The World of Data Science

Final Project

Fake News Detection

Presented By:

Hadar Asher

Ofir Sered

מכון טכנולוגי חולון
Holon Institute of Technology

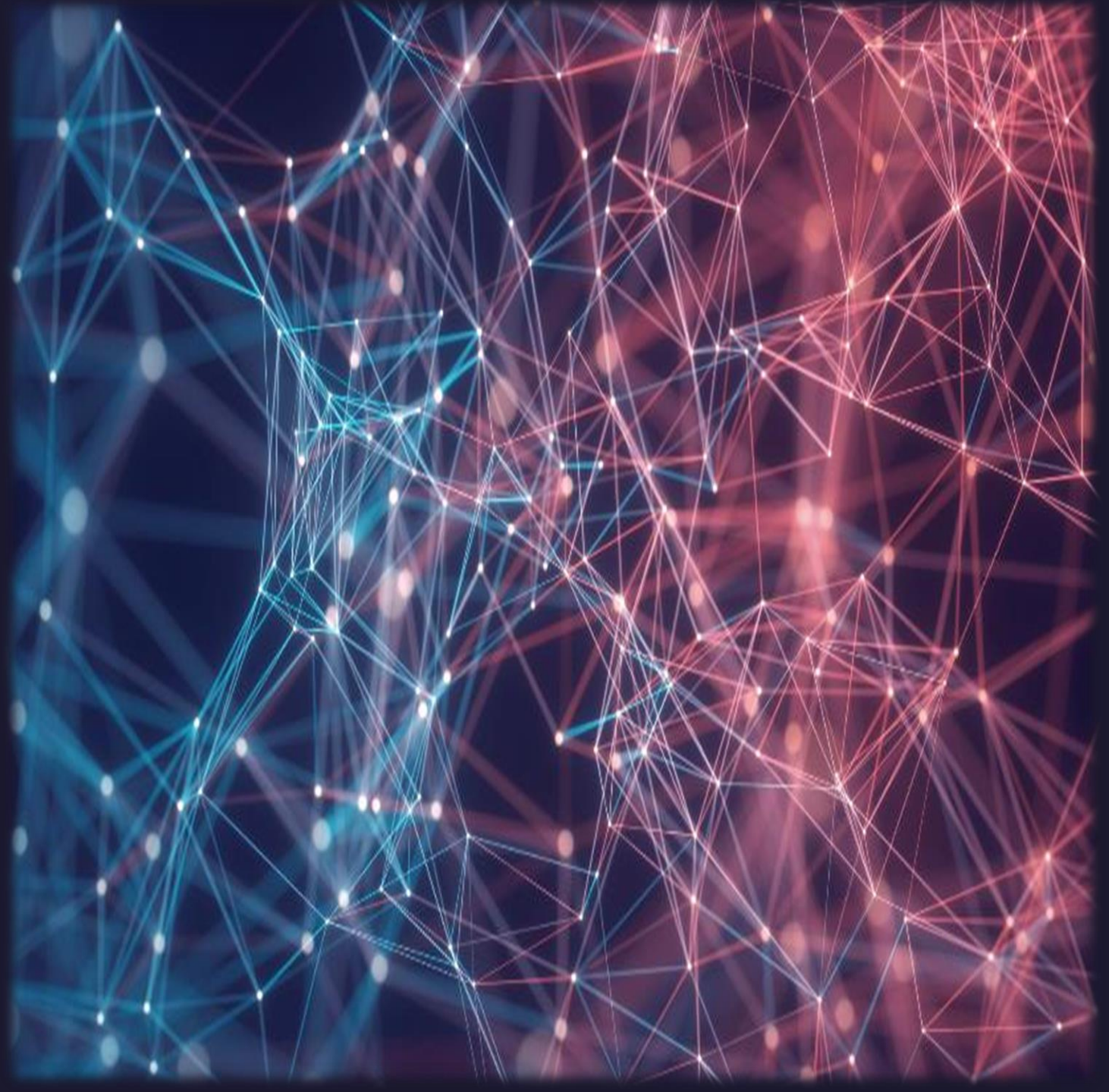
הקדמה:

- שאלת המחקר – האם ואיך ניתן לזהות אמינות כתבות?
- דרך – אסיפת מידע מאתרים שונים על פי הפרמטרים שקבענו.
- המטרה – לנצל טכניקות של מדעי הנתונים כדי לבנות מודל יציב ואמין.

מקורות:

- רשימת האתרים שביצענו Crawling,
- כחול הגדרנו כ – Reliable News
- צהוב הגדרנו כ - Unreliable News
- אדום הגדרנו כ – Fake News

- NBC News
- Fox News
- Middle East Monitor News
- Israel 365 News
- Empire News
- Info Wars News



דרך ההרכשה:

חלוקה לשתי פונקציות בכל אתר:

- איסוף כל כתובות URLS בכל אתר.
- איסוף כל המידע המוכל בכל כתבה.

שימוש בספריות:

- BeautifulSoup
- Selenium

הפרמטרים הראשיים הם:

כותרת ראשית	–	Headline
כותבים	–	Writers
ז'אנר	–	Genre
תאריך פרסום	–	Date
תוכן הכתבה	–	Content

טיוב נתונים:

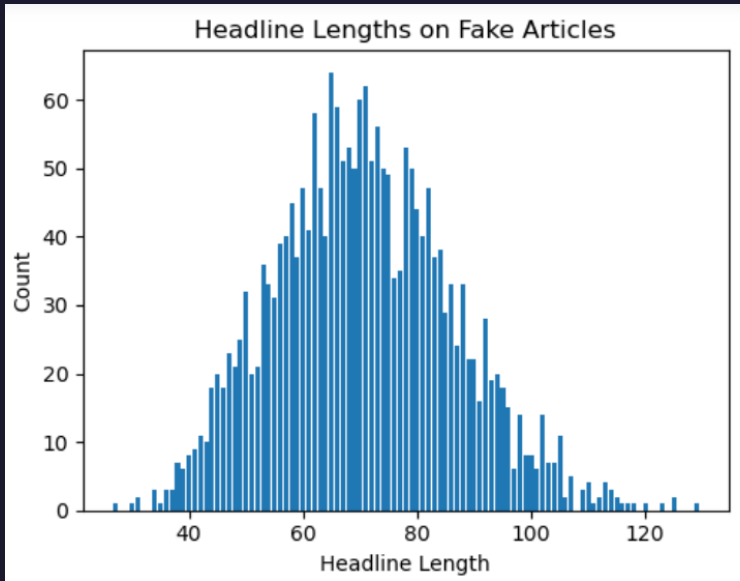
- עיבוד כל Dataframe בנפרד.
- הוספות עמודות נחוצות.
- מיזוג Dataframe אחד גדול.
- טיוב נתונים כללים.

Visualization & EDA

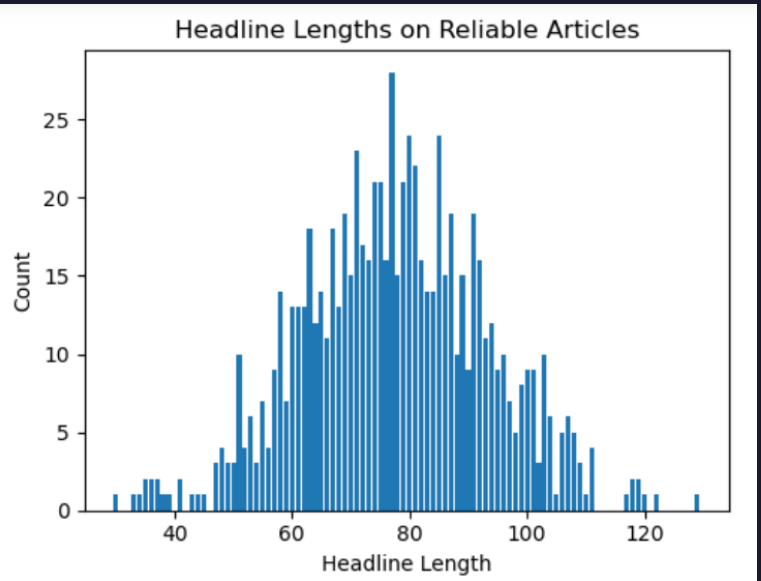
ויזואליזציות:

אורך כותרות הכתבות:

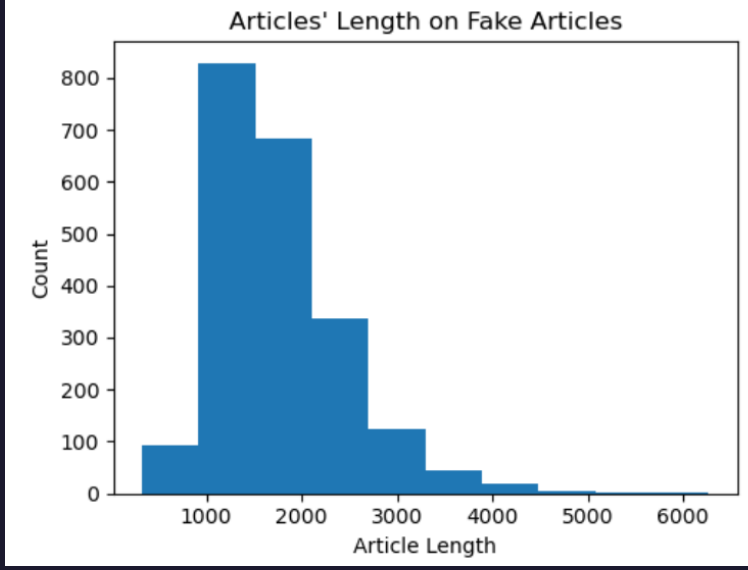
Fake

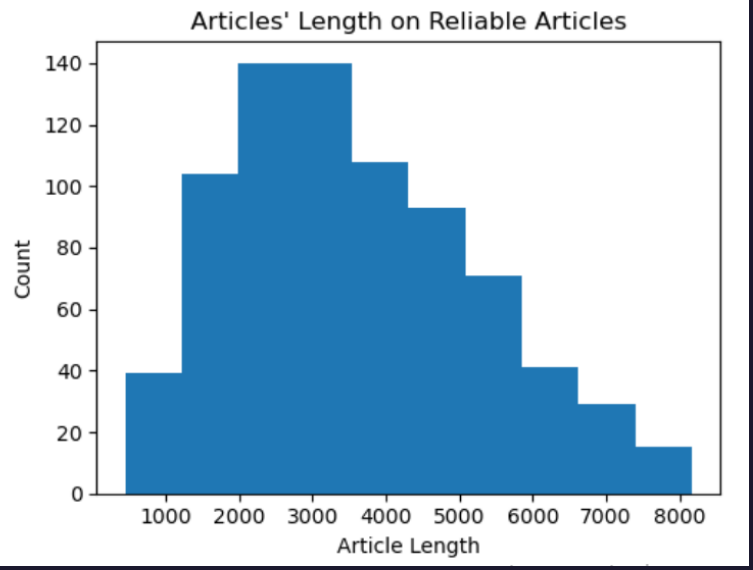


Reliable



אורך תוכן הכתבות:

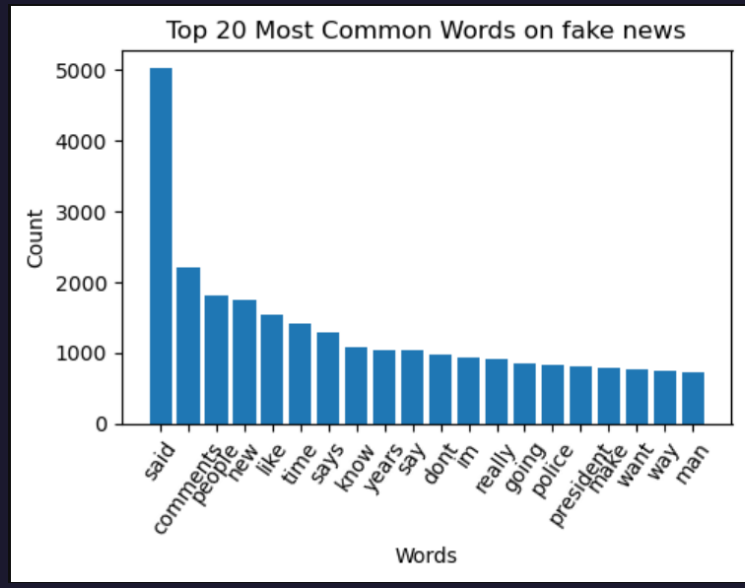




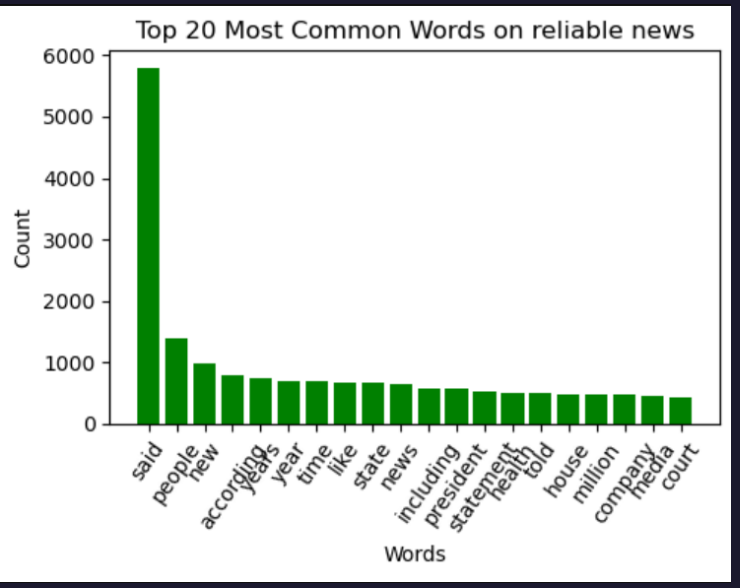
ויזואליזציות:

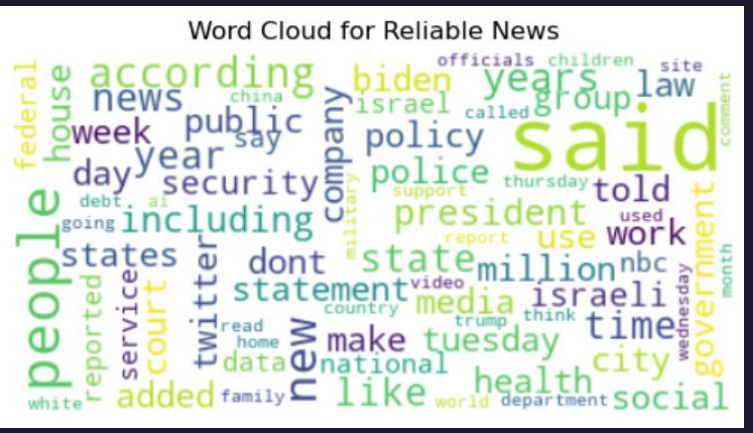
המילים הכי נפוצות:

Fake



Reliable





מודלי למידת המכונה ורמות דיוק:

- רגרסיה ליניארית -

```
Accuracy on training data = 0.9999999992038767
Accuracy on test data = 0.7203384881022445
R-squared: 0.7203384881022445
```

- עצי החלטה -

```
Accuracy on training data =  1.0
Accuracy on test data =  0.9828473413379074
f1-score: 0.9882629107981221
```

- נייב בייס -

```
Accuracy on training data: 0.9922813036020584
Accuracy on test data: 0.9571183533447685
f1-score: 0.9708963911525029
```

סיכום ומסקנות:

- מטרת המחקר בפרויקט הייתה האם ניתן לזהות מידע שגוי בתוך כתבה .
- ניסינו מגוון של ויזואליזציות והסקנו שקשה להשיג קשר בין רוב התכונות שאספנו לבין זיהוי מידע שגוי ולכן החלטנו לעשות ניתוח טקסט לקבלת תוצאות נוספות.
- ראינו שלכל המודלים יש דיוק יחסית גבוה ואנחנו מניחים שזה בגלל חוסר בנתונים.