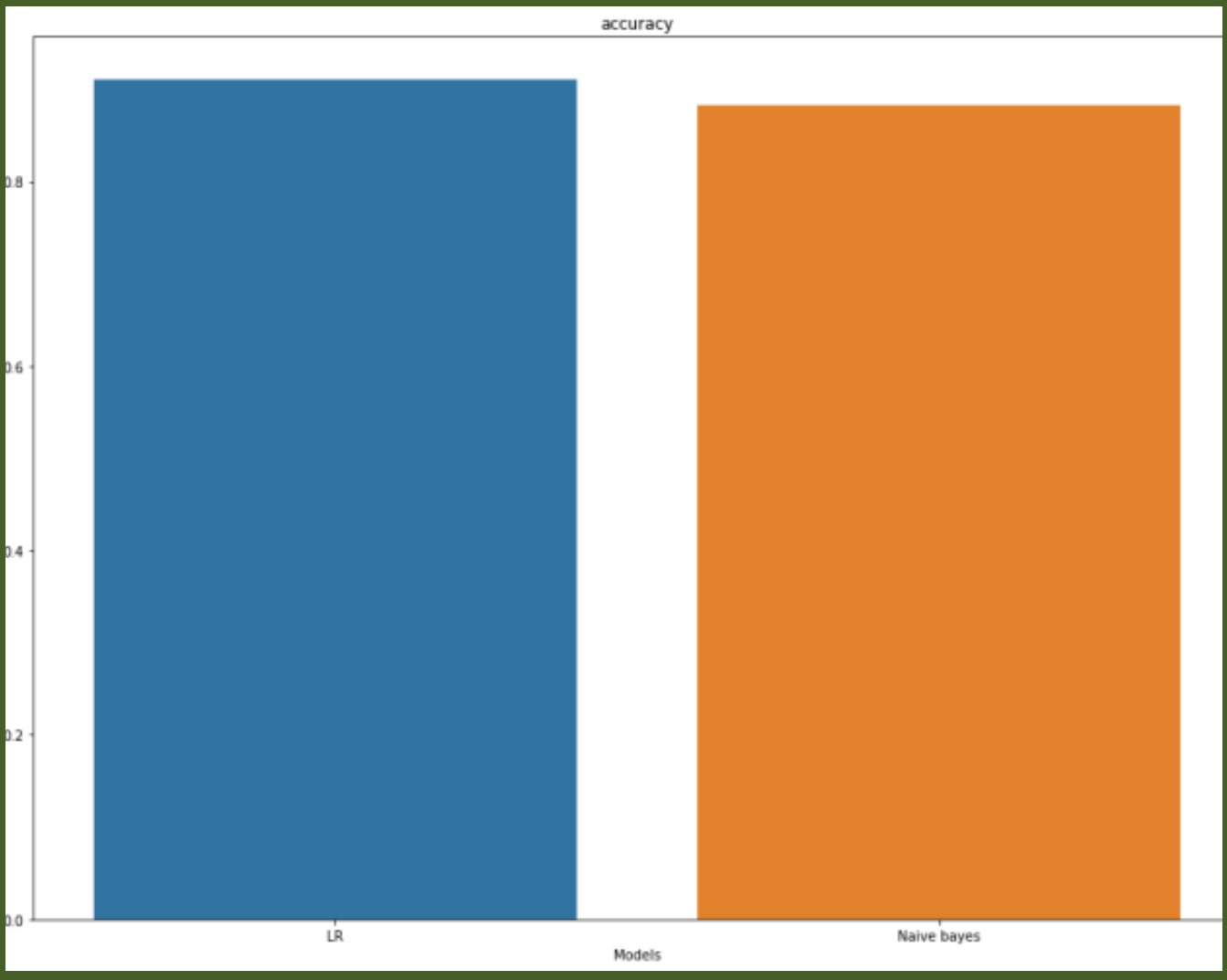
NLP DOCUMENTATION

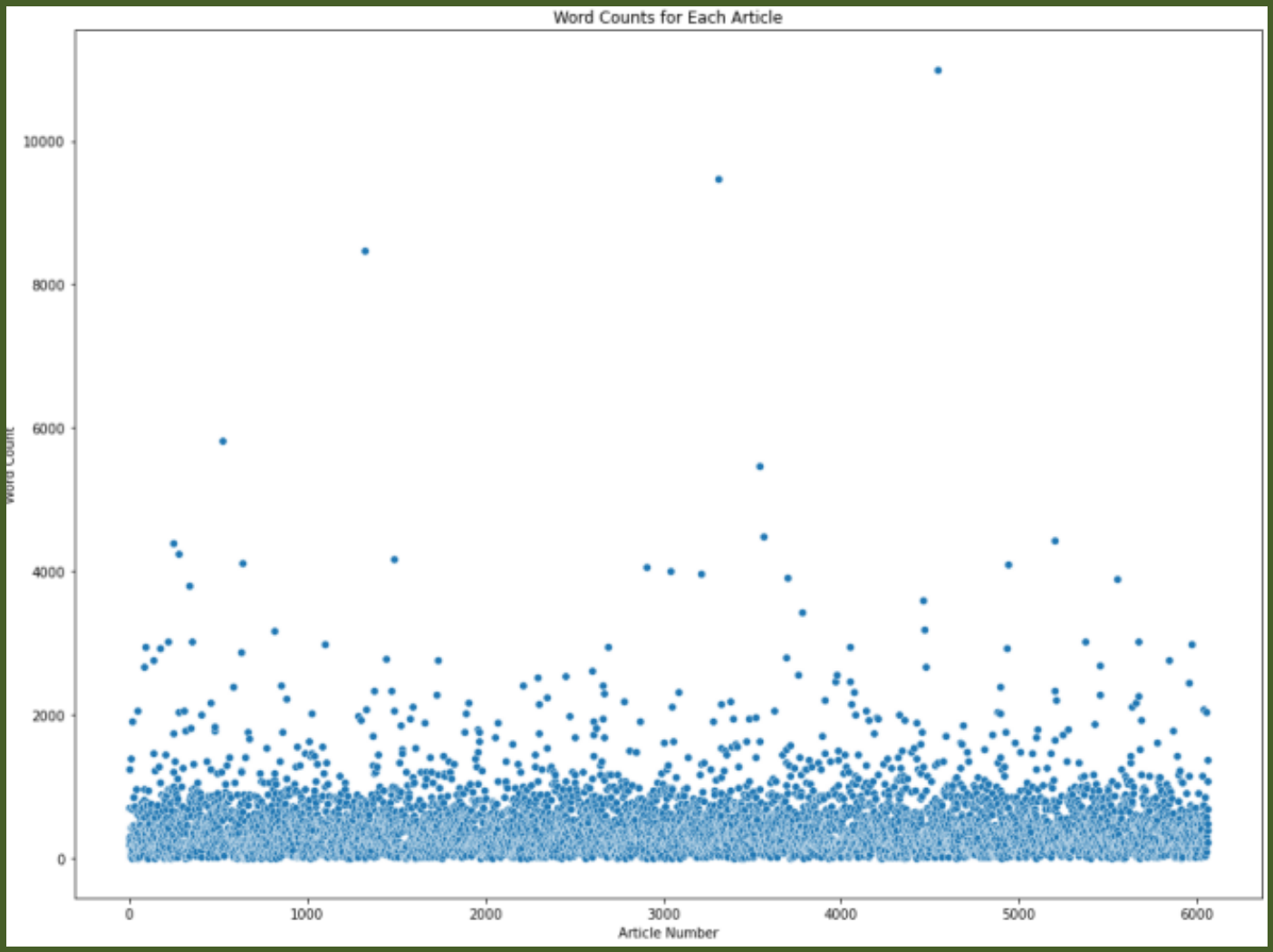
TEAM 9

اسامه علي عبدالسميع الشرقاوي 20201700106
أحمد بهاء الدين أحمد محمود خطاب 20201701508
احمد طارق امين علي 20201700055
كمال كريم فؤاد رمضان 20191700897.00
احمد محمد جلال سليم 20201700071
يوسف طلعت صابر محمد 20201701016

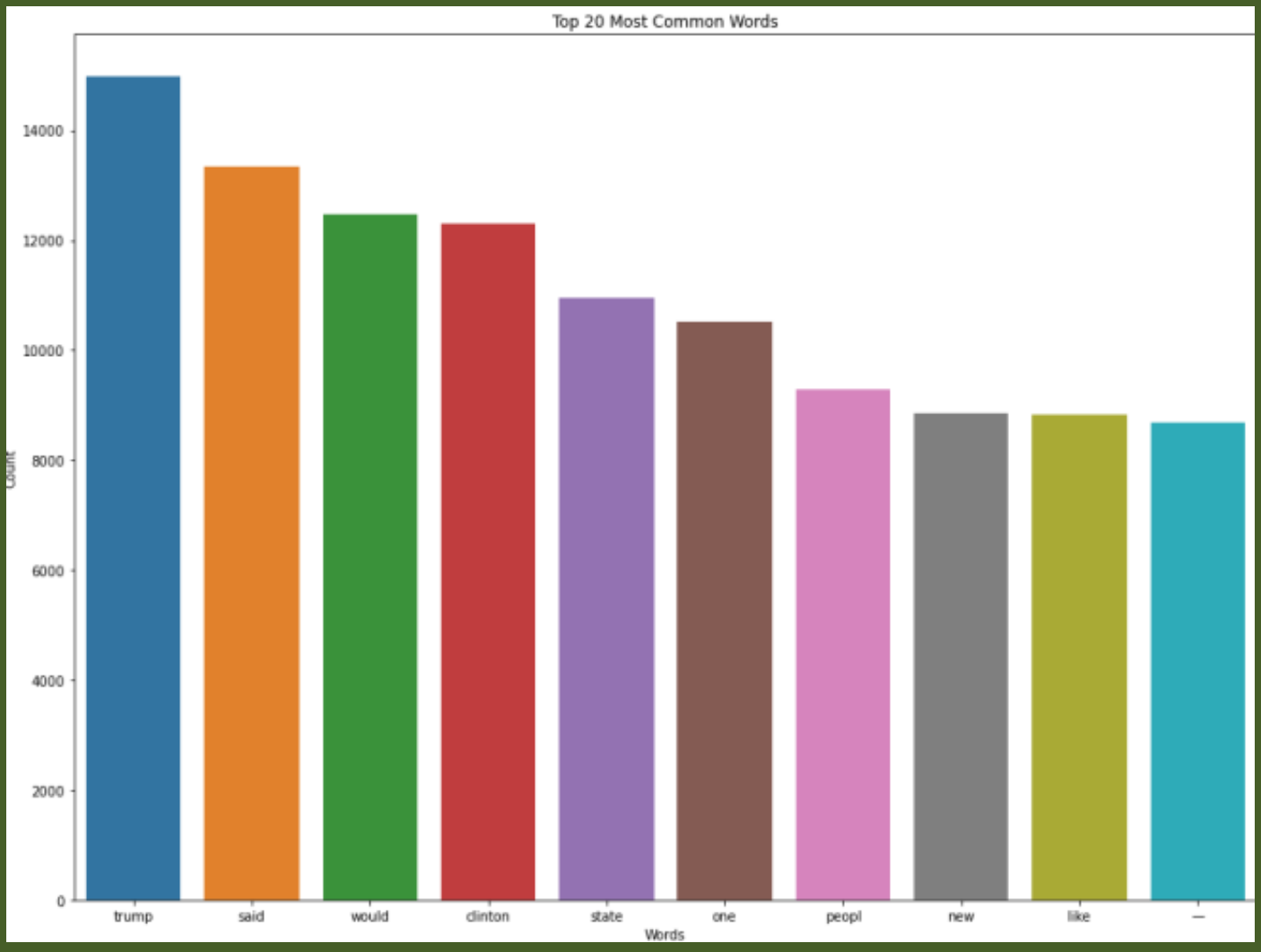
Models Accuracy



Word count for each news



Top 10 most common words



preprocessing

• Converting the text to lower case:-

• remove stopwords:-

- import stop words for English language and remove it from the news data frame

• Stemming:-

- imported porterstemmer and applied stemming to the data frame.

• lemmatization :-

• imported *WordNetLemmatizer* and applied to the data frame

• Parameters:-

- data: which is the data frame
- target_columns: the target class to be preprocessed

• Feature Selection:-

we used to columns in the data frame

- text: which is the news text.
- label: which tell if it's Fake or Real news.

MODEL TRAINING

TF_IDF

We used the TF_IDF vectorizer.
The TF-IDF algorithm assigns a weight to each word in a document based on its frequency in the document and its frequency in the corpus of all documents. so it's so helpful to use with many news to make the vector for the machine learning model.

Models

we used 2 models:-

. Logistic Regression .MultinomialNB

- we used the to models and got accuracy of 84% and 91% approximate, and saved the models using joblib for future testing.
- we also draw the accuracy plot as shown in the first page.