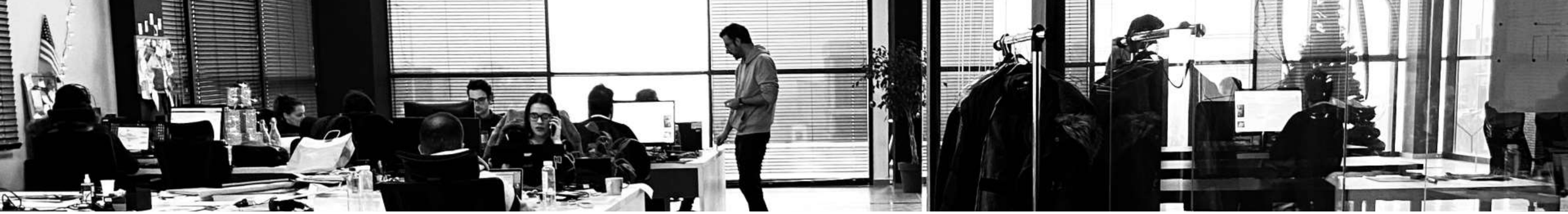


Detection of Sensational Language Features in News Headlines

3 September, 2024

Definition of Sensational Language

Sensational Language refers to a style of expression that uses striking or shocking details to quickly evoke intense, often superficial interest, curiosity, or emotional reactions from the audience.

1. The complexity of human language.

2. How far can we journey along this path?

3. What are our guild stones?

Research questions:

Our contribution:

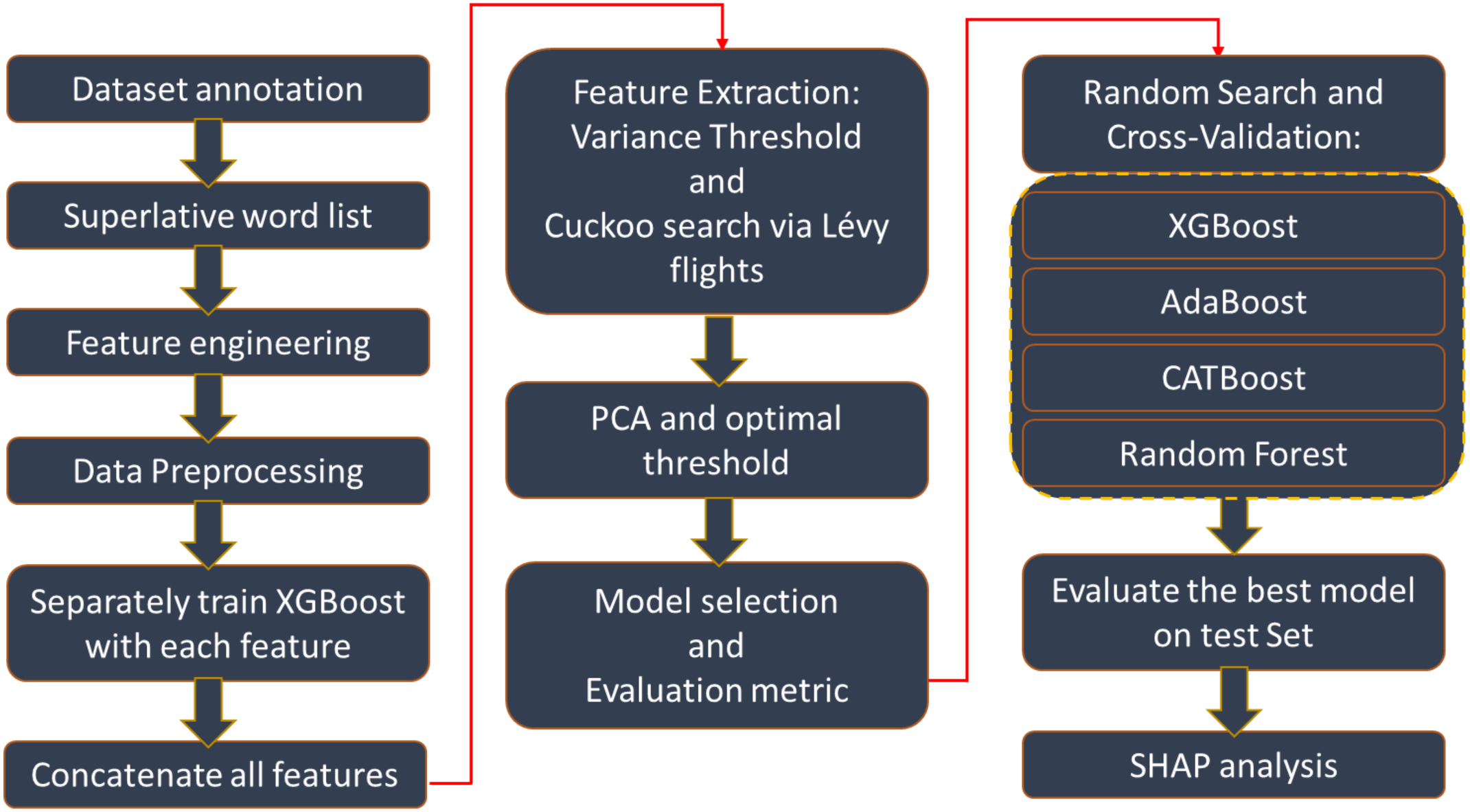
1. MIRUKU sensational news headline dataset
2. Multiple linguistic
3. Robust model
4. Effective features

Environment configuration

Google Colab Pro+

A100 GPU

Architectural flow



Dataset annotation

News Clickbait Dataset from Kaggle:

The clickbait corpus:
‘BuzzFeed’, ‘Upworthy’, ‘ViralNova’, ‘Thatscoop’, ‘Scoopwhoop’ and ‘ViralStories’.

The non-clickbait article:
‘WikiNews’, ’New York Times’, ‘The Guardian’, and ‘The Hindu’.

32,000 news headlines

Dataset annotation

OpenAI Text Models GPT-4o-2024-05-13

Likert Scale

Zero-shot prompting

Batch 10 request and delay request for 5 seconds

OpenAI API key in Colab Environment

Dataset annotation

The MIRUKU Sensational News Headline Dataset

headline	clickbait	sensation	sensation_score	sensation_reason	emotion	arousal	arousal_score	arousal_reason
1,100 evacuated due to massive Halifax brush fire	0	1	4	The use of terms like "massive" and "1,100 evacuated" creates a significant sense of urgency and drama, making the headline very sensational.	anger, fear, sadness	Yes	0.85	The headline involves a sudden evacuation and a massive fire, which implies immediate danger and distress, leading to high physiological and psychological activation.
2006 Formula 1 season starts with GP of Bahrain	0	0	0	Factual and straightforward statement without sensational language or tone.	neutral	No	0.343	The headline is an informative statement about the start of the sports season, which is relatively low in arousal and merely factual.
10 Celebrity Red Carpet Looks That Are Out Of This World	1	1	4	The phrase "Out Of This World" is hyperbolic and evokes a strong emotional response, suggesting an extraordinary impression.	neutral	No	0.353	The headline is descriptive and mentions a common topic in entertainment, which does not evoke significant physiological or psychological activation.
A Civil Rights Victory Party on the Mall	0	0	0	The language in the text is neutral and factual without any exaggerated or emotive expressions.	joy, anticipation, trust	Yes	0.6	The headline depicts a celebratory event related to a significant social achievement, likely generating moderate excitement and positive anticipation.
...								

Zero-shot prompting with likert scale:

- **Utilize Chat GPT-4o to annotate on <u>News Headline Clickbait dataset</u> (32,000 lines) :**

  Role: "You are a Chief Natural Language Processing and Linguistics Engineer."},

  Content: Analyze the sensational/arousal level of the following text: {text}.

  Sensational/Arousal score instruction:

  (0) Not at all (mean 0–0.75)
  (1) Not too much (mean 0.76–1.50)
  (2) Somewhat (mean 1.51–2.25)
  (3) Fairly (mean 2.26–3.25)
  (4) Very (mean 3.26–4)

- **Reply in four columns:**

  1.Sensational Score
  2.Reply in three columns: Sensational or Not (Yes or No), Sensational Scores (0-4), Reason
  3.Emotion(s) or neutral.
  4.Arousal or Not (Yes or No).
  5.Arousal Scores for the news headline (0-4).
  6.The reason for the score of the arousal of news headlines instead of for all the emotions.

Superlative word list

Internet

Brown

web_text

Reuters

movie reviews

Gutenberg

Superlative word list

General superlative adjectives:
(re.compile(r'\b\w+est\b'), 'ADJ'),
(re.compile(r'\b\w+iest\b'), 'ADJ'),

Multi-syllable superlatives:
(re.compile(r'\b(?:most|least|best|worst)\s+\w+\b'), 'ADJ'),

Superlative adverbs:
(re.compile(r'\b(?:most|least|best|worst)\s+\w+ly\b'), 'ADV'),

Special superlative adjectives:
(re.compile(r'\b(?:foremost|hindmost|inmost|innermost|nethermost|outmost|outermost|topmost|undermost|upmost|uppermost|utmost|uttermost)\b'), 'ADJ'),

List of common superlative adjectives:
(re.compile(r'\b(?:ablest|angriest|baldest|battiest|beadiest|bitterest|blackest|blandest|blankest|bleakest|blondest|bloodiest|bluest|…

Specific multi-word superlative phrases:
(re.compile(r'\b(?:most beautiful|most boring|most colorful|most comfortable|most complete|most cruel|most delicious|most difficult|most evil…

WHY WE DON'T DIRECTLY USE POS tag?

1. Spacy - NER
2. Stanza - POS

Superlative word list

Wrong tag in POS:

13 Animals Who Are Way More Gangster Than You
['13', 'NUM'], ['Animals', 'NOUN'], ['Who', 'PRON'], ['Are', 'AUX'], **['Way', 'NOUN']**, ['More', 'ADJ'], ['Gangster', 'NOUN'], ['Than', 'ADP'], ['You', 'PRON']

Way should be ADV

12 Mind-Blowing Ways To Eat Polenta
[('12', 'NUM'), **('Mind', 'NOUN'), ('-', 'PUNCT'), ('Blowing', 'NOUN')**, ('Ways', 'NOUN'), ('To', 'PART'), ('Eat', 'VERB'), ('Polenta', 'NOUN')]

Mind-Blowing should be ADJ

Wrong tag in NER:

Poll shows **Prévval** with clear lead, but ineligible candidate **Siméus** could have presented a challenge

Failed to identify Prévval and Siméus as PERSON

11 Quotes From Harry Potter To Help You Cope With Loss
[('11', 'CARDINAL'), ('**Harry Potter', 'PERSON**')]

Harry Potter could be WORK_OF_ART

unavailable features

POS 3-grams

NER

Feature engineering

Lexical Features:

1. Number of Words in the Headline
2. Number of stop words in the headlines
3. The ratio of the number of stop words to the number of content words
4. TF-IDF with stop words
5. TF-IDF without stop words
6. Superlative word list

Syntactic Features:

7. Syntactic 4-grams

Semantic Features:

8. Sentence Subjectivity and Objectivity Evaluation
9. Sentiment Analysis

Readability Features:

10. Informality (Flesch-Kincaid Readability)

Stylistic Features:

11. Elongated Words
12. Punctuation

Data Preprocessing

Data Loading and Feature Selection

- Preselected feature indices.

Feature Name Saving:

- Joblib for SHAP analysis.

Data Splitting:

- 80-20 ratio to split off the test set
- 75-25 ratio to split the training set and validation set.

Feature Scaling:

- MinMaxScaler
- Fitting only on the training set, then transforming the validation and test sets.

Handling Class Imbalance:

- SMOTE (Synthetic Minority Over-sampling Technique).

Separate Handling of Validation and Test Sets:

- Only transforming (not fitting) the validation and test sets.

Baseline model XGBoost and separately train with each feature

```
param_dist_xgb = {
'max_depth': [3, 6, 9],
'min_child_weight': [1, 3],
'n_estimators': [1000],
'learning_rate': [0.01, 0.1, 0.3],
'subsample': [0.8, 1.0],
'colsample_bytree': [0.8, 1.0],
'gamma': [0, 0.1]
}

# Create K-Fold cross-validation object
kfold = KFold(n_splits=5, shuffle=True, random_state=42)

random_search_xg = RandomizedSearchCV(
estimator=xgb_classifier,
param_distributions=param_dist_xgb,
n_iter=100,
cv=kfold,
n_jobs=-1,
verbose=2,
scoring='f1',
random_state=42
)
```

Research questions:

1. Can we utilize a single linguistic feature to effectively and stability identify the sensational language in news headlines?

Feature	Best CV score
Number of stop words	0.6304
Syntactic 4-grams	0.6233
TF-IDF with Stop words	0.6168

Feature	Accuracy
TF-IDF with Stop words	0.66
Syntactic 4-grams	0.64
Sentiment analysis	0.62

Feature	F1-score non-sensation
TF-IDF with Stop words	0.71
Elongated Words	0.69
Punctuation	0.67

Feature	F1-score sensation
Number of words	0.61
Number of stop words	0.61
The ratio of stop words to content words	0.59
Flesch-Kincaid Readability	0.59
Subjectivity and Objectivity	0.59
Flesch-Kincaid Readability	0.59
TF-IDF with Stop words	0.59
Sentiment analysis	0.57

1. Unfortunately, we can not.

Concatenate all features

52,190

Feature Extraction

1. Filter method:

2. Wrapper method:

Variance Threshold

Removing Constant/Nearly Constant Features:

- Reducing dimension
- Reducing the Risk of Overfitting
- Improving Model Interpretability

Threshold:

- Threshold: 0.0001, Features retained: 9437
- **Threshold: 0.001, Features retained: 1236**
- Threshold: 0.01, Features retained: 73
- Threshold: 0.1, Features retained: 8

Feature Extraction

1. Filter method:

2. **Wrapper method:**

- **Cuckoo Search via Lévy flights**

Advantage:

- Objective function
- Global search capability
- Balancing exploration
- Handle non-linear problems

616 features

Feature Extraction

1. Filter method:

2. **Wrapper method:**

- **Cuckoo Search via Lévy flights**

- **Cuckoo Search**

Imagine in a market:

- Search agent: Cuckoo bird
- Bird's Nest: stall(solution) and products(features)
- Egg: new or improved solutions.
- Parasitic behavior: lay eggs(new or better solution)
- The stall owner (host bird): if better

Feature Extraction

1. Filter method:

2. **Wrapper method:**

- **Cuckoo Search via Lévy flights**

Lévy flights

- Short-distance Hopping: nearby stalls
- Occasional Long-distance Flights: explore new regions
- Unpredictability: may do both, local optima
- Adaptability: stay a bit longer

PCA (Principal component analysis)

joy

surprise

anticipation

trust

anger

fear

sadness

disgust

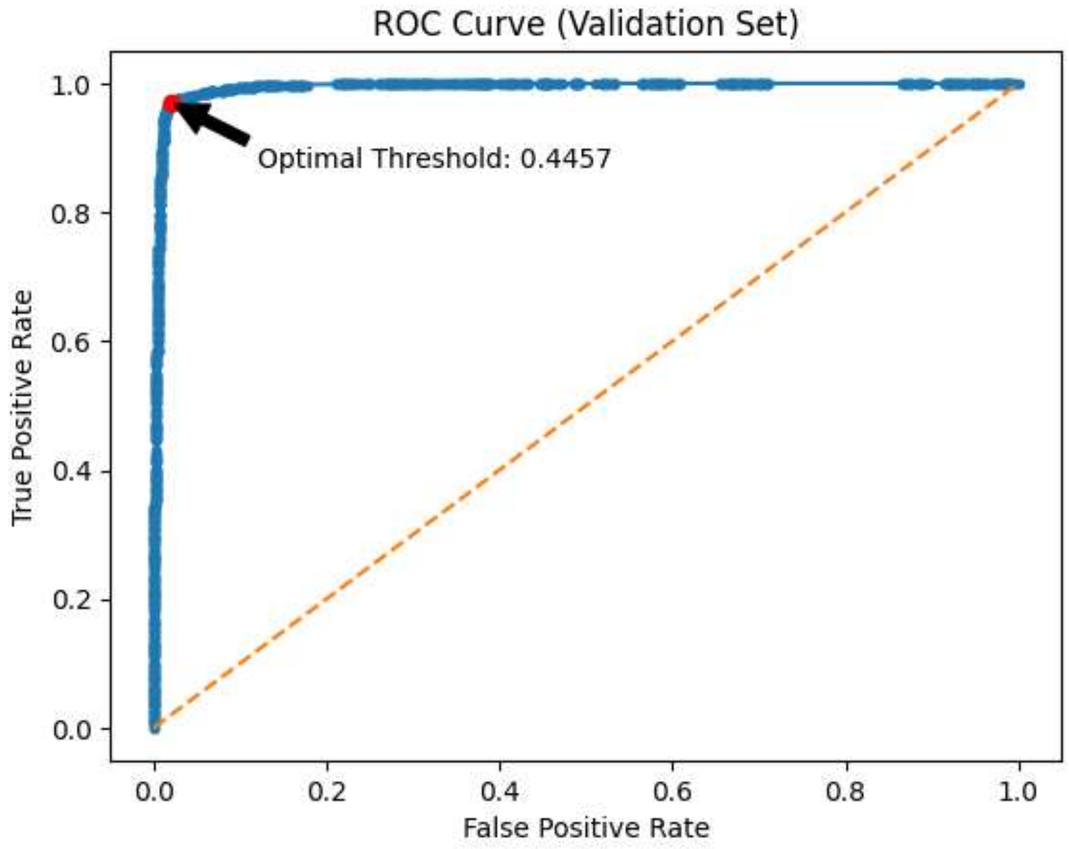
neutral

PCA and optimal threshold

Sensation score

Arousal score

Emotion after PCA



Model Selection

- XGBoost
- XGBoost **with Optimal Threshold**
- XGBoost **with superlative adjective word list**
- XGBoost **with the superlative adjective words list and Optimal Threshold**
- AdaBoost
- CATBoost
- Random Froest

Advantage:

- Ensemble Learning
- Handling Non-Linearity
- Handling High-Dimensional Data
- Noise Resistance
- Preventing Overfitting
- Explainable

Research questions:

2. What algorithm demonstrates robust performance in identifying sensational language within news headlines?

CATBoost				
Best cross-validation score				0.6697
	Precision	Recall	F1-score	Support
Non-sensation	0.68	0.68	0.68	3223
Sensation	0.64	0.64	0.64	2862
Accuracy			0.66	6085
Macro avg	0.66	0.66	0.66	6085
Weighted avg	0.66	0.66	0.66	6085

XGBoost with Superlative Adjective Words List				
Best cross-validation score				0.6644
	Precision	Recall	F1-score	Support
Non-sensation	0.68	0.67	0.68	3223
Sensation	0.64	0.65	0.64	2862
Accuracy			0.66	6085
Macro avg	0.66	0.66	0.66	6085
Weighted avg	0.66	0.66	0.66	6085

2. Yes, we can.

CATBoost				
Best cross-validation score				0.6697
	Precision	Recall	F1-score	Support
Non-sensation	0.68	0.68	0.68	3223
Sensation	0.64	0.64	0.64	2862
Accuracy			0.66	6085
Macro avg	0.66	0.66	0.66	6085
Weighted avg	0.66	0.66	0.66	6085

CATBoost on Test Set				
	Precision	Recall	F1-score	Support
Non-sensation	0.68	0.68	0.68	3223
Sensation	0.64	0.64	0.64	2862
Accuracy			0.66	6085
Macro avg	0.66	0.66	0.66	6085
Weighted avg	0.66	0.66	0.66	6085

2. Yes, we can.

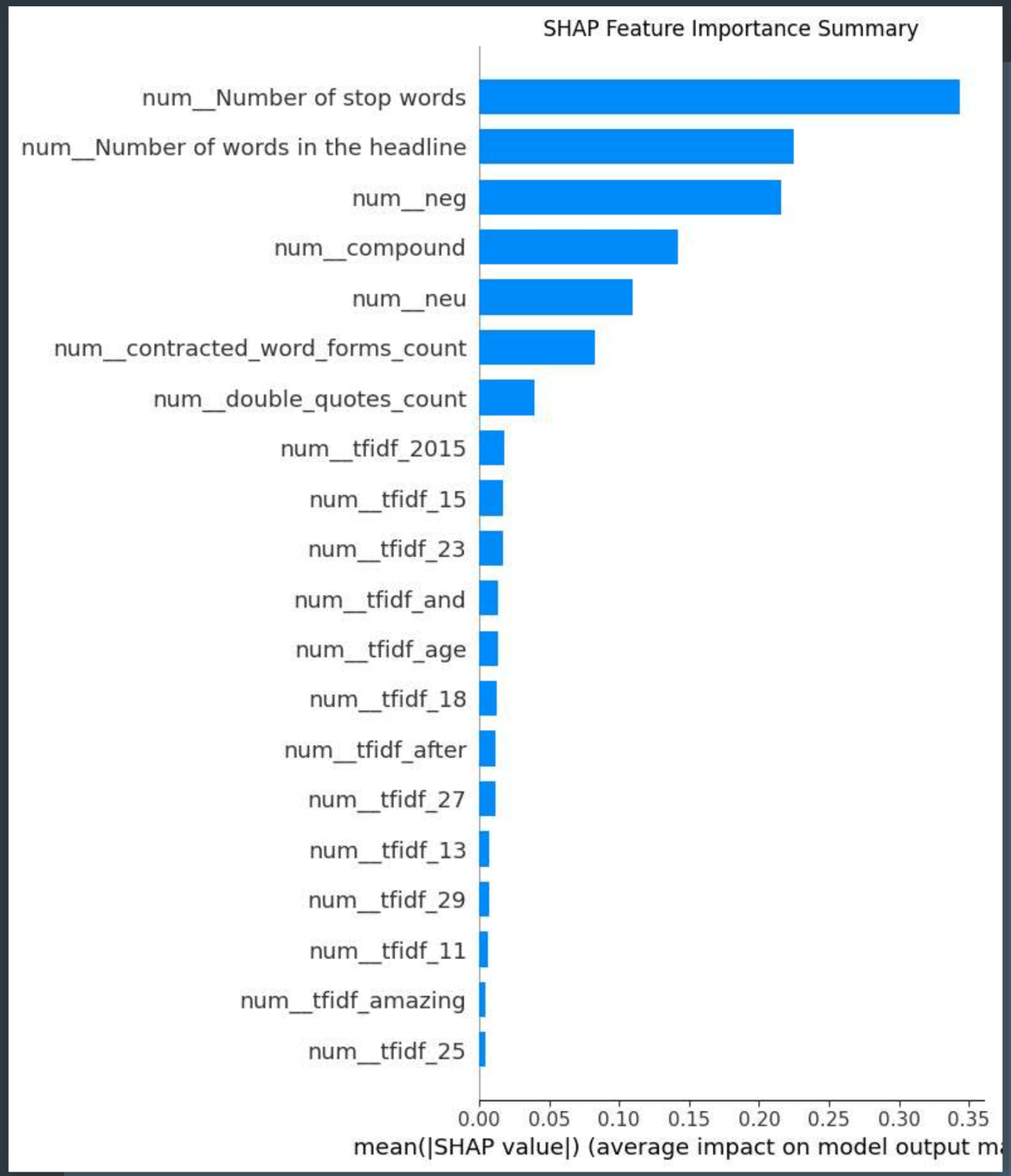
SHAP analysis

- **Lexical Features:**
  1. Number of Words in the Headline
  2. Number of stop words in the headlines
  3. TF-IDF with stopwords
  4. ~~Superlative word list)~~
     - ~~The ratio of the number of stop words to the number of content words~~
- **Syntactic Features:**
  - ~~Syntactic 4-grams~~
- **Semantic and Pragmatic Features:**
  5. Sentiment Analysis
     - ~~Sentence Subjectivity and Objectivity Evaluation~~
- ~~**Readability Features:**~~
  - ~~Informality (Flesch-Kincaid Readability)~~
- **Stylistic and Rhetorical Features:**
  6. Elongated Words
  7. Punctuation -
     - Double quotes
     - Contracted word forms

3. Which features most effectively identify sensation language in news headlines among selected features?

Research questions:

3. Top 20 important features.



4. Spacy stop word list, around 320 words:

{'call', 'upon', 'still', 'nevertheless', 'down', 'every', 'forty', '‘re', 'always', 'whole', 'side', "n't", 'now', 'however', 'an', 'show', 'least', 'give', 'below', 'did', 'sometimes', 'which', "'s", 'nowhere', 'per', 'hereupon', 'yours', 'she', 'moreover', 'eight', 'somewhere', 'within', 'whereby', 'few', 'has', 'so', 'have', 'for', 'noone', 'top', 'were', 'those', 'thence', 'eleven', '**after**', 'no', '’ll', 'others', 'ourselves', 'themselves', 'though', 'that', 'nor', 'just', '’s', 'before', 'had', 'toward', 'another', 'should', 'herself', '**and**', 'these', 'such', 'elsewhere', 'further', 'next', 'indeed', 'bottom', 'anyone', 'his', 'each', 'then', 'both', 'became', 'third', 'whom', '‘ve', 'mine', 'take', 'many', 'anywhere', 'to', 'well', 'thereafter', 'besides', 'almost', 'front', 'fifteen', 'towards', 'none', 'be', 'herein', 'two', 'using', 'whatever', 'please', 'perhaps', 'full', 'ca', 'we', 'latterly', 'here', 'therefore', 'us', 'how', 'was', 'made', 'the', 'or', 'may', '’re', 'namely', "'ve", 'anyway', 'amongst', 'used', 'ever', 'of', 'there', 'than', 'why', 'really', 'whither', 'in', 'only', 'wherein', 'last', 'under', 'own', 'therein', 'go', 'seems', '‘m', 'wherever', 'either', 'someone', 'up', 'doing', 'on', 'rather', 'ours', 'again', 'same', 'over', '‘s', 'latter', 'during', 'done', "'re", 'put', "'m", 'much', 'neither', 'among', 'seemed', 'into', 'once', 'my', 'otherwise', 'part', 'everywhere', 'never', 'myself', 'must', 'will', 'am', 'can', 'else', 'although', 'as', 'beyond', 'are', 'too', 'becomes', 'does', 'a', 'everyone', 'but', 'some', 'regarding', '‘ll', 'against', 'throughout', 'yourselves', 'him', "'d", 'it', 'himself', 'whether', 'move', '’m', 'hereafter', 're', 'while', 'whoever', 'your', 'first', 'amount', 'twelve', 'serious', 'other', 'any', 'off', 'seeming', 'four', 'itself', 'nothing', 'beforehand', 'make', 'out', 'very', 'already', 'various', 'until', 'hers', 'they', 'not', 'them', 'where', 'would', 'since', 'everything', 'at', 'together', 'yet', 'more', 'six', 'back', 'with', 'thereupon', 'becoming', 'around', 'due', 'keep', 'somehow', 'n‘t', 'across', 'all', 'when', 'i', 'empty', 'nine', 'five', 'get', 'see', 'been', 'name', 'between', 'hence', 'ten', 'several', 'from', 'whereupon', 'through', 'hereby', "'ll", 'alone', 'something', 'formerly', 'without', 'above', 'onto', 'except', 'enough', 'become', 'behind', '’d', 'its', 'most', 'n’t', 'might', 'whereas', 'anything', 'if', 'her', 'via', 'fifty', 'is', 'thereby', 'twenty', 'often', 'whereafter', 'their', 'also', 'anyhow', 'cannot', 'our', 'could', 'because', 'who', 'beside', 'by', 'whence', 'being', 'meanwhile', 'this', 'afterwards', 'whenever', 'mostly', 'what', 'one', 'nobody', 'seem', 'less', 'do', '‘d', 'say', 'thus', 'unless', 'along', 'yourself', 'former', 'thru', 'he', 'hundred', 'three', 'sixty', 'me', 'sometime', 'whose', 'you', 'quite', '’ve', 'about', 'even'}

SHAP Value

Positive to SHAP:

1. Number of stop words in the headlines
2. Sentiment Analysis – Negative
3. Punctuation – Contract word
4. Punctuation – Double quote
5. TF-IDF – and

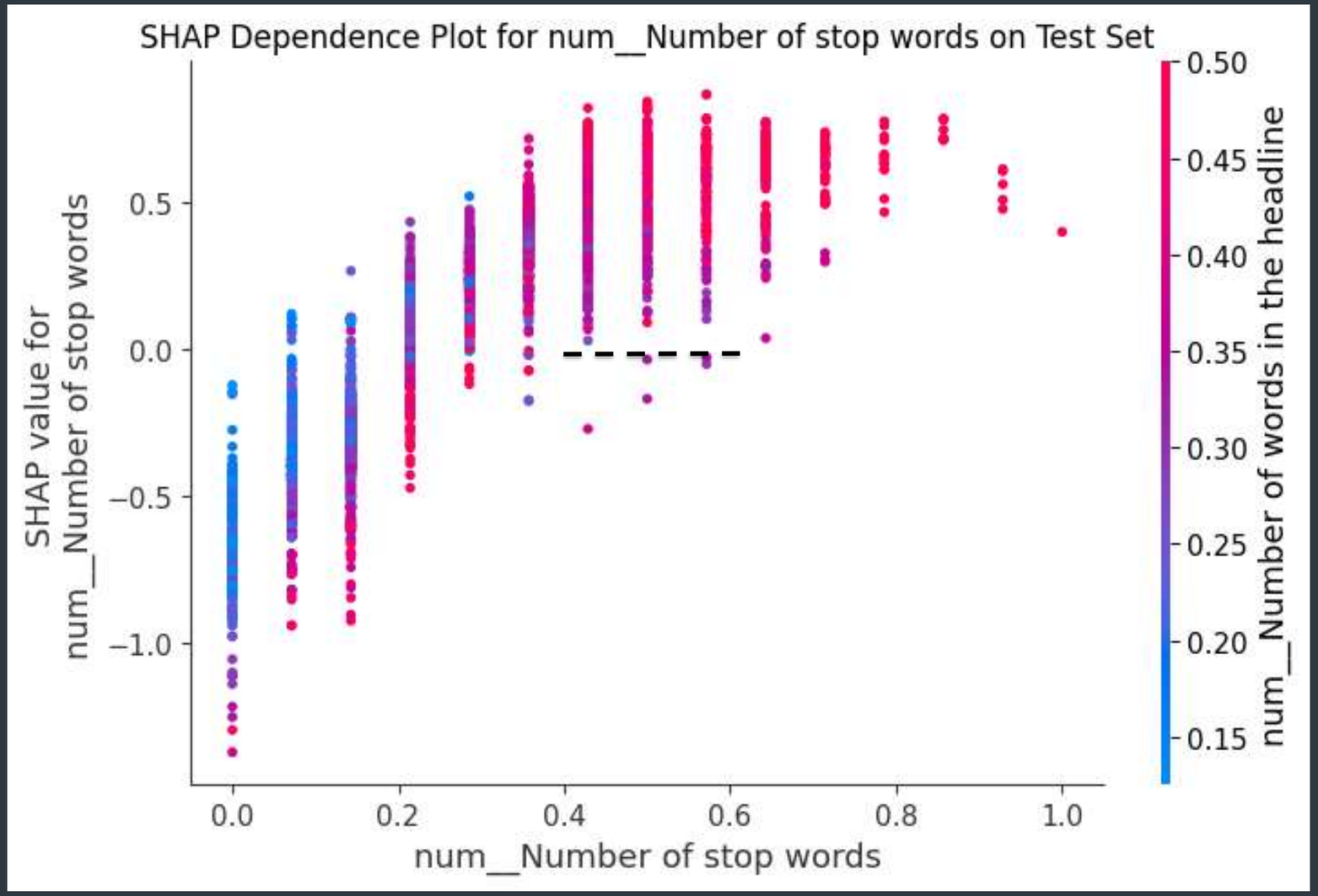
Negative to SHAP:

1. Sentiment Analysis – Compound
2. Sentiment Analysis – Neutral
3. TF-IDF – age
4. TF-IDF – after

Complexity:

1. Number of Words in the Headline

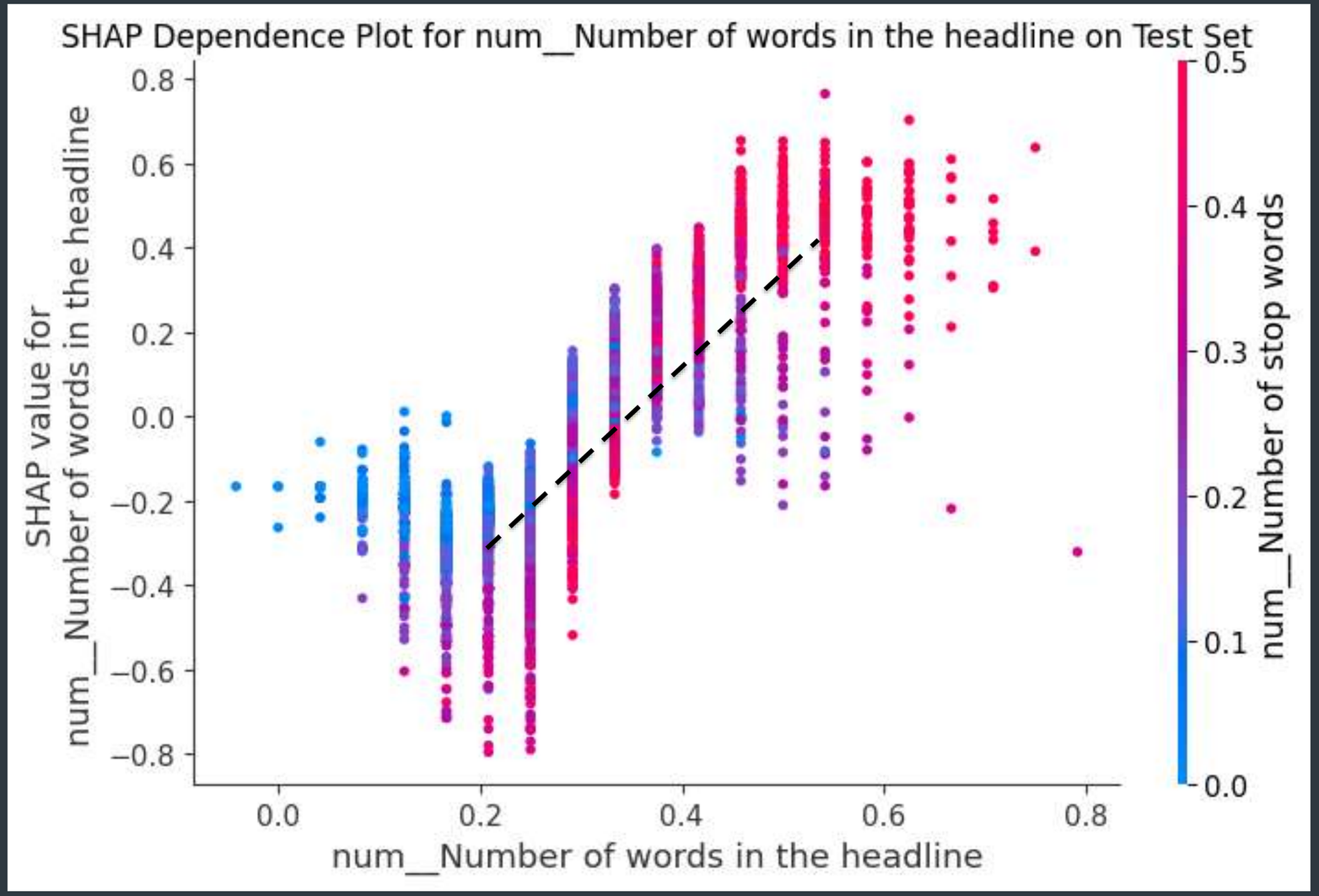
1.



Lexical Features- Number of stop words:

- Ascending
- Non-linear
- Positive impact 0.4~0.6

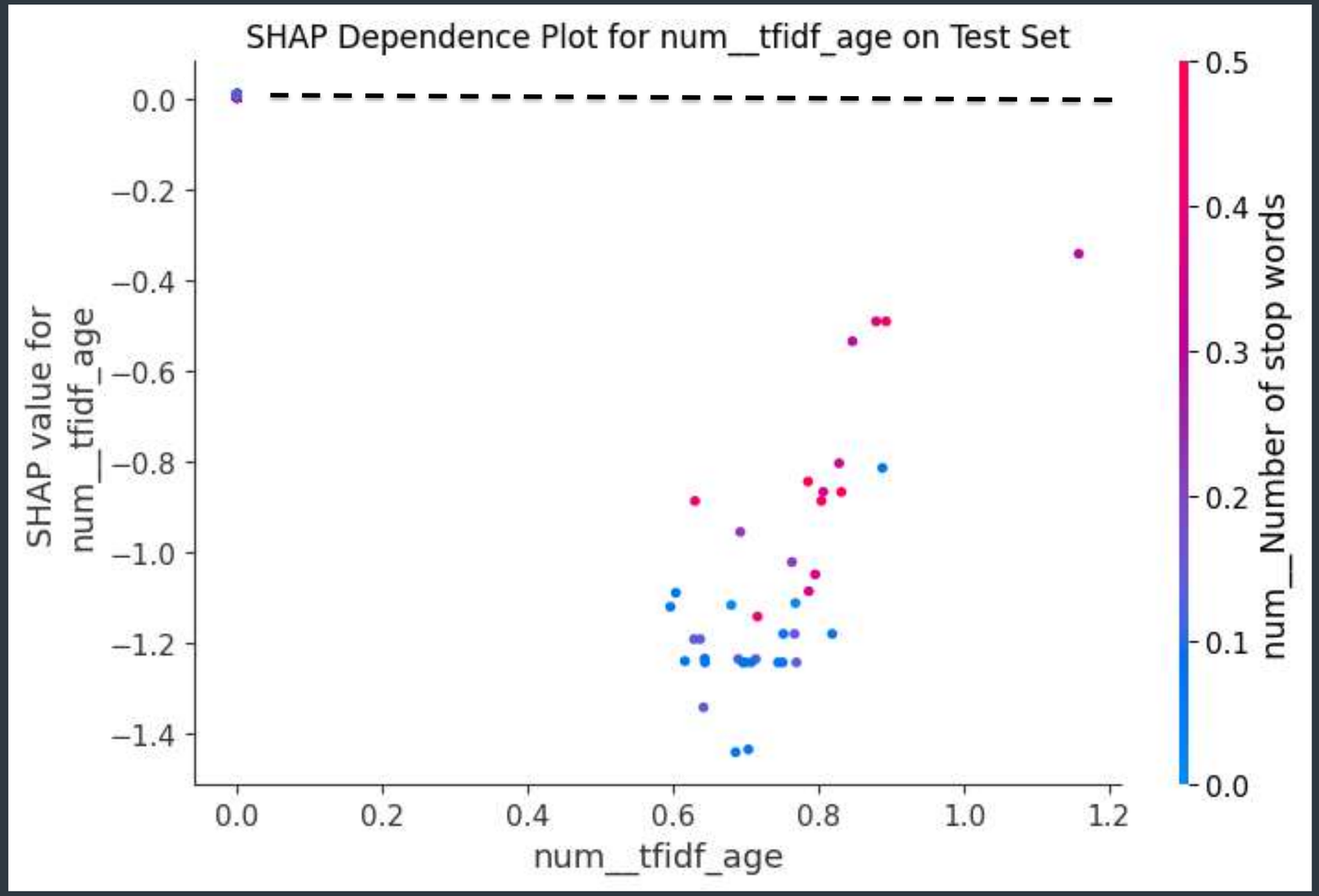
2.



Lexical Features- Number of words in headline:

- Non-linear
- Over 0.2
- Positive 0.4~0.6

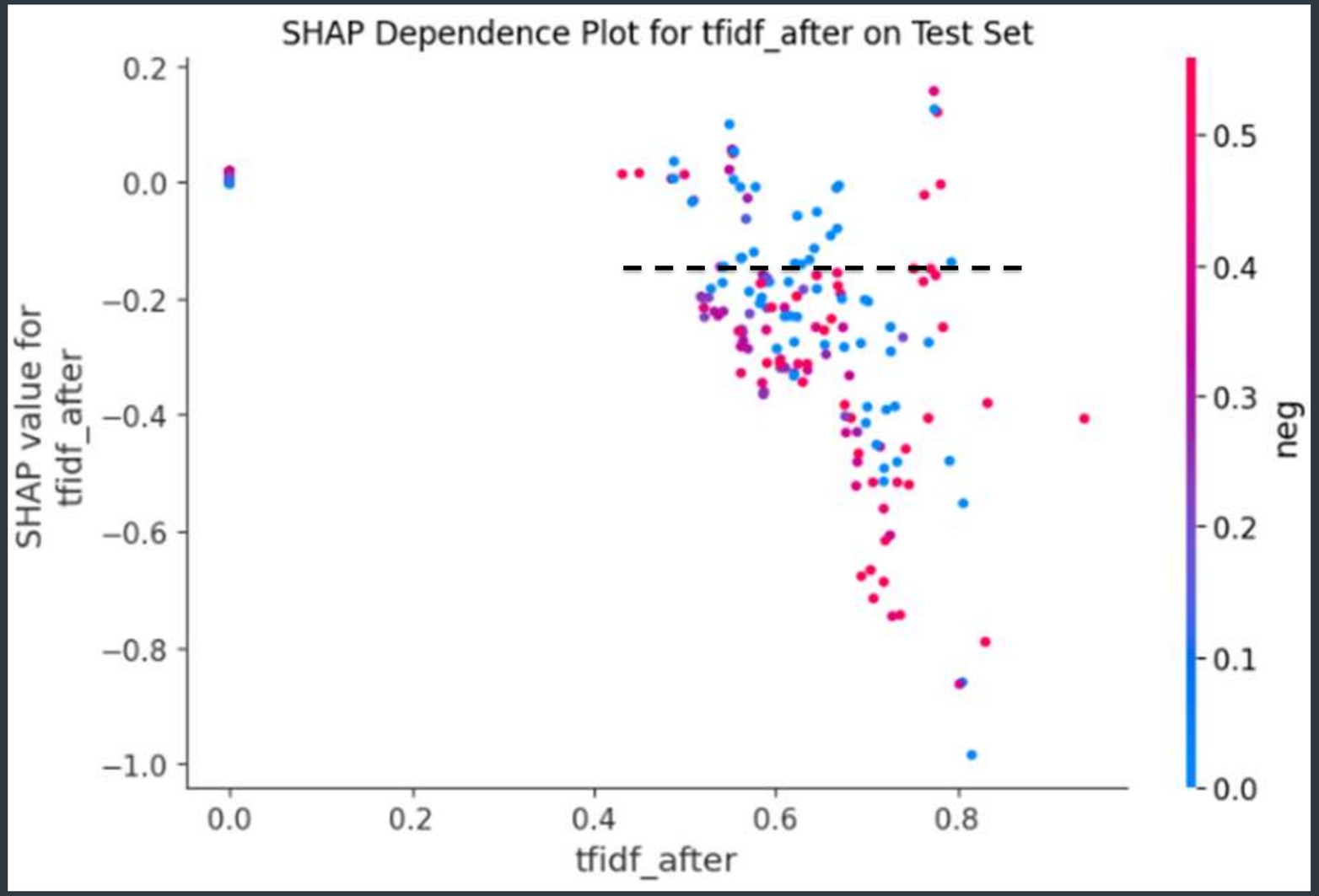
3.



Lexical Features - TF-IDF with stop words, “age”:

- Strong negative
- Non-linear
- High variability

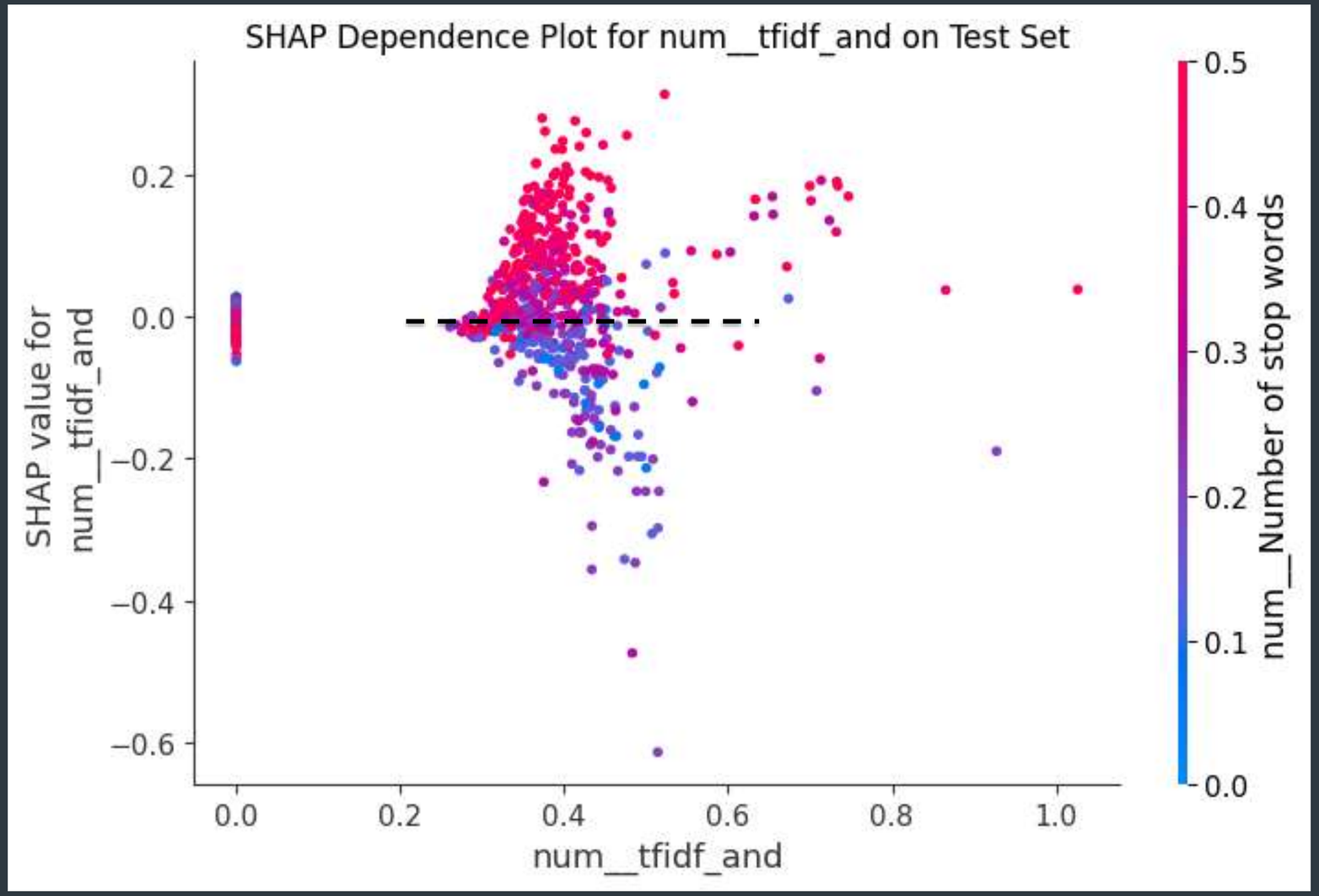
4.



Lexical Features - TF-IDF with stop words, “after”:

- Significant negative
- Non-linear
- Discrete pattern

5.

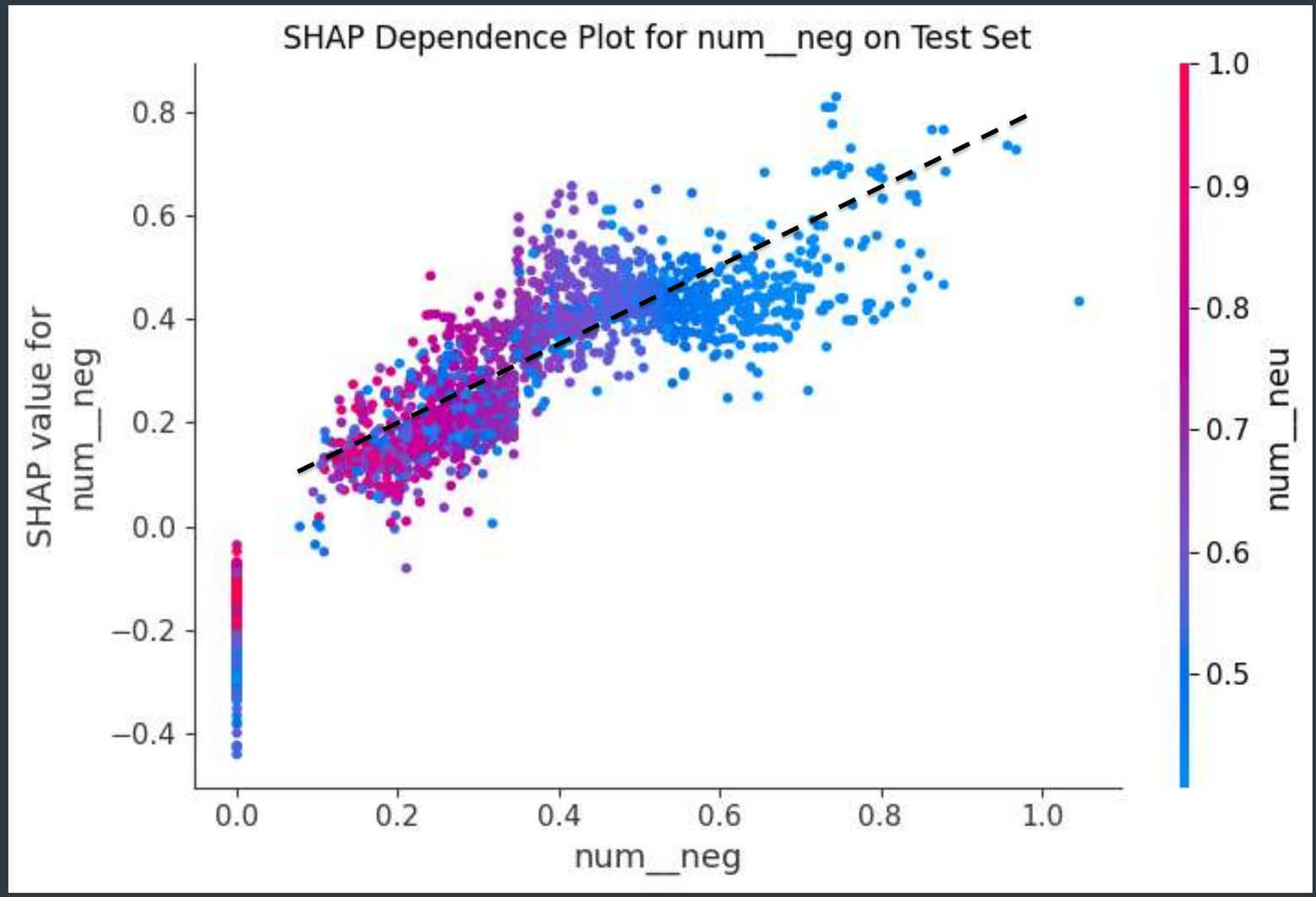


Lexical Features - TF-IDF with stop words, “and”:

- Non-linear
- Centralized, 0.3~0.5
- Red color, 0.3~0.4

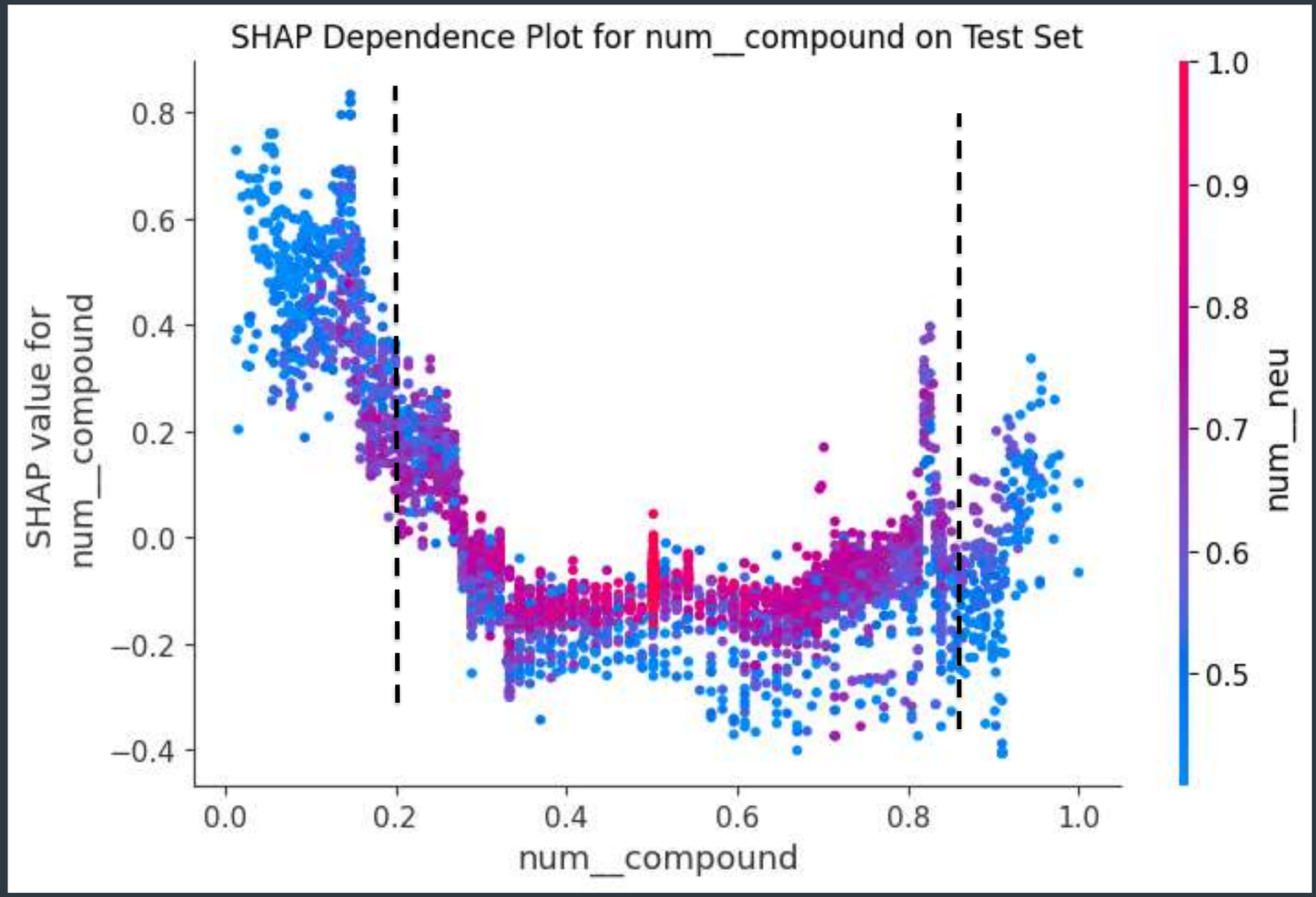
Semantic and Pragmatic Features-negative sentiment:

- Significant positive
- Ascending
- High variability



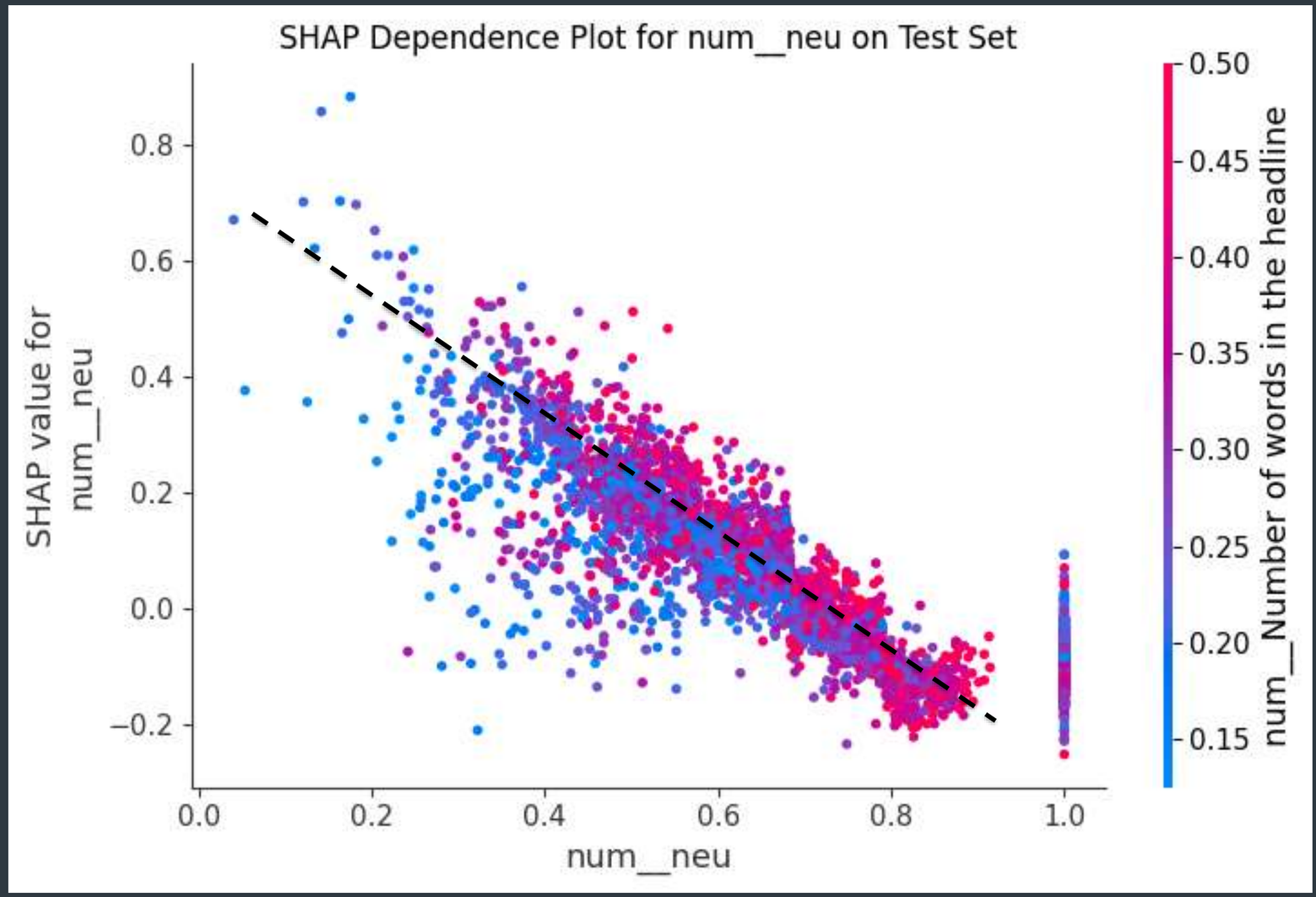
Semantic and Pragmatic Features- compound sentiment:

- Decreasing
- Inflection point, 0.2
- High fluctuations, 0.8

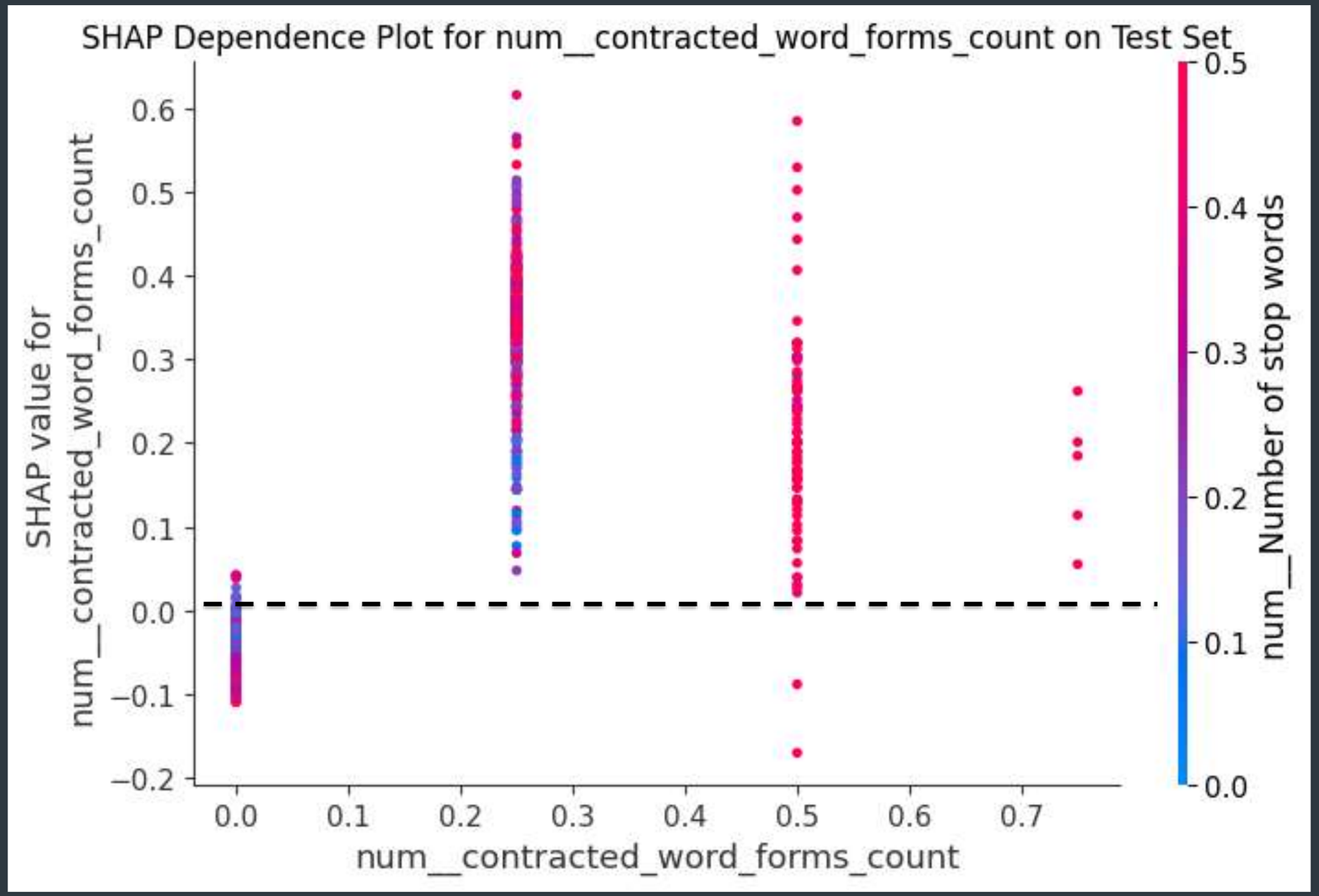


Semantic and Pragmatic Features- neutral sentiment:

- Negative impact
- Neutral, num. word in headline
- High variability



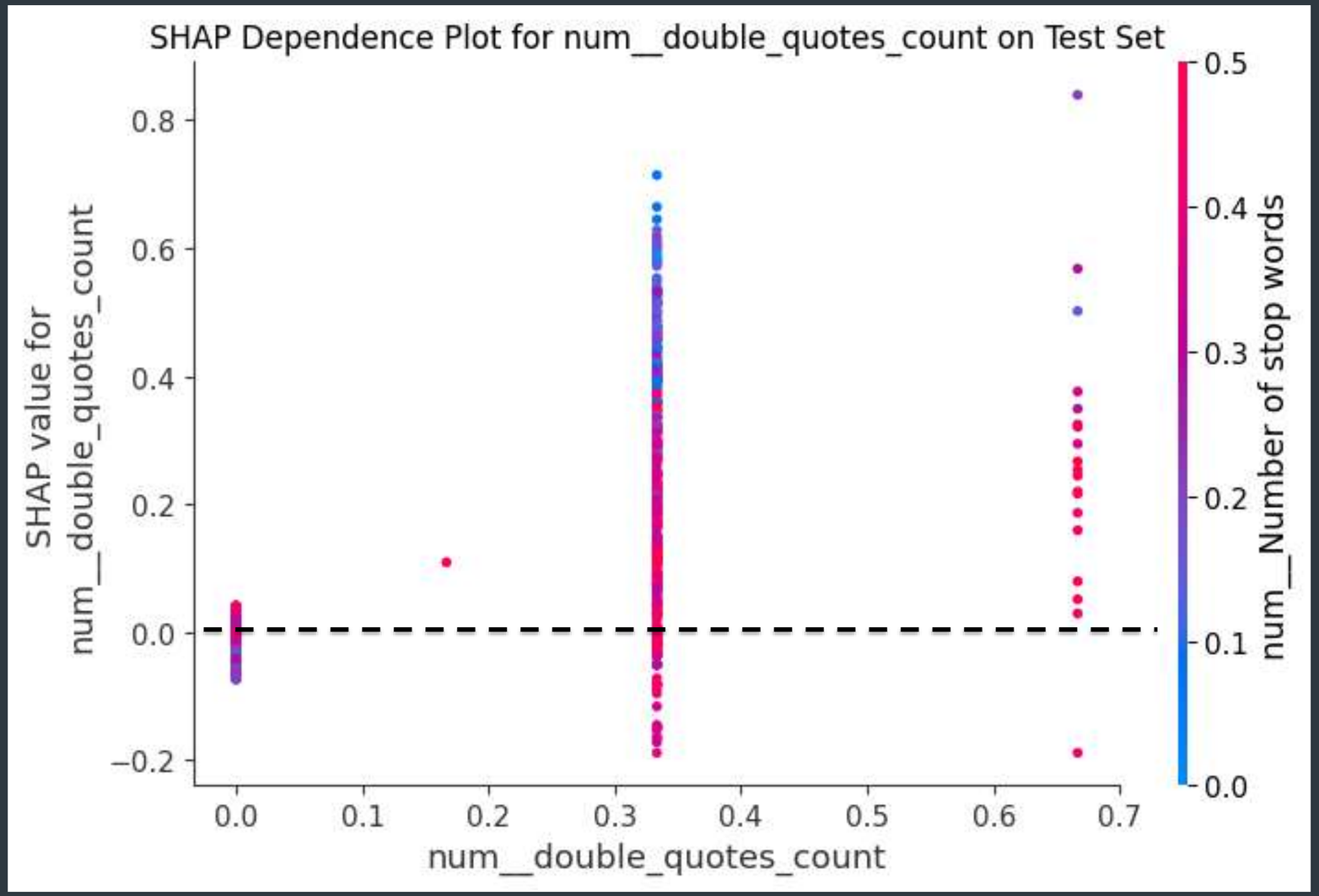
9.



Stylistic Features – Punctuation, contracted word count:

- Positive impact
- Especially 0.5
- Relationships with stop words

10.



Stylistic Features – Punctuation, double quotes count:

- Positive impact
- Relationships with stop words
- Complexity

Our contribution:

1. MIRUKU sensational news headline dataset.
2. Multiple linguistic
3. Robust model
4. Effective features

Future work

- Clustering algorithms
- Bias detection
- Attention mechanism explained