**הוצאת מידע מטוויטים**

**מקור הרעיון**

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.

**תהליך**

1. שליפת התאריכים של הציוצים של משתמש מתוך מאגר המידע של המשתמשים שברשותנו.
2. חישוב של הפרשי הזמן בימים בין כל שני ציוצים עוקבים שלו
3. שמירה של הפרשי הזמנים בווקטור ונרמול שלהם
4. הוספת הווקטור בתור עוד feature לפלט של הword2vec על מילות הציוצים.

**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.

**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 (<קישור להגדרה של GoldensteinScale>).

(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 <איך מחשבים את מדד ההקשר?>.
5. נשמור את מדדי ההקשר האלו של כל המשתמשים בווקטור אותו אנחנו מנרמלים ומוסיפים כעמודה נוספת לפלט של הfeature extracture כמידע נוסף על כל user.