Duluth at SemEval-2019 Task 4: Hyperpartisan News Detection using Logistic Regression and Convolutional Neural Network

Saptarshi Sengupta & Ted Pedersen

Department of Computer Science
University of Minnesota, Duluth

1. Introduction

- The growing popularity of social media has made the proliferation of fake news easily possible.
- ***Hyperpartisan News*** reports events with extreme bias for a certain party.
- The proposed models participated in Task 4 of SemEval 2019 where it was ranked 23rd out of the 42 participating teams, the objective of the task being to classify an article as Hyperpartisan or not.

Example of Hyperpartisan (H) News

“Are you sick of Republicans? Or just right-wingers in general? Do you want to send a message to Washington that you aren't going to buy into their racist, sexist, xenophobic, homophobic and classist nonsense for one second longer? Then do the very thing that Donald Trump unintentionally encouraged in a recent tweet: Encourage Hillary Clinton to run for president in 2020!”

Example of Mainstream (M) News

“Clinton, a Chappaqua resident, is set to write “What Happened,” her recounting of her loss to Donald Trump in the 2016 presidential election. The book will be published on Sept. 12 by Simon and Schuster. “In the past, for reasons I try to explain, I’ve often felt I had to be careful in public, like I was up on a wire without a net. Now I’m letting my guard down,” Clinton, who previously served as Senator and Secretary of State, said.”

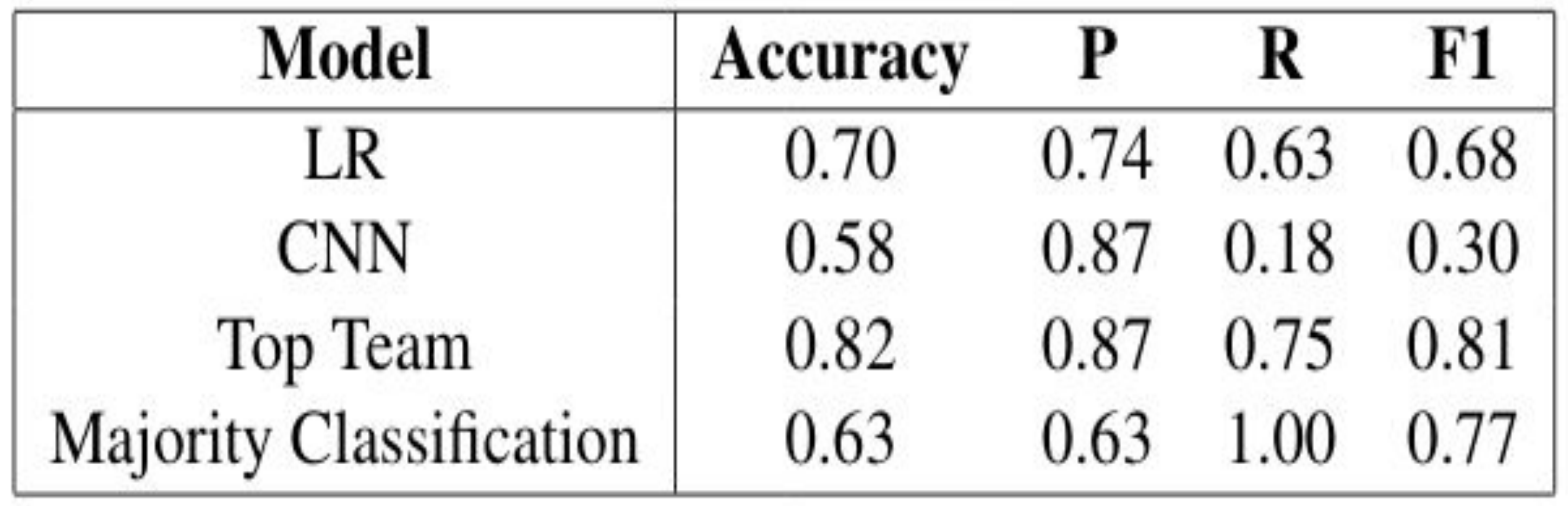
2. Methodology

Two models were developed:

- Logistic Regression (LR) classifier trained on simple unigram features.
  - *Terms with frequency less than 12 were discarded.*
- Convolutional Neural Network (CNN) trained on word embeddings learned from the training data.
  - *CNN was trained over 100 epochs.*
- Both models tuned using 10-fold cross validation.

3. Results

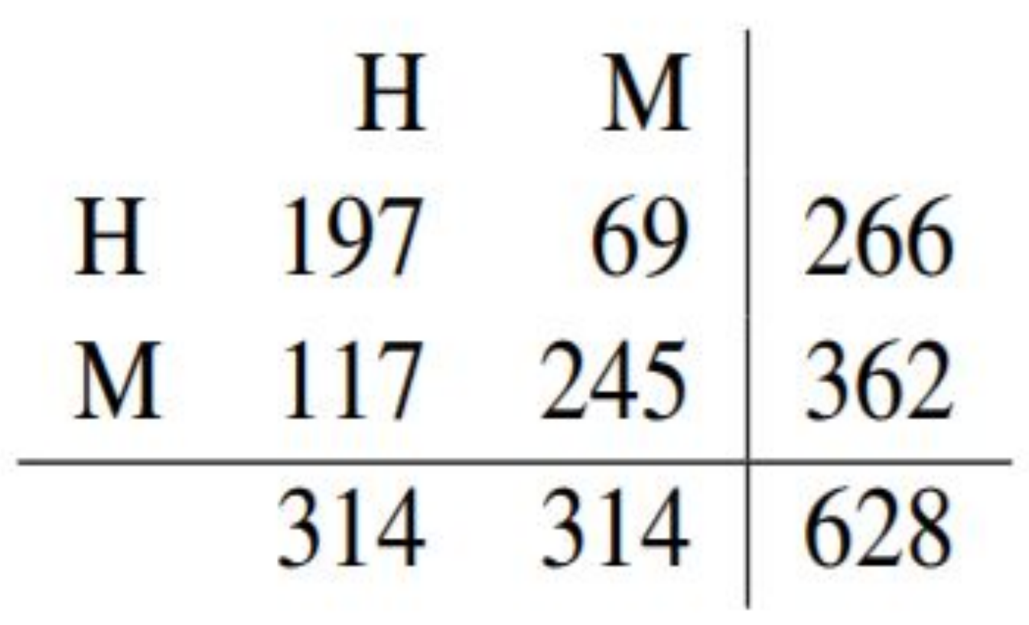
On the final held out test data, the models performed as follows:



Model	Accuracy	P	R	F1
LR	0.70	0.74	0.63	0.68
CNN	0.58	0.87	0.18	0.30
Top Team	0.82	0.87	0.75	0.81
Majority Classification	0.63	0.63	1.00	0.77

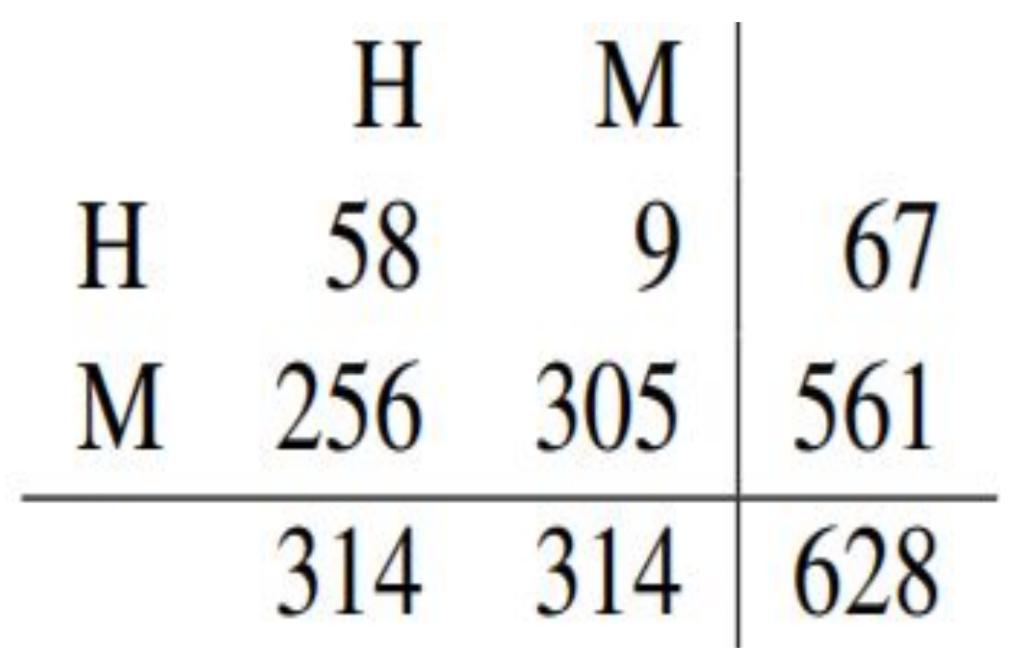
Table 1: Final Evaluation Results

- Low accuracy of CNN could be because of overfitting or the small size of the training dataset (**645 - 238 H & 407 M articles**).
- The **test data** had **314 articles** per class.
- CNN’s imbalanced predictions (cf. Table 3) vs LR’s balanced predictions (cf. Table 2) resulted in its high precision and low recall.



	H	M	
H	197	69	266
M	117	245	362
	314	314	628

Table 2: LR Confusion Matrix



	H	M	
H	58	9	67
M	256	305	561
	314	314	628

Table 3: CNN Confusion Matrix

4. Feature Analysis

The dataset contained articles pertaining to the 2016 US presidential campaign.

- It was observed that ***H articles contained a few more person names*** than M articles.
- Features which stood out for H news were **Hillary**; **Arpaio**; **Maria**; **Clintons** and **Hitler**.
- M articles rated features like **Donald** and **Twitter** as highly important. As Trump is an avid Twitter user, his tweets would likely find their way onto mainstream news.

5. Future Work

- Exploring larger datasets which may help tune the CNN model better.
- For preprocessing, we wish to perform **named entity recognition** as proper nouns could represent important information for this problem.