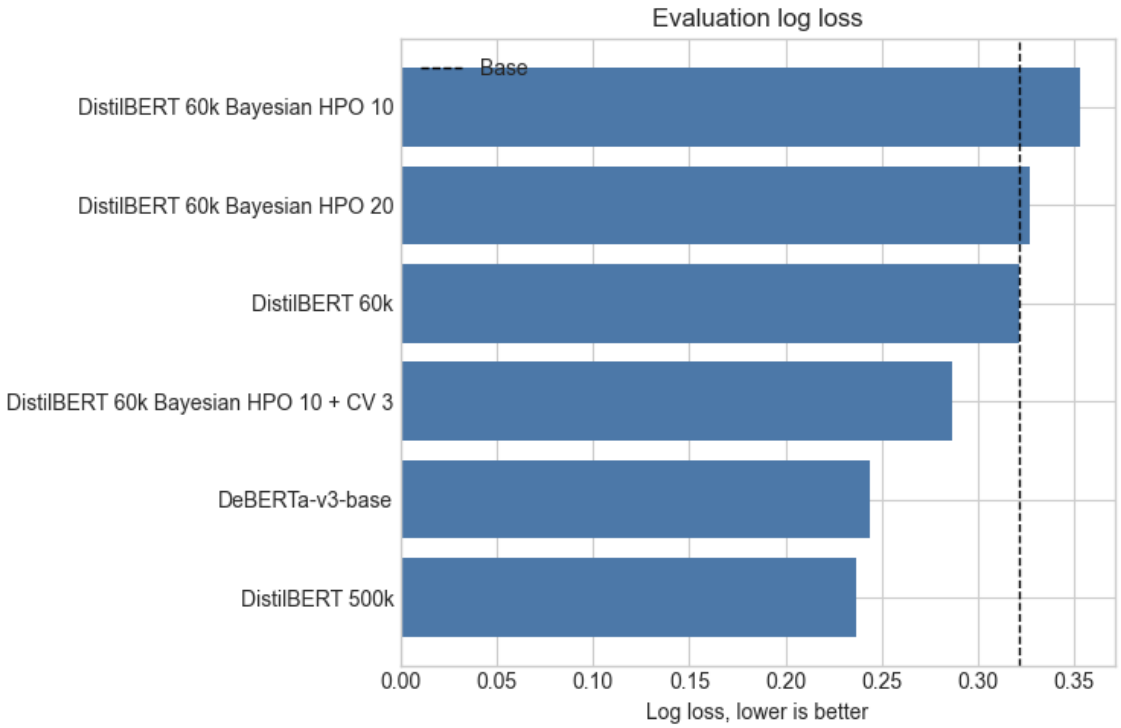


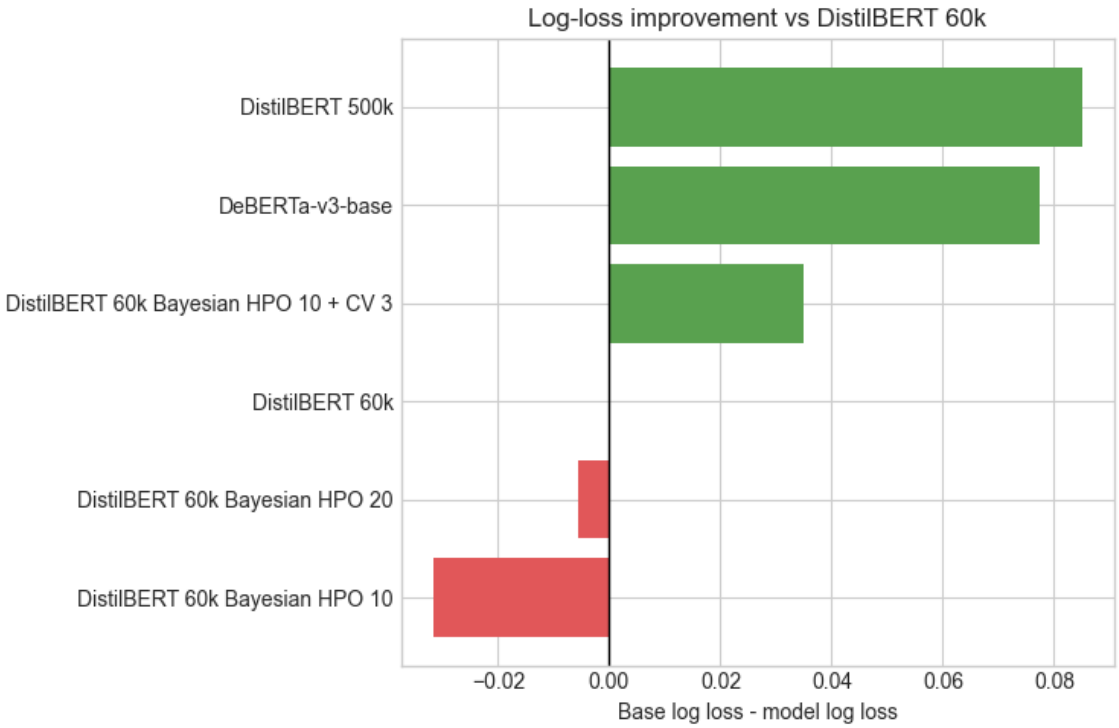
Results & Comparisons – Total

- LogLoss higher HPO, highest w/ HPO 10
- CV solves, but CV use "new" dataset only HPO use "old" dataset
- Counterpoint – Base use "old" dataset

Model	Accuracy	Delta_Accuracy	ROC_AUC	Delta_ROC_AUC	LogLoss	Delta_LogLoss
DistilBERT 500k	0.9029	0.0434	0.9677	0.0330	0.2368	0.0850
DeBERTa-v3-base	0.9007	0.0412	0.9651	0.0304	0.2443	0.0775
DistilBERT 60k Bayesian HPO 10 + CV 3	0.8799	0.0204	0.9510	0.0163	0.2869	0.0349
DistilBERT 60k Bayesian HPO 10	0.8669	0.0074	0.9415	0.0068	0.3534	-0.0316
DistilBERT 60k Bayesian HPO 20	0.8648	0.0053	0.9398	0.0051	0.3275	-0.0057
DistilBERT 60k	0.8595	0.0000	0.9347	0.0000	0.3218	0.0000

Results & Comparisons – Total





Results & Comparisons – destilBERT

- Challenges
  - Training time, google collab greedy with free GPU:'(
    - Consequences like small max length of tokens (128)
- What we would have done differently
  - Train everything on good processed data
  - Use all of dataset
  - Save many models while doing HPO
  - Record time to train

Conclusion

- Processing and cleaning large amounts of natural language data was an interesting challenge
- We successfully built and evaluated two classification pipelines
  - Room for improvement
- Would like to, time permitting:
  - Experiment with larger datasets, other models
  - Experiment with additional feature columns
    - URLs, authors
  - Scrape our own datasets from reliable / unreliable sources and see what our ML models predict