

**אוניברסיטת תל אביב**

**הפקולטה להנדסה**

**מדעים דיגיטליים להיי-טק**

**פרויקט סדנה ביישומי למידה - דו״ח מסכם**



מרצה: ד״ר אורן אלישע

מתרגל: מר עודד עובדיה

מגישים:

ניצן כהן - 209511187

ינאי איתם ברדוש – 209540517

עידן בר עוז – 206670853

05.03.2023

סמסטר א', תשפ"ג

**סיכום פרויקט**

**רקע**

במסגרת קורס סדנת יישומי למידה, פיתחנו מודל אשר מטרתו היא לחזות האם ציוץ ברשת החברתית ״טוויטר״ מתאר התרחשות של מצב קיצון-אסון (כדוגמת רעידת אדמה). בפרויקט חקרנו שני דאטה סטים של ציוצים בניסיון לחזות האם פרמטרים מסוימים של הציוץ, כדוגמת מספר לייקים גבוה באופן חריג או האם המשתמש מאומת, יכולים לשפר את יכולת החיזוי האם ציוץ מתאר אסון אמיתי.

תוצאות הפרויקט יוכלו לשמש כדי להתריע על אסונות ולעזור לזהות אסונות אמיתיים מהר יותר על ידי ניתוח נתוני מדיה חברתית. מידע זה יכול להיות בעל ערך לצוותי ניהול אסונות בזיהוי ותגובה לאסונות במהירות וביעילות רבה יותר.

הפרויקט כולל שימוש באלגוריתמים של למידת מכונה ולמידה עמוקה, ושיטות NLP על מנת לנתח את הציוצים ולחזות את הלייבלים.

**תיאור הפרויקט ותוצאותיו**

**בעיה 1- תחזית האם ציוץ עוסק באסון אמיתי**

תחילה, חקרנו את הדאטה הראשון הכולל ציוצים שונים שחלקם עוסקים באסונות. מצאנו כי יש פחות ציוצים שהתוכן שלהם קשור לאסון אמיתי (נספח 1). בנוסף, בדקנו את מילות המפתח והאורך של הציוצים לפי הלייבל הרלוונטי (נספח 2), עיבדנו את הטקסט בכך שהורדנו מילות stopwords ותווים (בעזרת regex), וביצענו נרמול מילים בעזרת שיטת lemmatization על מנת להוריד את ממדיות הבעיה.

לאחר מכן, השארנו רק את הטקסט המעובד והלייבל הרלוונטי. חילקנו את סט האימון לאימון וולידציה, הפכנו את המילים לווקטורים בעזרת tf-idf, וביצענו 3 מודלים של למידת מכונה: Logistic Regression, Random Forest, MultinomialNB ובנוסף מודל למידה עמוקה. בעזרת סט הוולידציה גילינו כי מודל הDL מביא את התוצאות הטובות ביותר 82.2% (נספח 3) ולכן השתמשנו בו על מנת לחזות על סט המבחן. בתוצאה על סט המבחן קיבלנו 79.5% accuracy (נספח 4).

**בעיה 2- תחזית האם ציוץ עשוי לקבל מספר לייקים גבוה באופן חריג, ובנוסף תחזית האם משתמש מאומת**

כעת, חקרנו את הדאטה השני הכולל ציוצים שעוסקים בקורונה. מצאנו כי יש 64 שפות שונות ובחרנו להתייחס לשפה האנגלית בלבד. בהמשך, מיינו את הציוצים לפי תאריך פרסומם על מנת להתאמן על הציוצים המוקדמים ולתת תחזית על המאוחרים יותר, כמו שלמדנו בבעיות מסוג timeseries. בחרנו לצמצם את הדאטה שלנו, לפי כמות עוקבים וסטטוסים מעל החציון משום שמטרתנו היא להפיץ את הידיעה על האסון בצורה המהירה ביותר, ולכן נתייחס רק למשתמשים הללו. גם כאן, עיבדנו את הטקסט באותה דרך כמו בבעיה הראשונה.

במהלך חקירת הדאטה, גילינו ש95% מהציוצים קיבלו מתחת ל58 לייקים (נספח 5). בחרנו להתייחס לבעיה זו מסוג outlier detection, על מנת לזהות את תוכן הציוצים שמקבלים כמות גבוהה של לייקים באופן חריג. נציין כי ציוץ חריג הינו בעל לייבל 1, וציוץ שאינו חריג הינו בעל לייבל 0 (נספח 6). אנחנו מניחים כי ציוץ שמקבל כמות לייקים גבוהה הוא ציוץ שמופץ בצורה רחבה. לשם כך, השארנו רק את הטקסט המעובד והלייבל הרלוונטי (ציוץ חריג או לא). חילקנו את סט האימון לאימון וולידזציה, ביצענו embedding וחזינו בעזרת מודל למידה עמוקה בצורת autoencoder.

כדי למדוד את תוצאתנו השתמשנו ב- recall מכיוון שמדובר בבעיית זיהוי חריגים וקיבלנו תוצאה שאינה מספקת- 26.5% (נספח 7). בנוסף, תחזית כמות קבלת הלייקים שציוץ יקבל הינה בעייתית מכיוון שהיא משתנה עם הזמן. בכל זאת בהמשך נרצה להשתמש בתחזית זו, על מנת לבדוק אם אנחנו מצליחים לשפר את תוצאות המודל הראשון.

בעקבות כך, החלטנו לנסות לשפר את תוצאות המודל הראשון בעזרת תחזית האם המשתמש מאומת ע״י טוויטר או לא. הכוונה במשתמש מאומת היא משתמש בעל תג כחול. התג הכחול בטוויטר מאפשר לאנשים לדעת שחשבון המשתמש בעל עניין ציבורי ואותנטי. אנו טוענים כי פרמטר זה עשוי להעיד על יכולת הפצת הציוץ בצורה רחבה.

כעת, עבור בעיה זו גילינו כי יש בערך שני שליש משתמשים לא מאומתים ע״י טוויטר ושליש מאומתים (נספח 8). בדקנו את אורך הציוצים וכמות הלייקים הממוצעת לכל סוג משתמש. קיבלנו כי משתמשים מאומתים כותבים ציוצים ארוכים יותר ומקבלים יותר לייקים בממוצע (נספח 9-10), מה שיכול לחזק את טענתנו. לשם כך, השארנו רק את הטקסט המעובד והלייבל הרלוונטי (סוג המשתמש). חילקנו את סט האימון לאימון וולידזציה. גם כאן ביצענו embedding וחזינו בעזרת מודל RNN מסוג LSTM. קיבלנו 84.36%accuracy על סט המבחן (נספח 11).

**בעיה 3- ניסיון שיפור המודל הראשון**

לבסוף, ביצענו שלושה מודלים על דאטה האסונות:

1. מודל שחוזה את אפיון המשתמש.
2. מודל שחוזה האם ציוץ עשוי לקבל מספר לייקים גבוה באופן חריג.
3. מודל שמשלב את שניהם.

מכיוון שאין לנו דרך לוודא זאת אנו מגדירים זאת כבעיית unsupervised.

מודל ראשון יוצר עמודה של האם המשתמש מאומת או לא, קיבלנו כי היחס בין המשתמשים המאומתים ללא מאומתים הוא דומה ליחס שבדאטה שעליו התאמן המודל(שני שליש לא מאומתים ושליש מאומתים, בקירוב). מודל שני יוצר עמודה שמציינת האם הציוץ עשוי לקבל כמות לייקים גבוהה באופן חריג. נציין כי אנחנו פחות מסתמכים על תוצאותיו משתי הסיבות שציינו לעיל, אך עדיין נבדוק את השפעתו על הבעיה הראשונה.   
מודל שלישי יוצר את העמודות שיצרו המודל הראשון והשני, על מנת לבדוק את ההשפעה של שילוב המידע המתקבל.

לאחר חקירה כיצד לשלב עמודת טקסט עם עמודה בינארית, גילינו כי אחת השיטות הינה לשרשר את הווקטור tf-idf של כל ציוץ (מהדאטה הראשון) עם העמודה הבינארית שיש לנו ולערום כל אחד מהווקטורים הללו זה על גבי זה כדי לקבל מטריצה.

אימנו את ה DLעל המטריצה הנ״ל, וביצענו תחזית האם ציוץ מדבר על אסון אמיתי או לא, וקיבלנו את התוצאות הבאות (נספח 12) :

1. מודל 1- עבור מודל זה קיבלנו 0.7977% accuracy.
2. מודל 2- עבור מודל זה קיבלנו 0.7968% accuracy.
3. מודל 3- עבור מודל זה קיבלנו 0.7974% accuracy.

נבחין כי תוצאות המודלים מעט גבוהות יותר, אך התוצאות הללו לא מתקבלות בכל ההרצות.

**מסקנות**

הפרויקט מראה את השפעת ההתחשבות בסטטוס האימות של המשתמש ו/או כמות לייקים גבוהה בחיזוי האם ציוץ קשור לאסון. ניתן להסיק מתוצאות המודלים כי הוספת העמודות הבינאריות הנ״ל אינה משפיעה בצורה משמעותית על החיזוי האם הציוץ עוסק באסון אמיתי.

ייתכן כי תוצאת הפרויקט הייתה משתפרת אילו היינו מצליחים לבנות מודלים מותאמים יותר לבעיה. בנוסף, לא מן הנמנע כי דאטות חופפים יותר היו יכולים לשפר את התוצאות.

**קשיים שנתקלנו במהלך הפרויקט**

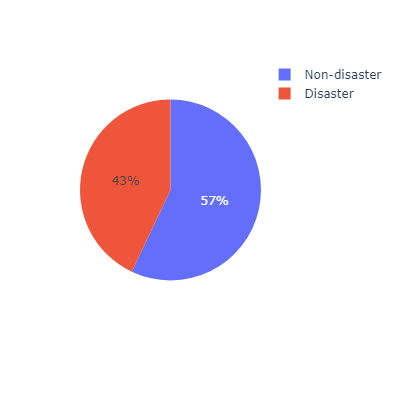
קושי מרכזי שחווינו במהלך הפרויקט היה חיזוי מספר הלייקים לכל ציוץ. בעיה זו מורכבת מהמון גורמים שונים, כגון מאפיינים שונים של המשתמש אשר צייץ, זמן הציוץ, תוכן הציוץ ועוד גורמים בלתי נצפים. לכן כאשר ניסינו לחזות זאת, קיבלנו תוצאות בדיוק נמוך ולא הצלחנו ליצור מודל מוצלח אשר מתחשב בכלל הגורמים.

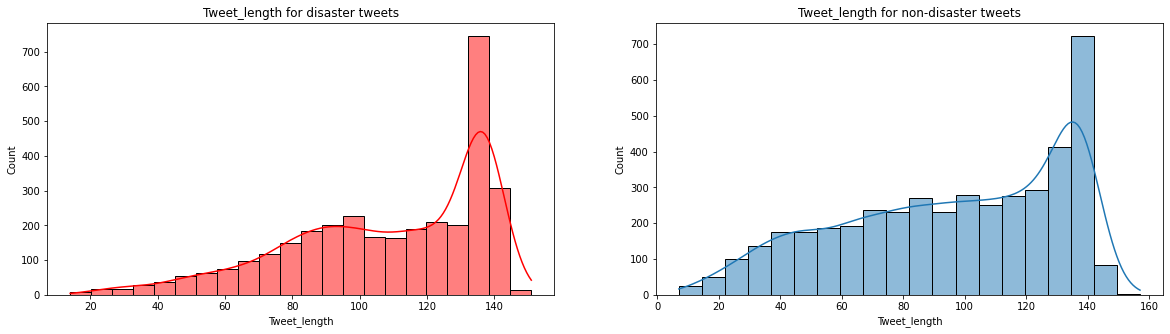
כתוצאה מקושי זה, בחרנו לשנות את סוג הבעיה לoutlier detection, כדי להפוך את הבעיה לפשוטה יותר. גם כאן נתקלנו בקושי שתיארנו, ולכן החלטנו לנסות לחזות את סוג המשתמש.

קושי נוסף הוא התמודדות עם שני דאטות שאין בהן חפיפה מלבד טקסט הציוץ. היינו רוצים שתהיה חפיפה בין שני הדאטות כדי לקחת פרדיקציה של מודל אחד כפיצ׳ר במודל השני. התמודדנו עם קושי זה כפי שהצגנו בבעיה השלישית.

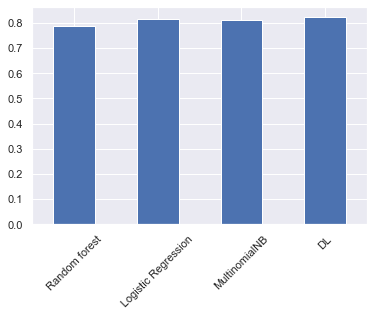
**נספחים**

1.

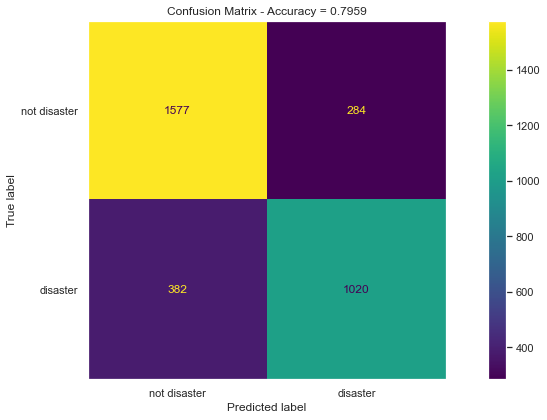


 2.

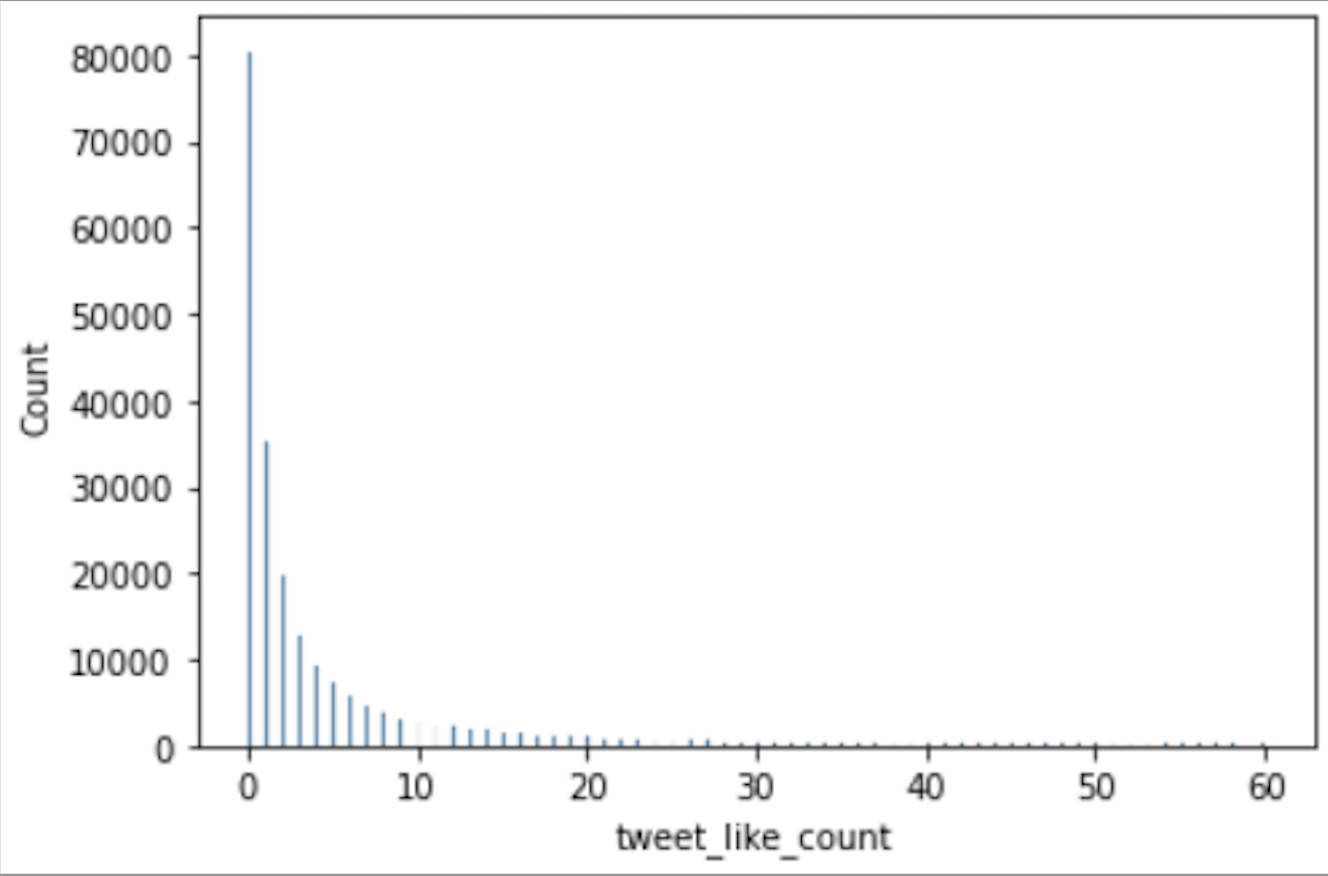
3.



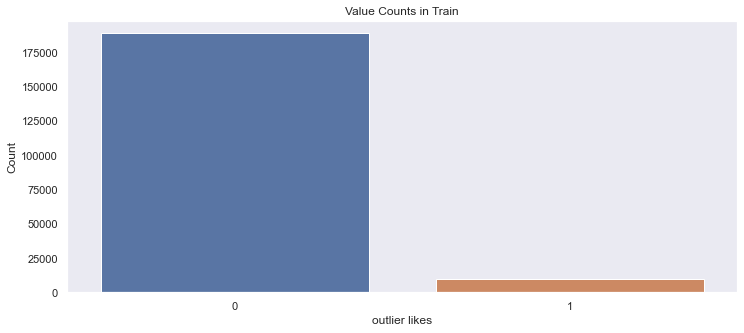
4.

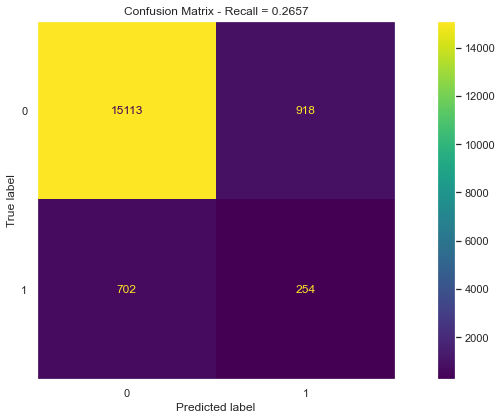


5.

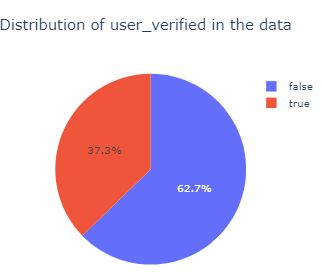


6.

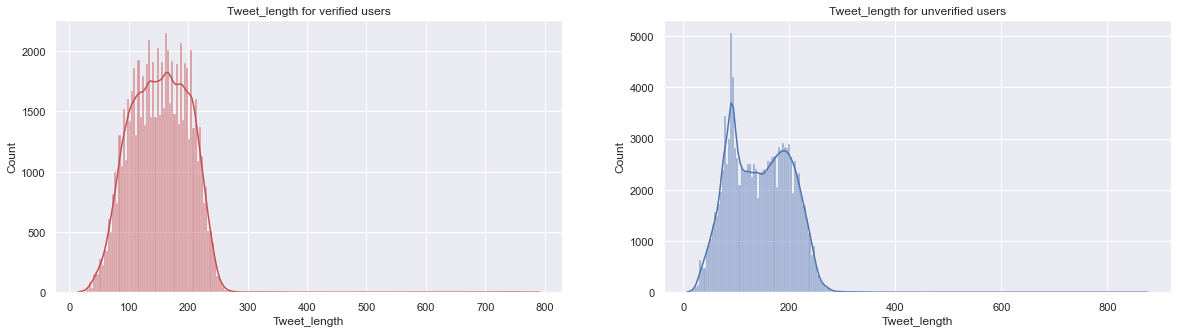


7.

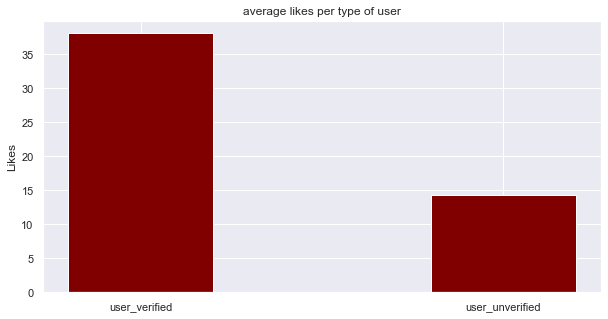
8.



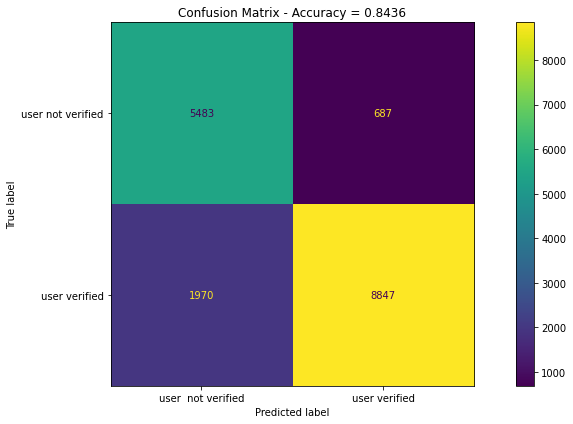
9.



10.



11.



12.

