Clickbait Detection

2025 NLP Course Project

Jakub Sawicki, Jędrzej Sokołowski, Wiktor Woźniak

What is Clickbait?

- Attention-grabbing headlines, often omitting key facts
- Sparks curiosity and encourages clicks using sensational language
- Difficult to detect automatically: can be subtle and context-dependent
- Exists on a spectrum, not just binary classification

Motivation & Challenges

- Manipulative clickbait undermines trust in news and social media
- Hard to draw the line between catchy and misleading
- Even humans disagree about what counts as clickbait
- Growing sophistication with AI-generated clickbait

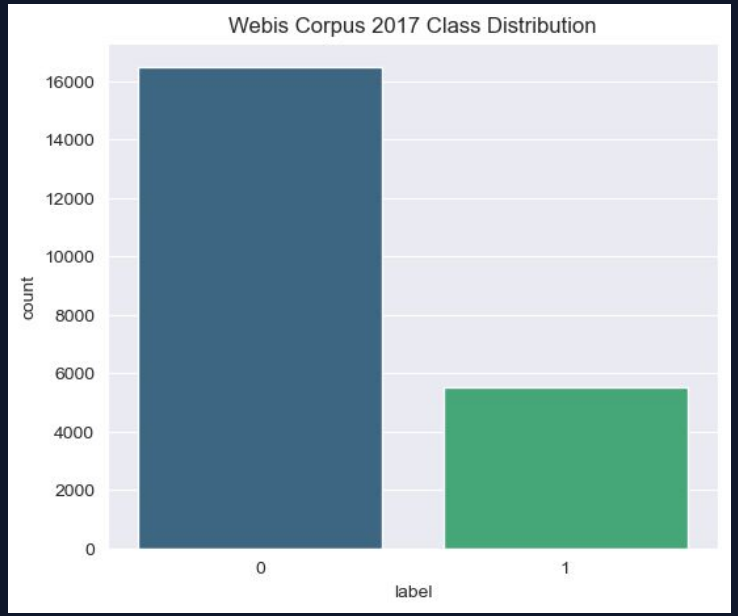
Related works, SOTA Approaches

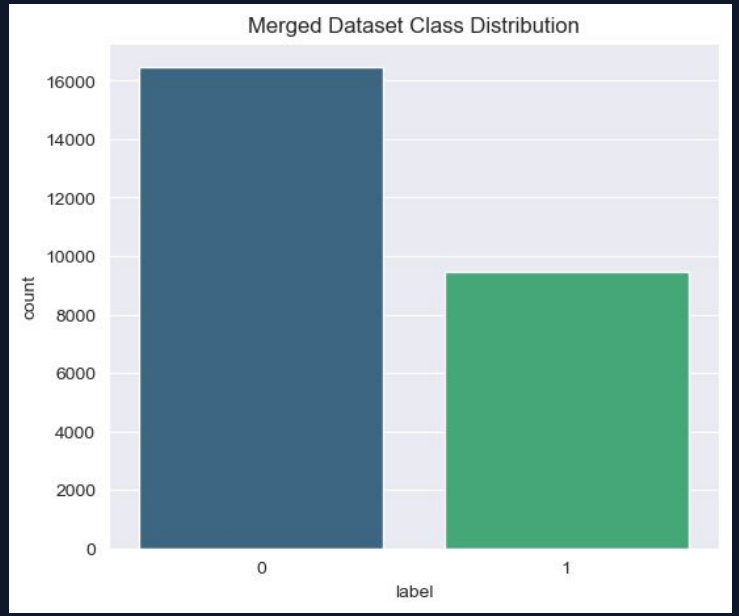
- Early: Hand-crafted features (punctuation, length, headline-article overlap)
- Classical ML: SVM, Random Forests, high accuracy with engineered features
- Deep Learning: CNNs, RNNs using word embeddings (Word2Vec, GloVe)
- Transformers: BERT, RoBERTa, current state-of-the-art for both detection and “spoiling”

Key Open Datasets

- Webis Clickbait Corpus 2017 (Twitter, 38,517 posts, 9,276 clickbaits)
- Webis Clickbait Spoiling Corpus 2022 (Twitter, Reddit, Facebook, 5,000 posts, all clickbaits)
- Wikinews Clickbait Corpus (crowdsourced English news headlines)
- Thai & Chinese Headline Datasets (expands multilingual, cross-domain evaluation)

Merging datasets



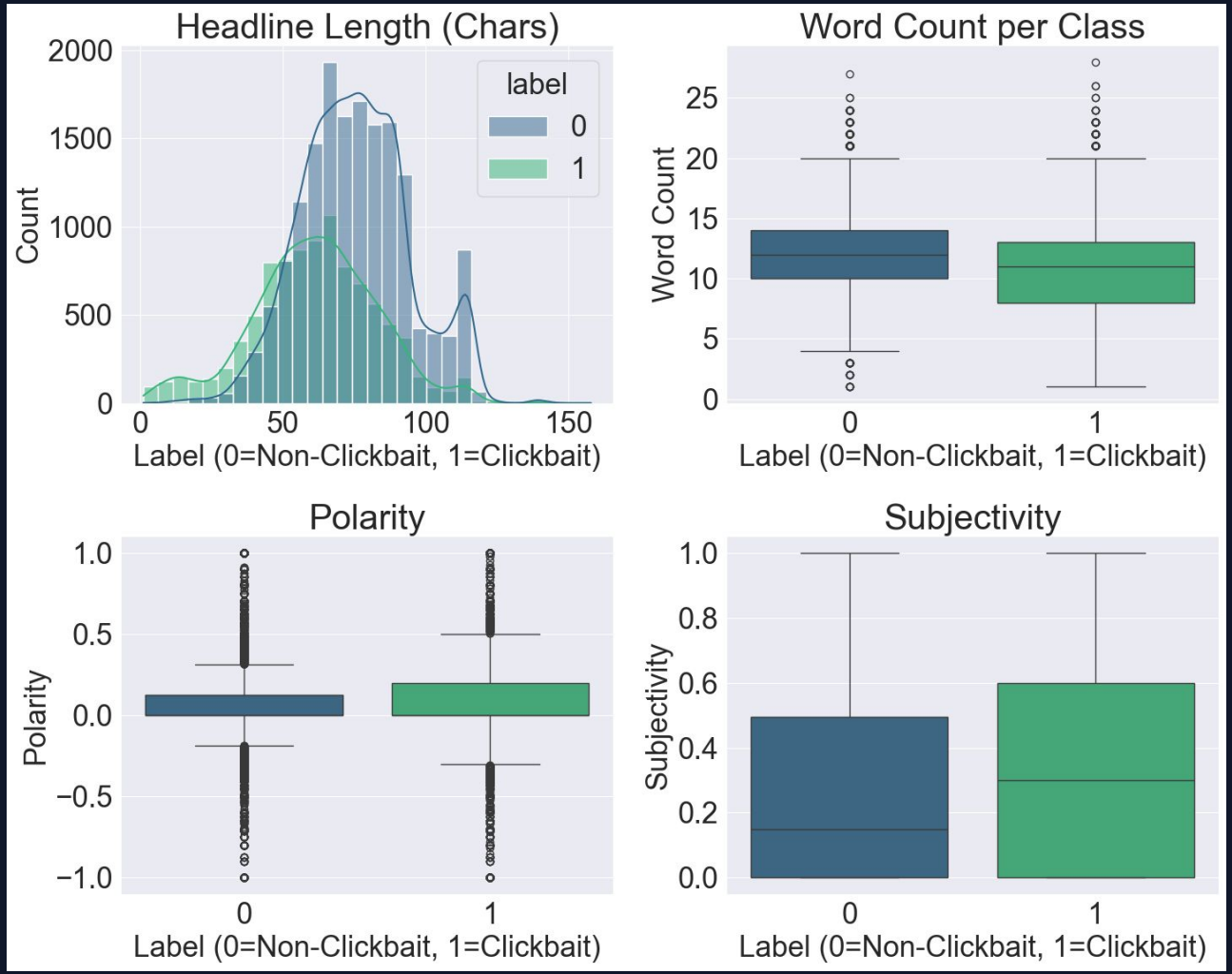


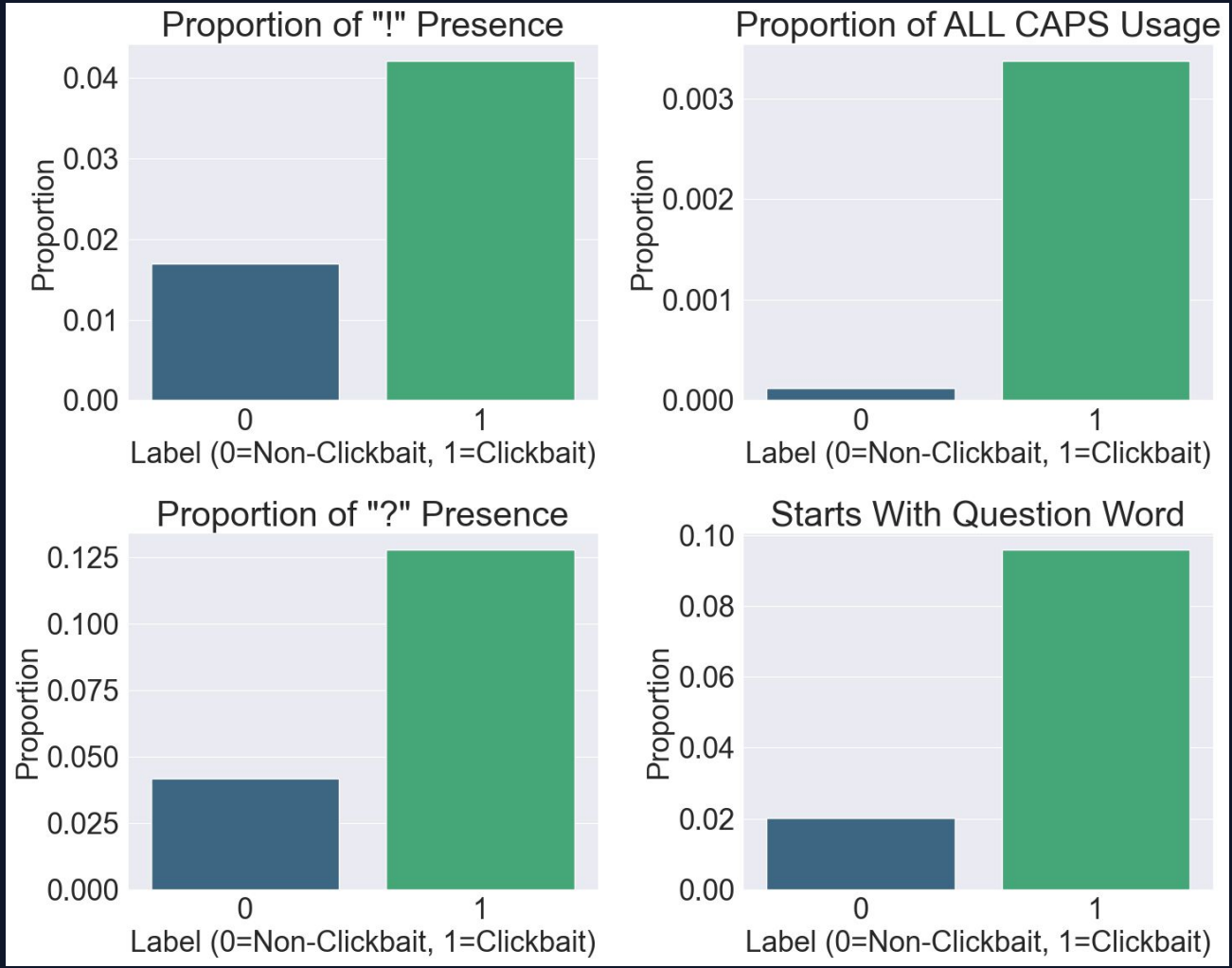
Project Goals & Methods

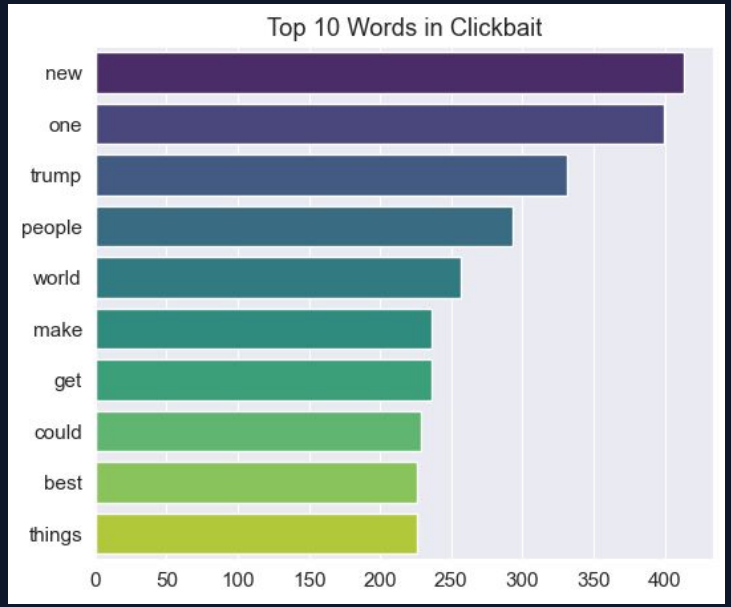
- Build interpretable, effective clickbait detection system using open benchmarks
- Compare classical ML, neural, and transformer models for detection
- Evaluate Precision, Recall, F1-score

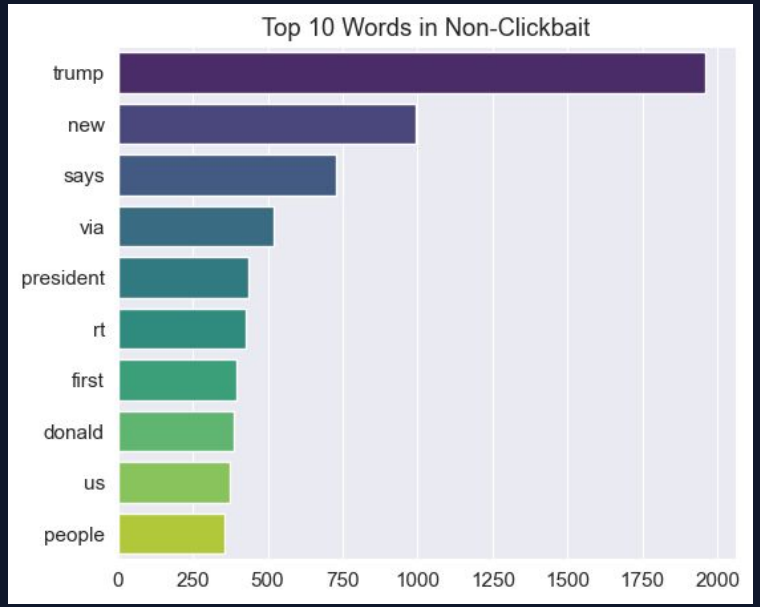
Data Processing

- Remove empty rows
- Train/Validation/Test split - 0.7/0.15/0.15
- Training dataset undersampling
- Add new features:
  - Character count
  - Word count
  - Exclamation marks (!)
  - Question marks (?)
  - ALL CAPS presence
  - Sentiment polarity
  - Sentiment subjectivity
  - Question word markers


Headline Length (Chars)
Count
label
0
1
Label (0=Non-Clickbait, 1=Clickbait)
Word Count per Class
Word Count
Label (0=Non-Clickbait, 1=Clickbait)
Polarity
Polarity
Label (0=Non-Clickbait, 1=Clickbait)
Subjectivity
Subjectivity
Label (0=Non-Clickbait, 1=Clickbait)


Proportion of "!" Presence
Proportion
0.04
0.03
0.02
0.01
0.00
0
1
Label (0=Non-Clickbait, 1=Clickbait)
Proportion of ALL CAPS Usage
Proportion
0.003
0.002
0.001
0.000
0
1
Label (0=Non-Clickbait, 1=Clickbait)
Proportion of "?" Presence
Proportion
0.125
0.100
0.075
0.050
0.025
0.000
0
1
Label (0=Non-Clickbait, 1=Clickbait)
Starts With Question Word
Proportion
0.10
0.08
0.06
0.04
0.02
0.00
0
1
Label (0=Non-Clickbait, 1=Clickbait)


Top 10 Words in Clickbait
new
one
trump
people
world
make
get
could
best
things
0
50
100
150
200
250
300
350
400


Top 10 Words in Non-Clickbait
trump
new
says
via
president
rt
first
donald
us
people
0
250
500
750
1000
1250
1500
1750
2000

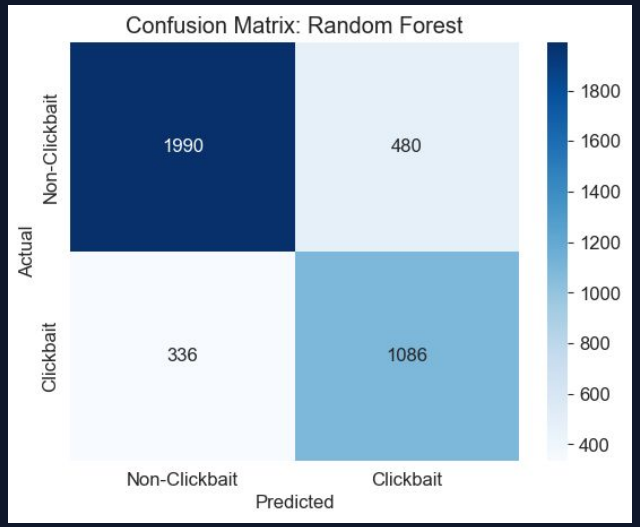
Feature & Model Summary

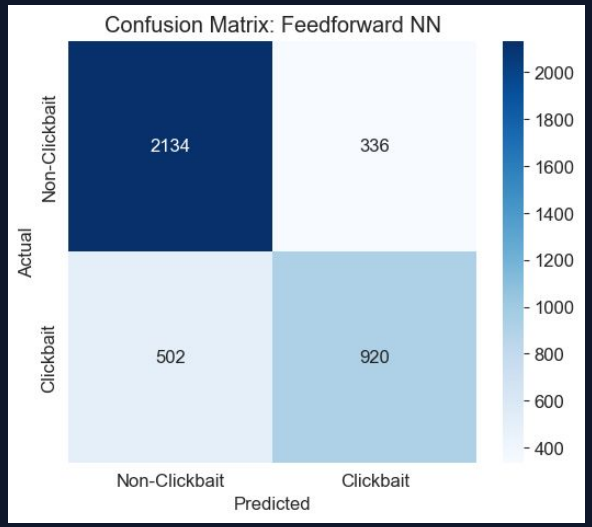
- Classical Baseline: Random Forest (TF-IDF)
- Feedforward Neural Network (3 hidden layers with ReLU)
- Transformer: DistilBERT fine-tuned for headline classification
- Metrics: Precision, Recall, F1-score

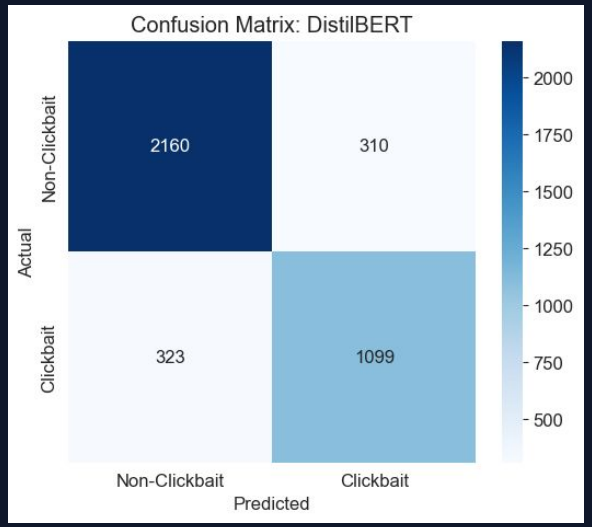
Random Forest Model Performance (Test Set)				
Class	**Precision**	**Recall**	**F1-score**	**Support**
0 (Non-Clickbait)	0.858 ± 0.003	0.821 ± 0.009	0.839 ± 0.005	2470
1 (Clickbait)	0.711 ± 0.010	0.765 ± 0.005	0.737 ± 0.006	1422

Feed-Forward Neural Network Performance (Test Set)				
Class	**Precision**	**Recall**	**F1-score**	**Support**
0 (Non-Clickbait)	0.827 ± 0.015	0.838 ± 0.023	0.832 ± 0.005	2470
1 (Clickbait)	0.713 ± 0.018	0.693 ± 0.039	0.702 ± 0.012	1422

DistilBERT Model Performance (Test Set)				
Class	**Precision**	**Recall**	**F1-score**	**Support**
0 (Non-Clickbait)	0.881 ± 0.014	0.861 ± 0.031	0.871 ± 0.011	2470
1 (Clickbait)	0.770 ± 0.032	0.797 ± 0.032	0.782 ± 0.008	1422


Confusion Matrix: Random Forest
Non-Clickbait
Clickbait
Actual
1990
480
336
1086
Non-Clickbait
Clickbait
Predicted
1800
1600
1400
1200
1000
800
600
400


Confusion Matrix: Feedforward NN
Non-Clickbait
Clickbait
Actual
2134
336
502
920
Non-Clickbait
Clickbait
Predicted
2000
1800
1600
1400
1200
1000
800
600
400


Confusion Matrix: DistilBERT
Non-Clickbait
Clickbait
Actual
2160
310
323
1099
Non-Clickbait
Clickbait
Predicted
2000
1750
1500
1250
1000
750
500

Clickbaitness measure

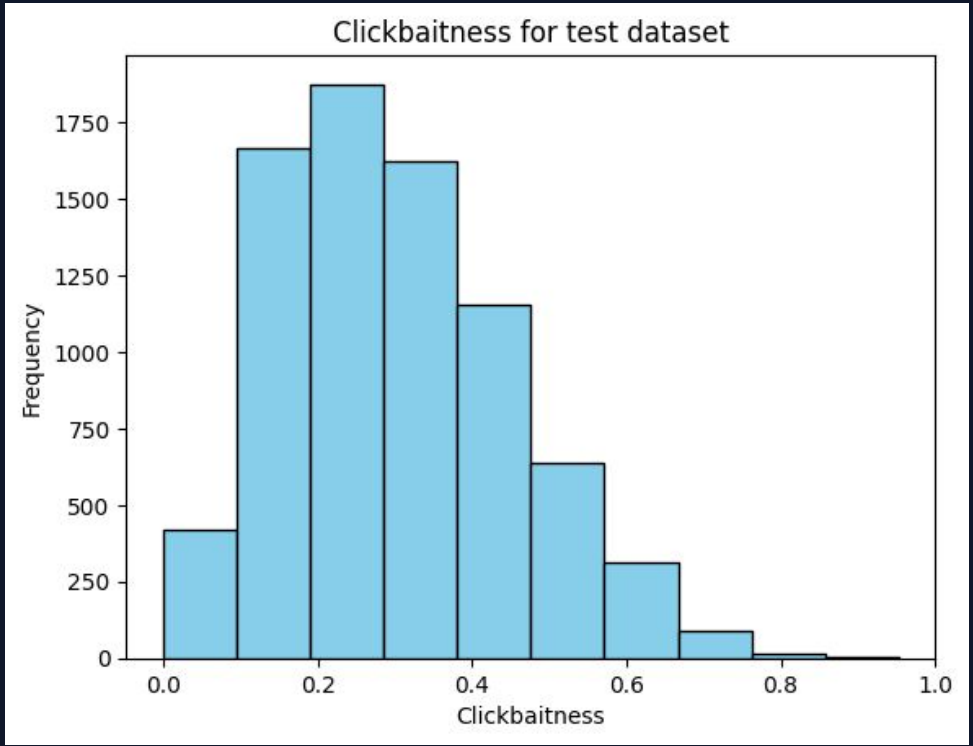
Baitness = (EC + C + S + EOT) / 4

EC = Eye catchness

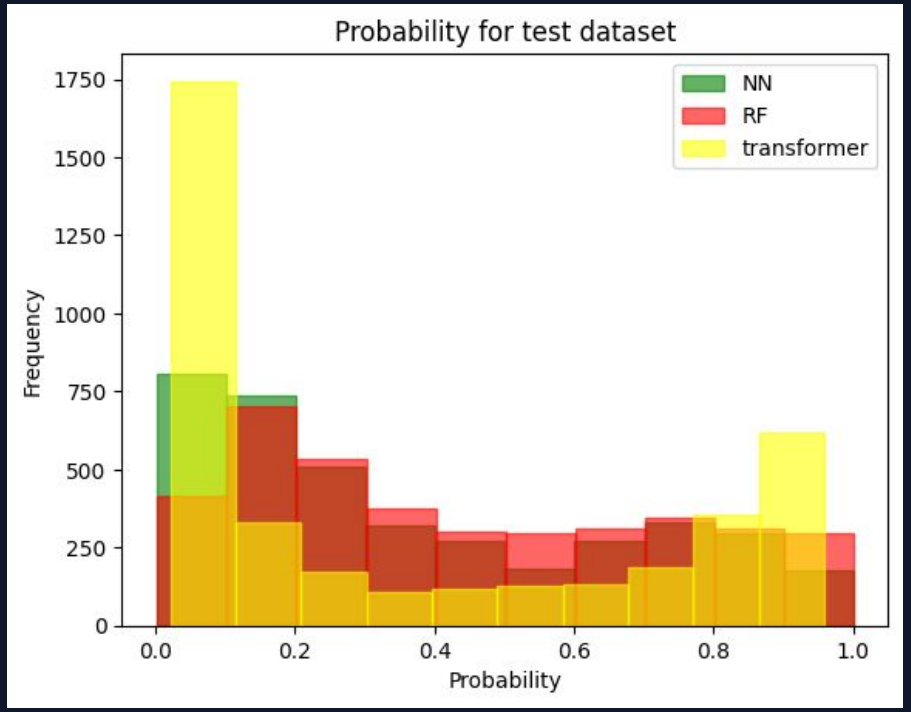
C = Curiosity

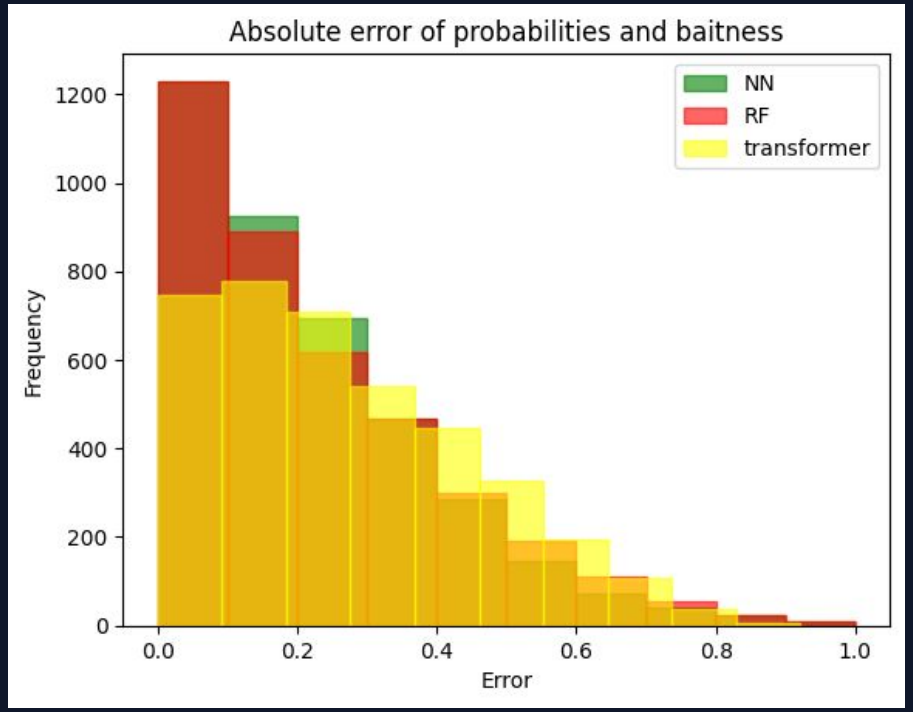
S = Sentiment

EOT = Ease of text



Clickbaitness measure





Clickbaitness measure

Model	MAE	MSE
Neural Network	0.217 ± 0.174	0.077 ± 0.118
Random Forest	0.226 ± 0.188	0.086 ± 0.130
Transformer	0.269 ± 0.184	0.106 ± 0.127

Summary

- DistilBERT was a superior model
- FF-NN and RF approaches achieved comparable performance
- Clickbait probabilities are not a good method for predicting baitness of the post

Questions?