PUBLISHER 4.0

END-TO-END NLP PIPELINE FOR FALSE NEWS ARTICLE DETECTION AND CONTENT SUMMARIZATION

Presented By: Saleh Alkhalifa
Group: 4

AGENDA

1. Executive Summary
2. Dataset and Libraries
3. NLP Pipeline
4. Proposed Timeline
5. Conclusion

ABSTRACT

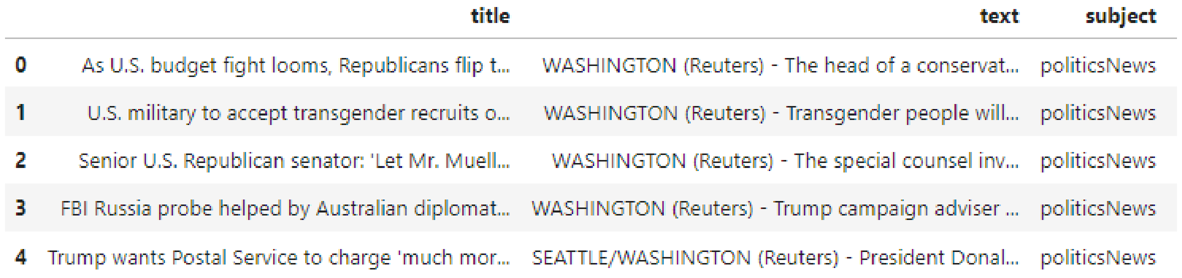
Reading and writing are integral parts to everyday life, especially in the digital and virtual age we live in today. Language today is one of our primary tools when it comes to expression and communication, allowing us to share thoughts, ideas and news around the globe. In recent years, articles containing false information have gained traction both in social media and on the news, with very few tools available to help limit or prevent these false narratives from gaining popularity. The proposed project is to develop an end-to-end pipeline that will comprise (1) classification model by which text, specifically articles, can be predicted as real or fake, (2) subject classification, and (3) text summarization for headline generation.

DATASET

Dataset Name: Fake and Real News Articles Dataset

Size: ~40,000 rows

Sample:



	title	text	subject
0	As U.S. budget fight looms, Republicans flip t...	WASHINGTON (Reuters) - The head of a conservat...	politicsNews
1	U.S. military to accept transgender recruits o...	WASHINGTON (Reuters) - Transgender people will...	politicsNews
2	Senior U.S. Republican senator: 'Let Mr. Muell...	WASHINGTON (Reuters) - The special counsel inv...	politicsNews
3	FBI Russia probe helped by Australian diplomat...	WASHINGTON (Reuters) - Trump campaign adviser ...	politicsNews
4	Trump wants Postal Service to charge 'much mor...	SEATTLE/WASHINGTON (Reuters) - President Donal...	politicsNews

Data Available: Title, Full Text, Subject, and Classification

[1] Ahmed H, Traore I, Saad S. "Detecting opinion spams and fake news using text classification", Journal of Security and Privacy, Volume 1, Issue 1, Wiley, January/February 2018.
[2] Ahmed H, Traore I, Saad S. (2017) "Detection of Online Fake News Using N-Gram Analysis and Machine Learning Techniques. In: Traore I., Woungang I., Awad A. (eds) Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments. ISDDC 2017. Lecture Notes in Computer Science, vol 10618. Springer, Cham (pp. 127-138).

LIBRARIES

- **Pandas – To organize and process data via Dataframe**
- **Numpy – To use various processing functions**

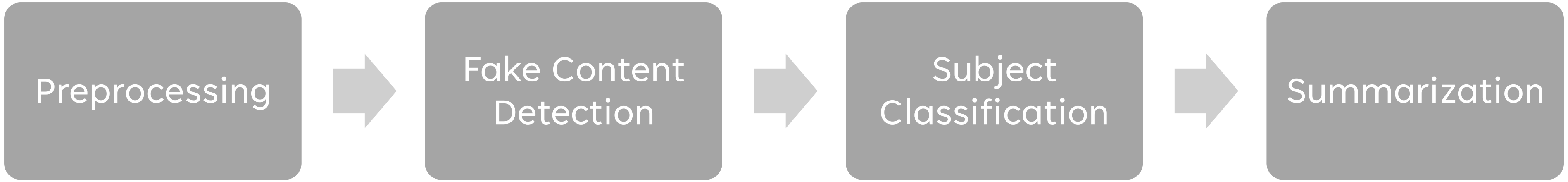
- **Nltk – To get stop-words and other NLP features**
- **Spacy – For various NLP functions and features**

- **Scikit-learn – For standard machine learning functions**
- **Keras – To develop classification and summarization models**
- **Tensorflow – To support several functions in Keras**

- **Matplotlib – To plot results and other input/output data**
- **Seaborn – To generate publication-style diagrams**

[3] Shervin Minaee, Nal Kalchbrenner, Erik Cambria, Narjes Nikzad, Meysam Chenaghlu, Jianfeng Gao. “Deep Learning Based Text Classification: A Comprehensive Review”, Arxiv.org
[4] Shervin Minaee et. al. “Deep Learning Based Text Classification: A Comprehensive Review”, Arxiv.org

E2E PIPELINE



Preprocessing

Fake Content Detection

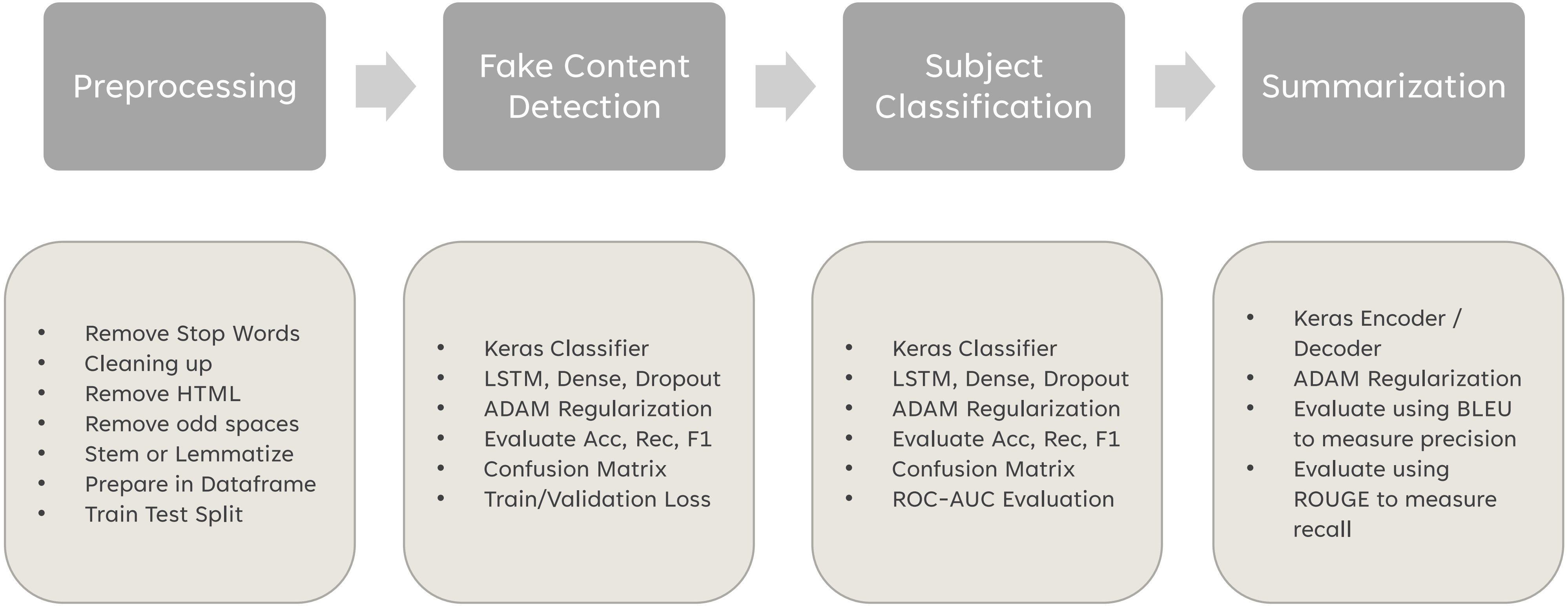
Subject Classification

Summarization

- Articles will be preprocessed and cleaned up to remove stop-words and other items
- Preprocessed articles will then be determined whether they are fake or true
- After subject classification, the articles will be classified to determine the subject
- Finally, the article will be summarized to determine an appropriate title

[5] Nathaniel Hoy, Theodora Koulouri. “A Systematic Review on the Detection of Fake News Articles”, Arxiv, **2021**.
[6] Mrinal Rawat, Diptesh Kanojia. “Automated Evidence Collection for Fake News Detection”, Arxiv, **2021**.

E2E PIPELINE



Preprocessing

- Remove Stop Words
- Cleaning up
- Remove HTML
- Remove odd spaces
- Stem or Lemmatize
- Prepare in Dataframe
- Train Test Split

Fake Content Detection

- Keras Classifier
- LSTM, Dense, Dropout
- ADAM Regularization
- Evaluate Acc, Rec, F1
- Confusion Matrix
- Train/Validation Loss

Subject Classification

- Keras Classifier
- LSTM, Dense, Dropout
- ADAM Regularization
- Evaluate Acc, Rec, F1
- Confusion Matrix
- ROC-AUC Evaluation

Summarization

- Keras Encoder / Decoder
- ADAM Regularization
- Evaluate using BLEU to measure precision
- Evaluate using ROUGE to measure recall

[5] Nathaniel Hoy, Theodora Koulouri. “A Systematic Review on the Detection of Fake News Articles”, Arxiv, **2021**.
[6] Mrinal Rawat, Diptesh Kanojia. “Automated Evidence Collection for Fake News Detection”, Arxiv, **2021**.

PROPOSED TIMELINE

Week	Items to Complete	Risks
Week 8	Explore and Preprocess Data	“Healthiness” of the dataset
Week 9	Subject Classification	40,000 rows of large text data might be too much for this laptop
Week 10	Legitimacy Classification	Distribution of Data
Week 11	Content Summarization	Abstractive vs Extractive?
Week 13	Prepare Report and Presentation	N/A
Week 14	Present	N/A

THANK YOU