Sarcasm Detection in News Headlines

Group 16 - Snack Overflow

Terezia Olsiakova - 12331438
Arian Etemadihaghighi - 12242985
Viktoriia Ovsianik - 12217985
Maximilian Scheiblauer - 11776651

Contents

Task

Goals of our project are:

1. Develop an approach to classify headlines as sarcastic & non-sarcastic (not fake news!)
2. Explore misclassified cases and understand reasons for incorrect classification
3. Understand if there are any patterns that “The Onion” website follows to create sarcastic headlines

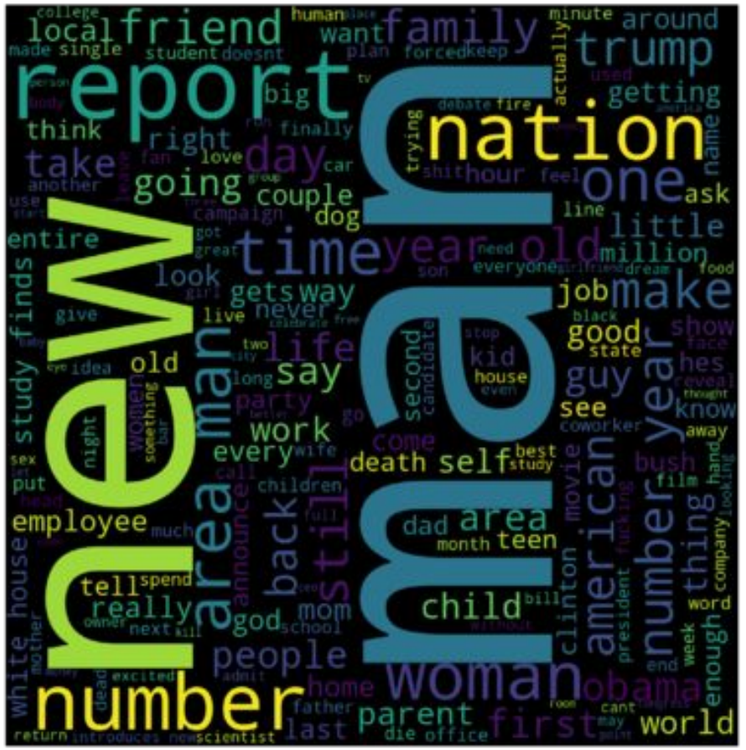
Data

1. **News Headlines Dataset for Sarcasm Detection***

- 28,619 annotated headlines
  - sarcastic from “The Onion”
  - non-sarcastic from “The Huffington Post”
- Balanced data
  - 13,635 sarcastic, 14,984 non-sarcastic

2. **Headlines Generated by ChatGPT (Additional data)**

Using ChatGPT we generated 1000 news headlines (sarcastic & non-sarcastic) and 1000 news headlines in a style similar to “The Onion” and “The Huffington Post”



Sarcasm headlines word cloud

* Data and Diagram Source: [News-Headlines-Dataset-For-Sarcasm-Detection]

Contents

Preprocessing

1. The quality of the dataset is quite good:

- No missing values, just a few duplicates
- Balanced classes

2. Not much of a preprocessing + did not exclude stopwords

3. Exported dataset in CoNLL-U & JSON format

Baselines

- Deep learning-based:
  - DistilBERT
  - Neural Network
- Non-deep learning based:
  - Bag of Words (Naive Bayes)
  - Logistic regression

Results:

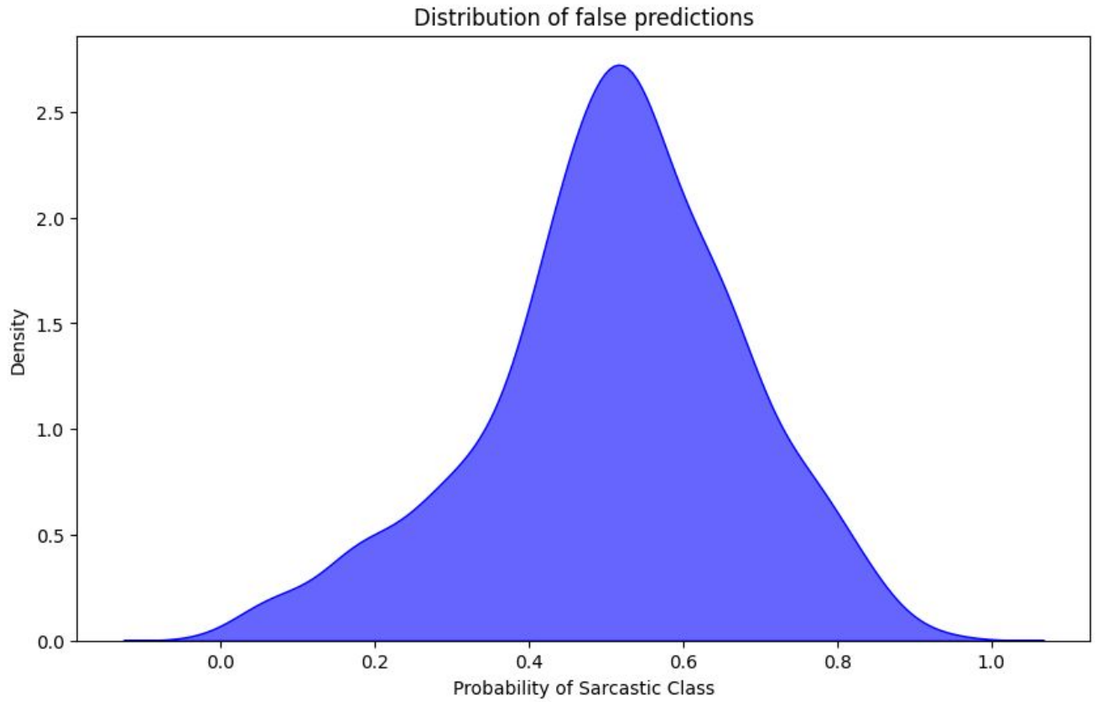
Model	Precision	Recall	F1-score
Naive Bayes	0.85	0.85	0.85
Neural Network	0.84	0.83	0.83
Logistic Regression	0.83	0.83	0.83
DistilBERT	0.82	0.75	0.78

Contents

Error Analysis

Error Distribution

Patterns in misclassified headlines



Wrong Classifications

Examples of baseline failures

Category	Index	Headline	True Class	Predicted Class	Probability
Egregious Error	341	pizza rat have return	0	1	0.935584
Egregious Error	190	pope visit one of italy 's most dangerous area	0	1	0.912140
Egregious Error	327	18 alternative to those play out dorm room poster	0	1	0.882580
Egregious Error	95	woman meet george w. bush while report for jury duty	0	1	0.854374
Close Call	402	retire research chimp be really enjoy their new home	0	1	0.500631
Close Call	264	congress pass natural disaster digital enhancement funding	1	0	0.499686
Close Call	172	radical islamist preacher anjem choudary find guilty of support isis	0	1	0.501011
Close Call	1	call from daycare can not be good	1	0	0.499253

Insights from Error Analysis and Baselines

Key Takeaways

- Class not obvious without real knowledge
- Some words have unintuitive weights
- Recognizable patterns

Performance Patterns

- BoW Methods perform very good
- Errors are well distributed
- Balanced predictions

Areas for Improvement

- Include more information than just words
- Are the found patterns writing style or ‘real’ sarcasm?

Choice of Advanced Approaches

Syntactic Analysis

- POS-tags, dependency trees
- Transitional probabilities

New Data

- Onion Headlines, tweets, other sarcastic headlines
- See if models learned sarcastic patterns or Onion writing style

Pattern Analysis

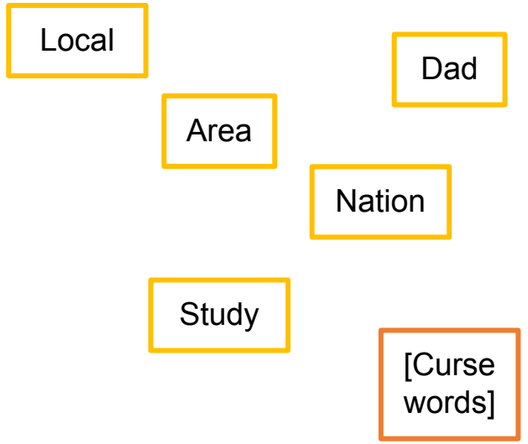
- Onion often parodies the tone and structure of serious journalism and traditional news outlets, while delivering over-the-top or ironic statements:

"7.1 Billion Demonstrate in Favor of Global Warming."

"Area Man Nervously Asks Girlfriend if She'll Settle."

"Guard in Video Game under Strict Orders to Repeatedly Pace Same Stretch of Hallway."

- There are certain patterns that are used often:
  - area man puts on some nice pants for once in his life
  - nation demands more slow-motion footage of syrup cascading onto pancakes
  - local woman has story about how she got these shoes
  - hillary clinton to nation: 'do not fuck this up for me'
  - study finds employees most productive when they can set their own salaries



- But these are a minority, Onion has a diverse style:
  - edge of table victorious over toddler
- And anyway HuffPost headlines can have similar patterns:
  - local officials grapple with trump's fearmongering on 'sanctuary city' policies

Pattern Analysis

- Onion's style hinders non-semantic pattern classification
- For instance:

"7.1 Billion Demonstrate in Favor of Global Warming." and "U.S. Upset After Aliens Land in Italy."

can become

"70,000 Demonstrate in Favor of [some cause]." and "U.S. Upset After [some politician] Lands in Italy."

and be actual non-sarcastic headlines, or

"Must-See TV Shows You Can't Miss This Fall."

can become sarcastic by removing a "not":

"Must-See TV Shows You Can Miss This Fall."

- So classification methods must take advantage of semantics
- And common patterns of Onion (like 'local man…') are already mostly captured by the Naive Bayes Bag of Words approach

Contents

Testing on New Data: ChatGPT

1. 1000 headlines:
   a. 500 similar to Onion

"Area Man Claims He's 'Listening' By Repeating Last Three Words You Said."

"New Study Finds Majority Of Clouds Prefer To Rain Over Picnics."

"Ghost Refuses To Haunt House With Open Floor Plan."

   b. 500 similar to HuffPost

"The Incredible Story of a Librarian Who Saved Hundreds of Books From Destruction."

"How to Actually Make Your New Year's Resolutions Stick."

"How I Found Myself After Losing My Job."

- Our Naive Bayes model from before, performs even better on this fake dataset!

Dataset	Precision (Sarcastic)	Recall (Sarcastic)	F1-score (Sarcastic)
Original	0.85	0.85	0.85
ChatGPT (faking Onion and HuffPost)	0.95	0.89	0.92

Testing on New Data: ChatGPT

1. 1000 headlines:
   a. 500 similar to Onion

"Area Man Claims He's 'Listening' By Repeating Last Three Words You Said."

"New Study Finds Majority Of Clouds Prefer To Rain Over Picnics."

"Ghost Refuses To Haunt House With Open Floor Plan."

   b. 500 similar to HuffPost

"The Incredible Story of a Librarian Who Saved Hundreds of Books From Destruction."

"How to Actually Make Your New Year's Resolutions Stick."

"How I Found Myself After Losing My Job."

- This happens because ChatGPT exaggerates their styles even further!
  - Four common patterns including words like *"nation"*, *"area man"*, *"local"*, and *"study"* cover the headlines with much higher recall

Dataset	Precision (Sarcastic)	Recall (Sarcastic)
Original	0.79	0.08
ChatGPT (faking Onion and HuffPost)	0.94	0.37

Testing on New Data: ChatGPT

2. 1000 headlines:
   a. 500 sarcastic news headlines

"Local Man Heroically Holds Door Open, Hopes for Medal of Honor."

"Facebook Argument Changes No Opinions, Shockingly."

"Remote Control Found in Couch Cushions After Missing for Six Years, Declared a Miracle."

   b. 500 news headlines

"The Surprising Link Between Gut Health and Mental Wellness."

"Why Journaling Is Becoming a Morning Ritual for Millions."

"What's Holding Back a Truly Wireless World?"

- Our Naive Bayes model from before, performs worse on this generic headlines!

Dataset	Precision (Sarcastic)	Recall (Sarcastic)	F1-score (Sarcastic)
Original	0.85	0.85	0.85
ChatGPT (faking Onion and HuffPost)	0.95	0.89	0.92
ChatGPT (faking generic headlines)	0.71	0.74	0.73

Testing on New Data: ChatGPT

2. 1000 headlines:
   a. 500 sarcastic news headlines

"Local Man Heroically Holds Door Open, Hopes for Medal of Honor."

"Facebook Argument Changes No Opinions, Shockingly."

"Remote Control Found in Couch Cushions After Missing for Six Years, Declared a Miracle."

   b. 500 news headlines

"The Surprising Link Between Gut Health and Mental Wellness."

"Why Journaling Is Becoming a Morning Ritual for Millions."

"What's Holding Back a Truly Wireless World?"

- And the patterns don't cover such a large fraction of the headlines anymore

Dataset	Precision (Sarcastic)	Recall (Sarcastic)
Original	0.79	0.08
ChatGPT (faking Onion and HuffPost)	0.94	0.37
ChatGPT (faking generic headlines)	0.71	0.14

Testing on New Data: Tweets

- Tweets annotated by their authors

- Preprocessing:
  - removed @user tags
  - same as headlines data

- Comparison to headlines:
  - short form content
  - more subjective, emotional
  - unprofessional writing

Results from Naive Bayes:

Class	Precision	Recall	F1-score
non-sarcastic	0.5	0.86	0.63
sarcastic	0.52	0.15	0.23

Are we training to detect sarcasm or Onion's writing style?

Syntactic Analysis

Workflow

1. Dependency Parsing to get POS Tags and syntax trees

2. Come up with syntactic features

3. Do feature extraction
    a. Recursive Feature Elimination (RFE)
    b. Random Forest

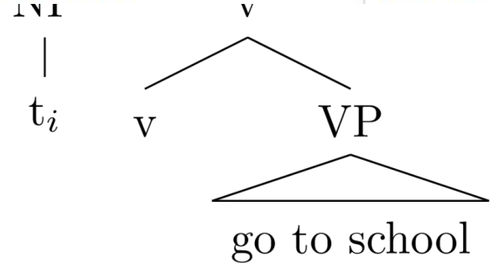
4. Train Models on the top 5 features
    a. Logistic Regression
    b. Random Forest

5. Evaluate the Syntactic Models

6. Combine via a Voting-Rule (hard-voting)

Feature Name	ANOVA_F	RFE Score	RandomForest Score	Average Score
ratio_verb	1.000000	1.000000	0.619594	0.873198
ratio_unique_pos	0.655548	1.000000	0.536356	0.730634
reversed_probability	0.000000	0.100000	0.979131	0.359710
syntactic_depth	0.513614	0.035714	0.281690	0.277006

Metric	Non-Sarcastic (Class 0)	Sarcastic (Class 1)	Macro Avg
Precision	0.73	0.73	0.73
Recall	0.76	0.71	0.73
F1-Score	0.74	0.72	0.73
Support	2218	2075	4293



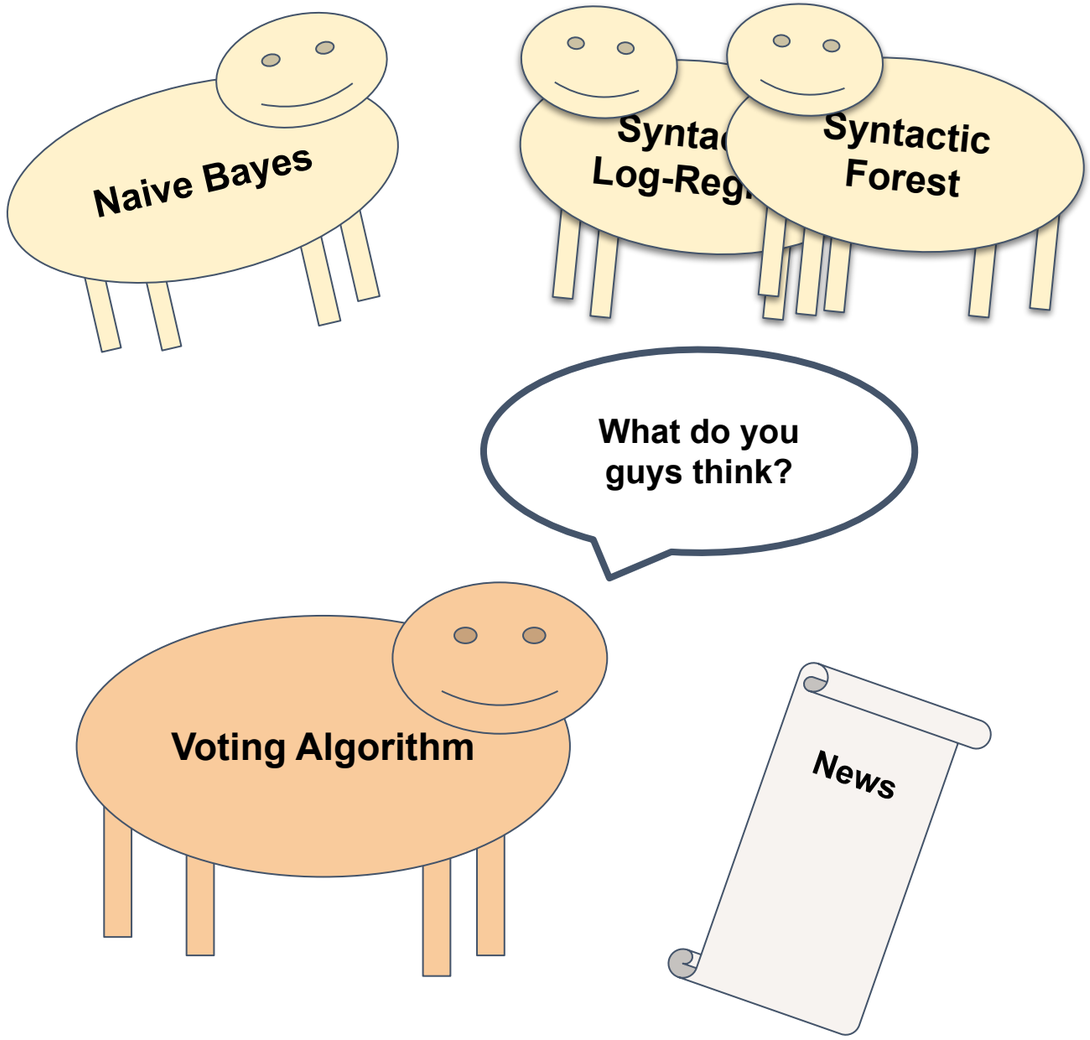
Syntactic Analysis - Takeaways

- Different models disagree on syntactic features

- Voting Algorithm
  - Ensemble of three Classifiers
  - Every Classifier predicts a class
  - Majority wins

- Does not perform as well a BoW Naive Bayes

- Possible improvements:
  - More Classifiers (like Random Forest, XGBoost)
  - Weighting the classifiers (Metamodel)



Voting - Performance on new data

- Performance on Tweets data is very bad
  - Syntactic Features do not apply to tweeting language

- Performance on Onion-Style and Huff Post-Style Headlines
  - Like Naive Bayes even better than validation data

- Performance on generic headlines has also worse than validation data

Dataset	Precision (Sarcastic)	Recall (Sarcastic)	F1-score (Sarcastic)
Tweets	0.62	0.59	0.60
ChatGPT (faking Onion and HuffPost)	0.88	0.87	0.87
ChatGPT (faking generic headlines)	0.68	0.68	0.68

Contents

Challenges and Insights

Challenges

- Bag of Words models like Naive Bayes are very powerful and it is hard to come up with something better
- Sarcasm is not well defined and very hard to generalize over multiple data sources

Insights

- When working on sarcasm detection topic including encoded real world information might be beneficial to improve the quality of classification
- Including sarcastic data from multiple sources would help to improve generalization of the model since websites and newspapers might have their unique style of headlines

Thank You!

Questions?