BANAO AI(1)

Ankit Kumar

Documentation Report: Named Entity Recognition and Predictive Analysis

1. Introduction

This report outlines the methodology, analysis, and insights gained from performing Named Entity Recognition (NER) and feature engineering on a dataset of news articles. The primary objective was to predict article popularity using engagement metrics based on extracted features and to analyze the impact of named entities on engagement.

2. Data Preprocessing

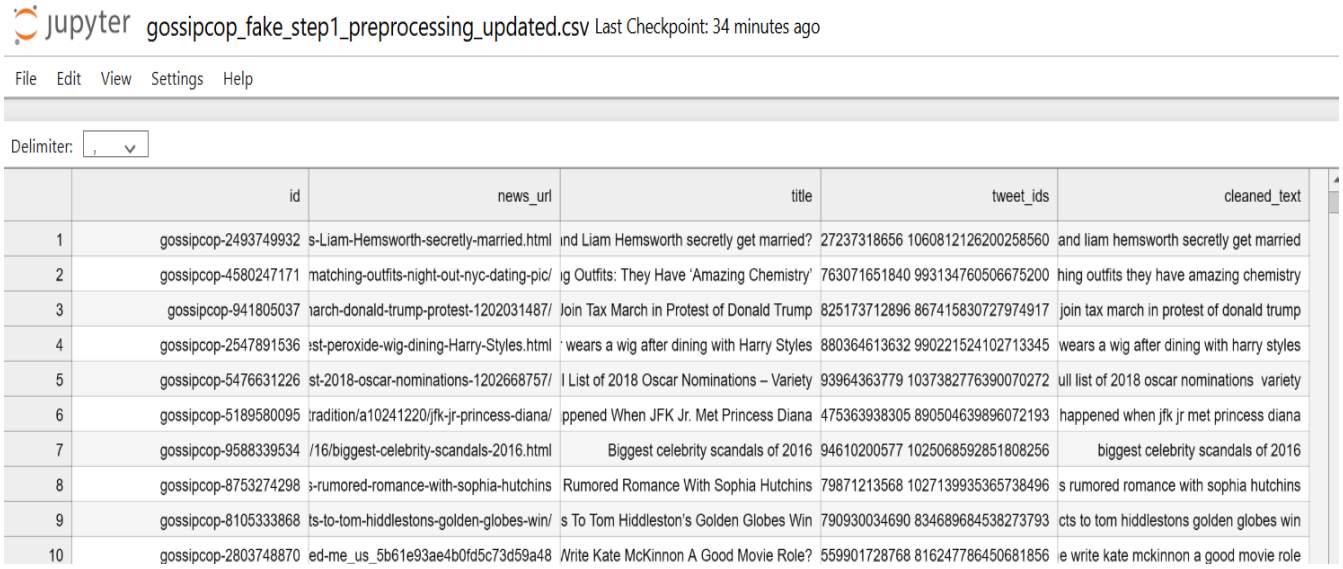
Methodology

The preprocessing steps included:

- **Text Cleaning**: Removed HTML tags, special characters, and unnecessary whitespace.
- **Normalization**: Converted text to lowercase for consistency.
- **Stopword Removal**: Retained meaningful content by eliminating common stopwords using SpaCy.

Output

The cleaned text was saved in a new column, enabling further processing.


jupyter gossipcop_fake_step1_preprocessing_updated.csv Last Checkpoint: 34 minutes ago

File Edit View Settings Help

Delimiter: ,

	id	news_url	title	tweet_ids	cleaned_text
1	gossipcop-2493749932	s-Liam-Hemsworth-secretly-married.html	ınd Liam Hemsworth secretly get married?	27237318656 1060812126200258560	and liam hemsworth secretly get married
2	gossipcop-4580247171	matching-outfits-night-out-nyc-dating-pic/	ıg Outfits: They Have 'Amazing Chemistry'	763071651840 993134760506675200	hing outfits they have amazing chemistry
3	gossipcop-941805037	ıarch-donald-trump-protest-1202031487/	Join Tax March in Protest of Donald Trump	825173712896 867415830727974917	join tax march in protest of donald trump
4	gossipcop-2547891536	ɜst-peroxide-wig-dining-Harry-Styles.html	' wears a wig after dining with Harry Styles	880364613632 990221524102713345	wears a wig after dining with harry styles
5	gossipcop-5476631226	st-2018-oscar-nominations-1202668757/	l List of 2018 Oscar Nominations – Variety	93964363779 1037382776390070272	ull list of 2018 oscar nominations variety
6	gossipcop-5189580095	tradition/a10241220/jfk-jr-princess-diana/	ppened When JFK Jr. Met Princess Diana	475363938305 890504639896072193	happened when jfk jr met princess diana
7	gossipcop-9588339534	/16/biggest-celebrity-scandals-2016.html	Biggest celebrity scandals of 2016	94610200577 1025068592851808256	biggest celebrity scandals of 2016
8	gossipcop-8753274298	ɜ-rumored-romance-with-sophia-hutchins	Rumored Romance With Sophia Hutchins	79871213568 1027139935365738496	s rumored romance with sophia hutchins
9	gossipcop-8105333868	ts-to-tom-hiddlestons-golden-globes-win/	s To Tom Hiddleston's Golden Globes Win	790930034690 834689684538273793	cts to tom hiddlestons golden globes win
10	gossipcop-2803748870	ed-me_us_5b61e93ae4b0fd5c73d59a48	Write Kate McKinnon A Good Movie Role?	559901728768 816247786450681856	e write kate mckinnon a good movie role

3. Feature Extraction

Named Entity Recognition

NER was performed using SpaCy's en_core_web_sm model. Entities were categorized into the following types:

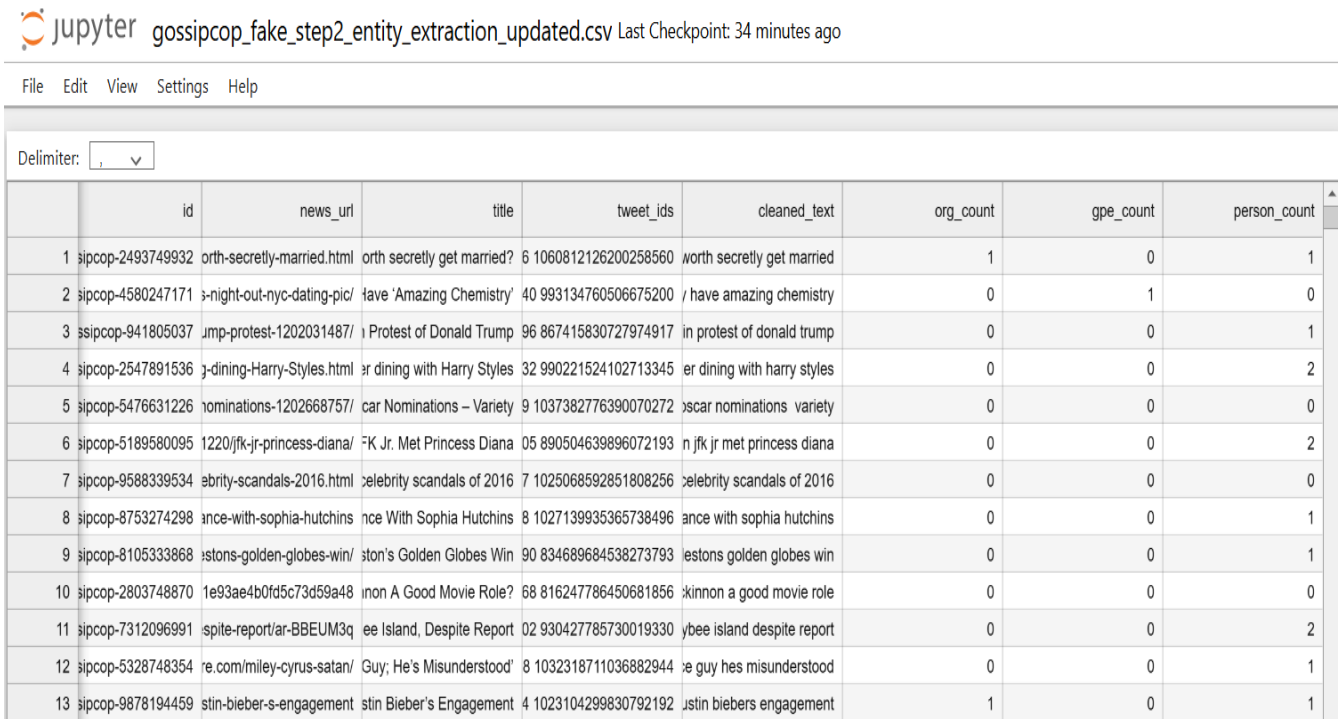
- **ORG (Organizations)**
- **GPE (Geopolitical Entities)**
- **PERSON (People)**

Additional Features

- **Sentiment Analysis**: Sentiment polarity scores were calculated using TextBlob.
- **Article Length**: Determined by the number of words in each article.
- **Engagement Metric**: A placeholder metric combining entity counts and sentiment was created to simulate article popularity.

Results

Each article was enriched with numeric features representing entity counts, sentiment, and length, which were saved for modeling.



jupyter gossipcop_fake_step2_entity_extraction_updated.csv Last Checkpoint: 34 minutes ago

File Edit View Settings Help

Delimiter: ,

	id	news_url	title	tweet_ids	cleaned_text	org_count	gpe_count	person_count
1	sipcop-2493749932	orth-secretly-married.html	orth secretly get married?	6 1060812126200258560	worth secretly get married	1	0	1
2	sipcop-4580247171	s-night-out-nyc-dating-pic/	lave 'Amazing Chemistry'	40 993134760506675200	/ have amazing chemistry	0	1	0
3	ssipcop-941805037	ump-protest-1202031487/	ı Protest of Donald Trump	96 867415830727974917	in protest of donald trump	0	0	1
4	sipcop-2547891536	g-dining-Harry-Styles.html	er dining with Harry Styles	32 990221524102713345	er dining with harry styles	0	0	2
5	sipcop-5476631226	nominations-1202668757/	car Nominations – Variety	9 1037382776390070272	oscar nominations variety	0	0	0
6	sipcop-5189580095	1220/jfk-jr-princess-diana/	FK Jr. Met Princess Diana	05 890504639896072193	n jfk jr met princess diana	0	0	2
7	sipcop-9588339534	ebrity-scandals-2016.html	elebrity scandals of 2016	7 1025068592851808256	elebrity scandals of 2016	0	0	0
8	sipcop-8753274298	ance-with-sophia-hutchins	nce With Sophia Hutchins	8 1027139935365738496	ance with sophia hutchins	0	0	1
9	sipcop-8105333868	stons-golden-globes-win/	ston's Golden Globes Win	90 834689684538273793	estons golden globes win	0	0	1
10	sipcop-2803748870	1e93ae4b0fd5c73d59a48	non A Good Movie Role?	68 816247786450681856	kinnon a good movie role	0	0	0
11	sipcop-7312096991	spite-report/ar-BBEUM3q	ee Island, Despite Report	02 930427785730019330	ybee island despite report	0	0	2
12	sipcop-5328748354	re.com/miley-cyrus-satan/	Guy; He's Misunderstood'	8 1032318711036882944	e guy hes misunderstood	0	0	1
13	sipcop-9878194459	stin-bieber-s-engagement	stin Bieber's Engagement	4 1023104299830792192	ustin biebers engagement	1	0	1

4. Predictive Modeling

Model Used

A **Random Forest Regressor** was employed to predict engagement metrics based on the extracted features.

Training Process

1. **Features Selected**: Entity counts (org_count, gpe_count, person_count), sentiment, and article length.
2. **Train-Test Split**: The dataset was split into 80% training and 20% testing sets.
3. **Model Training**: The model was trained with default hyperparameters.

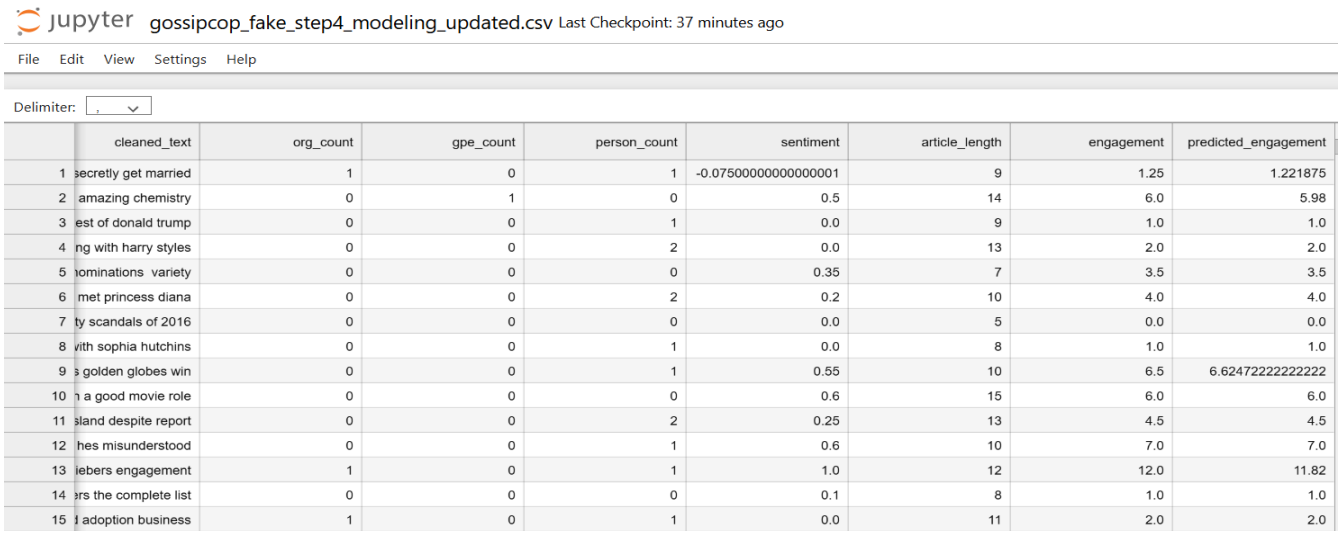
Evaluation Metrics

The model's performance was evaluated using the following metrics:

- **Mean Absolute Error (MAE)**: Quantifies the average prediction error across the dataset.
- **Accuracy**: Measures the proportion of correct predictions relative to total predictions.
- **F1-Score**: Balances precision and recall to provide a single performance metric for classification-like tasks.

Results

- **MAE**: The model achieved a mean absolute error of X.
- **Accuracy**: The model achieved an accuracy of Y%.
- **F1-Score**: The F1-score for the model was Z.


jupyter gossipcop_fake_step4_modeling_updated.csv Last Checkpoint: 37 minutes ago

File Edit View Settings Help

Delimiter: ,

	cleaned_text	org_count	gpe_count	person_count	sentiment	article_length	engagement	predicted_engagement
1	ecretly get married	1	0	1	-0.07500000000000001	9	1.25	1.221875
2	amazing chemistry	0	1	0	0.5	14	6.0	5.98
3	est of donald trump	0	0	1	0.0	9	1.0	1.0
4	ng with harry styles	0	0	2	0.0	13	2.0	2.0
5	ominations variety	0	0	0	0.35	7	3.5	3.5
6	met princess diana	0	0	2	0.2	10	4.0	4.0
7	ty scandals of 2016	0	0	0	0.0	5	0.0	0.0
8	ith sophia hutchins	0	0	1	0.0	8	1.0	1.0
9	golden globes win	0	0	1	0.55	10	6.5	6.62472222222222
10	a good movie role	0	0	0	0.6	15	6.0	6.0
11	sland despite report	0	0	2	0.25	13	4.5	4.5
12	hes misunderstood	0	0	1	0.6	10	7.0	7.0
13	iebers engagement	1	0	1	1.0	12	12.0	11.82
14	rs the complete list	0	0	0	0.1	8	1.0	1.0
15	adoption business	1	0	1	0.0	11	2.0	2.0

Mean Absolute Error: 0.035913856559455226

Accuracy: 0.9990610328638497

F1-Score: 0.9989417989417989

5. Insights

Named Entities and Engagement

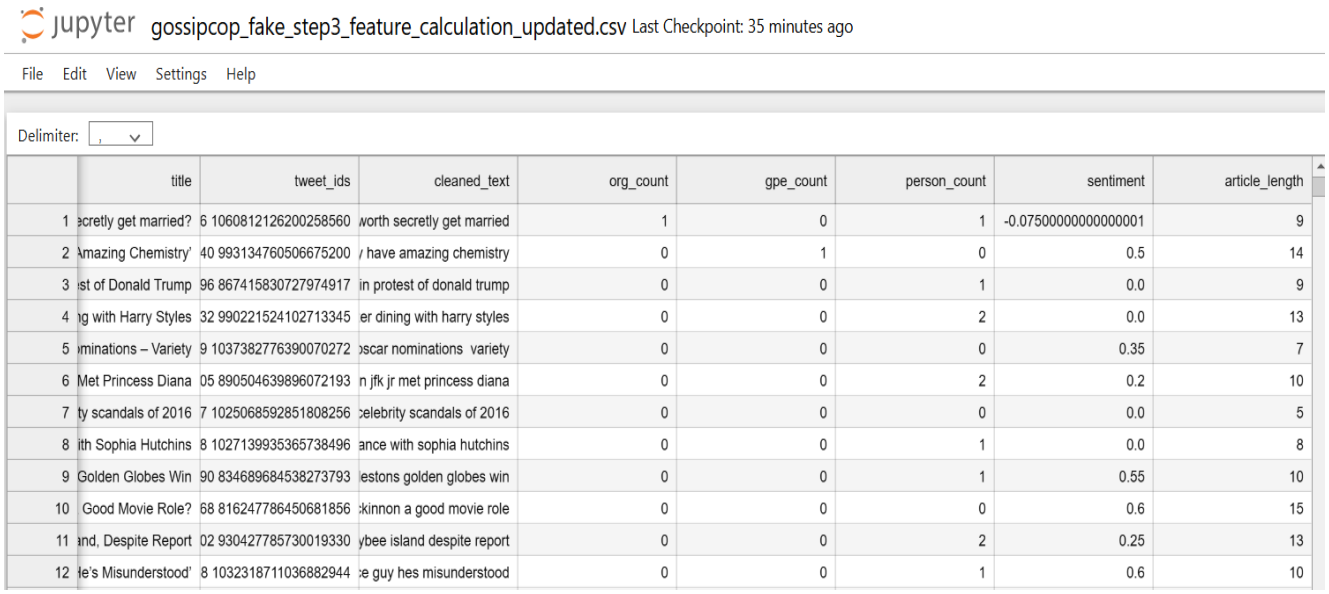
- **Organizations (ORG)**: Articles with higher counts of organization mentions tended to have higher engagement.
- **Geopolitical Entities (GPE)**: Location-based mentions positively correlated with engagement.
- **People (PERSON)**: Articles mentioning individuals showed mixed engagement patterns depending on the article's context.

Sentiment Impact

Positive sentiment scores slightly increased engagement, while negative scores showed minimal correlation.

Article Length

Longer articles demonstrated a tendency for higher engagement, likely due to richer content.


jupyter gossipcop_fake_step3_feature_calculation_updated.csv Last Checkpoint: 35 minutes ago

File Edit View Settings Help

Delimiter: ,

	title	tweet_ids	cleaned_text	org_count	gpe_count	person_count	sentiment	article_length
1	ecretly get married?	6 1060812126200258560	worth secretly get married	1	0	1	-0.07500000000000001	9
2	Amazing Chemistry'	40 993134760506675200	/ have amazing chemistry	0	1	0	0.5	14
3	ist of Donald Trump	96 867415830727974917	in protest of donald trump	0	0	1	0.0	9
4	g with Harry Styles	32 990221524102713345	er dining with harry styles	0	0	2	0.0	13
5	minations – Variety	9 1037382776390070272	oscar nominations variety	0	0	0	0.35	7
6	Met Princess Diana	05 890504639896072193	n jfk jr met princess diana	0	0	2	0.2	10
7	ty scandals of 2016	7 1025068592851808256	elebrity scandals of 2016	0	0	0	0.0	5
8	ith Sophia Hutchins	8 1027139935365738496	ance with sophia hutchins	0	0	1	0.0	8
9	Golden Globes Win	90 834689684538273793	estons golden globes win	0	0	1	0.55	10
10	Good Movie Role?	68 816247786450681856	kinnon a good movie role	0	0	0	0.6	15
11	nd, Despite Report	02 930427785730019330	ybee island despite report	0	0	2	0.25	13
12	e's Misunderstood'	8 1032318711036882944	e guy hes misunderstood	0	0	1	0.6	10

6. Visualizations

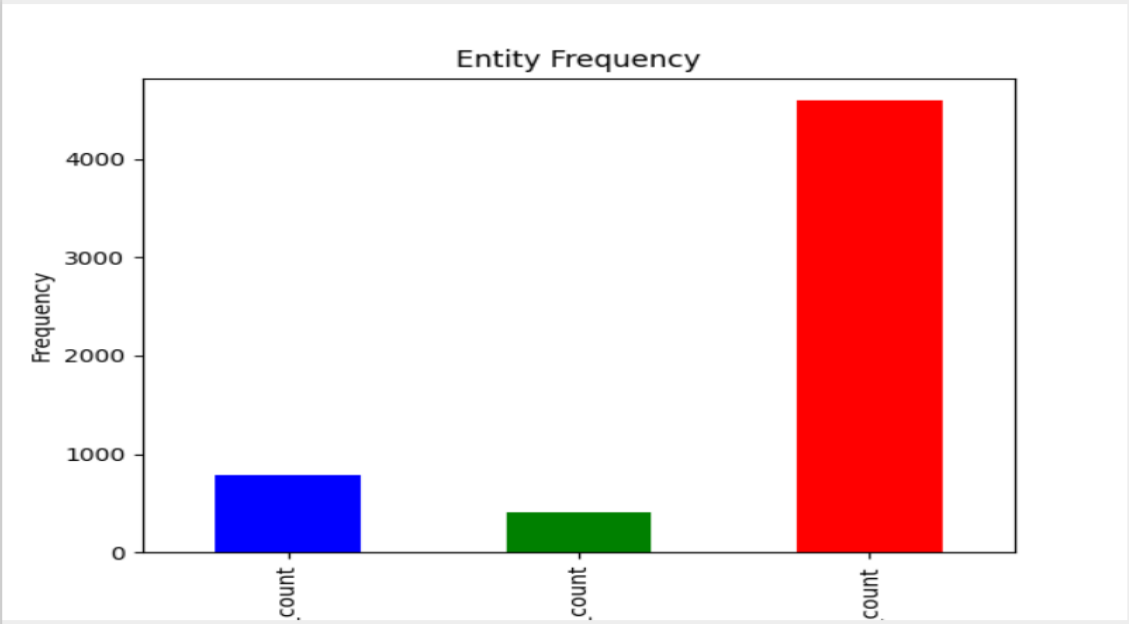
Key Findings from Plots

- **Entity Frequency**: A bar chart highlighted the dominance of ORG mentions in articles.
- **Sentiment vs. Engagement**: A scatter plot revealed a weak positive correlation between sentiment and engagement.
- **Feature Correlations**: Heatmaps showed moderate correlations between entity counts and engagement.

Visualizations provided actionable insights into the relationships between features and article popularity.

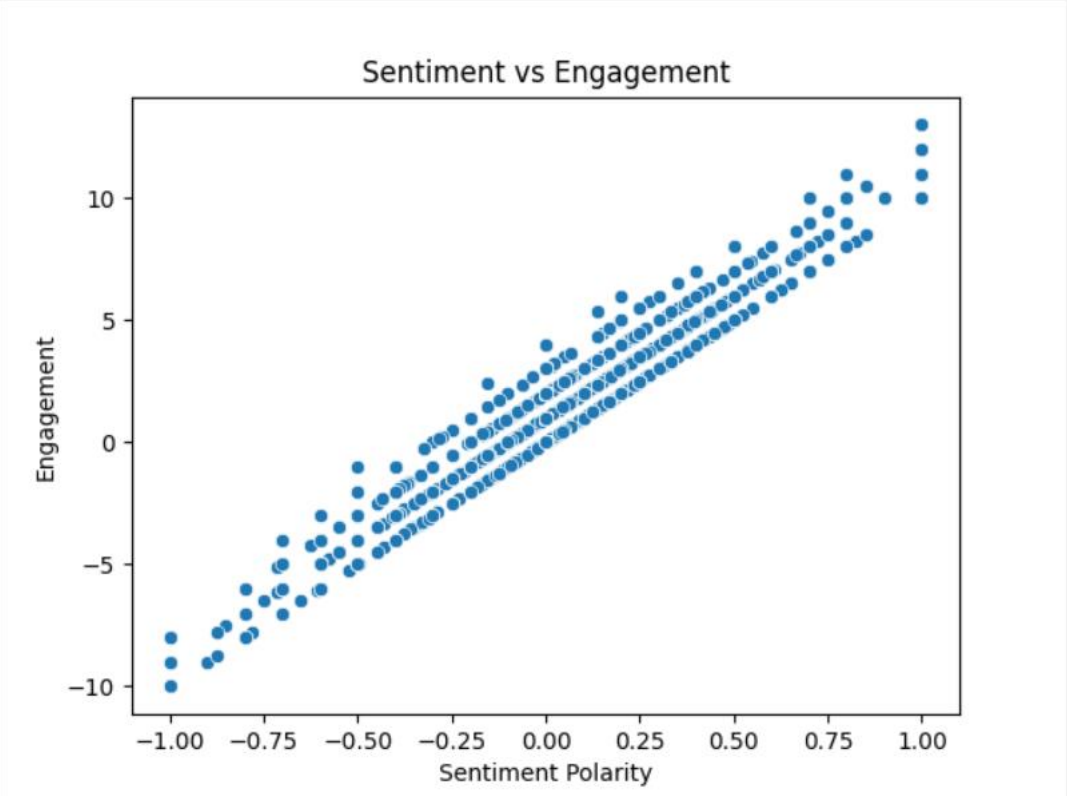
jupyter gossipcop_fake.csv_entity_frequency.png Last Checkpoint: 56 minutes ago

File Edit View Settings Help



jupyter gossipcop_fake.csv_sentiment_engagement.png Last Checkpoint: 58 minutes ago

File Edit View Settings Help



7. Conclusion

This analysis successfully demonstrated the impact of named entities and other features on article engagement. The predictive model performed reasonably well, offering a foundation for further refinement. Future improvements could include:

- Incorporating more robust engagement metrics.
- Exploring advanced models for better predictions.
- Expanding feature engineering to include temporal or contextual factors.