אוניברסיטת תל אביב

הפקולטה להנדסה

מדעים דיגיטליים להיי-טק

<u>פרויקט סדנה ביישומי למידה - דו"ח מסכם</u>

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

הרקע לפרויקט

הפרויקט נועד לחקור את האפשרות לחזות האם ציוץ מטוויטר מתייחס לאסון אמיתי או לא. בפרויקט חקרנו 2 דאטה סטים של ציוצים בניסיון לחזות האם מאפיינים מסוימים של הציוץ, כמו מספר לייקים גבוה באופן חריג או האם המשתמש מאומת, יכולים לשפר את יכולת החיזוי האם ציוץ קשור לאסון אמיתי או לא.

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 או לא

תחילה, חקרנו את הדאטה הראשון הכולל ציוצים שונים שחלקם עוסקים באסונות. מצאנו כי יש פחות ציוצים שהתוכן שלהם קשור לאסון אמיתי(ראו נספח 1). בנוסף, בדקנו את מילות המפתח והאורך שלהם לפי הלייבל הרלוונטי(ראו נספח 2). אחר כך, 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 ביותר82.2% (ראו נספח 3) ולכן השתמשנו בו על מנת לחזות על סט המבחן. בתוצאה על סט המבחן קיבלנו 79.5% accuracy.(ראו נספח 4)

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

שנית, חקרנו את הדאטה השני הכולל ציוצים שעוסקים בקורונה. מצאנו כי יש 64 שפות שונות ובחרנו להתייחס לשפה האנגלית בלבד. לאחר מכן, 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. בחרנו לעשות זאת 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(ראו 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 בעל לייבל 1, וציוץ שאינו חריג בעל לייבל 0(ראו 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.

קיבלנו תוצאה שאינה מספקת אותנו במדידת הrecall- 26.5%(ראו נספח 7). השתמשנו ב- recall מכיוון שמדובר בבעיית זיהוי חריגים. 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. לדעתנו 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(ראו נספח 8). בדקנו את אורך הציוצים וכמות הלייקים הממוצעת לכל סוג משתמש. קיבלנו כי משתמשים מאומתים כותבים ציוצים ארוכים יותר ומקבלים יותר לייקים בממוצע(ראו נספח 9-10), מה שמחזק את טענתינו. 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 accuracy 84.36% על סט המבחן(ראו נספח 11).

בעיה 3- ניסיון שיפור המודל הראשון

לבסוף, ביצענו 3 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 סומכים 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 הDL על המטריצה הנ״ל, וביצענו תחזית האם ציוץ מדבר על אסון אמיתי או לא, וקיבלנו את התוצאות הבאות(ניתן לראות בנספח מספר 12) :

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

נבחין כי תוצאות המודלים יחסית זהות, ולכן נסיק שההשפעה של הוספת העמודות הבינאריות אינה משפיעה בצורה משמעותית על החיזוי האם עוסק באסון אמיתי.

מסקנות

מסקנת הפרויקט הינה שהוספת עמודת משתמש מאומת או לא מאומת לעמודת הטקסט של הציוץ עשויה לשפר את יכולת החיזוי האם ציוץ קשור לאסון או לא. אנו יכולים להסביר שזה הגיוני משום שכאשר משתמש מאומת אז הציוץ שלו ככל הנראה יותר מהימן ממשתמש לא מאומת.

הפרויקט מראה את השפעת ההתחשבות בסטטוס האימות של המשתמש, בחיזוי האם ציוץ קשור לאסון או לא.
תוצאות הפרויקט יכולות 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.

קשיים שנתקלנו במהלך הפרויקט

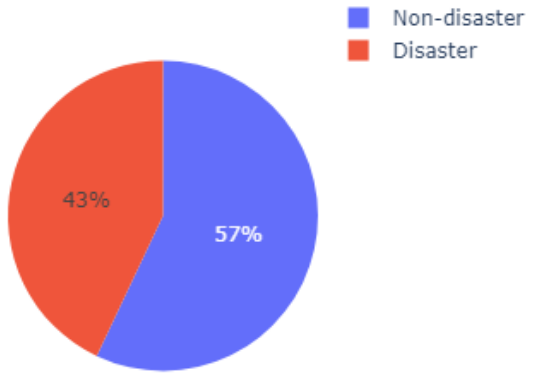
קושי מרכזי שחווינו במהלך הפרויקט היה חיזוי מספר הלייקים לכל ציוץ. בעיה זו מורכבת מהמון גורמים שונים, כגון מאפיינים שונים של המשתמש אשר צייץ, זמן הציוץ, תוכן הציוץ ועוד גורמים בלתי נצפים. לכן כאשר ניסינו לחזות זאת, קיבלנו תוצאות בדיוק נמוך ולא הצלחנו ליצור מודל מוצלח אשר מתחשב בכלל הגורמים.

כתוצאה מקושי זה, בחרנו להפוך את הבעיה לזיהוי מספר לייקים גבוה באופן חריג, כדי להפוך את הבעיה לפשוטה יותר. גם כאן נתקלנו בקושי שתיארנו, ולכן החלטנו לנסות לחזות את סוג המשתמש בכדי לשפר את תוצאות המודל הראשון.

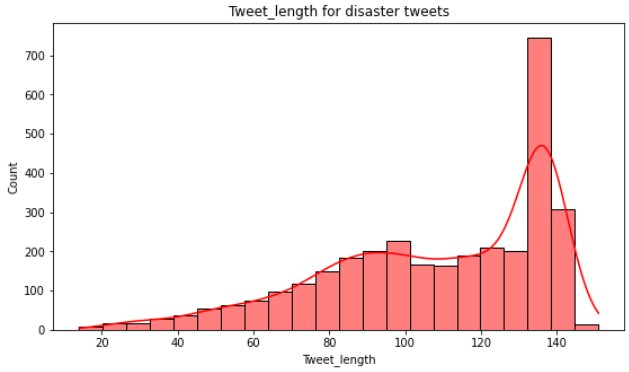
קושי נוסף הוא התמודדות עם 2 דאטות שאין בהן חפיפה מלבד טקסט הציוץ. היינו רוצים שתהיה חפיפה בין 2 הדאטות כדי לקחת פרדיקציה של מודל אחד כפיצ'ר במודל השני. התמודדנו עם הקושי כפי שהצגנו בבעיה 3.

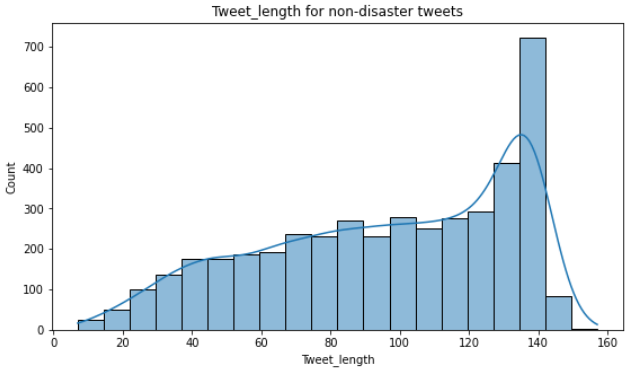
נספחים

.1

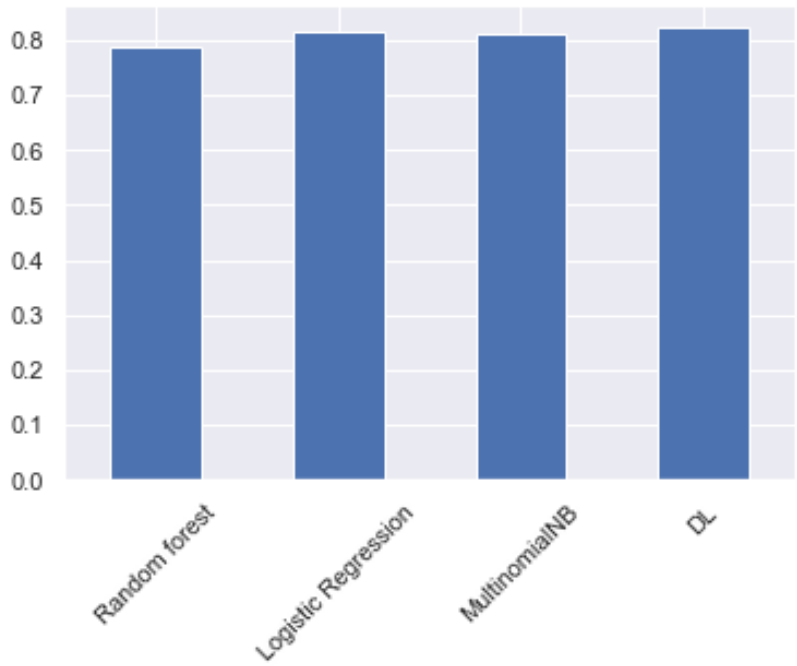


.2

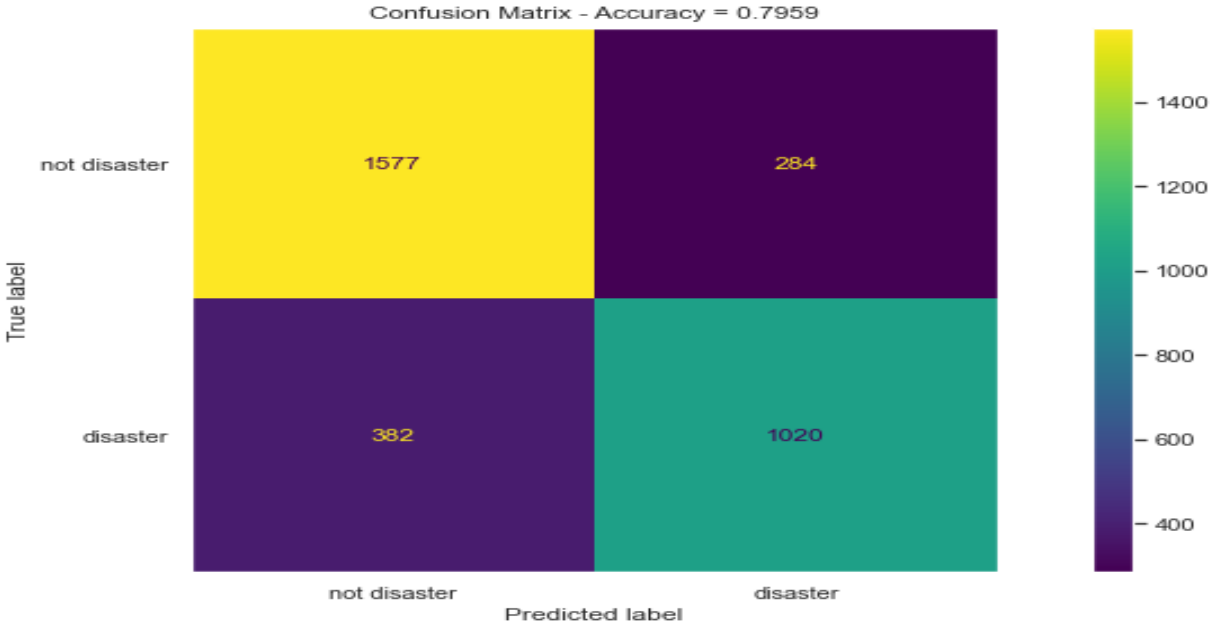




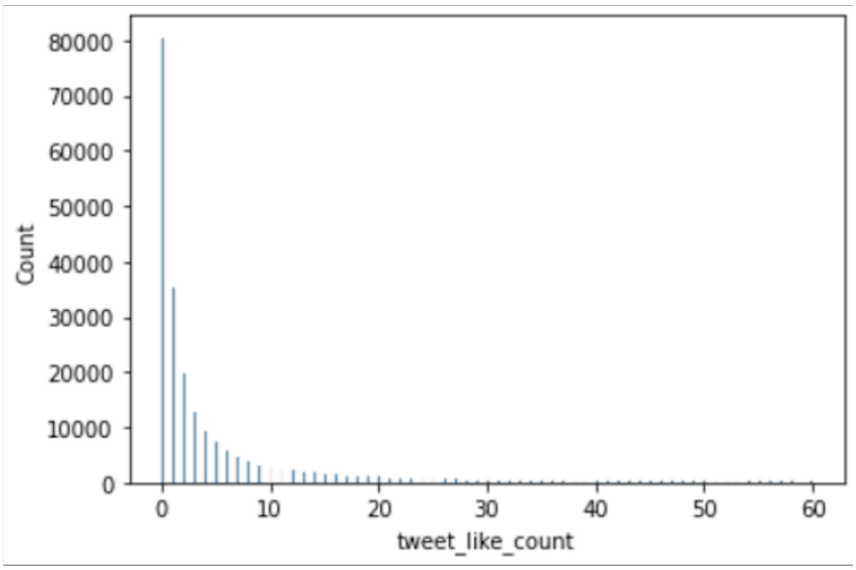
.3


0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
Random forest
Logistic Regression
MultinomialNB
DL

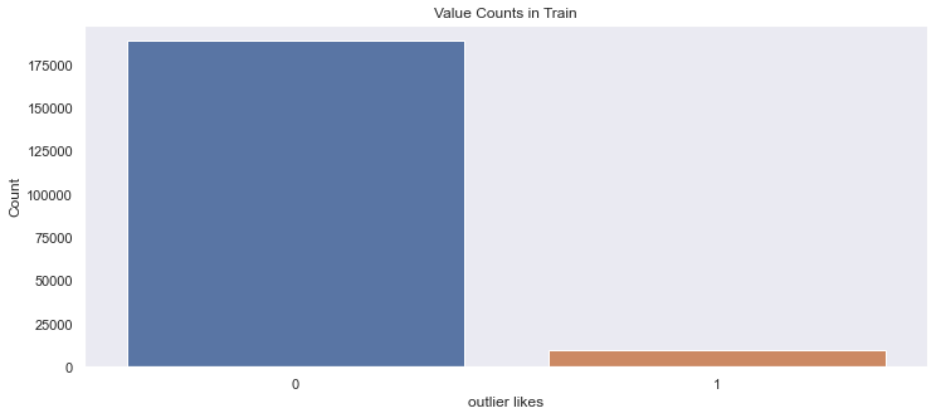
.4


Confusion Matrix - Accuracy = 0.7959
not disaster
disaster
True label
1577
284
382
1020
not disaster
disaster
Predicted label
1400
1200
1000
800
600
400

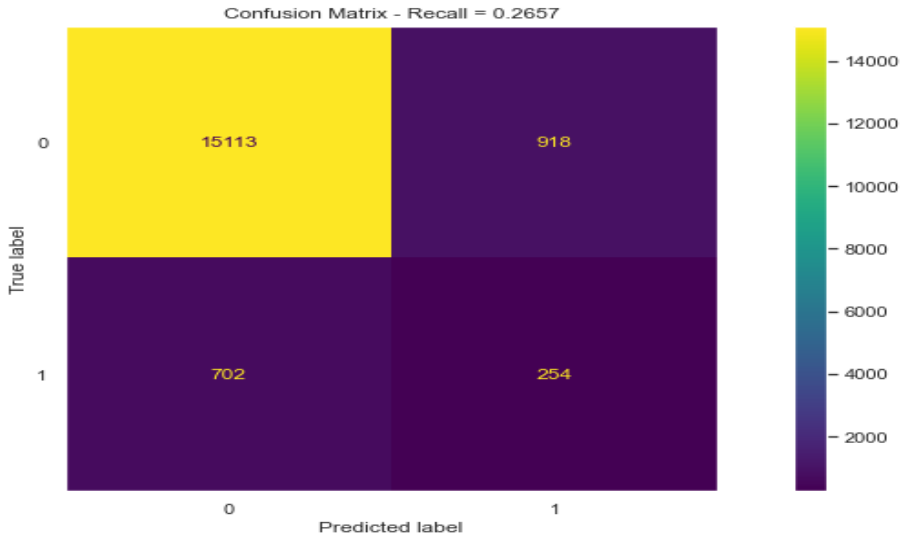
.5


80000
70000
60000
50000
40000
30000
20000
10000
0
Count
0
10
20
30
40
50
60
tweet_like_count

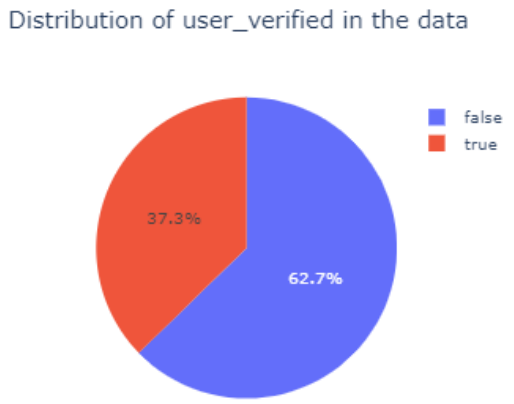
.6


Value Counts in Train
175000
150000
125000
100000
75000
50000
25000
0
Count
0
1
outlier likes

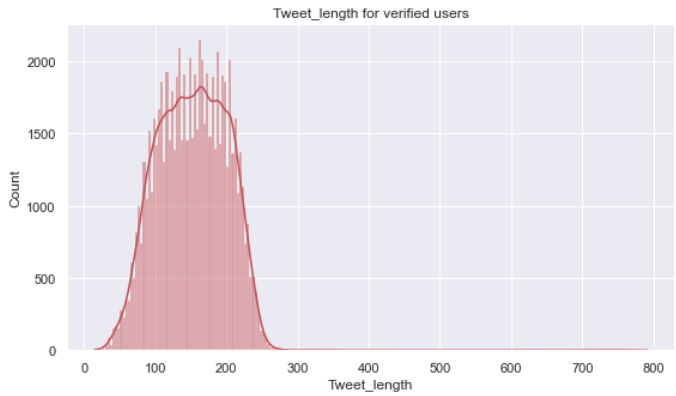
.7

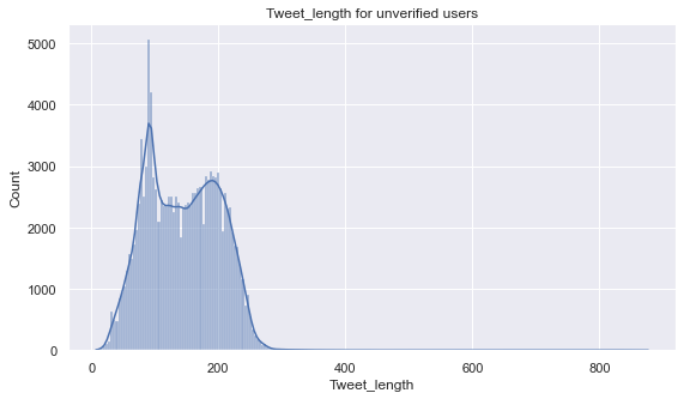

Confusion Matrix - Recall = 0.2657
15113
918
702
254
0
1
True label
0
1
Predicted label
14000
12000
10000
8000
6000
4000
2000

.8

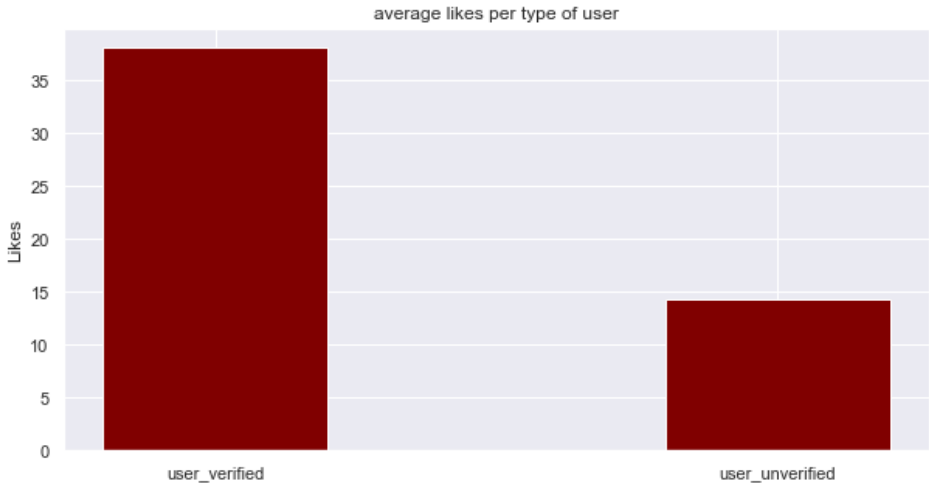

Distribution of user_verified in the data
false
true
37.3%
62.7%

.9

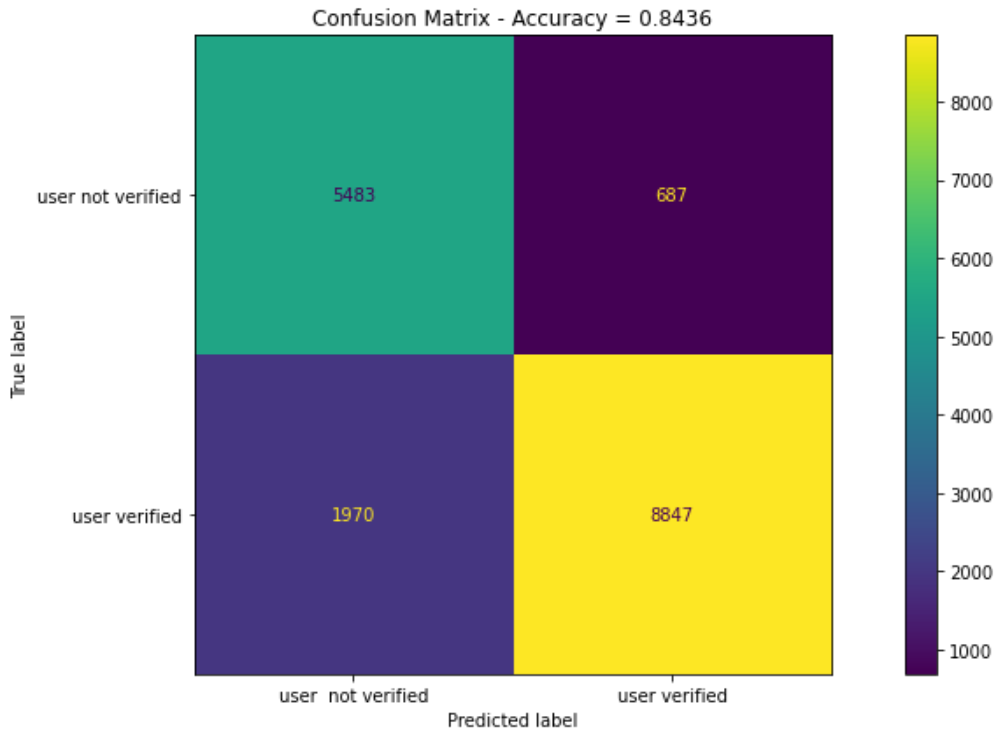

Tweet_length for verified users
Count
Tweet_length
0
500
1000
1500
2000
0
100
200
300
400
500
600
700
800


Tweet_length for unverified users
Count
Tweet_length
0
1000
2000
3000
4000
5000
0
200
400
600
800

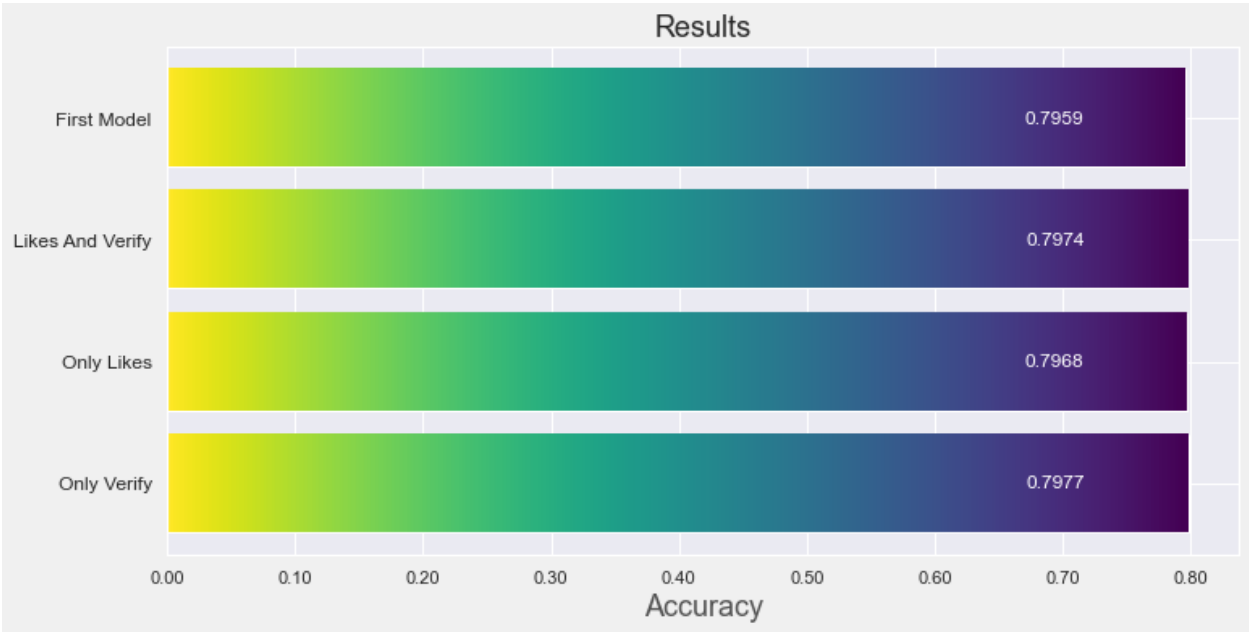
.10


average likes per type of user
Likes
0
5
10
15
20
25
30
35
user_verified
user_unverified

.11


Confusion Matrix - Accuracy = 0.8436
user not verified
user verified
True label
5483
687
1970
8847
user not verified
user verified
Predicted label
8000
7000
6000
5000
4000
3000
2000
1000

.12


Results
First Model
Likes And Verify
Only Likes
Only Verify
0.7959
0.7974
0.7968
0.7977
0.00
0.10
0.20
0.30
0.40
0.50
0.60
0.70
0.80
Accuracy