Fake News Classification Using NLP

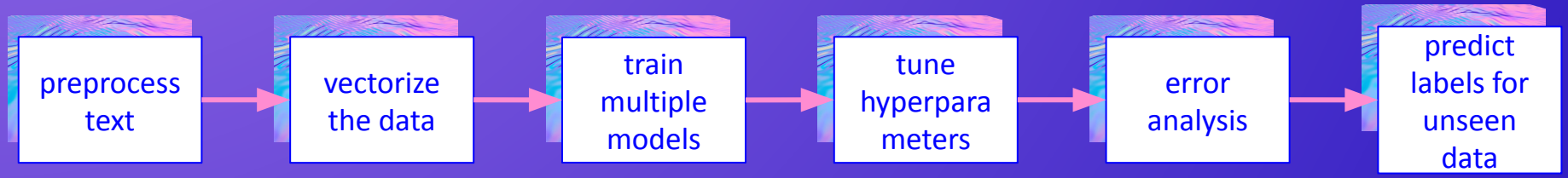
Predicting Real vs. Fake News Headlines

Hayley, Mohamad, Mara

Overview

Goal:
Build a classifier that distinguish ***fake news (0)*** from ***real news (1)*** using headliners.

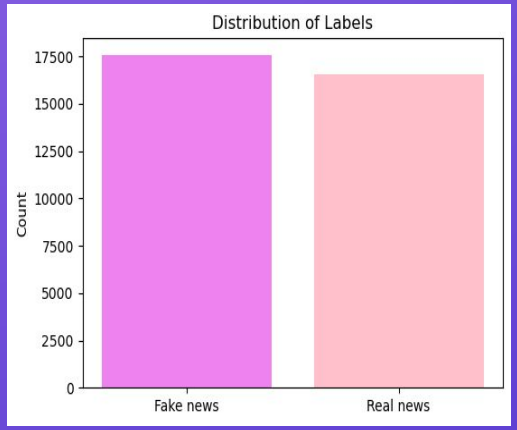
Workflow:

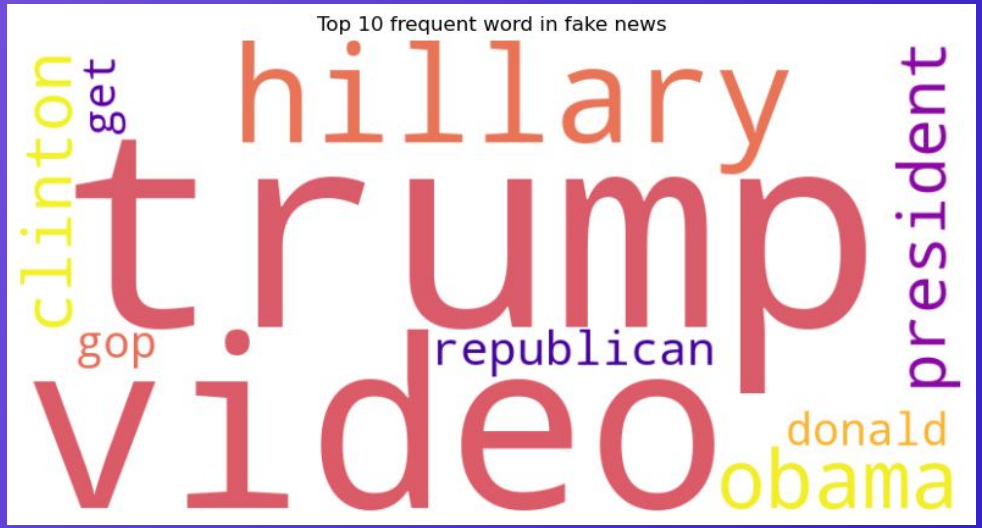


Dataset

Training Data:

- 34152 headlines
- 80/20 split into train & testing
- 14043 fake and 13278 real news (almost balanced)





Preprocessing

- lowercasing
- unicode normalization (NFKC)
- censored slur detection and replacement
- number normalization
- whitespace normalization
- tokenization via custom regex

additional processing

Legends

Colors	Links
Added	(f)irst change
Changed	(n)ext change
Deleted	(t)op

censored_slur

		Original			data_clean
f	1	brand-new	f	1	brand-new
	2	pro-trump		2	pro-trump
	3	ad		3	ad
	4	features		4	features
	5	so		5	so
	6	much		6	much
t	7	a**	t	7	censored_slur
	8	kissing		8	kissing
	9	it		9	it
	10	will		10	will
	11	make		11	make
	12	you		12	you
	13	sick		13	sick

rem punctuation

		Original			data_no_punc
n	1	brand-new	n	1	brandnew
	2	pro-trump		2	protrump
	3	ad		3	ad
	4	features		4	features
	5	so		5	so
	6	much		6	much
t	7	a**	t	7	censored_slur
	8	kissing		8	kissing
	9	it		9	it
	10	will		10	will
	11	make		11	make
	12	you		12	you
	13	sick		13	sick

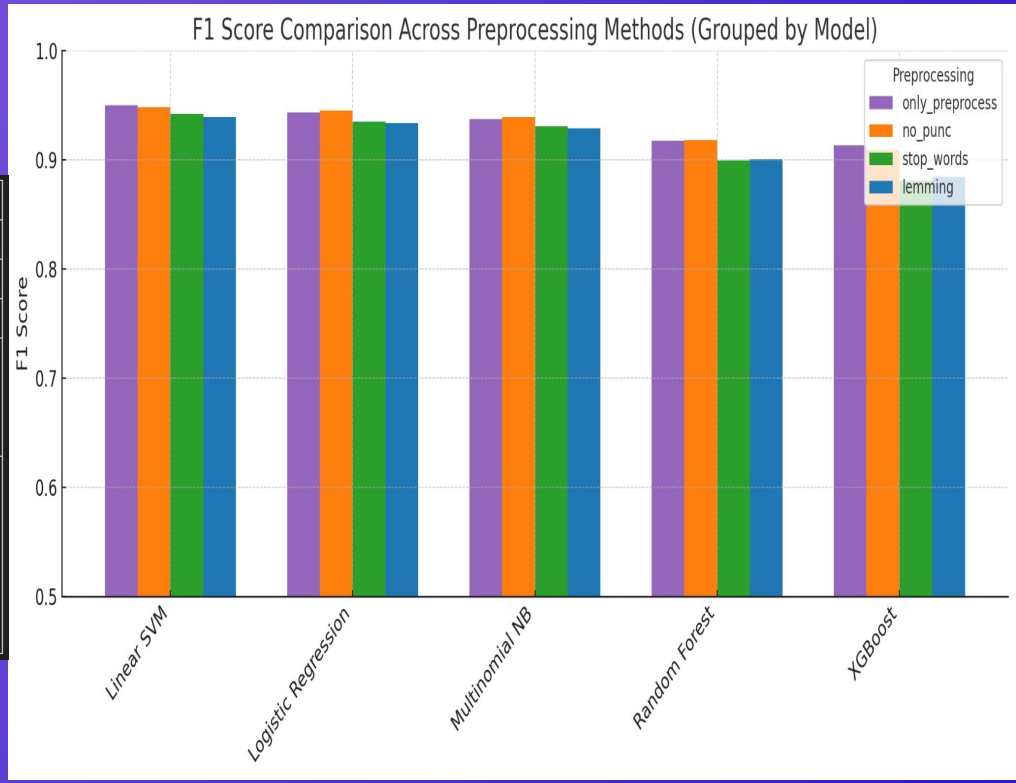
lemmatizing

		Original			data_lemm
f	1	brand-new	f	1	brand-new
	2	pro-trump		2	pro-trump
	3	ad		3	ad
n	4	features	n	4	feature
	5	so			
	6	much		5	much
n	7	a**	n	6	censored_slur
	8	kissing		7	kissing
n	9	it	n		
	10	will			
	11	make		8	make
t	12	you	t		
	13	sick		9	sick

Baseline Models and Data Selection

Model	Parameters
Linear SVC	`max_iter=5000`, `class_weight='balanced'`, `random_state=42`
Logistic Regression	`max_iter=2000`, `lbfgs`, `class_weight='balanced'`, `random_state=42`
Multinomial NB	`alpha=1.0`
Random Forest	n_estimators=200 max_depth=50 max_features='sqrt' n_jobs=-1
XGBoost	n_estimators=200 max_depth=6 eta=0.1 subsample=0.8 colsample_bytree=0.8 eval_metric='logloss' n_jobs=-1

Vectorizer: TF-IDF(ngrams_range(1,1)

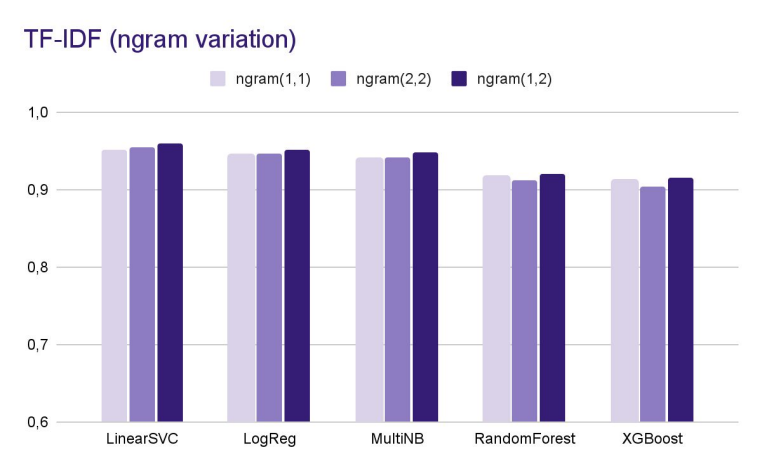


Vectorization Approach

1. **World-Level:**
   a. CountVectorizer (Bag of Words)
   b. TF-IDF

2. **Character-Level:**
   a. TF-IDF

Custom Tokenizer:
- handle censored words
- robust token regex

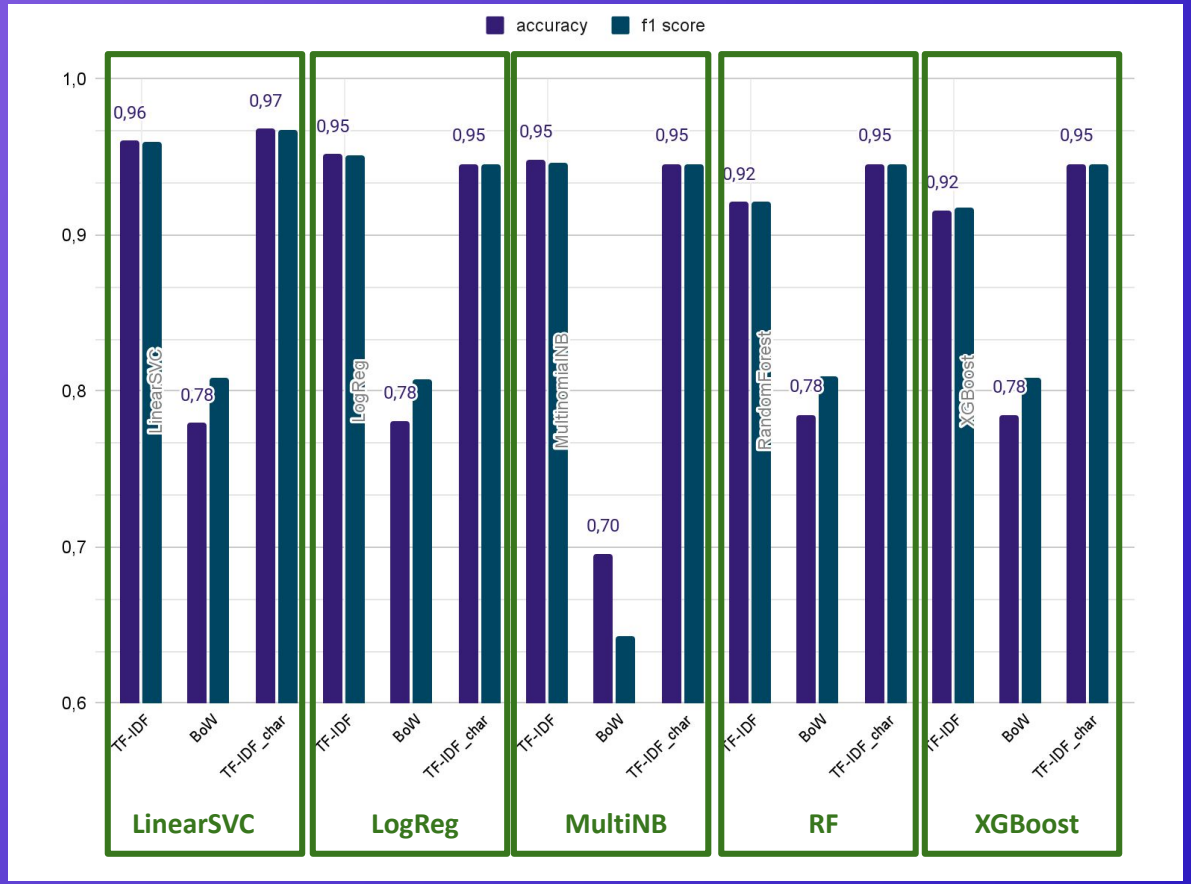


Vectorizer	Level	Performance
TF-IDF	word	medium
BoW	word	weakest
TF-IDF	char	best

Model & Vectorizer Comparison

Vectorizer Configurations

Vectorizer	Parameters
TF-IDF (word)	tokenizer=custom_tokenizer lowercase=False token_pattern=None ngram_range=(1,2) min_df=2 max_df=0.9
Bag-of-Words (BoW)	tokenizer=custom_tokenizer lowercase=False token_pattern=None ngram_range=(1,2) min_df=0.1 max_df=0.9
TF-IDF (char-level)	analyzer=char lowercase=False sublinear_tf=True ngram_range=(3,6) min_df=2 max_df=0.9



Vectorizer & Model Tuning – Grid Search

Model Pipeline

TF-IDF (character-level) → Linear SVC

- Analyzer: `char`
- Classifier: `LinearSVC(class_weight='balanced', max_iter=5000)`
- Sublinear TF scaling optional

Best Parameters Found

- `analyzer = 'char'`
- `ngram_range = (2,5)`
- `min_df = 5`
- `sublinear_tf = True`
- `C = 0.5`

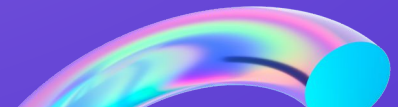
Grid Search Parameter Space

Component	Parameter	Values
TF-IDF	ngram_range	(2,5), (3,5), (3,6), (4,7), (5,8)
TF-IDF	min_df	3, 5, 10
TF-IDF	sublinear_tf	True, False
Linear SVC	C	0.5, 1.0, 2.0, 4.0

Performance Comparison

Metric	Before Tuning	After Tuning
Accuracy	**0.9707**	**0.9718**
F1 Score	**0.9698**	**0.9718**

F1 Improvement: +0.010



Error Analysis (Misclassifications)

Confusion Matrix
[[5050 245]
[162 4789]]

#misclassified samples: 407

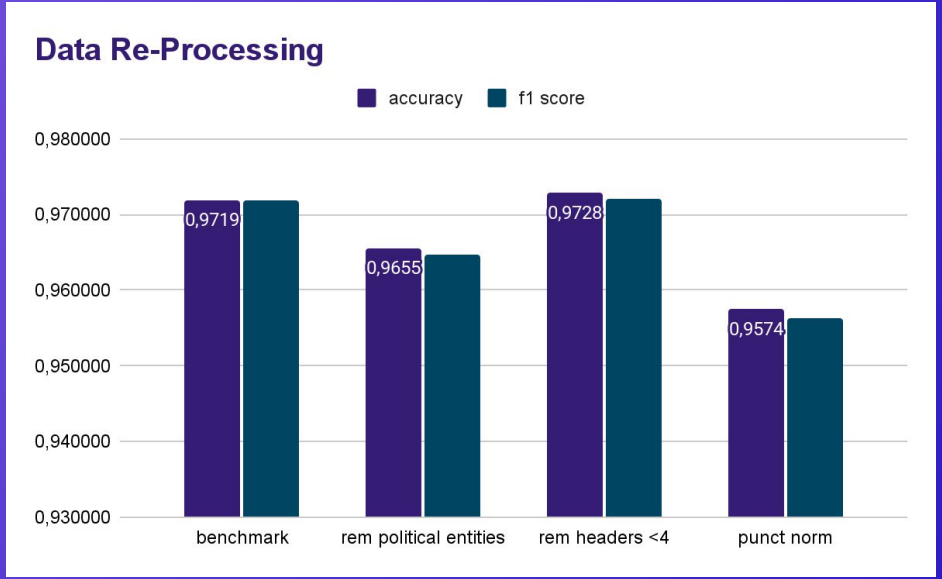
Accuracy: 0.9719
F1 Score: 0.9718

Headline	True label	Predicted label
trump hires son,s wedding planner to run new york housing department	0	1
facing tough re-election	0	1
obama to u.s.military on putin: ,he did not take my warnings, on syria,encourages russians to ,get a little smarter,	0	1
united nations chief 'very worried' by saudi-lebanon tensions	1	0
trump takes credit for republican ,win, in georgia	0	1
hurry	0	1
vietnam vet	0	1
trump praises putin while tweeting about a nuclear arms race	0	1
trashy	0	1
an obama\tnot the president\tbrings down the house at democratic convention	1	0

Data Re-Preprocessing

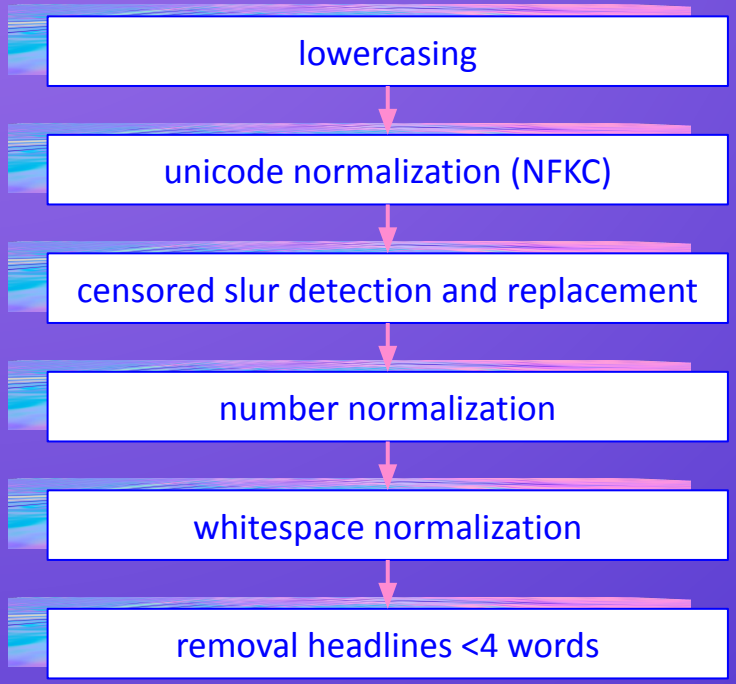
Preprocessing Enhancements

Step	Details
Remove political entities	Remove the following names: • trump • obama • clinton • putin • democrats • republicans • gop • senate • congress
Filter short headlines	Remove headlines with: • fewer than **4 words**
Normalize punctuation	Convert stylized quotes to standard ASCII: • “ ” „ ‟ → " • ‘ ’ ‚ → '



Best Pipeline

Optimized Preprocessing:



Optimized Vectorizer:

```
tfidf_best = TfidfVectorizer(
      analyzer='char',
      min_df=5,
      ngram_range=(2, 5),
      sublinear_tf=True)
```

Optimized Model:

```
svm_best = LinearSVC(
      class_weight='balanced',
      max_iter=5000,
      random_state=42)
```

Testing Data

📰 Headline 1

“Copycat muslim terrorist arrested with assault weapons”

Prediction: 🔴 **FAKE**

📰 Headline 2

“Mi school sends welcome back packet warning kids against wearing u.s. flag to school”

Prediction: 🔴 **FAKE**

📰 Headline 3

“Wow! chicago protester caught on camera admits violent activity was pre-planned: ‚it‚s not gonna be peaceful‚”

Prediction: 🔴 **FAKE**

📰 Headline 4

“‘France's macron says his job not 'cool' cites talks with turkey's erdogan”

Prediction: 🟢 **REAL**

📰 Headline 5

“North korea's kim jong un fetes nuclear scientists holds celebration bash”

Prediction: 🟢 **REAL**

📰 Headline 6

“Merkel names refugee expert as foreign policy adviser”

Prediction: 🟢 **REAL**

Random Testing Data

🔴 FAKE

🟢 REAL

Headlines	Label	Predicted (Our Model)	Predicted (Fake News BERT)
"NASA announces successful deployment of Earth-observing climate satellite"	🟢	🔴	🔴
"NOAA reports warmer-than-average ocean temperatures across the Atlantic",	🟢	🔴	🔴
"FDA approves new treatment for adults with chronic respiratory disease"	🟢	🟢	🟢
"NIH researchers identify protein linked to improved immune response",	🟢	🔴	🔴
"U.S. Geological Survey confirms magnitude 4.2 earthquake near California coast",	🟢	🟢	🔴
"Department of Energy funds new research into next-generation battery materials",	🟢	🟢	🔴
"CDC releases updated guidelines for preventing seasonal influenza",	🟢	🔴	🔴
"National Park Service reopens trail system following completion of safety repairs",	🟢	🟢	🔴
"Scientists confirm discovery of ancient underground city spanning 600 miles beneath Alaska",	🔴	🔴	🔴
"NASA whistleblower claims hidden mission found alien structures on Mars",	🔴	🔴	🔴
“Researchers warn that listening to certain radio frequencies can alter human DNA",	🔴	🟢	🔴
"Government developing technology to control weather using giant satellites",	🔴	🔴	🔴

Our Model

Accuracy: **0.54**

F1 Score:

0	0.50
1	0.57

Fake News BERT

Accuracy: **0.46**

F1 Score:

0	0.59
1	0.22

Learnings

Data Preprocessing Matters
Improvement through normalization & custom tokenizer (censored words).

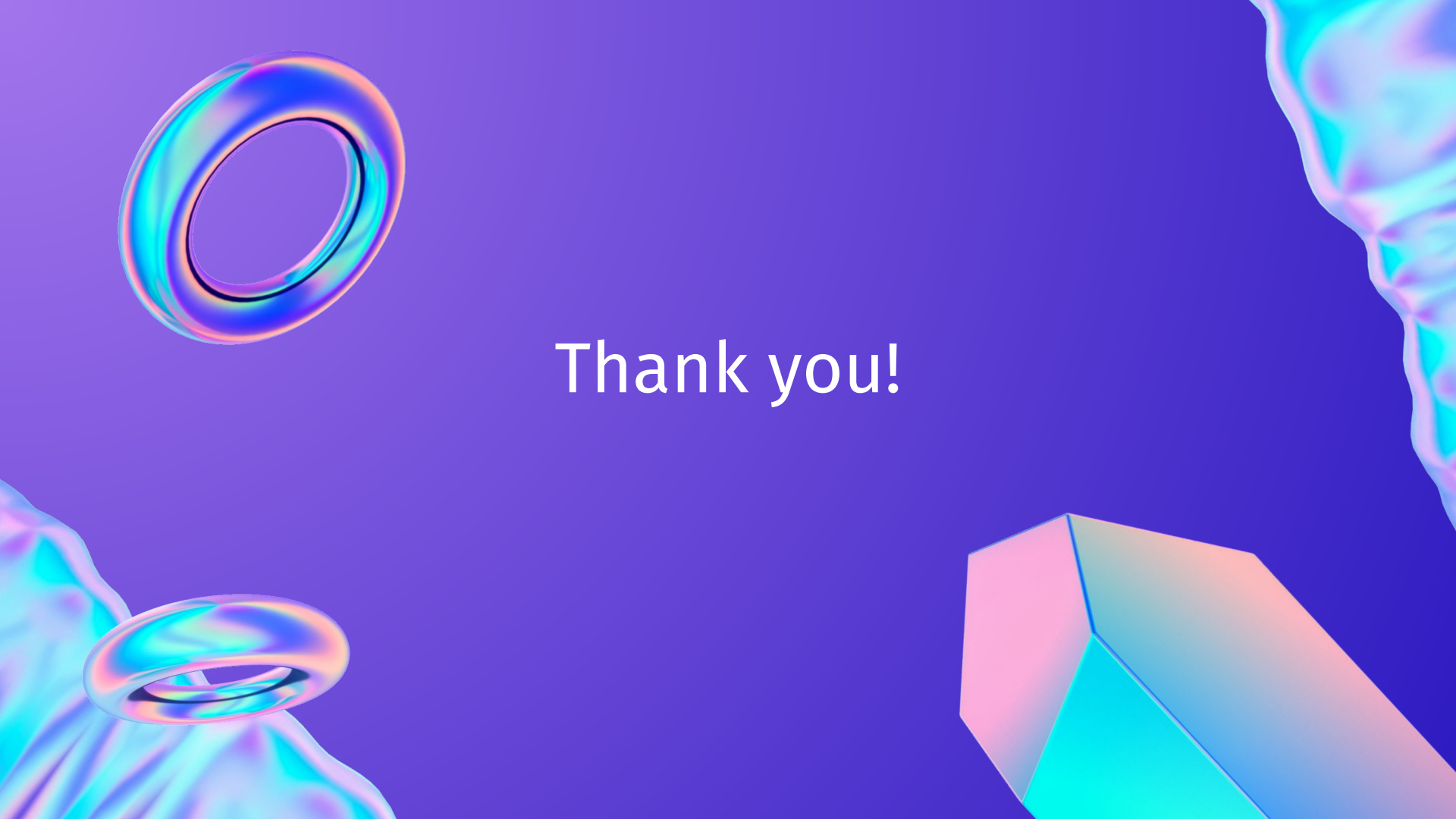
Vectorization
Char-level TF-IDF outperformed word-based methods.

Modeling
Linear SVC dominated above all tested variations.

Hyperparameter
Tuning further improved through finding best n-gram ranges.

Error Analysis
Confusion analysis revealed concrete preprocessing improvements.

Final Model
Robust & strong prediction on relatively simple model (accuracy & F1 Score: ~0.968).
With 54% accuracy on random tested data it could outperform a pre-trained transfer model.


Thank you!