**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.

**תהליך**

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.
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.

**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.

**תהליך**

1. לאחר הוצאת התכונות של הציוצים מהfeature 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. נוסיף את הווקטורים כעוד features לפלט של הtweet combiner.

**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.

**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.
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.