האוניברסיטה העברית בירושלים

הפקולטה להנדסה ומדעי המחשב

זיהוי אלימות ברשת במדיה החברתית

פרויקט סיכום בקורס עיבוד שפה טבעית בעברית 67680

מגישות:

ירדן טל, 203730700

מאי טל, 2033334735

תאריך: אוגוסט 2020

1. תקציר

לצד ההיבטים החיוביים של ההתפתחות המהירה של הטכנולוגיה ישנם גם לא מעט סכנות שזו טומנת בחובה. אחת מהן היא **בריונות רשת** (**Cyber-bullying**), מדובר בתופעה של השנים האחרונות, בה אנשים מאמצים את הרשתות החברתיות כבמה להתאכזר לאחרים, להשמיצם ברבים, להפיץ שמועות, לתקוף ולפגוע. אנשים משתמשים בבמה הרחבה שמאפשרת הרשת, באמצעותה יכולים להגיע למספר בלתי נתפס של אנשים, על מנת להפיץ שמועות, להפיץ סרטונים ותמונות מביכות, להוציא מהארון, לפרסם פרטים פרטיים, לפגוע באנשים בצורה הרסנית ובלתי הפיכה. המונים נפגעים מתופעת הבריונות ברשת, אשר הפכה למסוכנת ביותר.

במהלך השנים, פותחו שיטות במטרה להכריע האם טקסט הוא 'אלים' או לא, אך משימה זו איננה משימה פשוטה עקב הייחודיות בטקסטים אלו. הודעות מקוונות במדיה החברתית בדרך כלל קצרות מאוד ושופעות בסלנגים, ניבים, חיבורי מילים, ריבוי אותיות, שגיאות כתיב, סמלים וראשי תיבות. על מנת להתמודד עם מאפיינים אלו, במאמר זה נציע שיטה המשלבת עיבוד מקדים של טקסט בעברית יחד עם סיווג וניתוח אשכולות (clustering) להודעות של בני נוער ברשתות החברתיות. בשלב הראשון ביצענו עיבוד טקסט מעמיק המבוסס על מילונים לקסיקוגרפיים וסמנטיים. לאחר עיבוד הטקסט, ביצענו שלב של feature engineering בו איחדנו פיצ'רים על ידי ביצוע למטיזציה בעזרת YAP, בחנו פי'צרים (features) מבוססי מילה ותו, איחדנו כמות גדולה של פיצ'רים לפיצ'ר אחד חזק על ידי semantic lexicon. לבסוף את שלב הקלסיפיקציה ביצענו באמצעות אלגוריתם Random Forest ואת שלב הקלאסטרינג באמצעות אלגוריתם k-means שהשיגו את התוצאות הטובות ביותר. תוצאות הניסוי מצביעות בבירור כי השימוש בשיטה שהצענו יכול לספק באופן אפקטיבי שיפורים בביצועי המסווגים וניתוח האשכולות.

2. מבוא

2.1 רקע

האינטרנט הינו מקום פורה ומעצים לילדים, אך בו בזמן הוא בעל סכנות רבות אשר יכולות להשפיע על התפתחות הילדים.

בריונות ברשת היא סוגיה חמורה ונפוצה המשפיעה יותר ויותר על משתמשי האינטרנט ובעיקר על בני הנוער. היא מוגדרת כפעולה אגרסיבית, בה נעשה שימוש באינטרנט על מנת להשמיץ, להטריד, להביך, להפחיד או לתקוף אדם יחיד או קבוצה במרחב הסייבר ובמקרים רבים עלולה להוות עבירה פלילית. בריונות רשת יכולה להתקיים בבלוגים, אתרי אינטרנט, קהילות מקוונות, SMS, רשתות חברתיות ובאמצעות סרטוני וידאו אשר צולמו בטלפון הסלולרי ופורסמו באתר או קהילה מקוונת. היא באה לידי ביטוי בפרסום שמועות, איומים, הערות מיניות, מידע אישי של הקורבן ועוד ולרוב מונעת מיצרי כעס, תסכול, קנאה או נקמה. בחלק מהפעמים הבריונים אינם מודעים לחומרת מעשיהם ועושים זאת מתוך שעמום או לחץ חברתי המופעל עליהם על ידי ילדים אחרים.

להטרדות באינטרנט יש השפעות פסיכולוגיות עמוקות על ילדים, כמו דיכאון נפשי ויצור טרמואות שילוו את הילדים לאורך חייהם. הטרדה באינטרנט לרוב תגרום לילדים להתרחקות חברתית והסתגרות בבית ובמקרים קיצוניים אף להתאבדות.

האתגרים במאבק בבריונות ברשת כוללים גילוי בריונות ברשת בזמן שהיא מתרחשת, דיווח על כך לסוכנויות האכיפה וזיהוי טורפים וקורבנותיהם. כיום, אין קהילה מקוונת או חברתית המשלבת

מערכת לזיהוי אוטומטי של אלימות ברשת. למרות חומרת הבעיה, ישנם מעט מאוד מאמצים מוצלחים לגילוי התנהגות פוגענית, בגלל מספר מכשולים כמו דקדוק, שגיאות כתיב והקשר מוגבל למדי.

2.2. שאלת המחקר

מטרת עבודת מחקר זו היא להילחם באלימות ברשת באמצעות זיהוי מקרים של בריונות והתנהגות פוגענית במדיה החברתית על ידי איסוף מערך הנתונים, עיבוד וניקוי הנתונים לשיפור הדיוק, סיווג הטקסט, חלוקה לאשכולות (clustering) והערכה וניתוח של המודל הטוב ביותר.

בתחילת התהליך הגדרנו שתי שאלות מחקר עליהן נרצה לענות-

א. האם ניתן לתייג הודעות ברשת להודעות אלימות ולכאלה שאינן אלימות

ב. מהם סוגי האלימות ברשת

שאלות המחקר אותן בחרנו מהוות בעיה בתחום הnlp בשפה העברית. הן מצריכות עיבוד מקדים על הטקסט, הבנת הסלנג, הציניות, צורת הכתיבה שהתפתחה בקרב האנשים, ובני נוער בפרט, ברחבי הרשת. לשם כך נדרשנו במהלך המחקר להשתמש בכלים ומילונים של סמנטיקה, פרגמטיקה ומורפולוגיה הייחודיים לשפה העברית.

3. מאגרי מידע

3.1. איתור ותיאור הנתונים

מערכי נתונים לבריונות ברשת בדרך כלל מורכבים מתגובות של משתמשים, פוסטים, תמונות וסרטונים במדיה החברתית.

מאגר המידע בו השתמשנו נבנה לגילוי בריונות ברשת. חברת keepers[1] יצרה את מערך נתונים זה לשימוש באפליקציה מתפתחת המשמשת לבקרת הורים, ומציגה הודעות ספציפיות שזוהו כמסוכנות תוך שמירה על פרטיות הילדים. הטכנולוגיה אשר פיתחה החברה מספקת פתרון להורים המעוניינים לשמור על ילדיהם מפני הסכנות הטמונות בניידים שלהם. הטכנולוגיה מאתרת ומדווחת בזמן אמת על כל תוכן שעלול להיות פוגעני ולהעיד כי הילד חשוף לסכנה. המאגר מורכב מהודעות של בני נוער מהרשתות החברתיות, כאשר ההודעות מתויגות בסיווג בינארי להודעות אלימות או לא אלימות.

<u>מספר פרטים אודות המאגר:</u>

המאגר מכיל 9,287 ודעות, מתוכן 40.8% (3,786 הודעות) אלימות.

מאגר הנתונים נתון בשני קבצי טקסט (קובץ עבור הודעות המתויגות כאלימות וכקובץ עבור הודעות שאינן אלימות), כאשר כל שורה מציינת הודעה.

כאמור, מאגר הנתונים מכיל הודעות של בני נוער ומאופניות בהמון סלנג, שגיאות כתיב, ריבוי אותיות, מילים מחוברות. בנוסף, ההודעות מאוד קצרות בסגנון של הודעות WhatsApp.

[1] https://www.keeperschildsafety.net/careers.html

הנתונים האידאליים לשאלת המחקר הם נתונים ממגוון רחב יותר של אוכלוסיות ורשתות חברתיות שיכולים לייצג סוגי אלימות שונים ורבים יותר, מאגר דאטה פחות מאוזן מבחינת כמות ההודעה האלימות מול אלו שאינן אלימות ועל כן גם יותר מייצג את המציאות (במציאות קיימות בעולם הרבה יותר הודעות שאינן אלימות מאשר הודעות אלימות), נתונים המכילים גם מאטה-דאטה של משתני רשת, כמו פרטי השולח, זמן, מיקום. הודעות המכילות אימוג'ים.

3.2. ניקוי מאגר המידע

שלב זה מהווה שלב מקדים לשלבי האימון והסיווג שמטרתו להתאים את מבנה ותוכן ההודעות לשיטה בה המסווגים משתמשים. בשלב זה ניסנו לייצר טקסטים מתאימים יותר מטקסטים קצרים ורועשים על מנת לשפר את ביצועי המסווגים. על כן, שתי פעולות הניקיון אותן בעצנו היו-

א. הורדת סימני פיסוקי- הסרת כלל סימני הפיסוק מההודעות במאגר המידע

ב. הורדת stop words

נציין כי מאגר המידע שקיבלנו היה נקי למדי ואינו הכיל הרבה מידע מיותר. עם זאת, נדרשו הרבה פעולות לנורמליזציה של הדאטה, עליהן נפרט בסעיף השיטה (סעיף מספר 5).

4. סקירת ספרות

שגשוג הרשת החברתית הובילה להתפשטות נרחבת של בריונות ברשת, המהווה בעיה קשה למדי לילדים ולבני נוער. במהלך השנים האחרונות הוצעו מספר טכניקות למדידה ואיתור של תוכן פוגעני בפלטפורמות כמו אינסטגרם, יוטיוב, Yahoo ועוד.

על מנת להבין מה נעשה בתחום בשנים האחרונות, סקרנו מספר מאמרים שנראו לנו מייצגים. במאמרים השונים העוסקים בנושא השתמשו בתכונות טקסטואליות ומבניות כדי לחזות את יכולת המשתמש בהפקת תוכן פוגעני [1], הסתמכו על embedding כדי להבחין בין הערות פוגעניות לכאלה שלא [2], זיהו דברי שנאה על ידי שימוש ב supervised learning classification [3, 14, 13, 12].

באחד המאמרים גילו בריונות ברשת על ידי פירוק (decomposing) לנושאים רגישים ואספו הערות מסרטונים מעוררי מחלוקות וכך סיווגו את הטקסט [3]. בנוסף, ישנם כאלו שחקרו מאפיינים לשוניים בתחום הבריונות ברשת במטרה לאתר סוגים שונים של בריונות [4] ואף חקרו תמונות שפורסמו באינסטגרם ותגובות שקשורות אליהן כדי לאתר בריונות ברשת [5].

כמו כן, בהרבה מהעבודות הקודמות נעשה שימוש בתכונות (features) כמו פיסוק, URLs, part of speech, n-grams, bag of words כמו גם בתכונות לקסיקליות הנשענות על מילונים של מילים פוגעניות [14, 13, 12].

הרבה מהמחקרים בנושא השתמשו בגישות מפוקחות (supervised approaches) שונות המשמשות לזיהוי : מודל רגרסיה [5], Naive Bayes, SVM ועצי החלטה [6,1]. עם זאת, יש גם כאלו שעשו שימוש בגישה מבוססת גרפים על מנת לזהות התנהגות שלילית [5] ושימוש ב sentiment analysis של טקסט שעוזר בזיהוי האם התוכן פוגעני או לא. למשל, במאמר אחד הסתמכו על sentiment כדי לזהות קורבנות בטוויטר המהווים סיכון גבוה לעצמם או לאחרים תוך כדי שהתחשבו ברגשות ספציפיים כמו כעס, מבוכה ועצב [7].

5. השיטה

השיטה שלנו התבססה על עיבוד מקדים (preprocessing) על הדאטה שהיה ברשותנו שכלל ניקוי ונרמול של הדאטה, לאחריו ביצענו שלב של feature engineering ואז feature selection ולבסוף על מנת לענות את שאלות המחקר שלנו ביצענו classification בעזרת מסוגים שונים ו- clustering בעזרת אלגוריתם k-means.

נתאר את שלבי השיטה -

1. **עיבוד מקדים Preprocessing**- כחלק מהעיבוד המקדים של הדאטה בצענו מספר ניקויים בסיסיים כפי שתואר בסעיף ניקוי מאגר המידע (סעיף 3.2).

בנוסף, בשל המאפיינים הייחודים של הודעות רשת, שבעיקרן הודעות WhatsApp, בצענו נרמול לדאטה (text normalization) אשר כלל –

- **למטיזציה-** בכדי להתמודד עם בעיית הדלילות אותה הצגנו עבור הודעות רשת, ניסינו ליצור הומוגניות של המילים בטקסט ולהפחית את מספר הפיצ'רים שלנו (features). מטרת תהליך זה היא סיווג פשוט ונח, ולכן טוב יותר. לצורך הביצוע השתמשנו בכלי ייחודי לעברית, YAP[2], מודל חישובי לניתוח אוטומטי של טקסטים בעברית שנבנה במעבדה של פרופסור רעות צרפתי באוניברסיטת בר-אילן. כך, כל מילה בקורפוס שלנו הפכנו ללמה שלה. בכדי שתהליך יתבצע בצורה מוצלחת כלל האפשר בצענו נרמולים נוספים לדאטה שכללו תרגום מילות סלנג, תיקון שגיאות כתיב, הורדת ריבוי אותיות, זאת בעזרת מילונים לקסיקוגרפים וסמנטיים אותם בנינו בעצמנו.
- **תיקון ריבוי אותיות או תווים-** למשל כמו "שלייייי"
- **תרגום מילות סלנג-** בנינו מילון המכיל מילות סלנג רבות בעברית הנפוצות בעיקר בקרב בני הנוער ותרגומן
- **הפרדת מילים מחוברות-** למשל פירוק של מילים כמו "חיימשלי" ל"חיים שלי", "אמשך" ל"אמא שלך"

2. **Feature engineering**- את הנתונים הגולמיים שלנו הפכנו לתכונות המייצגות טוב יותר את הבעיה העומדת בבסיס, וכתוצאה שפרנו את דיוק המודלים איתם השתמשנו לצורך קלסיפיקציה וקלסטור.

לשם כך, ניסינו להשתמש בכמה סוגי ייצוגים (word embedding):

- **Bag of words + tf-idf-** מודל המשמש בדרך כלל לסיווג מסמכים. כאשר המודל מייצג את כמות המופעים של כל מילה בטקסט. באופן זה, כל מילה מהווה תכונה שעוזרת לסיווג המסמך. מודל זה הוא לרב מודל דליל, בהודעות רשת WhatsApp בפרט (בשל אורכן הקצר), אך לאחר עיבוד מקדים מוצלח, הערכתינו היא שהמודל יעבוד בצורה טובה ויתרום לסיווג ההודעות. במודל שלנו בחרנו להשתמש גם בunigram וגם ב-bigram תוך התעלמות מביטויים המופיעים פחות מ-3 פעמים בכל הקורפוס שכן אלו כנראה מהווים רעש ולא תורמים לסיווג, וכן מאלו שמופיעים ביותר מ-95% מהטקסטים בקורפוס. לאחר יצירת bag of words בצענו משקול בעזרת TF-IDF[3] שהוא מדד

[2] https://nlp.biu.ac.il/~rtsarfaty/onlp/hebrew/about
[3] https://en.wikipedia.org/wiki/Tf%E2%80%93idf

לחשיבותה של מילה במסמך בתוך אוסף מסמכים, ובכך לוקח בחשבון את תדירות התרחשות המילה בתוך הקורפוס כולו וכן בתוך כל מסמך בנפרד. ערך tf-idf עולה באופן יחסי למספר הפעמים שמילה מופיעה במסמך ומקוזז על ידי מספר המסמכים בקורפוס המכילים את המילה, מה שעוזר להתאים את העובדה שמילים מסוימות מופיעות בתדירות גבוהה יותר באופן כללי.

את הייצוג הנ"ל בצענו בשני אופנים -

א. **Word level**- כל מילה (או כמה מילים) הם יחידת הבסיס ומהווים פיצ'ר.

ב. **Character level**- שימוש בתווים או אותיות כיחידת הבסיס במקום מילים. Character level n-grams (shingles) יכול להיות מאוד שימושי בתחומים (domains) מסוימים, אלימות ברשת היא דוגמא טובה לכך. Character n-gram מאפשר הפשטה מסוימת מרמת המילים ומספק איתנות לשונות הכתיבה שמאפיינת את נתוני המדיה החברתית. כך למשל, ייצוג זה מאפשר להתמודד עם שגיאות כתיב, ריבוי אותיות וסלנג בצורה טובה. לדוגמא: n-gram של 3 על המילה "בחילה" ייתן "בחי", "חיל", "ילה" ואז גם אם כתובה המילה באחת מהצורות הבאות: "בחילהההההה", "בחילדה", "בחילע" אז עדיין מופיע לנו "בחי" למשל בכולם, ואז המודל עדיין יכול לקשר ביניהם ולהבין שהן על אותו משקל גם אם הן בווריאציות שונות.

- **Sentiment lexicon for negative/positive affect**- לקסיקון בעל רשימת מילים שליליות ורשימת מילים חיוביות. מכיוון שהודעות של בריונות ברשת כוללות בדרך כלל קללות או מילים מעליבות, מילים אלו מהוות אינדיקציות טובות לקיום בריונות. לכן, בחרנו רשימה של מילים מעליבות על סמך הידע הקודם שלנו וחלקן ממקורות חיצוניים[4]. רשימה המילים השליליות מכילה מילים המעידות על קללה או רגשות שליליים, כמו עצב, התאבדות, חרם, זיון, שרמוטה וכו' ואילו רשימת המילים החיוביות מעילה מילים המעידות על רגשות חיוביים, שמחה, אהבה, קבלה ועוד. כך, אני מייצרים שני פיצ'רים חזקים שמייצגים כמות גדולה של מילים המהוות אינדיקציה לאלימות או אי אלימות ברשת. בניית הפיצ'ר נעשה על ידי חישוב כמות הפעמים שמופיעה אחת מהמילים ברשימה בטקסט מסוים בקורפוס ביחס לכמות המילים בטקסט.
- **Word2Vec**- טכניקה לעיבוד שפה טבעית. האלגוריתם word2vec משתמש במודל רשת עצבי כדי ללמוד שיוך מילים ממספר גדול של טקסטים. לאחר הכשרה, מודל כזה יכול לאתר מילים נרדפות או להציע מילים נוספות למשפט חלקי. כפי שמשתמע מהשם, word2vec מייצג כל מילה כווקטור. הווקטורים נבחרים בקפידה כך שפונקציית דמיון מתמטית פשוטה (cosine similarity[5] בין הווקטורים) מצביעה על רמת הדמיון הסמנטי בין המילים המיוצגות על ידי אותם וקטורים. על מנת לייצג את המילים בקורפוס שלנו בעזרת מודל זה, השתמשנו במודל שאומן במעבדה של פרופסור יואב גולדברג באוניברסיטת בר אילן על דאטה בעברית מטוויטר[6]. השימוש במודל נעשה מתוך רצון

---

[4] https://www.kaggle.com/rtatman/sentiment-lexicons-for-81-languages

[5] https://en.wikipedia.org/wiki/Cosine_similarity#:~:text=Cosine%20similarity%20is%20a%20measure,to%20both%20have%20length%201.

[6] https://drive.google.com/drive/folders/1qBgdcXtGjse9Kq7k1wwMzD84HH_Z8aJt

לנסות לשפר את ייצוג הפיצ׳רים שלנו ובכך לשפר את משימת הסיווג ולהקלסטור. אך במקרה שלנו, ייצוג הפיצ׳רים בעזרת word2vec לא הביא לביצועים טובים יותר, ייתכן שהסיבות לכך נובעות מהנתונים עליהם המודל אומן או לחילופיו מכך שהנתונים שלנו אינם בעלי שפה עשירה ומגוונת, מועטים וקצרים.

3. **Feature selection-** תהליך בחירת תת-קבוצה של תכונות מתוך כל מכלול התכונות. טכניקה זו שימושית מכמה סיבות: פישוט המודלים, זמני אימון קצרים יותר, הורדת מימד, הכללה משופרת על ידי צמצום התאמת יתר. הנחת היסוד המרכזית בעת שימוש בטכניקת בחירת תכונות היא שהנתונים מכילים כמה תכונות שאינן מיותרות או לא רלוונטיות, ובכך ניתן להסיר אותן מבלי להיגרם לאובדן מידע רב.
בחרנו לבחון שלוש שיטות של בחירת פיצ׳רים –

- הורדת תכונות עם שונות נמוכה, שיטה זו מסירה את כל התכונות שהשונות שלהן אינה עומדת בסף כלשהו. כברירת מחדל, היא מסירה את כל התכונות עם שונות אפס, כלומר תכונות שיש להן ערך זהה בכל הדגימות.
- SelectKbest - בחירת התכונות הטובות ביותר על סמך מבחנים סטטיסטיים חד-משתנים.
- SelectFromModel – בחירת התכונות לפי החשיבות וההשפעה שלהם על מודל קאלסיפיקציה מסוים.

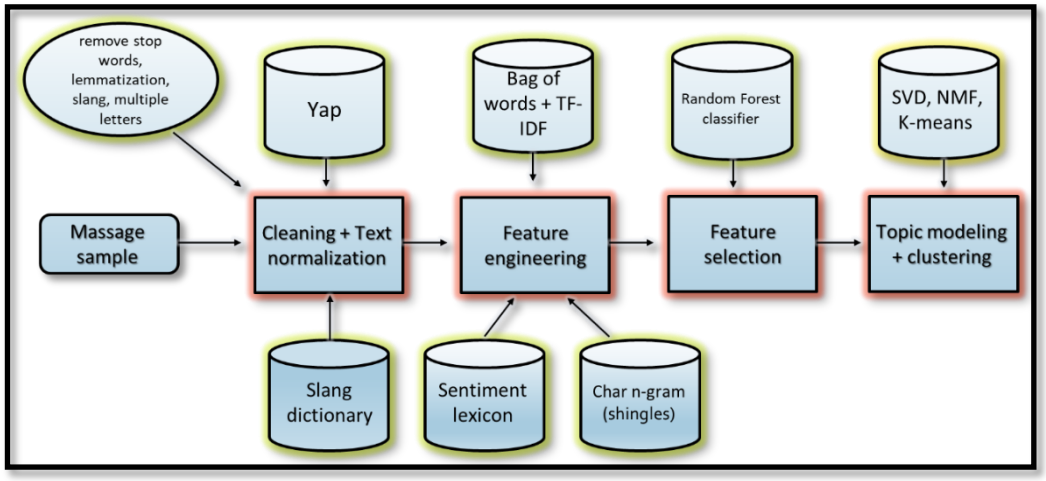
לבסוף, בחרנו להשתמש בשיטה השלישית של select from model ששפרה את הסיווג במקצת.

4. **סיווג המידע (classification)-** את סיווג ההודעות להודעות רשת אלימות או לא בצענו בעזרת מספר אלגוריתמי סיווג. ה- baseline שלנו היה אלגוריתם המסווג לפי רב (majority classifier). בתחילה בחרנו בשלושה אלגוריתמי סיווג בסיסיים מסוגים שונים על מנת להבין איזה סוג מודל מתאים ביותר למשימה שלנו: Naive Bayes, Logistic Regression, Random Forest. המודל שהשיג את התוצאות הטובות ביותר היה Random Forest, על כן החלטנו לבצע את הסיווג בעזרת שני אלגוריתמי סיווג מסוג עצים (xgboost Tree, GBM-Gradient Boosting Machine) ולבחון האם ניתן להשיג תוצאות טובות יותר, מה שהתגלה כלא נכון בעזרת שני אלו.

5. **Clustering-** על מנת לענות על שאלת המחקר השנייה שלנו - מהם סוגי האלימות הקיימות ברשת, ביצענו clustering לדאטה הנקי והמנורמל שלנו בעזרת אלגוריתם k-means לשלושה קלאסטרים (k=3). בכל אחד משלושת המחלקות ניתן לראות מאפיין של סוג אלימות מסוים, כאשר אנחנו סיווגנו אותם כ- כינויי גנאי, הערות מיניות ואיומים. ניסינו לבצע גם clustering על ידי אלגוריתמים שונים של Topic modeling כמו SVD ו-NFS, אך אלו השיגו תוצאות פחות טובות.

ניתן לראות כמה יתרונות בולטים במודל אותו בחרנו. ראשית, התהליך מסודר, פשוט וקצר, ומורכב ממספר מודלים פשוטים של למידה. לפיכך, גם אינו דורש כח חישובי גדול. בנוסף, בחירת הפיצ׳רים

למודל בצירוף נרמול וניקוי הדאטה הובילו להצלחת סיווג המידע על אף שמדובר בדאטה רווי בשגיאות כתיב, מילות סלנג, ריבוי אותיות.



תרשים 1- Pipeline - תהליך השיטה

6. הערכת התוצאות

את התוצאות בחנו בעזרת כמה מדדים מרכזיים. על מנת להבינם, נזכיר תחילה את הגדרות המושגים הבאים:

- True positive- tp: אחוז ההודעות עליהן דיווחנו כאלימות, והן אכן אלימות
- False positive- fp: אחוז ההודעות עליהן דיווחנו כאלימות, והן אינן אלימות
- True negative- tn: אחוז ההודעות עליהן דיווחנו כלא אלימות, והן אכן לא אלימות
- False negative- fn: אחוז ההודעות עליהן דיווחנו כלא אלימות, והן אלימות
-

כעת, את תוצאות המחקר בחנו בעזרת המדדים הבאים:

Accuracy: מודד את רמת הדיוק. כלומר, אחוז ההודעות שסווגו בצורה נכונה.

$$Accuracy = \frac{tp + tn}{tp + tn + fp + fn}$$

Recall: מודד את רמת הדיוק של סיווג הודעות אלימות. מה אחוז ההודעות האלימות שתויגו כאלימות.

$$Recall = \frac{tp}{tp + fn}$$

Precision: מודד את רמת הדיוק של הודעות שדיווחנו עליהן כאלימות. מה אחוז ההודעות שהן באמת אלימות, מתוך אלה שתייגנו כאלימות.

$$Precision = \frac{tp}{tp + fp}$$

Fb-measure: משלב את לprecision וה-recall. מדד זה הוא בערך הממוצע של השניים כאשר הם קרובים ובאופן כללי יותר הוא הממוצע ההרמוני.

$$F_\beta = (1 + \beta^2)\frac{Precision \cdot Recall}{(\beta^2 \cdot Precision) + Recall}$$

עבור הבעיה אותה בחרנו לחקור מדד הrecall יותר משמעותי, כיוון שטעות חמורה יותר במקרה זה הינה לתייג הודעה אלימה כלא אלימה, ולכן נרצה שאחוז ההודעות האלימות שתויגו כאלימות יהיה כמה שיותר גבוה, אך תוך התחשבות ב Precision. לכן, בחרנו למדוד את התוצאות בעיקר על פי F2 שנותן משקל רב יותר לrecall מאשר לprecision אך משלב את שניהם.

7. תוצאות ומסקנות

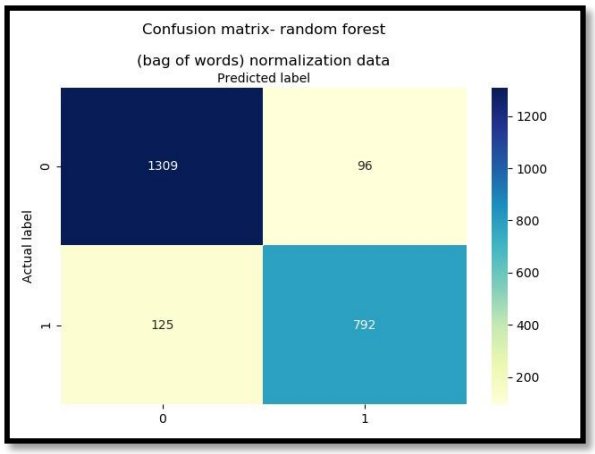
7.1 תוצאות

7.1.1 תוצאות הסיווג

התוצאות הראו כי המסווג שסיווג את הדאטה בצורה הטובה ביותר הוא ה-Random Forest. כאשר **שתי קומבינציות שונות של פיצ'רים הן אלו שהביאו לתוצאות הטובות ביותר**:

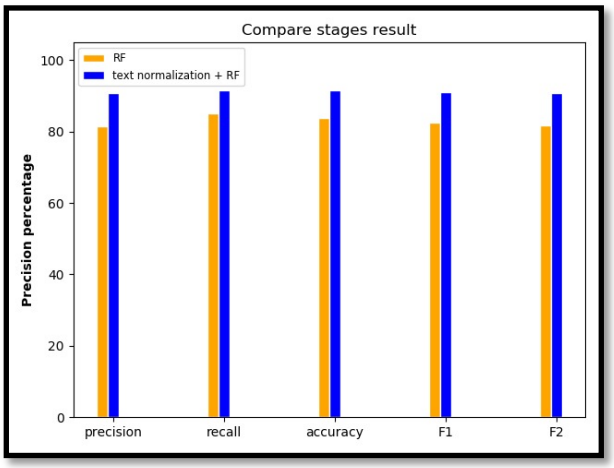
1. נורמליזציה (כפי שתואר בשיטה, סעיף 5.1) + למטיזציה בעזרת YAP+ sentiment lexicon features + פיצ'רים מבוססים מילים (bag of words – unigram or bigram)
2. ללא עיבוד מקדים + פיצ'רים מבוססי תווים\אותיות (Character level N-grams)

עבור שיטה מספר 1 המבוססת על פיצ'רים ברמת המילה – ניתן לראות בגרף המוצג כי בעזרת השיטה אותה הצענו 1309 מתוך 1405 הודעות לא אלימות סווגו נכון, כלומר הייתה שגיאה של כ- 7% . כמו כן, 792 מתוך 917 הודעות אלימות סווגו נכון, כלומר הייתה שגיאה של כ- 14% .



תרשים 2- Confusion matrix - כמות ההצלחות והטעויות של סיווג ההודעות לאלימות או לא עבור מסווג Random Forest עבור שיטה מספר 1.

בנוסף, בגרף הבא ניתן לראות כי כלל המדדים אותם בחנו עלו כאשר הוספנו לתהליך הסיווג את העיבוד המקדים שכלל ניקוי ונורמליזציה של הטקסט כפי שתיארנו בשיטה, כפי שציפינו.

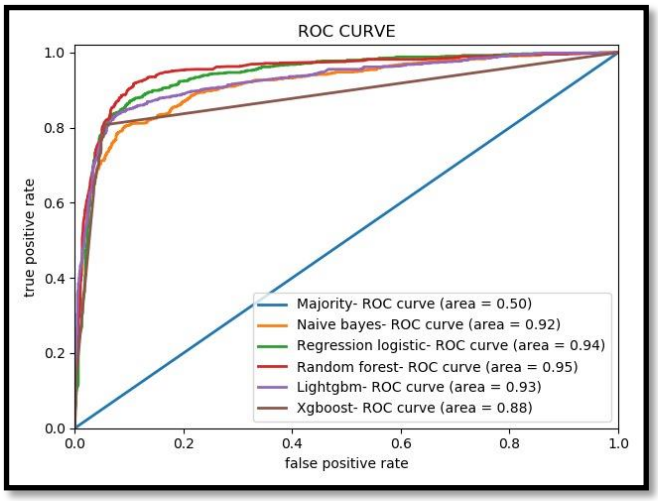


תרשים 3- השוואת תוצאות הסיווג של הRandom forest בכל המדדים אותם בחנו על דאטה נקי מול דאטה שעבר נורמליזציה

	Recall score	Precision score	Accuracy score	F1 score	F2 score
majority	0.5	0.3	0.5999	0.375	0.4412
naive bayes	0.8455	0.8749	0.8656	0.8551	0.8482
regression logistic	0.8635	0.8957	0.8833	0.874	0.8664
random forest	0.8965	0.9081	0.9065	0.9013	0.8982
lightgbm	0.8762	0.8923	0.8893	0.8825	0.8783
xgboost	0.8523	0.8859	0.8734	0.863	0.8553

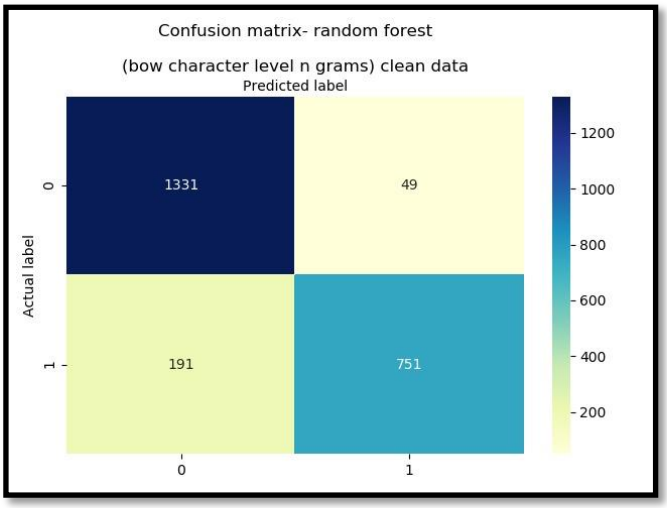
תרשים 4- תוצאות המדדים עבור המסווגים השונים. Feature engineering המתאים לשיטה מספר 1.

עקומת ROC מתארת את היחס שניתן לקבל בין TPR לFPR עבור כל אחד מהמסווגים שבחנו, במקרה שלנו נשאף לטעות כמה שפחות על הודעות אלימות לכן נרצה לבחור נקודה שנמצאת במקום גבוה בעקומה. על פי מדד ה-AUC ב-ROC Curve ניתן לראות כי מסווג הRandom Forest משיג את התוצאה הטובה ביותר (0.95).



תרשים 5- ROC CURVE

עבור שיטה מספר 2 המבוססת על פיצ׳רים ברמת התו\אות - ניתן לראות בגרף המוצג כי בעזרת השיטה אותה הצענו 1331 מתוך 1380 הודעות לא אלימות סווגו נכון, כלומר הייתה שגיאה של כ- 4% . כמו כן, 751 מתוך 942 הודעות אלימות סווגו נכון, כלומר הייתה שגיאה של כ- 20% .

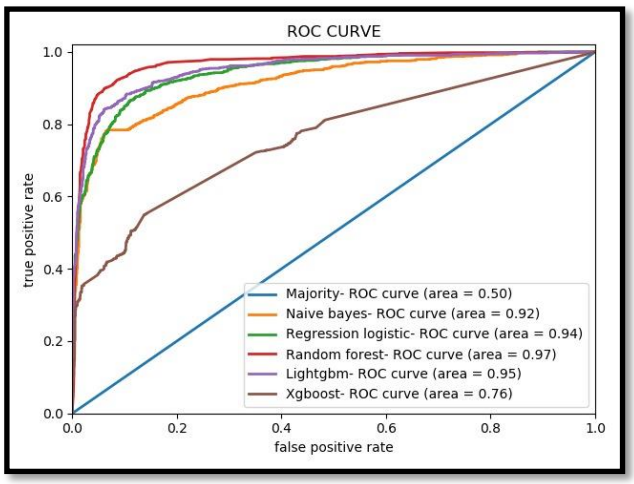


תרשים 6- Confusion matrix - כמות ההצלחות והטעויות של סיווג ההודעות לאלימות או לא עבור מסווג Random Forest עבור שיטה מספר 2.

	Recall score	Precision score	Accuracy score	F1 score	F2 score
majority	0.5	0.2883	0.5767	0.3657	0.436
naive bayes	0.8407	0.8778	0.8592	0.8498	0.8423
regression logistic	0.8525	0.8747	0.8661	0.8593	0.8543
random forest	0.9026	0.9202	0.9126	0.9088	0.9045
lightgbm	0.8869	0.8986	0.8953	0.8914	0.8884
xgboost	0.6857	0.8107	0.7308	0.6798	0.673

תרשים 7- תוצאות המדדים עבור המסווגים השונים. Feature engineering המתאים לשיטה מספר 2.

גם עבור שיטה מספר 2, על פי מדד ה-AUC ב-ROC Curve ניתן לראות כי מסווג ה Random Forest משיג את התוצאה הטובה ביותר (0.97).



תרשים 8- ROC CURVE

הטבלה הבאה (תרשימים 9 ו-10) מתארת את כלל הקומבינציות של הפיצ'רים בתהליך ה-feature engineering ולאחריו שלב ה- feature selection שניסינו עם ה- test set שהיה ברשותנו לאחר שבעצנו עליו את שלב הניקיון. התוצאות מוצגות עם מסווג הRandom Forest, כאשר התוצאות המודגשות הן התוצאות הטובות ביותר כפי שתיארנו מעלה (שיטות 1 ו-2).

Feature combination	Text normalization	Feature selection	Classification model	Scores				
				R	P	A	F1	F2
A + C	F, G		RF	0.895	0.909	0.904	0.899	0.897
A + C	F, G	V	RF	**0.903**	**0.911**	**0.909**	**0.902**	**0.899**
A	F, G		RF	0.871	0.882	0.880	0.874	0.871
A	F, G	V	RF	0.874	0.898	0.884	0.872	0.871
B	F, G		RF	0.872	0.891	0.880	0.876	0.873
B	F, G	V	RF	0.88	0.91	0.89	0.886	0.882
D	F, G		RF	0.820	0.849	0.838	0.828	0.827
D	F, G	V	RF	0.818	0.834	0.816	0.823	0.821
B			RF	0.86	0.89	0.88	0.87	0.820
B		V	RF	**0.892**	**0.919**	**0.909**	**0.898**	**0.894**
A + C			RF	0.881	0.893	0.891	0.875	0.874
A + C		V	RF	0.884	0.896	0.895	0.891	0.889
A			RF	0.839	0.802	0.85	0.847	0.841
A		V	RF	0.842	0.819	0.86	0.841	0.838
D			RF	0.825	0.858	0.849	0.835	0.828
D		V	RF	0.825	0.845	0.844	0.832	0.827

תרשים 9- ציון מסווג Random Forest על ה test set עם קומבינציות שונות של תכונות, לפי שיטות מדידה שונות (recall, precision, accuracy, F1 and F2)

A	word n-gram
B	char n-gram
C	sentiment lexicon
D	word2vec
F	normalization
G	YAP

תרשים 10- מיפוי קבוצות התכונות (תרשים 9)

7.1.2. תוצאות הקלאסטרינג (clustering) –

כאמור, הקליסטור נעשה על ידי שימוש באלגוריתם K-means עם k=3, וזו לאחר שבעצנו ניתוח לחלוקה הטובה ביותר בעזרת שיטה הנקראת "Elbow"[7], לקביעת מספר האשכולות למערך נתונים. את האלגוריתם הפעלנו על הדאטה לאחר שבצענו עלייה ניקויים ונרמול כפי שתיארנו בסעיף השיטה (סעיף מספר 5). ניתן לראות כי שלושת האשכולות התאימו במידה מסוימת לשלושה סוגי אלימות ברשת –

א. הערות מיניות (מילים כמו – זונה, שרמוטה, הזדיין, כוס)

ב. איומים (מילים כמו - התאבד, מת, חרם, בכה, שנא).

ג. כינויי גנאי (מילים כמו - סתום, דבע, מטומטם, טיפש, הומו)

[7] https://en.wikipedia.org/wiki/Elbow_method_(clustering)

תרשים 11 – הצגה בעזרת WorldCloud של המילים המשמעותיות ביותר בכל אחד משלושת הקלאסטרים אותם סיווג אלגוריתם k-means כאשר k=3.

7.2. איטרציה

במהלך הדרך היו מספר נקודות בהן שינינו גישה או הוספנו שלבים לשיטה -

1. לאחר ההרצות הראשוניות ושלב הEDA בצענו מספר צעדים -
   א. הוספנו מילים לרשימת הstop word הרלוונטיות לדאטה שלנו והורדנו כאלו לא רלוונטיות
   ב. יצרנו מילון של סלנג וחיבור מילים
   ג. בצענו תיקונים למילים שYAP לא תיקן כפי שצריך (לדוגמא - "מכות" הפך ל"מך" מלשון עוני ולא ל"מכה")
2. בהתחלה ניסינו לבצע סיווג של ההודעות עם ניקוי ונרמול של הטקסט אך ללא למטיזציה על ידי YAP. בהמשך הגענו להבנה שההודעות קצרות מאוד ולכן ישנו מספר נמוך של תכונות שניתן לייצא מהודעה בודדת. לכן, החלטנו להשתמש בYAP על מנת לאחד תכונות דומות בהטיות שונות לתכונה אחת, ובכך להבטיח שבמקרים רבים יותר – תימצא התכונה בהודעה.
3. כחלק מהפייפלין (pipeline) הראשוני תכננו לבצע שלב של feature selection לאחר השלב של feature engineering על מנת להקטין את מימד הפיצ'רים ולהתעלם ממילים שאינן חשובות לסיווג ולהבליט כאלו שיותר. כאשר בצענו מספר שיטות של feature selection הופתענו לגלות שהביצועים ירדו והגענו למסקנה שיתכן שייצוג אחר של הפיצ'רים יתאים לשיטה זו. להערכתנו כאשר ההודעות הן כל כך קצרות שלב בחירת התכונות גרם לכך שבהרבה הודעות אף מילה לא זוהתה שפיצ'ר כלל. בשל כך ניסינו להשתמש בייצוג של מודל word2vec שאומן על דאטה של טוויטר בעברית וכן ייצוג שכלל bigram וייצוג ברמת התוו\אות, כאשר השניים האחרונים באמת עזרו ושלב הfeature selection שיפר את הביצועים במקצת.
4. כדי לענות של שאלת המחקר השנייה אודות סוגי האלימות ברשת ניסינו תחילה להשתמש באלגוריתמים של Topic modeling לטקסט. אלו, לא עבדו בצורה מוצלחת על הדאטה שלנו והיה קשה מאוד לדלות מהם את סוגי האלימות הקיימים בדאטה שלנו. מתוך הבנה שמדובר

בהודעות קצרות מאוד וייתכן כי הסוג הזה של האלגוריתמים פחות מתאים לבעיה, עברו להשתמש באלגוריתם k-means שעבד בצורה טובה הרבה יותר עבור שאלת המחקר שלנו.

5. בשלב בו ניסינו לבדוק איפה אלגוריתם הסיווג טועה ואילו משפטים הוא אינו מסווג נכון הבחנו מהרבה מהמקרים הם משפטים שלאחר ניקוי ונרמול הפכו לקצרים מאוד (מילה או שניים) - על כן, החלטנו להוסיף פיצ'ר מרכזי וחזק בעזרת sentiment lexicon (ראה סעיף מספר 5 של השיטה) שתפקידו להבליט את העובדה שהמשפט מכיל מילים בודדות אך בעלי אופי אלים.
6. הוספת שני מסווגים נוספים - לאחר שהבנו שמסווג ה-Random Forest משיג עבורנו את התוצאות הטובות יותר עבור מספר ניסויים שונים, חשבנו לנכון לנסות עוד שני סוגי מסווגים מסוג עצים ולכן בחנו את תוצאות הסיווג גם של המסווגים - xgboost וGBM.

ניתן לראות כי השתמשנו במספר טכניקות שלא נלמדו במפורש בקורס, ביניהן - Character level N-grams (shingles), feature by sentiment lexicon, word2vec, k-means, GBM and xgboost.

7.3. <u>מסקנות</u>

ראשית, התוצאות מראות כי על ידי עיבוד מקדים בהתאם לתחום (domain) של הבעיה ובחירת מודל למידה מתאים ניתן לסווג הודעות רשת של בני נוער על אף המאפיינים הייחודיים והמאתגרים שלהם. בתוך כך, גישת העיבוד המקדים של הטקסט אותה אנו בחרנו, תוך התייחסות למאפיינים הייחודים של הודעות בקרב בני נוער אכן תרמה לסיווג וניתוח אשכולות (clustering) ההודעות.
זיהינו שתי דרכים מרכזיות להתמודד עם הבעיה של סיווג הודעות רשת לאלימות או לא –

- עיבוד מקדים ייחודי לבעיה + פיצ'רים מבוססים מילים (unigram or bigram)
- ללא עיבוד מקדים + פיצ'רים מבוססי תווים\אותיות (Character level N-grams)

כחלק מבחינת התהליך, השווינו את התוצאות עם ובלי שימוש בלמטיזציה על ידי YAP. ניתן להסיק כי כל עוד הכלי YAP מקבל כקלט מילים תקניות מהשפה העברית הוא מבצע את הלמטיזציה בצורה טובה, ומהווה כלי חשוב לאיחוד פיצ'רים שיכול מאוד לעזור לסיווג הודעות קצרות ובעיות בהן יש הרבה מאוד פיצ'רים המוגדרים כ"מילה" בקורפוס.
מצד שני, כאשר לא השתמשנו בעיבוד המקדים שלנו וב-YAP אך מודל ייצוג הפיצ'רים שלנו היה מבוסס char כפי שתארנו בשיטה, גם קבלנו תוצאות טובות. על כן, מסקנתנו היא שמודל מבוסס char מתאים אף הוא לסוגי בעיות כאלו המאופיינות בהודעות קצרות, עם הרבה שגיאות, סלנג, ריבוי מילים וחיבור אותיות. ויותר מכך, בחירת הפיצ'רים בצורה כזו מאפשרת התמודדות עם שגיאות כתיב, ריבוי אותיות, חיבורי מיילים ולכן איננה מצריכה עיבוד מקדים רחב.
לסיכום, זיהינו שתי גישות טובות לפתרון הבעיה שכל אחת מהן מתמודדת בדרכים קצת שונות עם האתגרים המגוונים בבעיה.

8. ניתוח שגיאות

ניתוח שגיאות של התוצאות מראה כי במקרים בהם הבריונות ניכרת ובוטה הסיווג יחסית פשוט למודל. מקרים כאלה מכילים צורות נפוצות של התעללות, גסויות או ביטויים המעידים על שליליות. את המקרים בהם המודל לרב טועה ניתן לחלק לכמה קבוצות:

א. מילות סלנג השזורות בשפה היומיומית וניתן להשתמש בהן גם לשלילה וגם כמילים ניטרליות\חיוביות. לדוגמא: "היא חתיכת שמרטוט", המילה "שמרטוט" היא מילה ניטרלית המתארת חפץ הנועד לניקיון, אך ברור כי במשפט הזה המילה היא מילה פוגענית. כמו גם המשפט "הוא ספורטאי רצח", המילה "רצח" היא מילה שלילית אך בסלנג הישראלי המילה יכולה גם לתאר תופעה או דבר חיובי, וברור כי זו הכוונה פה.

ב. מילים שליליות בעלות הטיה, שגיאות כתיב, ריבוי אותיות (שלא הוסר בשלב הנורמליזציה).

ג. טעויות תיוג או תיוג הניתן לפרשנות.

9. עבודות המשך

להלן מספר רעיונות לעבודות ומחקרי המשך שניתן לעשות על מנת לשפר את הפתרון לבעיה אותה הצגנו-

- **תוסף תיקון אוטומטי של שגיאות כתיב -** בהודעות בקרב בני נוער ישנן פעמים רבות שגיאות כתיב לא מכוונות, תוסף של תיקון אוטומטי של שגיאות כתיב בשלב העיבוד המקדים יכול לשמש ככלי לאיחוד פיצ'רים והורדת השונות ובכך לשפר את התוצאות.
- **מתן ציון אלימות -** הוספת פיצ'ר של חלוקת ההודעות האלימות לפי רמות אלימות שונות ועל ידי כך סיווג הודעה לפי ציון המצביע עד כמה ההודעה אלימה. תוסף זה יכול להיות מעניין ורלוונטי בייחוד למטרות פיקוח, על ידי כך שיאפשר לטפל במקרים דחופים המצריכים התערבות ומעקב. שימוש אפשרי נוסף - מתן ציון בזמן אמת לכותב ההודעה המצביע על חומרת האלימות בהודעה שלו.
- **אימון המודל על עוד מאגרי מידע מסוגים שונים -** למשל מאגרי מידע בקרב נוער בוגר ומבוגרים, מאגרי מידע של אוכלוסיות מיעוט, מאגרי מידע הקשורים לפוליטיקה. לכל אלו מאפיינים שונים מאשר למאגרי מידע בקרב בני נוער צעירים.
- סיווג הודעות על ידי סוג האלימות
- תיקון אוטומטי של הודעות אלימות ללא אלימות
- **איתור המשתתפים המעורבים בדרך כלל בבריונות ברשת-** זה יאפשר לנתח את ההקשר של אירוע בריונות ברשת ומכאן להעריך את חומרתו.

10. כלי NLP חסרים

במהלך הפרוייקט נתקלנו בכמה כלים שהיינו רוצות להכניס בתהליך, אך אינם היו קיימים עבור השפה העברית כלל או לפחות לא עבור הצרכים שלנו.

מסקירת הספרות שביצענו, נראה כי מחקרים קודמים שניסו לפתור בעיה דומה השתמשו במספר כלים אשר להערכתנו היו משפרים מאוד את הביצועים, בינהם - מילון סלנג אנגלי (NoSlang: Translator & Internet Slang Dictionary), מאגרים שנקראים WordNet או BableNet

שמטרתם חילוץ ואיחוד פיצ׳רים על ידי יחסים סמנטיים בין מילים וכן אלגוריתמים למציאת תבניות דומות בטקסט בשפה האנגלית במטרה להתמודד עם שגיאות כתיב וצורת הכתיבה הנהוגה ברשתות חברתיות (חיבור מילים, ריבוי אותיות וכו׳).
בנוסף, מודלים של רשתות עמוקות על מנת לבצע transfer learning כמו BERT אומנו ברובם על טקסטים בעברית מויקיפדיה שאינם מתאימים לשפה של הטקסטים המרכיבים את הדאטה שלנו.

11. דיון וסיכום

המשימה של סיווג וניתוח לפי אשכולות (clustering) של הודעות אלימות ברשתות החברתיות היא עדיין אתגר אמיתי בימינו, ובפרט בשפה העברית בה הכלים הנחוצים למשימה פחות מפותחים. שני נושאים עיקריים מקשים על יישום אלגוריתמי סיווג במשימה זו- המספר הנמוך של התכונות הניתנות לחילוץ מהודעה בודדת והעובדה שההודעות מלאות סלנג, ציניות, ניבים, ביטויים, שגיאות כתיב, חיבורי מילים וסמלים. על מנת למלא את החסר, הצענו גישה לעיבוד טקסט הכוללת נרמול של הטקסט ויצירת פיצ׳רים מתאימים. עיבוד מקדים והנדוס פיצ׳רים זה נועד לשפר את הביצועים של טכניקות סיווג להודעות טקסט קצרות ורועשות אלה. שיטת הנרמול והנדוס הפיצ׳רים מבוססת על לקסיקוגרפיה ומילונים סמנטיים יחד עם טכניקות לניתוח סמנטי. בנוסף, השיטה כוללת איחוד פיצ׳רים, על מנת להבטיח שבמקרים רבים יותר – תימצא התכונה בהודעה.
הערכנו את הגישה המוצעת בעזרת מערך נתונים אמיתי ולא מקודד יחד עם מספר אלגוריתמים מתחום ה- Nlp והלמידה.
תוצאותינו מצביעות בבירור כי השימוש בשיטה שהצענו יכול לספק באופן אפקטיבי שיפורים בביצועי המסווגים ואלגוריתמי הקלאטסור. לפיכך, מסננים מסורתיים הנמצאים כיום בשימוש יכולים להגדיל את הביצועים שלהם על ידי השימוש בטכניקה המוצעת על ידנו.
כמחקר עתיד, אנו מציעות לבחון את השיטה שלנו על מאגר מידע נוסף של הודעות, בדגש על מספר רב יותר של הודעות המחולקות באופן המייצג יותר את המצב בעולם, מסוגים וממקורות שונים. להערכתנו, למרות שהצעה זו הוערכה בהקשר של הודעות של ילדים ובני נוער ברשתות חברתיות, יש לנו ראיות המובילות אותנו להאמין שניתן ליישם את הטכניקה שלנו, ייתכן עם מעט שיפורים או שינויים קלים, גם כדי להתמודד עם הודעות מסוגים שונים ושל קהל יעד שונה.

הפניות

1. K. Dinakar, R. Reichart, and H. Lieberman. Modelling the Detection of Textual Cyberbullying. *The Social Mobile Web, 11, 2011.,7*
2. An Effective Approach for Cyberbullying Detection and avoidance Divyashree, *Vinutha H, Deepashree N S, Vol. 4, Issue 4, April 2016*
3. Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, Yi Chang: Abusive Language Detection in Online User Content, *Yahoo Labs Sunnyvale, CA, USA*
4. C. Van Hee, E. Lefever, B. Verhoeven, J. Mennes, B. Desmet, G. De Pauw, W. Daelemans, and V. Hoste. Automatic Detection and Prevention of Cyberbullying. *In Human and Social Analytics, 2015.,8*
5. H. Hosseinmardi, S. A. Mattson, R. I. Rafiq, R. Han, Q. Lv, and S. Mishra. Analyzing Labelled Cyberbullying Incidents on the Instagram Social Network. *InSocInfo, 2015,1*
6. I. Kayes, N. Kourtellis, D. Quercia, A. Iamnitchi, and F. Bonchi. The Social World of Content Abuser in Community Question Answering. *In WWW, 2015.,5*
7. J. M. Xu, X. Zhu, and A. Bellmore. Fast Learning for Sentiment Analysis on Bullying. *In WISDOM, 2012.,10*
8. https://www.kaggle.com/rtatman/sentiment-lexicons-for-81-languages - sentiment lexicon
9. https://drive.google.com/drive/folders/1qBgdcXtGjse9Kq7k1wwMzD84HH_Z8aJt - word2vec model
10. Homa Hosseinmardi , Sabrina Arredondo Mattson, Rahat Ibn Rafiq, Richard Han, Qin Lv, and Shivakant Mishra: Analyzing Labeled Cyberbullying Incidents on the Instagram Social Network, *Springer International Publishing Switzerland 2015*
11. Ying Chen , Sencun Zhu, Yilu Zhou, Heng Xu: Detecting Offensive Language in Social Media to Protect Adolescent Online Safety
12. Rui Zhao, Anna Zhou, Kezhi Mao: Automatic Detection of Cyberbullying on Social Networks based on Bullying Features, *Conference Paper · January 2016*
13. Karthik Dinakar, Roi Reichart, Henry Lieberman: Modeling the Detection of Textual Cyberbullying, *Massachusetts Institute of Technology Cambridge, MA 02139 USA*
14. Cynthia Van Hee, Gilles Jacobs, Chris Emmery, Bart Desmet, Els Lefever, Ben Verhoeven, Guy De Pauw, Walter Daelemans, and V´eronique Hoste: Automatic Detection of Cyberbullying in Social Media Text, *LT3, Ghent University CLiPS, University of Antwerp*