FAKE AND REAL NEWS CLASSIFICATION

Sons of Analysts

Authors

Wendy Mwiti
(Slides)

Monicah Iwagit
(Slides & Notebook)

Femi Kamau
(Preprocessing & Deployment)

Teofilo Gafna
(Modelling & Deployment)

Table of contents

Overview

- The prevalence of fake news has gradually increased over the years
- Internet has made it possible to publish news with hardly any restrictions
- Media trust dropped by 8% every year
- It’s been difficult for the audience to distinguish real and fake news

Problem Statement

Fake news has a high potential to change opinions, facts and can be the most dangerous weapon in influencing the society negatively. So we intend to develop a fake and real news classifier using natural language processing and machine learning algorithms.

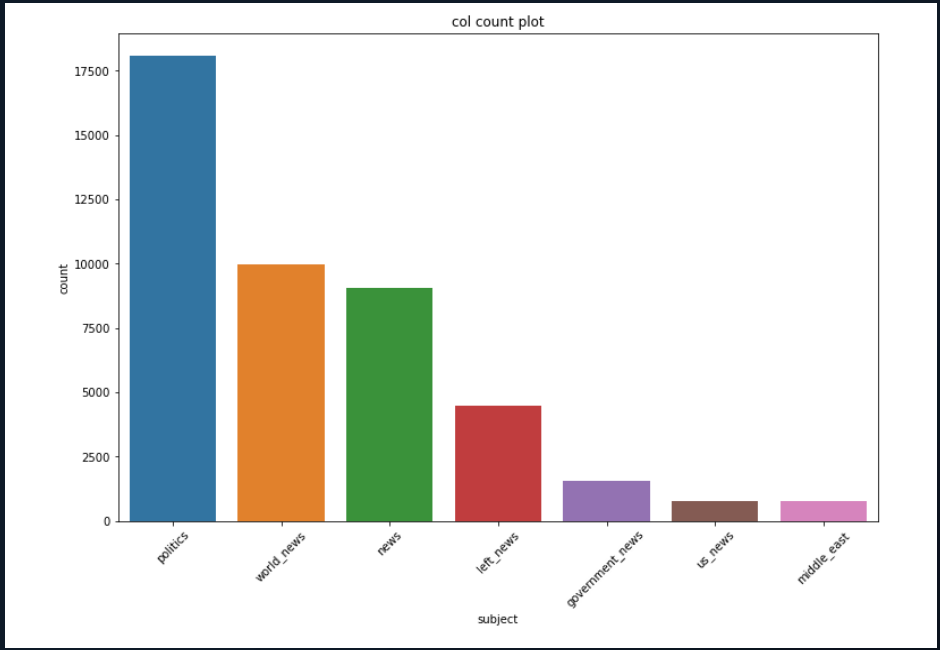
Data Understanding

- Data was extracted from news agencies such as Reuters covering articles between 2015 to 2018.
- Two datasets involved: one of fake news, another of real news which we concatenated.
- The dataset had 44898 rows, 5 columns and 209 Duplicates.
- Date column only had 2397 unique values out of 44898 rows
- Columns used are text and category (target).

Data Preparation

- Changed date column to datetime format and extracted months and years.
- Casted our target variable to a categorical datatype.
- Inspected and dropped duplicates.
- Grouped different systems referring to the same value together e.g politics News and politics.

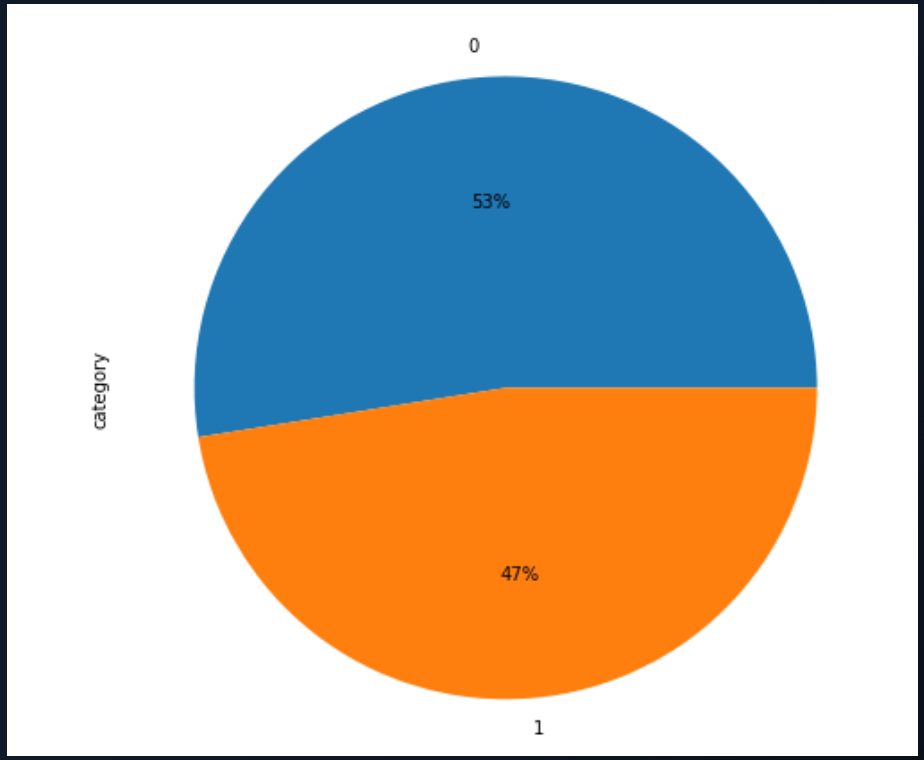
News Subject Analysis



Most articles written are on politics.

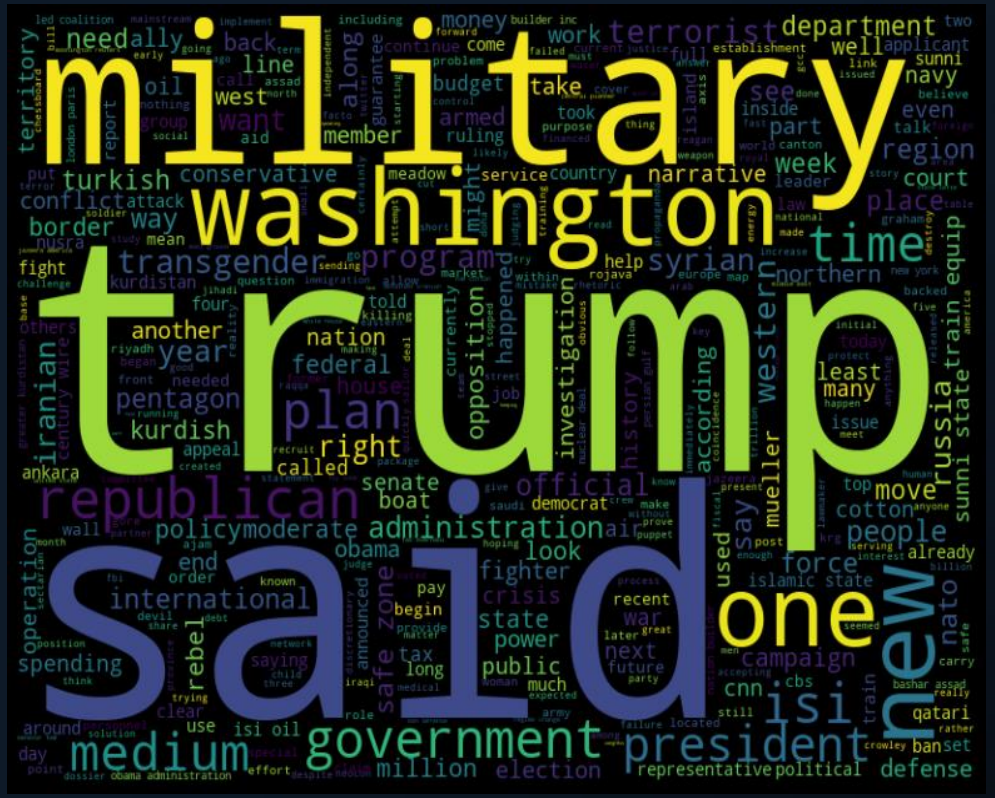
The Middle East is the least talked about in news.

Target Variable Analysis



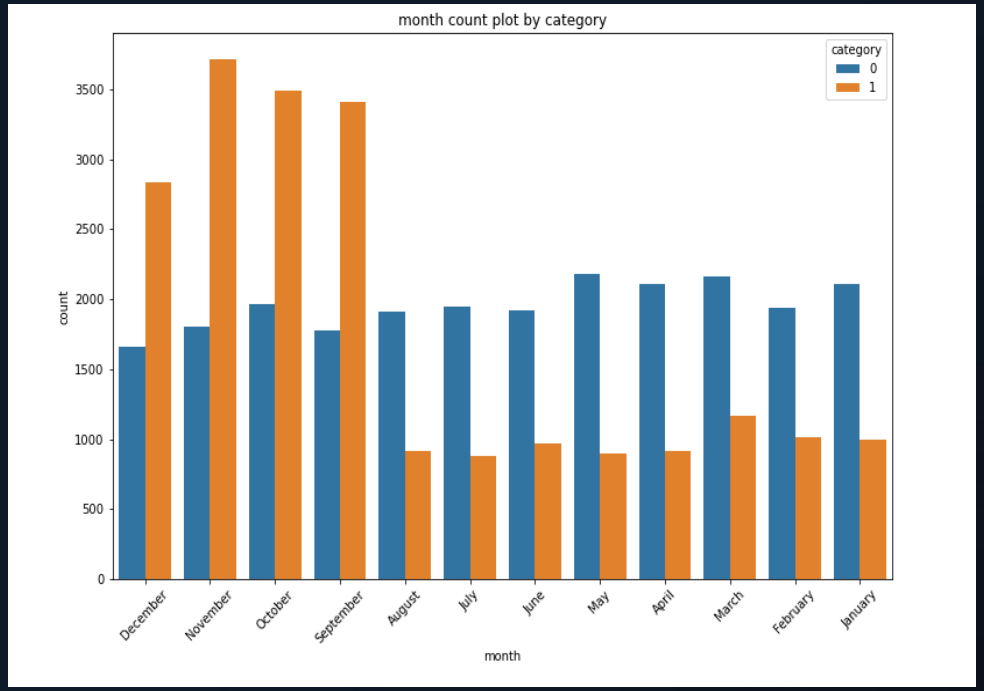
The data set is fairly balanced. However, fake news is slightly more than real news.

Fake News WordCloud



The trending topics covered in fake news include military, Washington and trump.

News versus Month Analysis



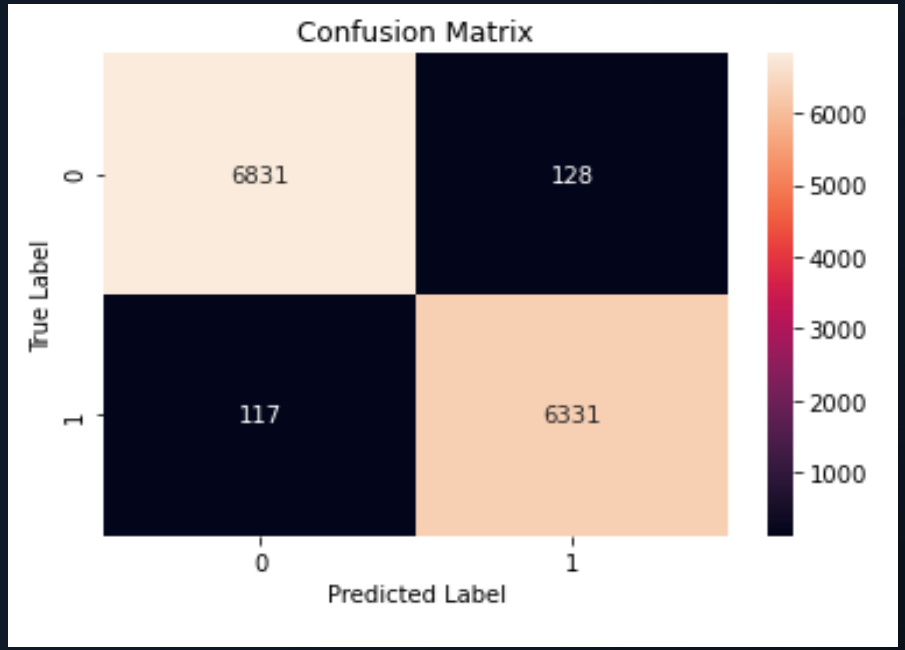
Most legitimate news were published during the last four months of the year.

Slightly balanced distribution of fake news throughout the year.

Modelling

Model	Precision (%)	Recall (%)	F1 Score (%)	Accuracy (%)
Logistic Regression	98.2	98.3	98.2	98.3
Random Forest	98.0	98.2	98.1	98.2
AdaBoost	99.4	99.6	99.5	99.5
Gradient Boosting	99.1	99.7	99.4	99.4
XG Boost	99.6	99.8	99.7	99.7

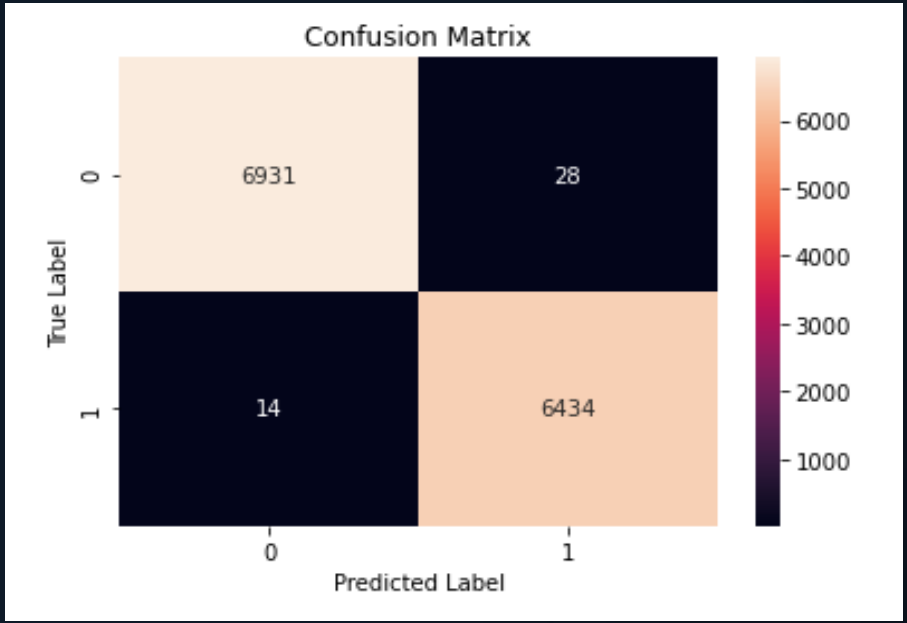
Random Forest confusion matrix



117 samples were predicted fake yet true.

128 samples predicted true yet false.

XGBoost confusion matrix



14 were classified fake yet true.

28 samples were classified as true yet false.

Deployment

Conclusion

Every single news has different characteristics so there is a need for a system that can check the content of the news in depth.

The results suggested that the approach is highly favorable since the application helps in classifying fake news and identifying key features that can be used for fake news detection.

Recommendations

- The use of the system classifier to detect whether an article posted is legitimate to avoid misinformation to the readers.

Future Work

- Build an automated fact-checking system that combines data looking at different aspects to help non-experts in classifying news.
- Use data that covers a wide range of time focusing on world news.
- Use PySpark to process data so as to reduce computation complexity.
- Use Twitter-API to get current news

Questions Section

Thank you