**סיווג חשבונות בוטים ב-Twitter**

**מגישים:**

**שם:** אסף חאייק ברוך **ת.ז:** 206783441  
**שם:** בן בנוז **ת.ז:** 207570573

**מנחים:**

אלעד נחמיאס ושאול מרקוביץ'

**הקדמה**  
<>

**תיאור הבעיה:**  
כיום הרשתות החברתיות הן חלק משמעותי ובלתי נמנע מחיי היומיום שלנו בעולם המערבי.  
יש להן השפעה רבה על האופן שבו העולם מתנהל מבחינת פרסומות, ארגון קהילתי, פוליטיקה ודברים חשובים אחרים.

דבר שיוצר בעיות ביחד עם ההשפעה הזאת הוא קיומם של חשבונות מזויפים שעושים פוסטים באופן אוטומטי ומנסים להטות את דעת הקהל.  
מדברים קטנים כמו יצירת פרסום חינם למוצר עד לדברים מהותיים ביותר כמו השפעה על בחירות - הבוטים האלה גורמים לערעור היציבות של המרקם החברתי.

המטרה שלנו היא שבהינתן משתמש ב-Tweeter נרצה לסווג אם אותו המשתמש מופעל ע"י אדם אמיתי או ע"י בוט.

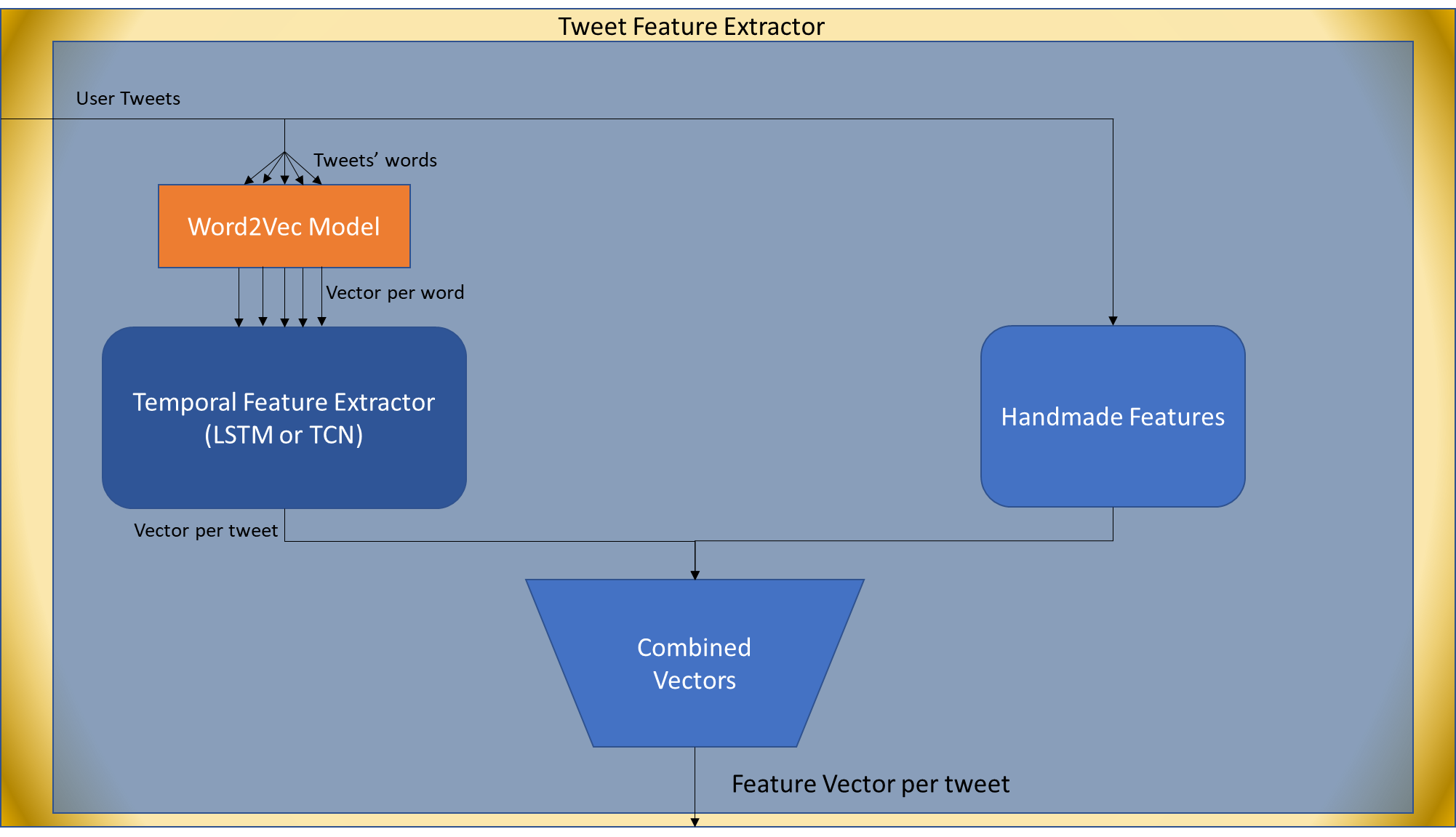
**מאגר המידע**  
<מאגר משתמשים + wikidata + gdelt>

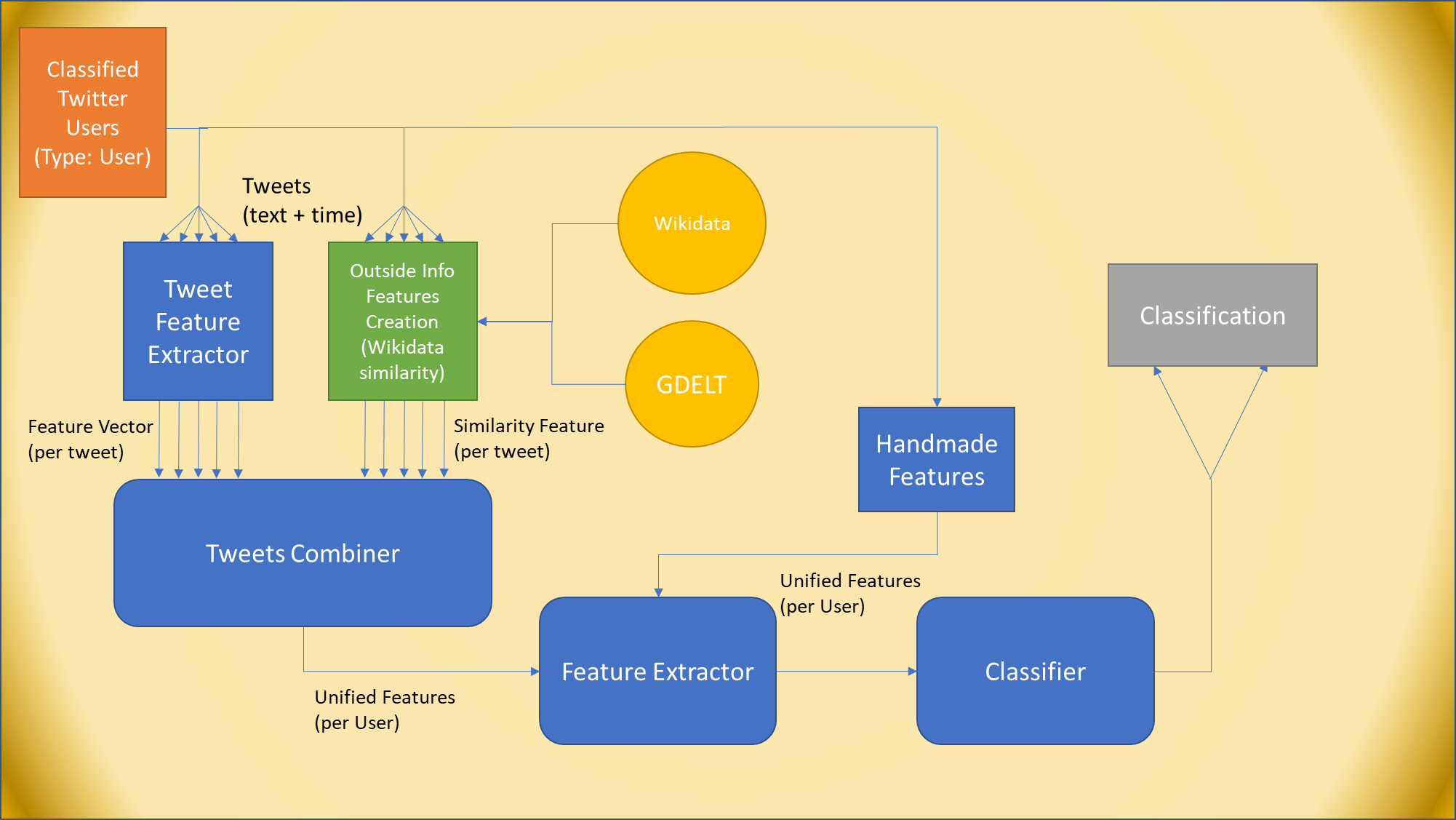
**עיבוד מקדים של המידע**  
<בחירת רק משתמשים באנגלית, שימוש ב-word2vec>

**פתרון הבעיה:**  
הדרך שחשבנו שאיתה אפשר לפתור את הבעיה היא הדרך הבאה:

1. שימוש במאגר נתונים של משתמשים מסווגים.
2. ניקח מכל משתמש את האינפורמציה הרלוונטית עליו ועל כל הפוסטים האחרונים שלו.
3. נמיר את אובייקט המשתמש לטנזור פיצ'רים באמצעות Feature Extractor שנגדיר וכנראה נלמד, כולל המרת מילים לוקטורים באמצעות ספריה מסוג Word2Vec והוספת הקשר חיצוני מ-wikidata ומ-GDELT.
4. לאחר עוד כמה שלבי עיבוד מקדים (כולל סטנדרטיזציה ו-feature selection) נעביר את הטנזור ואת הסיווגים לאלגוריתם למידה (כנראה רשת עמוקה).

**תלויות (ספריות) ואופן התקנה:**  
<להכניס מהקובץ>

**תיאור המערכת:**

****

למערכת יהיו את הרכיבים הבאים:

* מאגר נתונים מסווג\מאסף מידע (כולל מידע חשוב לפי GDELT וקטגוריות Wikidata).
* יוצר אובייקטים (מוודא קיום של פורמט אחיד לכל הדוגמאות).
* המרת אובייקט לטנזור כללי (ספציפית מילים ל-embedding באמצעות Word2Vec).
* Feature Extractor.
* Feature Selection.
* אלגוריתם למידה\מודל.

**מאגר נתונים + יוצר אובייקטים:**  
מאגרי המידע שלנו מורכבים ממשתמשים אמיתיים (כלומר שמאחוריהם עומדים אנשים אמיתיים) ובוטים ברשת החברתית טוויטר. הם מכילים את מזהה המשתמש בטוויטר ואת סיווגו כבוט או לא בוט.

**שליפת מידע ואחסונו**:  
כדי לגשת למידע המלא על כל משתמש כמו הציוצים שלו והפרופיל שלו היינו צריכים להשתמש בממשק של טוויטר tweepy . זה הוא api שבהינתן מזהה משתמש בטוויטר נותן גישה מלאה לפרופיל שלו ולציר הזמן שלו ולציוצים שלו. לכל משתמש במאגר ביצענו אחסון של כל המידע שיכולנו לקחת ממנו באובייקטים מסוג user שיצרנו שמאפשרים גישה נוחה ומהירה לכל המידע על המשתמש שנרצה לעשות בו שימוש. לאחר מכן לקחנו את האוביקטים והסיווג שנוצרו לכל המשתמשים ושמרנו בקובץ כדי שלאחר מכן נוכל לטעון את המידע במהירות ובפשטות בזמן ריצת המודל.

**סינון מידע:**  
בעת הגישה למשתמשים באמצעות tweepy קיימים משתמשים שהוסרו מטוויטר או שהפכו את הגישה לפרופילים שלהם לפרטיים ולכן לא ניתן לגשת אליהם ולקחת מהם מידע ולכן הסרנו אותם מהמאגר.

בנוסף הניתוח שהמודל שלנו יעשה בהמשך על המידע של המשתמשים דורש ששפת הציוצים שלו תהיה באנגלית ולכן משתמשים שציוציהם לא היו באנגלית הוסרו גם הם מהמאגר.

**עיבוד מקדים של הציוצים לפני השימוש ב-word2vec:**\*הסבר קצר על word2vec- מדובר ברשת נוירונים רדודה שמקבלת סט רחב של מילים ושולח כל אחת מהמילים לווקטור שונה משלה בעל מימדים גדולים כאשר מילים שחולקות הקשר נשלחות לווקטורים שהמרחק ביניהם יותר קטן.

הציוצים אותם אנחנו מקבלים מהמשתמשים חייבים לעבור "ניקוי" כדי ש-word2vec יוכל לעבד אותם בצורה טובה ומועילה.

ראשית ציוצים מכילים בדרך כלל תיוג של משתמשים אחרים ברשת על ידי התו '@' ולאחר מכן שם המשתמש , לכן כדי שword2vec יוכל לתפוס אזכור של משתמש בתור קונספט במשפט ולא בתור מילה ספציפית ומיוחדת עבור כל אזכור שונה בכל ציוץ החלפנו את אזכורי המשתמשים בהופעת התו '@' בלבד.

לאחר מכן שימוש בקישורים בציוצים צריך לעבור אותו תהליך כי אינינו רוצים ש-word2vec יסתכל על כל קישור כמושג נפרד ושונה בעולם אלא כקונספט הכללי של לשים קישור בציוץ ולכן החלפנו הופעה של כל קישור בתו '$'.

לבסוף הסרנו את כל שאר התווים המיוחדים מציוצים שכן הם לא תורמים ואף גורעים מההבנה של word2vec את ההקשר של המילים.

לאחר מכן המילים של כל ציוץ עוברות tokenization כלומר הציוץ מפורק לרשימה של המילים במרכיבות אותו כאשר כל מילה או פתגם או ביטוי הופך לאיבר ברשימה זו כלומר לtoken. את הtokens ניתן להכניס לword2vec בתור קלט שהוא יכול להתאמן עליו וגם לפלוט ווקטורים מתאימים בזמן ריצת המודל המלא.

**יצירת ה-features:**

**המרת מילים לוקטורים באמצעות word2vec:**  
נמיר כל מילה (לאחר העיבוד המקדים שתואר קודם) לוקטור באמצעות מודל word2vec של genism.  
<פירוט אקדמאי על מה זה וקישור למאמר המתאים>

**Tweet Feature Extractor:**מכיוון אנו יכולים להשתמש ב-word2vec רק בכדי לקודד מילים, עלינו להשתמש במודל שיכול לעבד וקטורים במימד הזמן (סדר המילים) בכדי להשיג וקטור מידע על *הציוץ*.  
ישנן 2 גישות מרכזיות לטפל בעיבוד מידע לאורך מימד הזמן:

1. הגישה הראשונה היא שימוש ברשתות נוירונים מסוג RNN (Recurrent Neural Network), שבהן מכניסים לשכבה כלשהי את הוקטורים אחד אחרי השני, עם שמירת מצב חבוי שחושב מהוקטור הקדם ומהמצב החבוי שלפניו (אפסים לראשון).  
   מתוך אלה בחרנו להשתמש ב-LSTM, מכיוון שהמודל הזה נוטה לתת את הפתרונות הכי איכותיים.
2. הגישה השנייה היא להשתמש בקונבולציות זמן (t-convolutions) על הוקטורים.  
   אלו רשתות קונבלוציה שעובדות על קלט חד מימדי, כאשר התכונות השונות מיוצגות כערוצים והערכים לאורך הזמן כמימד הוקטור (כלומר עבור מימד זמן T ו-F תכונות זה יהיה וקטורים T-מימדיים עם F ערוצים).  
   בפרט בחרנו להשתמש במודל ה-TCN (Temporal Convolutional Network) שהופיע במאמר שהציג לראשונה את הקונספט של קונבולוציות זמן <הכנס קישור למאמר>.

לכל אחד מהפתרונות יש יתרונות וחסרונות ביחס לאחרים:

<השוואת זמני ריצה>

<השוואת מספר פרמטרים>

<השוואת אורך שרשרת הגרדיאנטים>

הוצאת מידע מציוצים

**מקור הרעיון:**מקור המידע הראשי שממנו נרצה לשאוב מידע על המשתמש הוא ציוציו, לציוצים קיימים מספר תכונות שהשתמשנו בהם על מנת להוסיף עוד מידע למודל כדי שיוכל להשתמש בו לסיווג.

מאפיין מאוד בולט של בוטים ברשתות חברתיות הוא מספר גדול של ציוצים בהפרשי זמנים מאוד קצרים, ולכן משתמש שהפרשי הזמן בין ציוציו נורא קטנים יותר סביר להיות בוט, בנוסף הוספנו עוד תכונות של ציוצים כמו האם הציוץ הוא ציטוט או לא כערכים 0 או 1 , מספר הקישורים או אזכורים של משתמשים אחרים בציוץ, האם היה retweeted status לציוץ או לא כערכים 0 או 1, ומספר המשתמשים ששמו את הציוץ במועדפים שלהם.  
חישבנו את התוחלות ואת סטיות התקן של התכונות האלו בכדי לעשות סטנדטיזציה.  
הערכים האלה הם:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **תכונה:** | מס' אנשים שסימנו כמועדף | האם זה ציטוט | האם זה ציוץ מחדש | מס' קישורים ואזכורים |
| **תוחלת:** | 1.0901 | 0.0135 | 0.8929 | 0.8929 |
| **סטיית תקן:** | 17.6291 | 0.1156 | 0.3092 | 1.9421 |

**התהליך:**

1. שליפת התאריכים של הציוצים של משתמש מתוך מאגר המידע של המשתמשים שברשותנו, בנוסף נשלוף את היותו של ציוץ ציטוט או לא, מספר הקישורים והאזכורים של משתמשים אחרים בציוץ, והאם היה retweeted status, ומספר המשתמשים ששמו את הציוץ במועדפים שלהם.
2. חישוב של הפרשי הזמן בימים בין כל שני ציוצים עוקבים שלו.
3. שמירה של הפרשי הזמנים וכל אחד משאר הנתונים בווקטור מתאים משל עצמו לכל הציוצים.
4. הוספת הווקטורים בתור עוד features לפלט של הword2vec על מילות הציוצים, לאחר סטנדרטיזציה לפי תוכלת וסטיית תקן.

הוצאת מידע על user:

מהמשתמש עצמו ניתן להוסיף עוד תכונות כמו כמות העוקבים שלו וכמות החברים שלו, כאשר אינטואיציה אומרת שעבור בוט מספר העוקבים והחברים יהיה נמוך יותר למרות שקיימים בוטים שמתחזים טוב לאדם אמיתי והם בעלי מספר עוקבים גדול אך הנחנו כי הם חריגים, בנוסף מהציוצים ניתן להסיק מידע כללי על המשתמש ולכן בחרנו לקחת את ממוצע הפרשי הזמנים בין ציוצים עוקבים של המשתמש כאשר בוטים יהיו בעלי ממוצע יותר נמוך של הפרשי זמנים.  
גם לערכים האלה שמרנו תוחלות וסטיות תקן בכדי לעשות סטנדרטיזציה.  
התוחלות וסטיות התקן הן:

|  |  |  |  |
| --- | --- | --- | --- |
| **תכונה:** | מס' עוקבים | מס' חברים | הפרש זמן ממוצע בין ציוצים (בשניות) |
| **תוחלת:** | 3.2708e+03 | 6.6564e+02 | 5.6675e+04 |
| **סטיית תקן:** | 2.1951e+04 | 1.3241e+03 | 2.2061e+04 |

**התהליך:**

1. לאחר הוצאת התכונות של הציוצים מה-feature extractor נחשב את ממוצע הפרשי הזמנים בין ציוצים שונים של המשתמש ונשלוף לכל משתמש את מספר העוקבים והחברים שלו ונאחסן כל תכונה בווקטור מתאים משלו.
2. נבצע סטנדרטיזציה על וקטורי התכונה האלה.
3. נוסיף את הווקטורים כעוד features לפלט של ה-tweet combiner.

**Combiner + Extractor:**  
<פירוט על המבנה>

שימוש במקורות מידע חיצוניים

**מקור הרעיון:**בוטים ברשתות חברתיות כמו facebook ובמקרה שלנו twitter נוטים להיות בוטים שמצייצים בנושאים שהם כרגע "חמים" במדיה ובתקשורת במטרה לפגוע בנושאים שהם כרגע רגישים ולהטות דעת קהל של משתמשים אחרים ברשתות החברתיות. לכן משתמש שהציוצים שלו הם בעלי הקשר גבוהה לנושאים שמוזכרים הרבה בתקשורת באותה תקופה הוא משתמש שסבירותו להיות בוט פוליטי עולה, יש לציין שסדרת ציוצים כזו מאופיינת במרווח נורא קטן של זמן בין הציוצים השונים וגם בכך נעשה שימוש בנוסף לעוד מידע מהציוצים.  
על המדד הזה גם נעשה סטנדרטיזציה, יש לו תוחלת של 0.66 וסטיית תקן של 2.8933.

**כלים:**

1. Wikidata- מאגר מידע שמחזיק הגדרות, הסברים ומידע כללי עבור מונחים ואישים חשובים, באמצעותו אנחנו יכולים לחפש מילה במאגר המידע הזה ולשלוף כל מיני הגדרות והסברים עליה ובאמצעות כך להרחיב את ההקשר שלה במשפט.

(שליפת המידע ממאגר זה בוצעה באמצעות בוט שמתחבר לאתר ושולף ממנו את המידע המבוקש).

1. Gdelt- מאגר מידע שמחזיק מידע על אירועים כלל עולמיים שדובר עליהם בעיתונים ובתקשורת מסביב לעולם. מאגר המידע מסודר לפי תאריך ומכיל שמות של גורמים שמעורבים באירוע כמו מדינות, ממשלות, פוליטיקאים, ארגונים ומוסדות, ומכיל גם מדד לחשיבות הנושא בשם GoldensteinScale שמצביע על כמה שנושא קריטי למדינה או לארגון בו האירוע קרה, בשביל עוד מידע עליו ניתן לקרוא בראש העמוד בקישור הבא:

<http://web.pdx.edu/~kinsella/jgscale.html>

(שליפת המידע ממאגר מידע זה בוצעה על ידי הורדה של מידע מהמאגר באירועים שנמצאים רק בתאריכים שקרובים לציוצים של אחד המשתמשים שלנו, ולאחר מכן מיון ושליפת מידע מהקבצים שהורדו)

1. SequenceMatcher- כלי שמספק מדד להקשר בין מילה למשפט לפי חיפוש הופעות של המילה ומילים מאותו שורש במשפט, בסופו של דבר בהינתן מילה ומשפט מחזיר מספר שככל שגודלו גדול יותר הוא מצביע על קשר טוב יותר בין המילה למשפט.

**התהליך:**

עבור כל משתמש ביצענו את התהליך הבא:

1. לקחנו את תקופת הזמן שבו הוא צייץ בתדירות הכי גבוהה ומשם לקחנו 5 מהציוצים שלו (נלקחו רק 5 בגלל חסכון בזמן ריצת המודל)
2. באמצעות הכלי gdelt לקחנו את 10 הנושאים שהיו בעלי החשיבות הגבוהה ביותר באותה תקופה וזאת באמצעות מדד לחשיבות הנושאים שgdelt מספק בשם GoldsteinScale.
3. לאחר מכן לקחנו את הטקסט של כל אחד מ5 הטוויטים ולקחנו עבור כל מילה בציוצים את הגדרתה והרחבתה באמצעות הכליwikidata כאשר מילים שהם מילות הקשר או עצירה לא השתתפו בהרחבה הזאת מכיוון שהם לא יספקו מידע משמעותי שרלוונטי לזיהוי נושאים פוליטיים.
4. לאחר מכן השתמשנו בכלי בשם sequenceMatcher כדי לאמוד את טיב הקשר בין הנושאים שהיו חמים בתקופת הציוץ לבין הטקסט של ציוצים כולל ההרחבות שלהם שהתווספו לטקסט מwikidata, פשוט סכמנו את הפלט של הsequenceMatcher עבור כל אחד מהציוצים של משתמש עם הנושאים החמים שלו לכל אחד מהמשתמשים.

נשמור את מדדי ההקשר האלו של כל המשתמשים בווקטור אותו אנחנו מנרמלים ומוסיפים כעמודה נוספת לפלט של ה-feature extracture כמידע נוסף על כל user.

אימון:  
<אימון בלמידה עמוקה>

סיווג:  
<שימוש במסווג הלינארי המקורי>  
<סוגי מסווגים אחרים>

**ניסויים:**

* ~~נרצה לבדוק את ההשפעה של הקודקס שאיתו נלמד את מנגנון ה-Word2Vec.  
  לדוגמא האם כדאי ללמד רק על הציוצים במאגר, על ציוצים באופן כללי או שילוב של ציוצים וטקסטים אחרים.~~
* ~~נרצה לבדוק את ההשפעה של מספר הפיצ'רים שאנו רוצים להשתמש בו מתוך כולם.~~
* נרצה לבדוק השפעה של מסווגים שונים עם היפר פרמטרים שונים.
* נרצה לבדוק את ההשפעה שיש בכלל להקשר החיצוני שאנו מוסיפים לציוצים.
* השוואה בין LSTM ו-TCN כ-Temporal Extractor.
* בדיקת מסווגים אחרים על ה-features שניתן לקבל לאחר סוף האימון.
* בדיקת אלג' clustering על המידע בכדי שנוכל אולי להשיג עוד אינפורמציה על המשתמשים, בפרט על הפרדה בין סוגי בוטים.

מטריקות לתוצאות הסיווג:  
<פירוט החלוקה ל-train,valid,test (אולי stratified)>  
<המטריקות השונות (דיוק, recall, f1-score, מטריצת בלבול וכו'...)>

פרטי הניסויים:  
<פרטים על כל ניסוי>

תוצאות הניסויים:  
<תוצאות לכל ניסוי>

**פרטים טכניים:**

אנחנו נכתוב את הפרויקט הזה בשפה **python**, ונשתמש במנהל החבילות **anaconda**.  
בכדי לשלוף מידע מטוויטר נשתמש בחבילה **tweepy** שמאפשרת לנו לעבוד עם הממשק של טוויטר ולהוריד ממנו את המידע הרלוונטי.  
נרצה להשתמש במימוש **Word2Vec** של **gensim** בכדי ליצור קידוד של מילים לוקטורים.  
נשתמש במאגר הנתונים **GDELT** בכדי להשיג הקשר לאירועים וב-**wikidata** וקטגוריות ויקיפדיה בכדי להשיג הקשר כללי.  
נשתמש ב-**scikit-learn** בשביל דברים כלליים של למידה ומסווגים פשוטים וב-**pytorch** בכדי לבנות מסווג מורכב יותר.

**סיכום:**  
<סיכום>