Creating Natural Language Processing Models for Detecting Fake News with Limited Data

By Lakshya Mehta,
Summer Ventures 2023
Ethical Data Science
Marco Scipioni

UNIVERSITY OF NORTH CAROLINA
CHARLOTTE
CENTER FOR
STEM EDUCATION

Introduction

- Fake news is information spread with the intent to manipulate and misinform.
- It has affected the trust in democratic processes and the mitigation of the COVID-19 pandemic.
- It must be automatically moderated on the social media sites where it is spread.

(Meta, 2023)

Research Goal

- Create a successful, non-deep, machine learning model that can identify an article as fake or real solely based on the text or title of said article.

Literature Review

- Models - Deep Learning Neural Networks
- Text processing - TF-IDF & N-grams (Suhasini & Vimala, 2021) or TwIDw (Nagy & Kapusta, 2023)
- Sentiment analysis (Kumari et al. 2022)

Algorithm Selection

Joyce George (2020) tested three algorithms:

- Multinomial Naive Bayes (MNB)
- Support Vector Machine (SVM)
- Passive Aggressive Classifier (PAC)

Three additional algorithms:

- Logistic Regression
- Random Forest
- Extreme Gradient Boost (XGB)

Materials and Methods

Dataset description

Training Dataset

- 70k labeled article titles
- 34k fake and 35k real articles

(Steven, 2022)

Testing Dataset

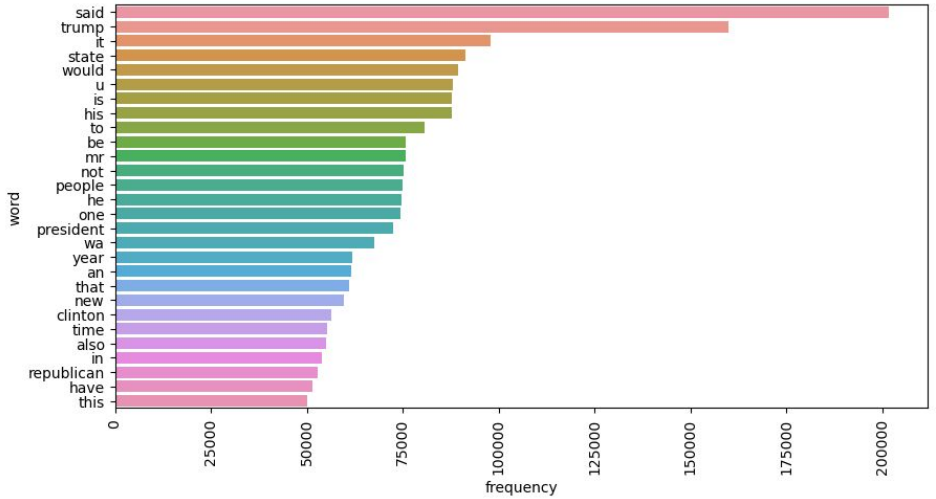
- 6000 labeled article titles and texts
- 50/50 split fake and real articles

(Jillani, 2022)

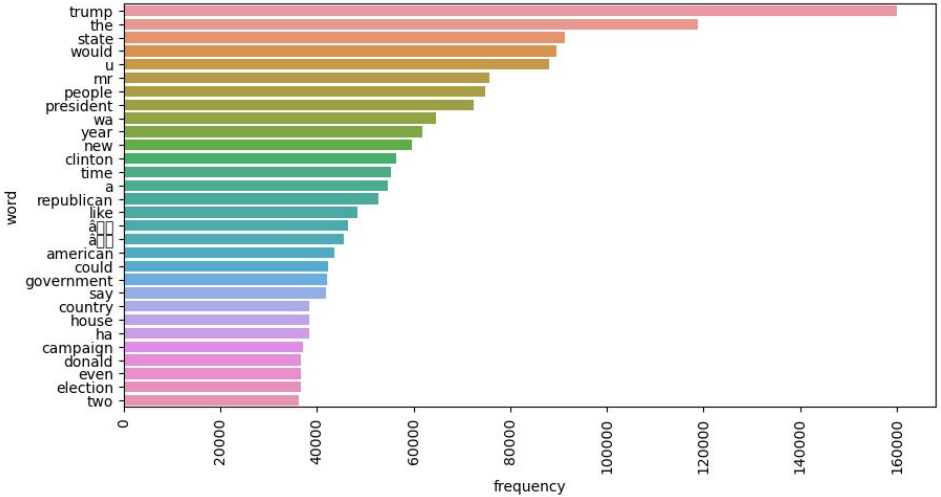
Data Preprocessing

- Compile clean data
- Lowercase and remove punctuation
- Tokenize text
- Remove common stop words
- Lemmatize text
- Visualize most common words and add any “useless” words back to stop words list

Before



After



Training

- Reprocess with new stop words list
- N-gram analyze - store every sentence as every set of consecutive words within
- Vectorize
- TF-IDF transform - provide rare items with more importance than common items
- Save models locally

N-Grams

"plata o plomo means silver or lead"

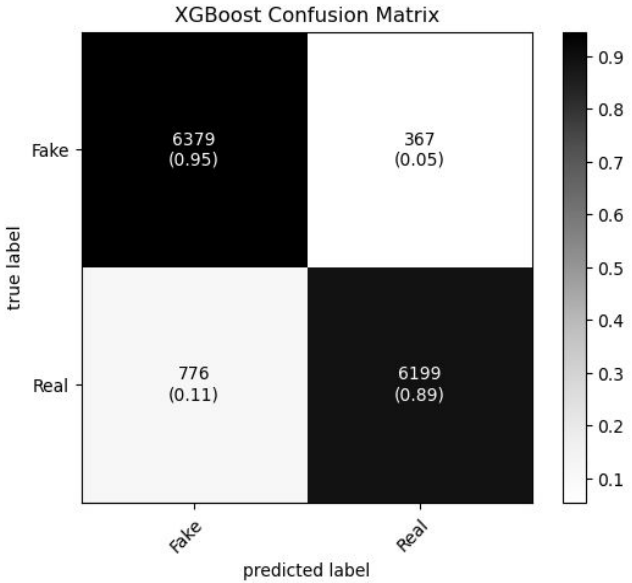
n	Name	Tokens
2	bigram	["plata o", "o plomo", "plomo means", "means silver", "silver or","or lead"]
3	trigram	["plata o plomo", "o plomo means", "plomo means silver", "means silver or ", "silver or lead"]

Shetty & Koss (2022)

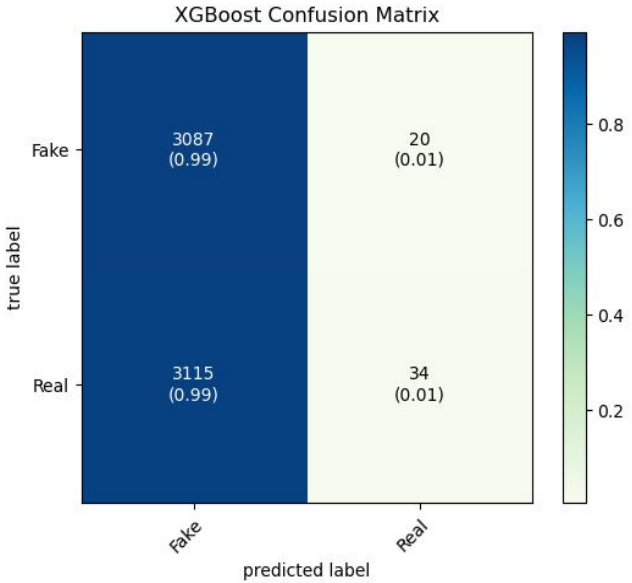
Results

Confusion Matrices

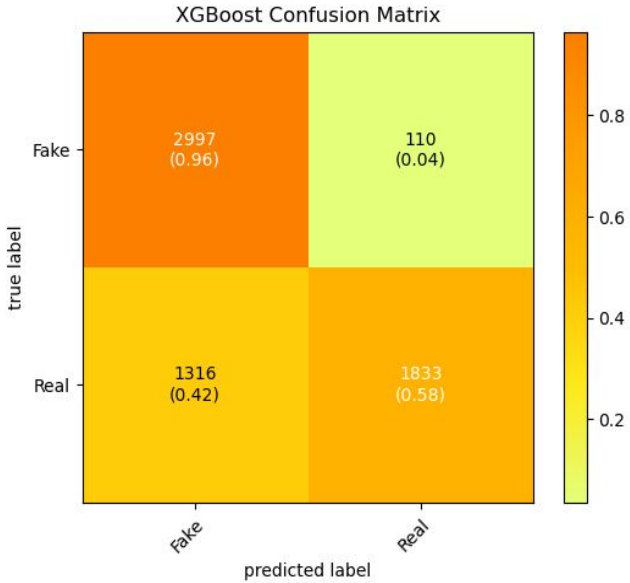
Training



Titles



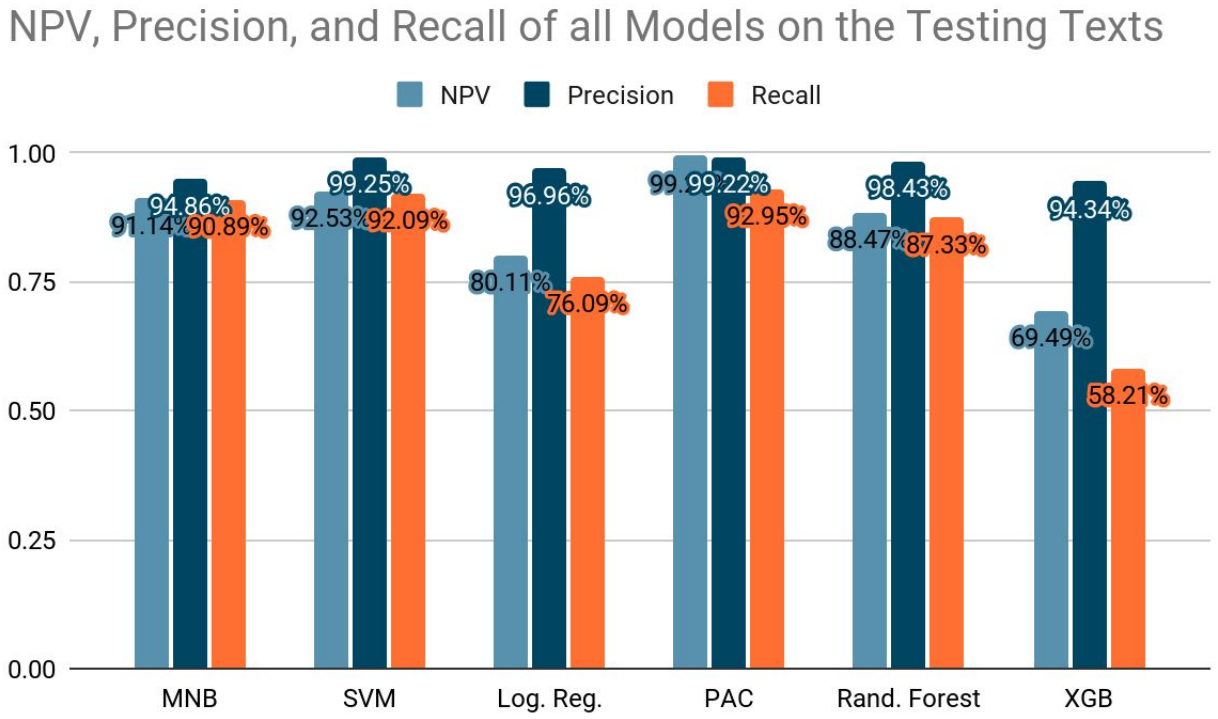
Texts



Accuracy Scores

Test Subset	Algorithms					
	MNB	SVM	Log. Reg.	PAC	Rand. Forest	XGB
Training	.91305	.95578	.93412	.96028	.93186	.91670
Titles	.69997	.53245	.53117	.55195	.49760	.49888
Texts	.92925	.95668	.86765	.96084	.92919	.77206

Other metrics



Discussion and Conclusion

Deployment

- Models train on a smaller and similarly enough structured dataset
- Less time can be spent training the models
- Requiring only the text of an article allows the models to be able to analyze as many articles as possible

Ethical Considerations

- Raw text input allow for “fairer” classifications
- Censorship associated with automated content moderation must be addressed

Future Directions

- Test the PAC model with data scraped directly from the Internet.
- Implement model optimizations such as grid-search and cross-validation.
- Experiment with feature extraction on the texts (eg. capital letter frequency, word/sentence length, etc.).

References

George, J. A. (2020, June 14). Fake News Detection using NLP techniques | by Joyce Annie George | Analytics Vidhya. Medium. Retrieved July 19, 2023, from https://medium.com/analytics-vidhya/fake-news-detection-using-nlp-techniques-c2dc4be05f99

Jillani. (2022). Fake or Real News. Kaggle. Retrieved July 18, 2023, from https://www.kaggle.com/datasets/jillanisofttech/fake-or-real-news

Kumari, R., Ashok, N., Ghosal, T., & Ekbal, A. (2022, January). What the fake? Probing misinformation detection standing on the shoulder of novelty and emotion. Information Processing & Management, 59(1), 102740. Science Direct. https://doi.org/10.1016/j.ipm.2021.102740

Nagy, K. S., & Kapusta, J. (2023, May 25). TwIdw—A Novel Method for Feature Extraction from Unstructured Texts. Appl. Sci., 13(11), 6438. https://doi.org/10.3390/app13116438

Shetty, B., & Koss, H. (2022). NLP Machine Learning: Build an NLP Classifier. Built In. Retrieved July 18, 2023, from https://builtin.com/machine-learning/nlp-machine-learning

Steven. (2022, July). Misinformation & Fake News text dataset 79k. Kaggle. Retrieved July 17, 2023, from https://www.kaggle.com/datasets/stevenpeutz/misinformation-fake-news-text-dataset-79k

Suhasini, V., & Vimala, N. (2021, June 7). A Hybrid TF-IDF and N-Grams Based Feature Extraction Approach for Accurate Detection of Fake News on Twitter Data. Turkish Journal of Computer and Mathematics Education, 12(6), 5710-5723. ProQuest. https://www.proquest.com/docview/2640416454?parentSessionId=OHIdFJoiZTk8KnB2WfNLzuxlhJtoU6i6EgoQhqQvDdY%3D

Zhang, X., & Ghorbani, A. A. (2020, March). An overview of online fake news: Characterization, detection, and discussion. Information Processing & Management, 57(2), 102025. Science Direct. https://doi.org/10.1016/j.ipm.2019.03.004

References Continued