**דו"ח פרוייקט סיום למידה עמוקה**

מבוא

במסגרת פרוייקט הגמר בקורס ניסינו לענות על כמה שאלות שעניינו אותנו, הנוגעות לפופולריות של תגובות לסרטונים המופיעים ברשת הסרטונים הפופולארית – YouTube.  
כידוע כמעט לכל משתמשי הפלטפורמה, בין אם המשתמש הינו יוצר ופעיל ברשת ובין אם הינו משתמש בסיסי באתר, יוטיוב כיום הינה פלטפורמה ששימושיה הינן מעבר להעלאת סרטונים בלבד, ויש הרואים בה רשת חברתית של ממש המזכירה את פייסבוק או את רשת הפורומים reddit. אפשר לראות הקבלות מסויימות ביכולת של משתמשים לסמן "לייק" לסרטונים והתגובות השונות, אלגוריתמים מתוחכמים שמטרתם להביא לצופה את הסרטונים והתגובות שיענינו אותו יותר על מנת למקסם את זמן הצפייה, דבר הדומה לנעשה ברשתות החברתיות שמעוצבות בצורה כזו שתגרום למשתמשים לרצות להשתמש במשך שעות ארוכות במערכת.  
בנוסף, יוטיוב יוצר קהילות של משתמשים אשר להם נושאים משותפים, ומשתמשים בו בתור מקום בו יכולים להביע את דעתם בנושא הסרטון בו הם מתעניינים – שבא לידי ביטוי במדור התגובות שמתחת לסרטון הרלוונטי, בדומה לקבוצות בפייסבוק כפי שאנו מכירים אותן.

כידוע – ברשת ישנה מערכת לספירת לייקים לכל תגובה. ולשאלתנו – האם מתוך הנתונים שמוצגים לנו על התגובה (התוכן שלה, שם הכותב, שם הסרטון, מספר צפיות וכו') נוכל להעריך האם זוהי תגובה שהקהל הצופה בסרטון יסכים ויאהב וכתוצאה מכך יקבל הרבה "לייקים", או שלחלופין זוהי תגובה שלא תזכה לתשומת לב רבה מצד צופי הסרטון.

על פניו – זו עלולה להיות משימה קשה מכיוון שלתגובות מיוטיוב יכול להיות מידע שמעבר לנתונים היבשים, כזה שקשה לנו להביא לידי ביטוי בתור דאטה. למשל, שתגובה תגרור הרבה לייקים מכיוון שהינה ציטוט מתוך הסרטון – מה שלא נגיש לנו מבחינת הדאטה שבחרנו להשתמש. בכל זאת, אנו רוצים לגלות האם ישנם דפוסים מסויימים שחוזרים על עצמם בתגובות המוצלחות והאם ניתן להשיג תוצאות סבירות בסיווג התגובות השונות.

בחירת הדאטה אותו נסווג

למשימה זו היה באפשרותנו לבחור בין כמה datasets שונים כגון דאטה על הסרטונים הכי נצפים בחודש האחרון, דאטה על הסרטונים הכי נצפים באופן כללי, דאטה על סרטונים אקראיים וכו'. לבסוף, החלטנו להשתמש בדאטה על תגובות לסרטונים הנצפים ביותר באופן כללי ולהחלטה זו כמה סיבות.

1. לסרטונים עם צפיות רבות יהיו גם תגובות רבות שלהן הקשר דומה, שלא זמין לנו מבחינת הנתונים היבשים, כפי שציינו במבוא, וקיווינו כי ניתן ללמוד מידע זה משאר התגובות לגבי סרטון ספציפי, כך שלדעתנו עדיף להשתמש במעט סרטונים עם הרבה תגובות מאשר בהרבה סרטונים עם מעט תגובות ובכך לנסות ולקבל רעיון טוב יותר על הקונטקסט של כל סרטון ולקוות כך לתוצאות טובות יותר.
2. לסרטונים הנצפים ביותר יש צופים מכלל אוכלוסיית העולם – מה שבמובן מסויים מנטרל את הבדלי התרבות והבדלי השפה שעלולות להיות בתגובות לסרטונים עם פחות צפיות. כלומר, לא סביר שבסרטון עם מיליארד צפיות התגובה המובילה תהיה ברוסית או שתכוון רק חלק קטן מצופי הסרטון, אלא יותר סביר שתגובה מובילה תופנה לרוב צופי הסרטון כך שיהיה למסווג קל יותר לזהות דפוסים במובן הזה.
3. מכיוון שלסרטונים אלה יש כמות צפיות עצומה, אזי גם כמות התגובות והלייקים עצומה אף היא באופן יחסי.  
   זהו יתרון מכיוון שאם היינו משתמשים בסרטונים עם מספר צפיות קטן יותר היינו מקבלים (אולי) סרטונים עם מעט מאוד תגובות שרבים מהן כלל לא קיבלו לייקים, והתגובות המובילות היו בעלות מספר לייקים קטן מה שהופך את המשימה לקשה הרבה יותר. קשה הרבה יותר להפריד בין תגובה שקיבלה 30 לייקים לתגובה שקיבלה 100 לייקים לעומת תגובה שקיבלה 30 לייקים בהשוואה לתגובה שקיבלה 6000 לייקים. מכיוון שבסרטונים אלו יש תגובות מכל הסוגים נוכל לקחת את הדאטה מהם בצורה כזו שתהפוך את המשימה שלנו לקלה יותר לסיווג.

מה שמוסיף על הקושי של המשימה הוא שיש לזכור כי יוטיוב מראה קודם כל את התגובות שקיבלו הכי הרבה לייקים, ורק לאחר מכן יראה את התגובות הפחות פופולאריות, על כן אם תגובה כבר פופולארית היא תזכה לחשיפה רבה יותר, לעומת זאת יכול להיות שתגובה תהיה בעלת פוטנציאל להיות "מוצלחת" אך כתוצאה מכך שלא פורסמה בזמן טוב או שדומה מדי לתגובה אחרת היא לא תזכה לאותה פופולאריות, על כן ישנו גם עניין של מזל בקבלת לייקים, מה שיבוא לידי ביטוי ששתי תגובות דומות (אולי אפילו זהות) מבחינת תוכן יקבלו כמות לייקים שונה מאוד.

לינק לדאטה שהשתמשנו:

[MOST LIKED COMMENTS ON YOUTUBE](%20https:/www.kaggle.com/nipunarora8/most-liked-comments-on-youtube)

קלט ופלט

אחרי שבחרנו את הדאטה בו אנו רוצים להשתמש אנו צריכים להחליט כעת כיצד אנו רוצים לבצע עליו סיווג.

פלט – צריך לזכור כי הסרטונים בהם בחרנו להשתמש על מנת להשיג תגובות לדאטה הינם סרטונים בעלי מיליוני צפיות ביום – מה שאומר שכך גם התגובות עליהם, כלומר כמות הלייקים על כל תגובה הינו נתון מאוד דינמי שיכול להשתנות בכל רגע. על כן, רצינו לחזות סדר גדול של כמות לייקים לכל תגובה, ולא מספר מדוייק. כלומר, אנו נרצה את כל אחת מהתגובות לסווג לאחת מהקבוצות הבאות:

0 לייקים

1-10 לייקים

11-100 לייקים

101-1000 לייקים

1001-10000 לייקים

10001-100000 לייקים

100001+ לייקים

הסיבות שבגללן הגודל של כל סדר גודל גדל באופן אקספוננצלי היא מכיוון שתגובות עם מעט לייקים הרבה יותר נפוצות מאשר תגובות עם כמות גדולה של לייקים מהסיבות שצויינו לעיל. באמצעות חלוקה זו אנו מחלקים את הדאטה שבידינו לקבוצות שדומות בגודלן אחת לשניה, מה שיעזור לנו למצוא דפוסים בדאטה לסיווג. בנוסף לכך, כפי שצויין, יוטיוב מראה למשתמש ראשית את התגובות בעלות כמות לייקים רבה ועל כן ההבדל במספרים הגדולים בלייקים פחות משמעותי מאשר במספרים הקטנים מכיוון שהתגובות שקיבלו כבר כמות לייקים גדולה נחשפים ליותר אנשים ועל כן ההבדל בין 20000 לייקים ל30000 פחות משמעותי במקרה הזה בין 0 ליקיים ל-5 לייקים, מכיוון שהתגובה עם 20000 לייקים נצפתה על ידי כמעט כל צופי הסרטון, לעומת התגובה בעלת 0 לייקים שנצפתה על ידי בודדים בלבד לרוב.

קלט – המידע שאנו נשתמש בו בתור הקלט לרשת הנוירונים שלנו הינם הנתונים הבאים שאנו יודעים על התגובה:

תוכן התגובה – נשתמש בתגובה כפי שהיא, כולל סימנים, אמוג'ים וכו' (כדי להפיק מקסימום מידע על התגובה). עליה נבצע מעיין "sentiment analysis", עליו נפרט בהמשך.

שם כותב התגובה – נבצע על השם עיבוד באמצעות LSTM על תווים בתקווה שנוכל לקבל מידע על שם המעלה וכך אולי להסיק מאיפה מגיעה התגובה מה שיכול לתת אינדיקציה לכמות לייקים (שם אמריקאי סביר יותר לקבל כמות לייקים רבה יותר משם משמתש ישראלי בגלל רמת אנגלית טובה יותר, הפרשי שעות, הזדהות עם יותר אנשים וכו'). פרט מידע זה יכול רק להוסיף מידע נוסף למסווג שיכול לעזור לסיווג טוב יותר.

שם הסרטון – כפי שצויין לעיל, הקונטקסט ילמד טוב יותר לגבי כל סרטון אם המסווג ידע על איזה סרטון התגובה התקבלה, מכיוון שלתגובות לאותו סרטון יש סבירות גבוהה להיות דומות זו לזו מבחינת ההקשר שלהן.

שם הערוץ המעלה – מכיוון שיתכן שהערוץ המעלה העלה כמה סרטונים שבהם השתמשנו לסיווג (כמו למשל Ed Sheeran) יתכן והתגובות יהיו דומות בין הסרטונים השונים של אותו ערוץ וכך אנו יכולים להבין טוב יותר מה יביא לתגובה מוצלחת על הסרטונים אותו העלה הערוץ.

תאריך העלאת התגובה – סביר שתגובות ישנות יותר יזכו לחשיפה רבה יותר ויזכו לכמות לייקים גדולה יותר.

ארכיטקטורה ותכנון המודל

לצורך בניית מודל המתאים לנתונים בידינו, היינו צריכים לבנות מודל מיוחד המתאים לדאטה שיש לנו.

דאטה זה מכיל את הנתונים הבאים:

תוכן התגובה, שם הסרטון, שם הערוץ, שם המשתמש של המגיב, תאריך העלאת התגובה וכמות הלייקים אותה קיבלה התגובה.

נתונים אלו מתחלקים ל-2 קבוצות, נתונים הבנויים בצורת רצפים התלויים בזמן, ונתונים שקבועים בזמן.

הנתונים שבנויים בצורת רצפים:

* תוכן התגובה – זהו משפט בעל משמעות בין המילים, וכיודע במקרים כאלה נבחר להתייחס למשפטים בתור רצפים.
* שם המשתמש של המגיב – מכיוון ששמות המשתמשים אינו קבוע אולי נוכל ללמוד על המגיבים אם נסתכל במקום על שם המשתמש ככלותו, אלא נראה בו כרצף של תווים, וכך נוכל להסיק משהו על המשתמש (כגון מוצא או קשר לסרטון).

הנתונים שאינם תלויים בזמן:

* שם הסרטון – נוכל להתייחס לכל סרטון שבידינו בתור אינדקס המייצג את הסרטון שיתקבל בתור קלט למסווג ובכך נוכל להסיק מידע על התגובה בהתאם לסרטון בו מופיעה.
* שם הערוץ – בדומה לשם הסרטון נוכל לייצג כל ערוץ עם מספר סידורי שתינתן כקלט למסווג.
* תאריך העלאת התגובה – תאריך זה הינו קבוע ונוכל להעביר אותו בצורת מספר כפי שאנו יודעים למסווג.

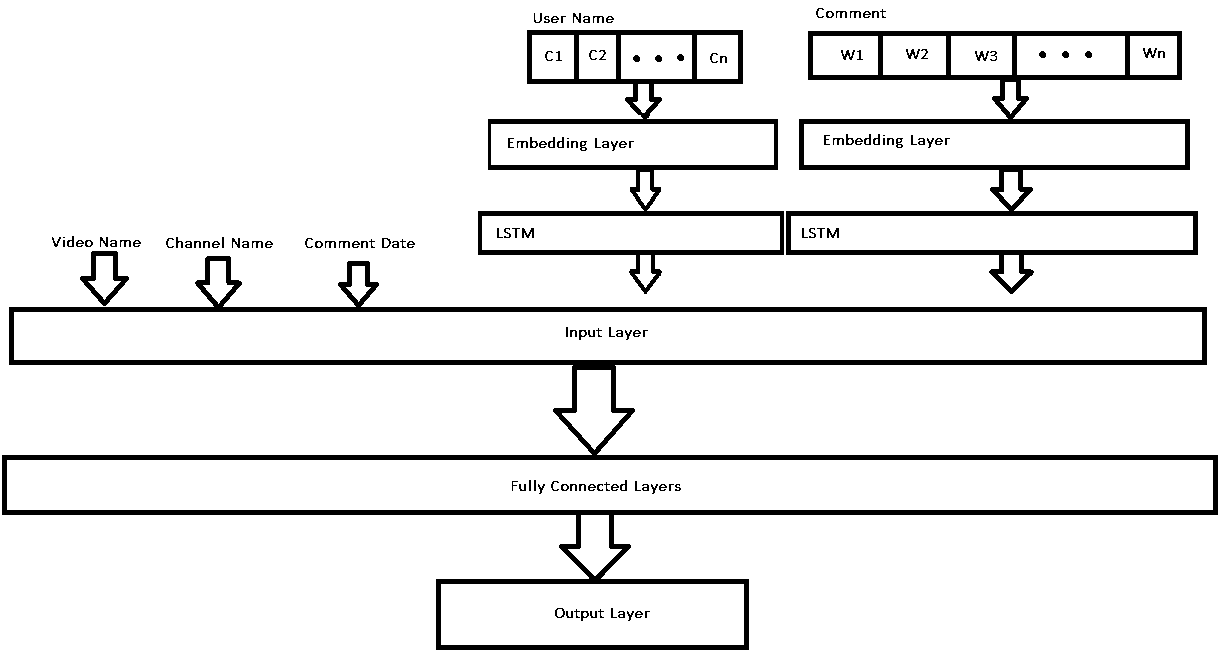
אנו רוצים שהסיווג יעשה על ידי כל הנתונים שבידינו, אך העובדה שחלק מהנתונים בצורת רצפים וחלק לא מצריך מאיתנו ליצור מודל מתוחכם יותר על מנת להתמודד עם קלט מהסוג הזה.

ראשית, אנו נרצה לקחת את תוכן התגובה, שזהו המידע המשמעותי ביותר בקלט, ולהמיר אותו ממידע על רצפים למידע שאינו תלוי בזמן על מנת שיתאים לשאר הפונקציה. על כן, נעביר את הקלט לאינדקסים, ולאחר מכן שכבת Embedding שיצרנו, משם נעבור בLSTM ונקח את הפלט של שכבה זו בתור הקלט של המסווג הכללי שלנו.

נוסף על כן, נעשה דבר דומה לרצפי השם של מעלה התגובה, כאשר ההבדל העיקרי הוא שהתוכן עליו נבצע סיווג רצפים במקרה זה הינו לתווים ולא למילים, בניגוד לאיך שעשינו לתוכן התגובות כפי שהוסבר.

את הפלטים של השכבות שציינו לעיל, נחבר לשאר הקלטים שאינם תלויים בזמן ואת מה שקיבלנו נוכל לסווג באמצעות רשת Fully Connected, אשר שכבת הפלט שלה בגודל 7, כמספר הבינים שקבענו מראש לכמות לייקים.

להלן איור המתאר את הרשת לנוחות:



פירוט השכבות:

שכבת Embedding של תוכן התגובות: ממד ה-embedding – 128  
LSTM תוכן התגובות – ממד כניסה כגודל ממד ה-embedding – 128, hidden – 128, מספר השכבות – 5, ממד היציאה – 32.

שכבת Embedding של שם המגיב: ממד ה-embedding – 16  
LSTM שם המגיב – ממד כניסה כגודל ממד ה-embedding – 16, hidden – 16, מספר השכבות – 3, ממד היציאה – 4.

סה"כ גודל הקלט שהתקבל עד כה – 36, כעת נשרשר אליו את הקלט שאינו בצורת רצפים.

תאריך העלאת התגובה – בנוי מ-6 נתונים – יום, חודש, שנה, שעה, דקה, שניה.  
שם הסרטון – מיוצג על ידי אינדקס שנקבע לפני תהליך האימון.  
שם הערוץ – מיוצג על ידי אינדקס שנקבע לפני תהליך האימון.

סה"כ התקבל שכבת קלט בגודל 32+4+6+1+1 = 44.

את שכבת הקלט הזאת נסווג באמצעות 3 שכבות Fully Connected שגודלן להלן:

שכבה ראשונה – ממד כניסה 44, ממד יציאה 128.

שכבה שנייה – ממד כניסה 128, ממד יציאה 32.

שכבת פלט – ממד כניסה 32, ממד יציאה 7 (כמניין הבינים).

עבור כל אחת משכבות אלה שכבת אקטיבציה בגודל המתאים, בחרנו להשתמש ב-logsoftmax, מכיוון שהסיווג שנעשה כאן אינו בינארי, אלא ישנם 7 אפשרויות פלט שונות.

את האימון החלטנו לעשות ב-batches מגודל 32.

את כל חלקי הרשת שציינו חיברנו יחד לרשת אחת שתתאים לקלט כפי שאנו משתמשים בו.