NLP Final Project

Gilad Tsfaty, Roy Segal, Eilon Aharoni

בחרנו את פרוייקט מספר 2 לעבודה (ציוצים), הורדנו ~90 אלף

ציוצים משני אתרי חדשות מרכזיים CNN,BBC.

בגיט יש את הקובץ פייתון.
קובץ טקסט output שמכיל את כל ההדפסות.
קובץ טקסט של דרישות עבור הרצת הפרויקט.
קבצי הציוצים.
קובץ WORD נוכחי.

מבנה העבודה:
1 - חילוץ כמות ציוצים פרקטית לעבודה.

2 – EDA.

3 – Sentiment Analysis.

4 – Summarization.

5 – Data PreProcess, TF-IDF.

6 – Word2Vec, Auto encoder.

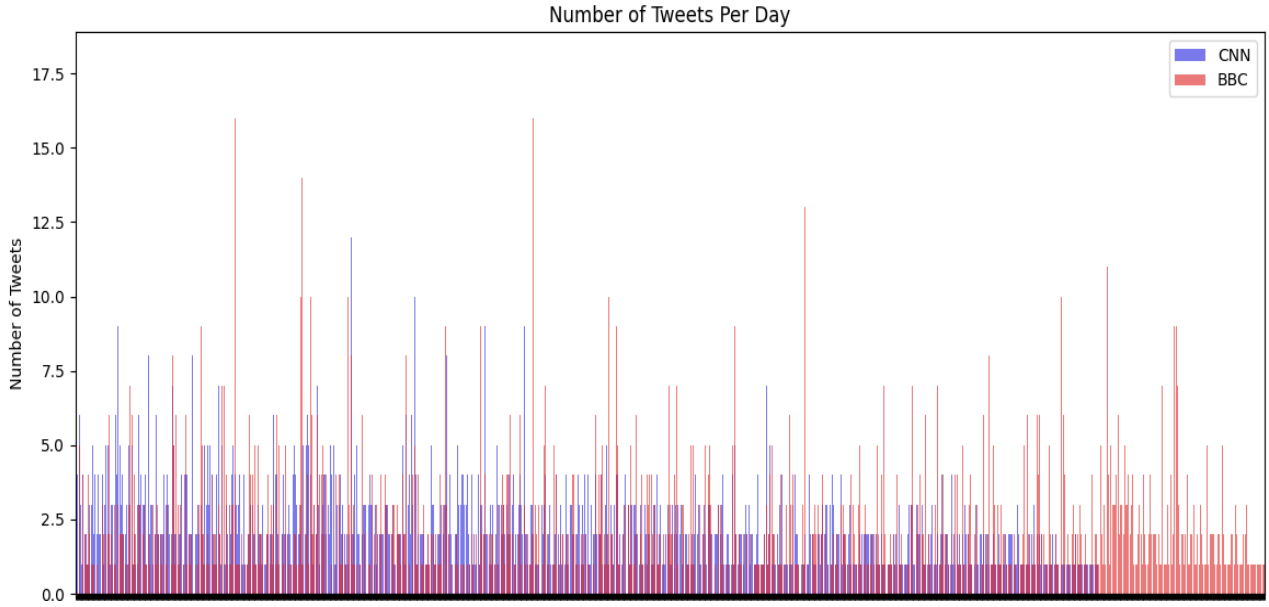
7 – NER.

8 – RNN,GPT לצורכי ג'נרוט.

.1

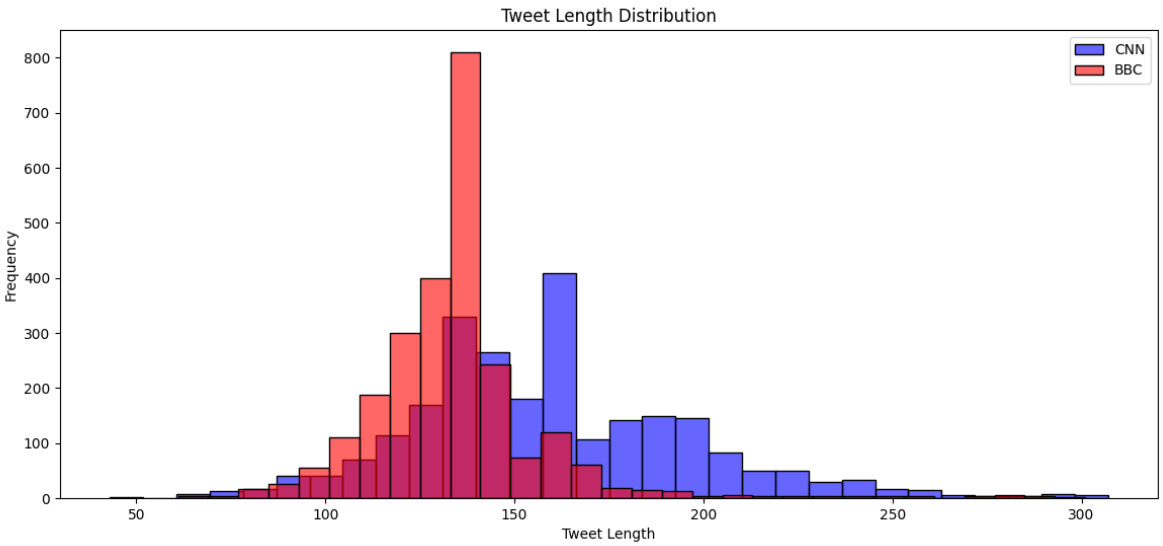
במסגרת הפרויקט, בחרנו לקיים ניתוח מעמיק של ציוצים מאתר חדשות, בשנים 2010-2020. לצורך כך, אספנו כ-90,000 ציוצים משני אתרי חדשות מובילים CNN,BBC. בחרנו מ2016 והלאה בשל אירועים משמעותיים כגון נשיאותו של דונלד טראמפ ויציאת בריטניה מהאיחוד האירופי ,(Brexit) אשר עוררו עניין ציבורי רב ויצרו דיון סוער ברשתות החברתיות. מתוך כלל הציוצים, דגמנו באופן אקראי 2,500 ציוצים מכל אתר(בחרנו לעבוד עם 5000 ציוצים כדי שזמן ריצת התוכנית יהיה נוח בזמן ריצה), כך שנוכל לבצע ניתוח מקיף תוך התמקדות בתקופה ובארועים הרלוונטיים.

.2
הדבר הראשון שבדקנו בEDA הוא את כמות הציוצים היומית ואת אורכם הממוצע, תוצאות:
Average number of CNN tweets per day: 2.25
Average number of BBC tweets per day: 2.47



מספר הציוצים היומי הממוצע של CNN ו BBC-הפתיע אותנו. בהתחשב שהם אתרי חדשות מובילים, ציפינו לכמות ציוצים גבוהה יותר. ייתכן כי תהליך הדגימה האקראי השפיע על התוצאה.
בגרף הראשון אפשר לראות שיש ימים ש CNN-לא צייצה בכלל .זה כנראה קשור לבחירה של הציוצים שנבדקו במחקר .בכל מקרה ,הגרף הזה תומך בממצאים האחרים ,כי נראה ש BBC-צייצו יותר כל יום.
ניתן להסיק מהגרף שלכל אתר חדשות יש מין דפוס קבוע של כמות ציוצים ביום אלא אם כן קורה אירוע חריג.

Average tweet length for CNN: 159.52 characters
Average tweet length for BBC: 134.18 characters



ניתוח אורך הציוץ הממוצע העלה כי אין הבדל משמעותי בין CNN ל BBC,
ממצא זה מעניין, שכן הוא עשוי להצביע על כך ששני האתרים מנסים להגיע לקהל דומה, המכיל משתמשים בעלי העדפות דומות לצריכת תוכן קצר וממוקד.
ייתכן כי אורך הציוץ הממוצע מושפע מגורמים כמו אילוצי הפלטפורמה (טוויטר) והרצון להגיע לקהל רחב ככל האפשר.

.Sentiment Analysis – 3

תוצאות:

CNN Sentiment Distribution:

negative : 55.08%

neutral : 23.16%

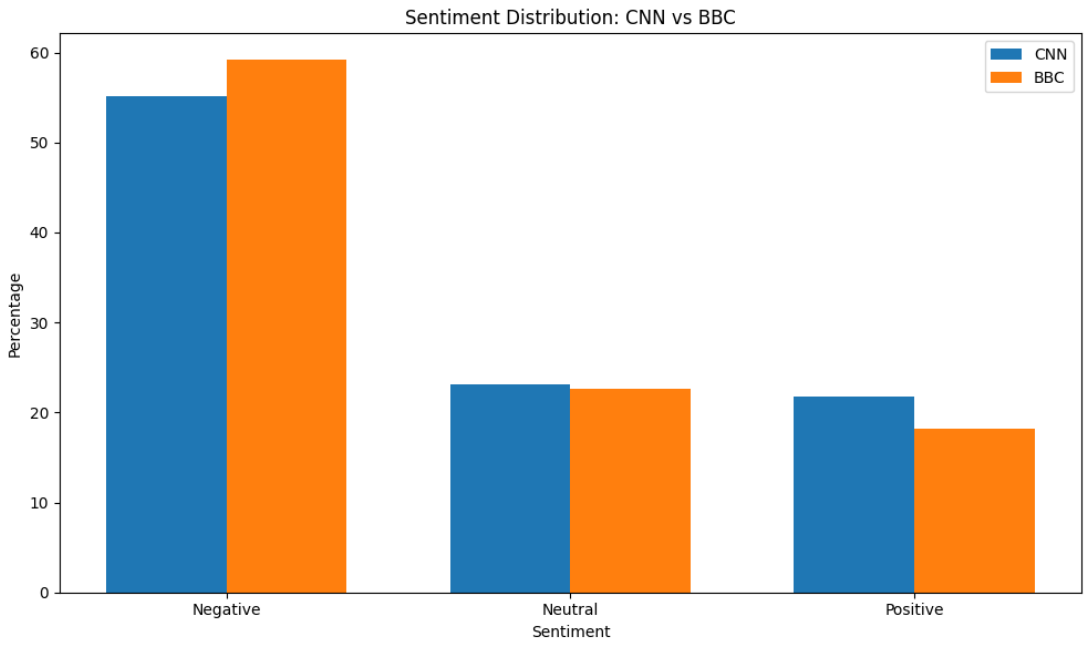
positive : 21.76%

BBC Sentiment Distribution:

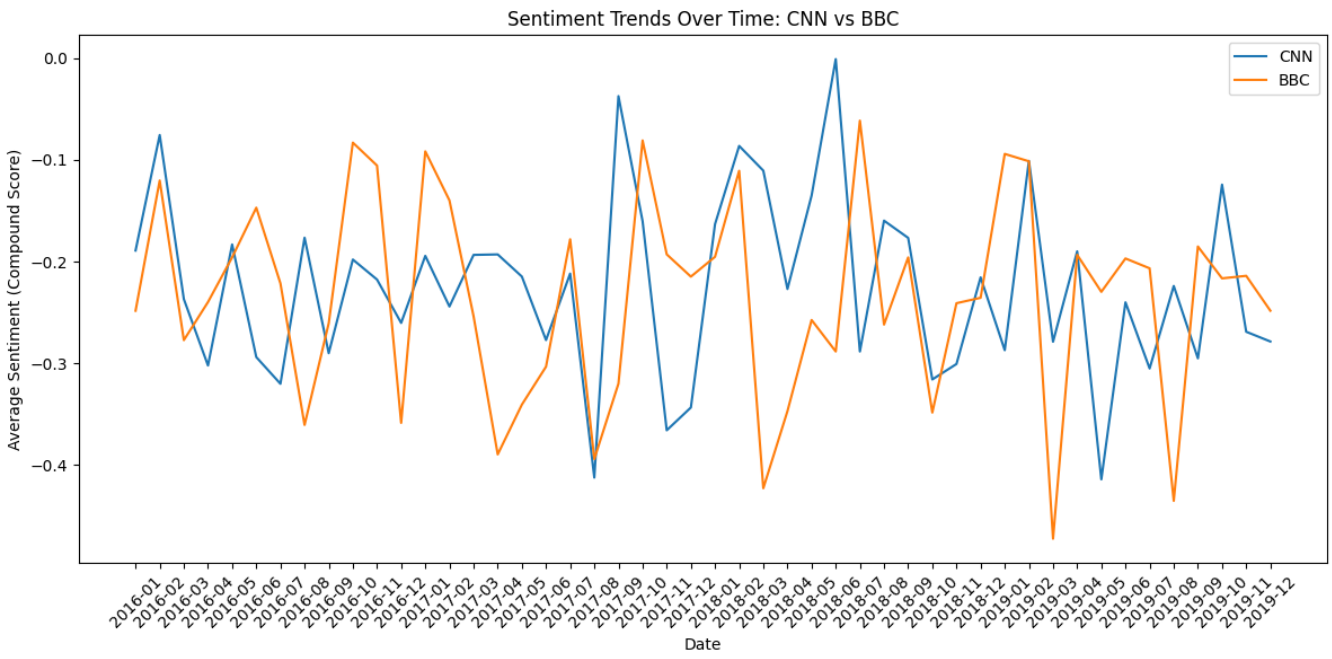
negative: 59.16%

positive: 22.64%

neutral: 18.20%



ניתוח הציוצים מגלה כי אתרי חדשות נוטים להתמקד בעיקר בחדשות שליליות. נראה כי סיפורים שליליים מושכים יותר תשומת לב ומייצרים יותר אינטראקציות ברשתות החברתיות. תוצאה זו מצביעה על כך שאתרי החדשות מעדיפים לפרסם חדשות רעות, גם אם הן אינן משקפות את התמונה המלאה של המציאות.



גרף זה מסמל את ערכו הסנטימנטלי של הציוצים באותו יום, ניתן לראות שלגרפים מגמה דומה מה שמראה לנו ששני האתרים למרות שמייצגים מדינות שונות, מגיבות באותו אופן ולאותן חדשות.

בשלב זה ביצענו Summarization לציוצים, כפי שדברנו כבר חשבנו שציוצים בהגדרתם מקוצרים, ולכן זה מסביר את העובדה שלא קיבלנו תקצירים משמעותיים לרוב הציוצים.

דוגמא לציוץ שהתקצור הניב תוצאה שונה:

Original Tweet 2:

Defense Secretary James Mattis and Homeland Security Secretary John Kelly have been sworn in to their Cabinet jobs https://t.co/y9xspIiJsx https://t.co/7yIB4Ho2Zs

Summarized Tweet 2:

Defense Secretary James Mattis and Homeland Security Secretary John Kelly have been sworn in.

Original Tweet 3:
President #BarackObama opens remarks in #Cuba by pledging solidarity in response to #BrusselsAttack. https://t.co/PNZghLCqGM

Summarized Tweet 3:
President Obama pledges solidarity in response to Brussels attack.

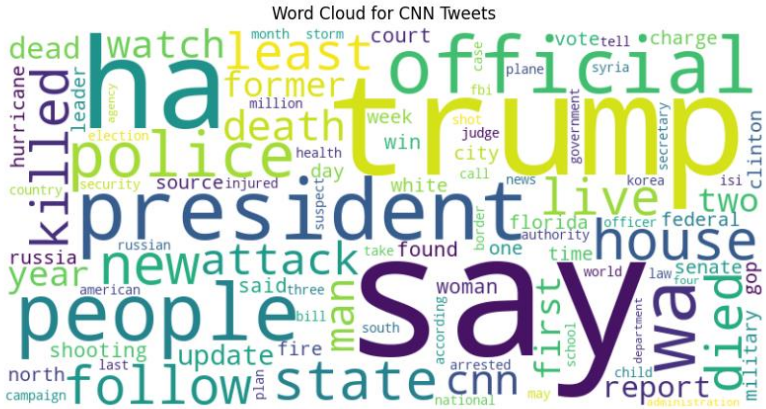
- 5

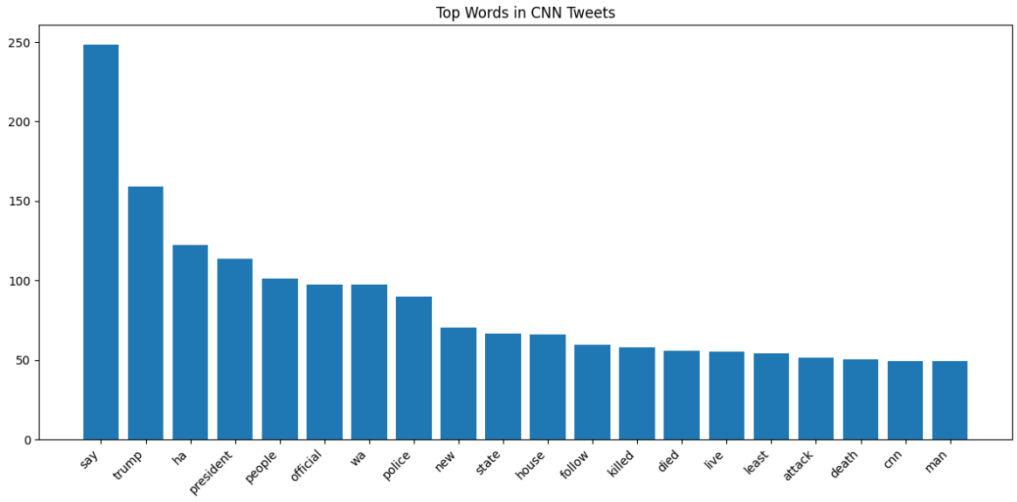
בשלב זה ביצענו Preprocess שכולל:

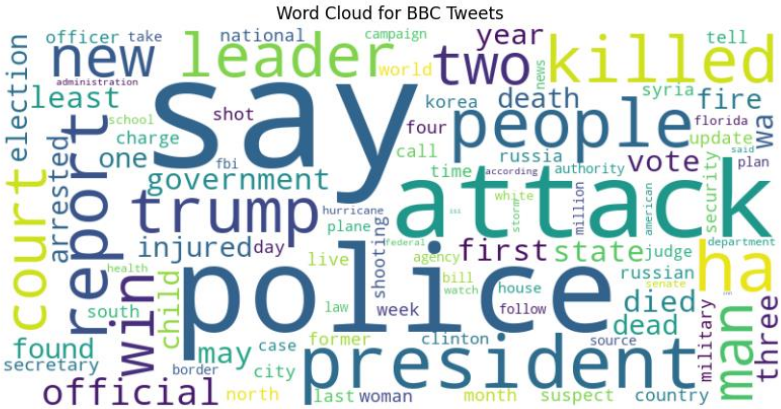
- Tokenization
- Lemmatization
- Stop words removal

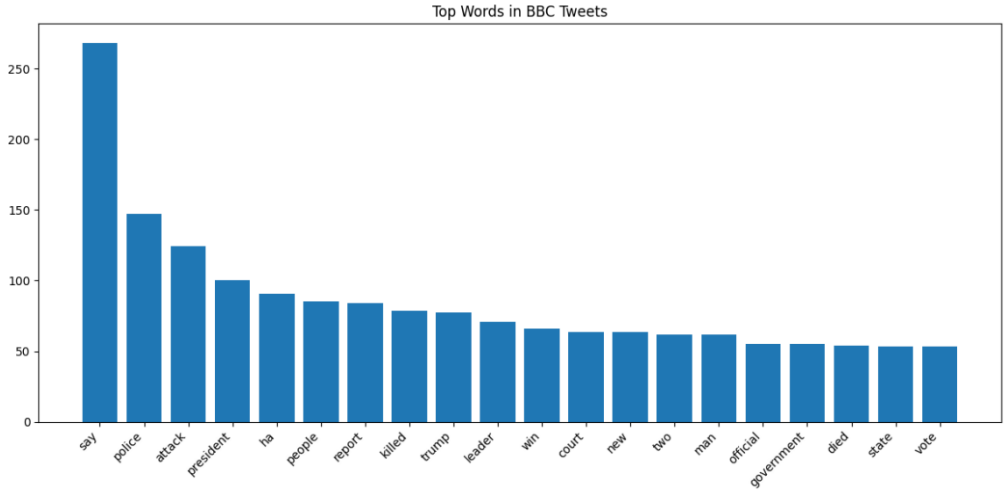
הפעלנו TF\IDF כדי למצוא את 20 המילים הנפוצות ביותר, והשוונו ביניהם.

תוצאות:









ניתוח מילים הכי נפוצות בשני האתרים העלה דמיון מפתיע. למרות ההבדלים בין CNN ו BBC, -שתי הרשתות משתמשות במילים דומות בתדירות גבוהה. כמו שחשבנו המילה trump מופיעה בשני האתרים אך יותר באתר CNN כי הוא מדווח על חדשות בארה"ב, המילה say מופיעה הכי הרבה בשניהם מכיוון שזאת מילה שממומשת בדרך שמצייצים לדוגמא:

Sources Say, Spokesman says... etc.

- 6
כאן השתמשנו במודל Word2Vec בהתחלה ניסינו להשתמש בגודל וקטור של 1024, אך התוצאות לא היו טובות, ניסינו כמה אפשרויות ו100 הניב את התוצאות הטובות ביותר.
מצאנו את 10 המילים החשובות ביותר לפי אתר, והראנו את המילים הקרובות אליהן לפי Word2Vec.

Top 20 words by Word2Vec similarities:
CNN

say: ['said', 'wa', 'day', 'near', 'amp']

trump: ['u', 'former', 'president', 'national', 'republican']

ha: ['u', 'wa', 'year', 'charge', 'day']

president: ['russia', 'u', 'vote', 'tell', 'national']

people: ['injured', 'dead', 'killed', 'one', 'shooting']

official: ['syria', 'show', 'amp', 'day', 'service']

wa: ['charge', 'amp', 'former', 'woman', 'home']

police: ['shooting', 'wa', 'woman', 'report', 'london']

new: ['ha', 'first', 'call', 'government', 'win']

state: ['u', 'ha', 'win', 'report', 'charge']
BBC

say: ['said', 'wa', 'day', 'near', 'amp']

police: ['shooting', 'wa', 'woman', 'report', 'london']

attack: ['amp', 'london', 'home', 'wa', 'near']

president: ['russia', 'u', 'vote', 'tell', 'national']

ha: ['u', 'wa', 'year', 'charge', 'day']

people: ['injured', 'dead', 'killed', 'one', 'shooting']

report: ['state', 'charge', 'police', 'airport', 'near']

killed: ['injured', 'dead', 'people', 'least', 'one']

trump: ['u', 'former', 'president', 'national', 'republican']

leader: ['call', 'plan', 'first', 'win', 'u']

AutoEncoder – מציאת המילים החשובות ביותר.

Top 20 words by Autoencoder:

['discount', 'flagging', 'costa', 'afraid', 'tfl', 'lam', 'torrance', 'lakdim', 'patron', 'trumpkim', 'farmer', 'someone', 'affecting', 'paving', 'hearing', 'marjory', 'monitor', 'reclaiming', 'gerry', 'justintrudeau']

השוואה בין:

- AutoEncoder
- Word2Vec
- TF/IDF

ניתן לראות שTf\IDF וWord2Vec הניבו תוצאות דומות והגיוניות.

לעומת זאת 20 המילים החשובות ביותר של AutoEncoder יצאו שונות, הגיוני שהחשיבות של המילים לא תתכתב באופן חד משמעי עם שכיחותן, אך המילים כמו costa, tfl לא ברורה חשיבותן.

- 7
השתמשנו בNER, NameEntityRecogintion.

Top 20 entities (NER):

CNN: [(('US', 'GPE'), 207), (('Trump', 'ORG'), 177), (('Trump', 'PERSON'), 148), (('CNN', 'ORG'), 105), (('first', 'ORDINAL'), 77), (('Russia', 'GPE'), 64), (('Florida', 'GPE'), 61), (('House', 'ORG'), 55), (('two', 'CARDINAL'), 54), (('GOP', 'ORG'), 52), (('Syria', 'GPE'), 48), (('2', 'CARDINAL'), 48), (('Russian', 'NORP'), 46), (('White House', 'ORG'), 43), (('Senate', 'ORG'), 43), (('FBI', 'ORG'), 42), (('ISIS', 'ORG'), 40), (('U.S.', 'GPE'), 34), (('North Korea', 'GPE'), 34), (('3', 'CARDINAL'), 32)]

BBC: [(('UK', 'GPE'), 258), (('US', 'GPE'), 246), (('Brexit', 'PERSON'), 89), (('London', 'GPE'), 77), (('Donald Trump', 'PERSON'), 75), (('EU', 'ORG'), 64), (('first', 'ORDINAL'), 56), (('BBC', 'ORG'), 50), (('Russian', 'NORP'), 50), (('Two', 'CARDINAL'), 48), (('French', 'NORP'), 44), (('British', 'NORP'), 44), (('Trump', 'PERSON'), 43), (('Syria', 'GPE'), 43), (('two', 'CARDINAL'), 42), (('England', 'GPE'), 40), (('Brussels', 'GPE'), 39), (('#', 'CARDINAL'), 36), (('three', 'CARDINAL'), 36), (('Russia', 'GPE'), 35)]

למשל טרמפ הינו גם שם וגם ארגון.

CNN מתמקדת בעיקר בפוליטיקה האמריקאית ובארגונים כמו בית הנבחרים והסנאט ,בעוד ש BBC-מפגינה פרספקטיבה גלובלית יותר עם אזכורים של בריטניה ,האיחוד האירופי ומדינות שונות.

RNN:

Generated CNN text: Breaking news senator vote change change change change marriage legislation project daily access say title ge2017 royalwedding follow live update throughout evening cnnsotu say always cnnsotu cnnsotu always cnnsotu starting pointless cnnsotu nuclear deal follow live update throughout cnnelection cnnsotu nuclear test say always cnnsotu starting pointless cnnsotu nuclear test say always

Generated BBC text: Latest update killed three others wounded shooting attack police say correcting earlier report say suspect dead 150 people board official say teen wa condition complex wa anniversary 2014 anniversary 2 inch rain tension murder three year ago wa fuel condition complex say court court court paper assault say court coastguard say follow

GPT:

The latest news from CNN:The FBI is investigating the death of a man who was shot and killed by a police officer in a shooting that left one person dead and another wounded. The man was shot in the head and died at the scene. The suspect is being held on $100,000 bail.The suspect is being held on $100,000 bail. The suspect is being held on $100,000 bail. The suspect is

אמנו את המודל RNN בהתחלה במשך 5 EPOC, דבר זה לקח זמן רב, והתוצאות שלו היו לא טובות בכלל, לכן החלטנו לעלות ל15 EPOC, דבר שהעלה פי 3 את משך הריצה של התוכנית, אבל קיבלנו תוצאות יותר טובות, אנחנו מניחים שבהנתן מחשבים יותר חזקים היינו יכולים להריץ עם יותר EPOC והיינו מקבלים תוצאות יותר טובות.
הג'ינרוט של הGPT התחיל טוב, ולאחר מכן נכנס לסוג של לולאה ונעצר.
ההבדלים בין הRNN לGPT הם משמעותיים בגלל שוב, כמות הEPOC שאימנו את הRNN, כי לקחנו מודל מאומן של 2GPT.

לסיכום:

בפרויקט הגמר שלנו בניתוח שפה טבעית (NLP), ניתחנו 5000 ציוצים של BBC, CNNו-תוך התמקדות באירועים מרכזיים משנת 2016 ואילך, כמו נשיאותו של טראמפ וברקזיט. ביצענו ניתוח נתונים ראשוני (EDA), ניתוח סנטימנט ותקצור ציוצים, וגילינו ששני האתרים מפרסמים בעיקר חדשות שליליות. לאחר מכן, ביצענו עיבוד מקדים, יישמנו מודלים של TF-IDF, Word2Vecו Autoencoder-כדי לזהות מילות מפתח, וביצענו זיהוי ישויות (NER)כדי להשוות בין מוקדי הסיקור של CNN ו .BBC-לבסוף, אימנו מודלים של RNN ו GPT-ליצירת טקסט, עם תוצאות מעורבות שהראו את הפוטנציאל והמגבלות של כל מודל.