פרוייקט NLP בעברית -
ניתוח עמדות לפי נושאים בפוסטים של חברי כנסת

אלה פאליק 208882134

1 בספטמבר 2020

1 רקע והגדרת הבעיה

לחברי כנסת יש נוכחות גבוהה במדיה, בין היתר במדיה החברתית. פוסטים בפייסבוק הם היום כלי עוצמתי שמאפשר תקשורת ישירה ולא מצונזרת בין חברי הכנסת לציבור. הפוסטים שחברי הכנסת מפרסמים מייצגים את הנושאים בהם הם מתעסקים והעמדות שהם מחזיקים לגביהם. אם היינו יכולים לנתח את הפוסטים האלו ולבדוק על מה הם כותבים ובאיזה אופן, נוכל אולי לקבל תובנות לגבי העמדות שהם מתקשרים לציבור.

הבעיה שעומדת לפנינו היא שכמות הטקסטים האלו גדולה מאוד. ח"כים מפרסמים לעיתים מעל 50 פוסטים ביום, וסקירה ידנית של פוסט דורשת זמן רב ויכולה להיות מוטית לפי דעות והנחות הקורא. היינו רוצים בין היתר לאפשר ניתוח של השינויים בייצוג נושאים ודעות לאורך זמן, מה שדורש בדיקה של אלפי פוסטים.

על כן, שימוש בכלים של NLP על מנת לאפשר אוטומטיזציה של ניתוח כזה יכולה, בהינתן תוצאות טובות, לייעל מאוד את התהליך. נרצה ליצור כלי שמאפשר ניתוח כזה באופן מהיר ושאינו דורש תיוג ידני בכמות גדולה.

יאיר לפיד, כחול לבן, 7 בנובמבר 2018, 12:54 הנצחון הדמוקרטי במירוץ לקונגרס בארה"ב יוצר לא מעט אתגרים לממשלה הנוכחית. כבר יותר משנתיים שכולם אומרים לנתניהו שזה יקרה, אבל הוא חשב שהוא יודע יותר טוב. קודם כל מפני שזה יקשה על טראמפ להמשיך לעבוד. לא היה לנו ואולי גם לא יהיה נשיא אמריקאי ידידותי יותר לישראל. בשנתיים האחרונות הוא עשה סדרה של צעדים מבורכים: העברת השגרירות, ביטול הסכם הגרעין והחזרת הסנקציות, ביטול התקציב לאונר"א, המלחמה הגלויה באו"ם. מה שמחמיר את המצב היא העובדה שנתניהו מזוהה היום באופן מוחלט עם המפלגה הרפובליקנית. שיתוף הפעולה הצמוד שלו ושל השגריר שלו רון דרמר עם הרפובליקנים ריסק לרסיסים כל נסיון להעמיד פנים שישראל נמצאת מחוץ למגרש הפוליטי בארה"ב. "אני לא מוכנה שהוא ייכנס אצלי בדלת", אמרה לי על דרמר חברת קונגרס דמוקרטית משפיעה במיוחד. "הוא לא ישראלי, הוא רפובליקני," אמרה לי סנטורית חשובה והדגישה שזה נכון לגבי נתניהו ודרמר...

איור 1: דוגמא לחלק מפוסט שפורסם ע"י יאיר לפיד.

שאלת המחקר: מהם הנושאים שאפשר למצוא בפוסטים של חברי כנסת ומה העמדות שהם מבטאים לגביהם?

נשים לב שניתן לחלק את השאלה לשתי משימות:

1. **Topic Analysis**: אילו נושאים ניתן למצוא באוסף כללי של פוסטים?
2. **Sentiment Analysis**: עבור פוסטים מנושא מסויים, אילו עמדות מבטאים בהם?

עם זאת, ממעבר על פוסטים, ניתן לראות שלעיתים קרובות הם מכילים יותר מנושא אחד (לדוגמא, איור 1). כך, בפוסט יחיד הכותב יכול להציג דעות חיוביות או שליליות על מספר נושאים, וניתוח העמדה עבור הנושא עליו אנחנו מסתכלים יהיה מוטה. על כן נעבור להסתכלות יותר מפוקסת, וננתח את הפוסטים לפי משפטים, בהנחה שמשפט מכיל לא יותר מנושא יחיד.
כלומר, נרצה למצוא דרך לתייג משפטים מפוסט למספר נושאים, ועבור כל משפט, לנתח את מידת התמיכה בנושא אליו הוא מתוייג. בסוף התהליך נרצה לקבל עבור לכל פוסט אפיון של נושאים שמופיעים בו, ואת אחוז התמיכה של הכותב בכל אחד מהנושאים האלו. הפרקטיקה הזו ידועה בתור **Aspect-based sentiment analysis (ABSA)**.

2 הדאטה

2.1 scraping

הדאטה מגיע מדאטה בייס של כיכר המדינה [1], אתר של הסדנא לידע ציבורי. האתר מכיל פוסטים של חברי כנסת בצירוף מידע כמו שם הכותב, שם מפלגתו, תאריך ושעת הפרסום ועוד. בנוסף יש אפשרות לחפש פוסטים על פי מילות חיפוש.
עבור המשימה הראשונה שהגדרנו נרצה לקחת אוסף כללי של פוסטים מהאתר ולחפש בו נושאים. עם זאת, על מנת לנתח את התוצאות של השלב הזה ולקבל הערכה מספרית, נרצה לנתח דאטה סט שמכיל פוסטים שהנושאים בהם ידועים, בהנחה שכלי שישיג תוצאות טובות על דאטה סט כזה יתן חלוקה טובה גם בדאטה סט הכללי.
נתמקד בשלושה נושאים : **"בנימין נתניהו", "חוק הלאום" ו "דונאלד טראמפ"**. נושאים אלו נבחרו מכיוון שמסקירה ראשונית של הדאטה סט נראה שיש פוסטים רבים שכוללים את הנושאים האלו, ובנוסף העמדות לגביהם מגוונות, כך שהם יוכלו לשמש גם עבור המשימה השנייה. נאסוף ~ 100 פוסטים מכל אחד מהנושאים האלו באמצעות שימוש באפשרות באתר לחפש פוסטים לפי מילות חיפוש. מילות החיפוש היו "נתניהו", "חוק הלאום" ו "טארמפ" בהתאמה.
לאחר scraping נקבל לכל פוסט את שם חבר כנסת הכותב, שם המפלגה, תאריך ושעת הפרסום והטקסט (לדוגמא, איור 1).

2.2 עיבוד הדאטה

2.2.1 חלוקה למשפטים, נקיון והפרדה למורפמות

נחלק את הפוסטים למשפטים לפי כל סימני הפיסוק, כולל פסיקים.
ננקה כל משפט מסימנים מיוחדים. נזרוק משפטים של פחות ממילה אחת, בהנחה שהם לא מייצגים טוב נושא או עמדה. לבסוף נבצע הפרדה למורפמות באמצעות ספריית YAP [2].

יאיר לפיד, כחול לבן, 7 בנובמבר 2018, 12:54
['ה נצחון ה דמוקרטי ב ה מירוץ ל ה קונגרס ב ארהב יוצר לא מעט אתגרים ל ה ממשלה ה נוכחית', 'כבר יותר מ שנתיים ש כולם אומרים לנתניהו ש זה יקרה', 'אבל הוא חשב ש הוא יודע יותר טוב', 'קודם כל מפני ש זה יקשה על טראמפ להמשיך לעבוד', 'לא היה ל אנחנו ו אולי גם לא יהיה נשיא אמריקאי ידידותי יותר ל ישראל', 'ב ה שנתיים ה אחרונות הוא עשה סדרה של צעדים מבורכים', 'העברת ה שגרירות', 'ביטול הסכם ה גרעין ו החזרת ה סנקציות', 'ביטול ה תקציב ל ה אונרא', 'ה מלחמה ה גלויה ב ה אום', ...]

איור 2 : דוגמא לתוצאת עיבוד חלק מפוסט שפורסם ע"י יאיר לפיד (הפוסט המקורי באיור 1).

2.3 תיוג

עבור המשימה הראשונה נקבל דאטה סט המכיל משפטים שהגיעו מפוסטים בשלושה נושאים: "בנימין נתניהו", "חוק הלאום" ו "דונאלד טראמפ". משפטים רבים מתוכם לא מתאימים לאף אחד מהנושאים הנ"ל. נרצה למצוא כלי שיפריד את המשפטים האלו לפי שלושת הנושאים + נושא "אחר". מכיוון שיש הרבה משפטים שקשורים לשלושת הנושאים העיקריים, התקווה היא שהם יופרדו ושכל שאר המשפטים יקבלו הסתברות אחידה בין הנושאים פחות או יותר. נתייג ידנית משפטים בדאטה סט לפי שלושת הנושאים + נושא "אחר".

עבור המשימה השנייה נשתמש רק במשפטים שמתוייגים (על פי התיוג הידני) לנושא מסויים מתוך השלושה, בהתעלמות מהנושא "אחר". כלומר נקבל שלושה דאטה סטס של משפטים, כל אחד מכיל משפטים מנושא אחד. בכל דאטה סט כזה נתייג משפטים לפי תמיכה $(+1)$, נייטרליות (0) או התנגדות (-1) לנושא של הדאטה סט.

	סה"כ	טראמפ			נתניהו			חוק הלאום			"אחר"
מספר פוסטים (לפי מילות חיפוש)	303	100			103			100			-
מספר משפטים (לפי מילות חיפוש)	2696	871			862			964			-
מספר משפטים (לפי תיוג נושא)	2696	314			442			348			1592
		1	0	-1	1	0	-1	1	0	-1	
מספר משפטים (לפי תיוג דעה)	1104	154	104	56	72	130	240	103	103	142	-

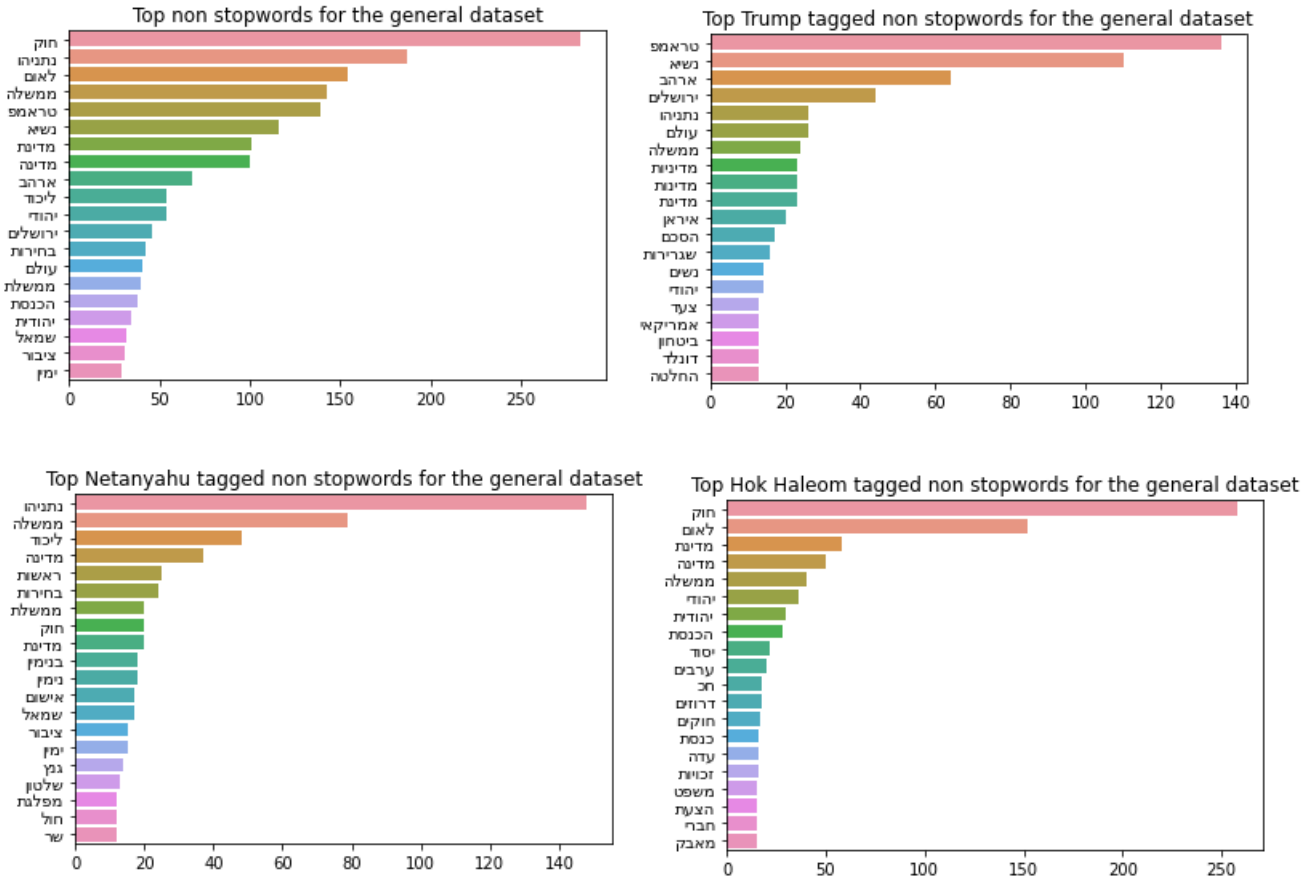
טבלה 1: הרכב הדאטה סטס הסופיים למשימות Topic Analysis ו Sentiment Analysis.

2.4 ויזואליזציות של הדאטה

נרצה לחקור את הדאטה שקיבלנו לאימון כדי לראות האם ניתן (לכל הפחות) לראות בעין אפשרויות להפרדה.

2.4.1 משימה ראשונה: Topic Analysis

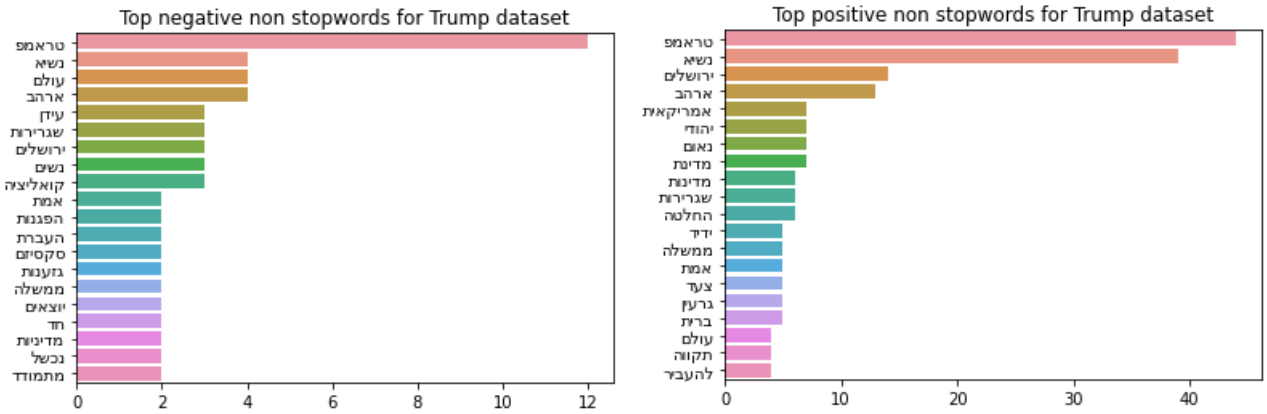
במשימה זו אנחנו מעוניינים לבצע למידה לא מפוקחת. כלומר קבוצת האימון שלנו היא אוסף כל 2696 המשפטים, שמגיעים מפוסטים שנאספו ע"י שימוש בשלוש מילות חיפוש: "נתניהו", "חוק הלאום" ו "טארמפ". נתבונן במילים נפוצות במשפטים על פי החלוקה לפי מילות חיפוש (איור 3). ניתן לראות הפרדה ברורה בין דאטה סטס שונים.



איור 3: 20 המילים הנפוצות ביותר (ללא Stopwords) בדאטה סט הכללי (שמאל למעלה), ובקבוצות המתוייגות לפי נושאים: "טראמפ" (ימין למעלה), "נתניהו" (שמאל למטה) ו"חוק הלאום" (ימין למטה).

2.4.2 משימה שנייה: Sentiment Analysis

במשימה זו אנחנו מעוניינים לבצע למידה מפוקחת. יש לנו שלושה דאטה סטס, אחד לכל נושא, וכל המשפטים מתוייגים לפי תמיכה $(+1)$, נייטרליות (0) או התנגדות (-1) לנושא. נתבונן לדוגמא בקבוצת האימון עבור דאטה סט של הנושא "טראמפ" (איור 4). ניתן לראות שהמילים הנפוצות ביותר ("טראמפ", "נשיא",...) הן משותפות ומעידות על הנושא בדאטה סט, אך ישנן מילים שיכולות להעיד על תמיכה ("ידיד", "תקווה",...) או התנגדות ("גזענות", "נכשל",...) בכל אחת מהקבוצות בהתאם.



איור 4 : 20 המילים הנפוצות ביותר המתוייגות לפי תמיכה (ימין) או התנגדות (שמאל) בדאטה סט "טראמפ". Stopwords לא נספרו מכיוון שהן מפיעות באופן יחסית אחיד בשתי הקבוצות.

3 סקירת ספרות

3.1 Embedding

ישנם מספר סוגי Embedding אפשריים עבור קורפוס.

Term Frequency או **Tf-idf**.

Bert Embedding: Bert הוא מודל שפה שמשלב מודל Pre-trained על קורפוס גדול מאוד עם תהליך fine-tuning על דאטה מתוייג. המודל ה Pre-trained של Bert לומד ייצוג של השפה עליה הוא מאומן (זמין בשפות שונות), ואפשר להשתמש בחלק הזה של Bert כדי לקבל ייצוג של הקורפוס. בייצוג זה צריך להשתמש כאשר מבצעים את תהליך ה fine-tuning על דאטה מתוייג [3].

3.2 Topic Analysis

ישנם מודלים רבים עבור זיהוי לא מפוקח של נושאים [4].

LSA - Latent Semantic Analysis: בהינתן מטריצה A המתארת קשר בין המשפטים בקורפוס למילים באוצר מילים (ייצוג מאלו שתיארנו בסעיף הקודם), נשים לב שהיא בדרך כלל מאוד דלילה, רועשת, ומפוזרת בצורה מיותרת על פני מימד גבוה מאוד. לכן נרצה לבצע הורדת מימדים כדי למצוא מספר נושאים שמייצגים היטב את הקשר בין מילים למסמכים.

הטכניקה הבסיסית היא Truncated SVD: נמצא פירוק לערכים סינגולריים (SVD) $A = USV$ ונוריד מימדים על ידי בחירה של K הערכים העצמיים הגדולים ביותר, ושמירה על K הוקטורים הראשונים מכל מטריצה (כאשר K = מספר הנושאים שנקבל). כך נקבל את K המימדים הכי משמעותיים של המרחב ש A פורשת. לאחר הורדת מימדים, U תייצג את הקשר בין משפטים לנושאים, ו V תייצג את הקשר בין נושאים למילים באוצר מילים.
ישנן עוד טכניקות שונות שנותנות קירוב של A ע"י הורדת מימדים וייצוג באמצעות כפל מטריצות: Non-nagative matrix factorization, PCA, Random SVD ועוד. הטכניקות מתבססות על סוגים שונים של פירוקים.

PLSA - Probabilistic Latent Semantic Analysis: הרעיון המרכזי הוא למצוא מודל הסתברותי שיכול לג'נרט את הדאטה שקיבלנו במטריצה A, כלומר לכל משפט d ומילה w, היינו רוצים לקבל $\mathbb{P}(d,w) \approx A_{d,w}$. על מנת לעשות זאת נשתמש בפירוק הבא: עבור משפט d ומילה w,

$$\mathbb{P}(d,w) = \mathbb{P}(d) \sum_{z \text{ topic}} \mathbb{P}(z|d)\,\mathbb{P}(w|z)$$

כאשר אינטואיטיבית, $\mathbb{P}(d)$ = כמה סביר לראות את המשפט d, והסכום = כמה סביר לראות את המילה w במשפט d.
במודל הזה, $\mathbb{P}(d)$, $\mathbb{P}(z|d)$, $\mathbb{P}(w|z)$ הם פרמטרים. $\mathbb{P}(d)$ ניתן ללמוד מהקורפוס, ואת שני האחרים ניתן למדל באמצעות התפלגויות מולטינומיות ולאמן אותן באמצעות אלגוריתם EM.
דרך נוספת לפרק את $\mathbb{P}(d,w)$ היא

$$\mathbb{P}(d,w) = \sum_{z \text{ topic}} \mathbb{P}(d|z)\,\mathbb{P}(z)\,\mathbb{P}(w|z)$$

כאשר את הפרמטרים ניתן לשערך באמצעות אלגוריתם EM, או באמצעות בייס מהפרמטרים שכבר שיערכנו בפירוק הקודם.
ניתן להשתמש בפירוק זה כדי לקבל ייצוג דומה למה שמקבלים ב LSA.

LDA - Latent Dirichlet Allocation: גרסא בייסיאנית של PLSA, כאשר משתמשים בפריורים בצורת דריכלה עבור ההתפלגויות $\mathbb{P}(d|z)$, $\mathbb{P}(w|z)$. כלומר במקום ללמוד את ההתפלגויות האלו מהדאטה בלבד, נשתמש בהנחה על ידע מוקדם שהמשתנים שלהן מתקבלים מהתפלגויות בצורת דריכלה (כאשר הפרמטרים של הפריורים הם הייפר פרמטרים של המודל).
זו הגרסא הכי נפוצה של PLSA.

3.3 Sentiment Analysis

למידת סנטימנט היא תחום מפותח ב NLP בשפה האנגלית, אך בעברית מעט מחקר קיים בנושא.

Lexicon-Based: בשפה האנגלית ישנם אלגוריתמים רבים ללמידה מפוקחת/ לא מפוקחת, המסתמכים על לקסיקונים של מילים עם משקל ידוע של סנטימנט חיובי/ שלילי. בעברית אמנם קיימים לקסיקונים כאלו [5] אך הם יחסית מצומצמים. בנוסף, ניתוח סנטימנט תלוי באופן משמעותי בהקשר, ועבור זיהוי תמיכה בנושא מסויים, משפט התומך בנושא יכול להביע דווקא סנטימנט שלילי מבחינת המילים המופיעות בו. על כן שיטה זו פחות מתאימה עבור פרוייקט זה.

Machine Learning Techniques: עבור למידה סנטימנט מפוקחת ניתן להשתמש באלגוריתמים קלאסיים של למידת מכונה, למשל, Naive Bayes, Regression, Random Forest וכו'. אלגוריתמים כאלו הם מהירים יחסית, לא דורשים כמויות יוצאות דופן של דאטה ולעיתים רבות משיגים תוצאות טובות בלמידת סנטימנט [6].

Neural Sentiment Analysis: אפשרות נוספת היא שימוש במודלים של למידה עמוקה (CNN, RNN, LSTM וכו'). מודלים כאלו נמצאים בשימוש רחב בניתוח סנטימנט בשפה האנגלית, וכן משיגים תוצאות טובות עבור קורפוסים מסויימים בעברית [7]. עם זאת, מודלים כאלו דורשים כמות גדולה מאוד של דאטה לאימון, ועבור כמות קטנה יחסית בדאטה סטס שלנו סביר שלא ישיגו תוצאות טובות.

Pre-Training with Fine-Tuning (Bert): כדי להתגבר על כמויות הדאטה הגדולות הדרושות עבור למידה עמוקה, ניתן להתבסס על המודלים שעברו אימון מקדים על כמות גדולה של טקסט עבור משימות NLP כלליות, ולהתאים אותם למשימה שלנו על ידי Fine-tuning שמתבצע באופן מפוקח על פי הדוגמאות המתוייגות. מודל קיים המתאים, בין היתר, לשפה העברית הוא Bert [3].

4 המודל

4.1 המשימה הראשונה: Topic Analysis

אנחנו מעוניינים לחלק את המשפטים לשלושה נושאים + נושא "אחר".
ראשית נוריד מהמשפטים stopwords, לפי רשימה מצורפת.
נשווה בין מספר מודלים שונים ומספר אלגוריתמים שונים על פי מדד Accuracy.

4.1.1 אלגוריתמים

נשתמש ב - 5 סוגי אלגוריתמים שראינו בקורס עם סוגי Embedding שונים (סה"כ 6 אלגוריתמים):
SVD, Random SVD, PCA, Non-Negative Matrix Factorization (NMF), LDA all with TF vectorization, NMF with TF-IDF vectorization

כל אלו מאפשרים לנו קלסיפיקציה ל K נושאים, ובנוסף מחזירים:

- ציון התאמה בין משפט לנושא.
- ציון התאמה בין מילה לנושא.

4.1.2 מודלים

1. **חלוקה ל 4 נושאים:** נריץ את האלגוריתמים עבור $K = 4$. נקבל סיווג של המשפטים ל 4 נושאים.

2. **חלוקה ל 3 נושאים + נושא "אחר":** נריץ את האלגוריתמים עבור $K = 3$. לאחר מכן נשנה סיווג של חלק ממשפטים לפי רמת הביטחון בציון התאמה בין משפט לנושא:

Certainty levels threshold:

T = אחוז המסמכים עבורם נאפשר בדיקה של "נושא אחר".
$W_{d,i}$ = ציון התאמה בין משפט לנושא שהתקבל מהאלגוריתם.
עבור משפט d בקורפוס ונושא $i \in [K]$, נגדיר:

$$\begin{aligned}
\text{certainty level}\,(d) &= \sum_{i=1}^{K} W_{d,i} \\
\text{certainty threshold}\,(d) &= \lambda \\
& s.t. \text{for } T\text{\% of docs, certainty level}\,(d) < \lambda \\
W_{d,K+1} &= \text{certainty score for doc } d \text{ to be in topic "other"} \\
&= \begin{cases} 1 - \text{certainty level}(d) & if \text{ certainty level}\,(d) < \lambda \\ 0 & else \end{cases}
\end{aligned}$$

אז נקבל מסווג:

$$\hat{f}\,(d) = \arg \max_{i \in [K+1]} W_{d,i}$$

נקבל סיווג של המשפטים ל 4 נושאים. T נקבע ל 30%.

3. הרצה בשני שלבים: נחלק את המשימה לשני שלבים:

1. נמצא את המשפטים שמתאימים לנושא "אחר":

(א) נריץ את מודל 2 ונקבל סיווג של המשפטים ל 4 נושאים.

(ב) ניקח את הנושא עם הכי הרבה סיווגים להיות נושא "אחר". עבור משפטים שסווגו כך נקבע את התווית להיות 0.

2. נסווג את שאר המשפטים ל 3 נושאים:

(א) נריץ את האלגוריתמים עבור $K = 3$ על כל המשפטים שקיבלו סיווג אחר בשלב הקודם, נקבל סיווג שלהם ל 3 נושאים.

3. בסה"כ קיבלנו סיווג של כל המשפטים ל 4 נושאים.

4.2 המשימה השנייה: Sentiment Analysis

עבור כל דאטה סט של נושא מסויים, אנחנו מעוניינים לחלק את המשפטים לפי תמיכה $(+1)$, נייטרליות (0) או התנגדות (-1) לנושא.
נשווה בין מספר אלגוריתמים שונים על פי מדד Accuracy.

4.2.1 אלגוריתמים

נשתמש ב - 4 סוגי אלגוריתמים שראינו בקורס:
Naive Bayes (MultinomialNB), Logistic Regression, Random Forest Classifier, Bert with supervised fine tuning

Logistic Regression, Random Forest Classifier נבחנו עם ובלי משקולות לקבוצת האימון (משקול לפי גודל כל מחלקה בקבוצת האימון) - Balanced\ Unbalanced.
כל אלו מאפשרים לנו קלסיפיקציה ל 3 מחלקות.

4.3 שקלול לפי פוסטים

המטרה הסופית שלנו היא לקבל עבור פוסט את הנושאים שמופיעים בו (מתוך הנושאים שבחרנו) ואת ציון התמיכה בהם:
לכל פוסט ולכל נושא i, אם קיים בפוסט משפט שתוייג לנושא i, נאסוף את כל המשפטים בפוסט שקיבלו את התיוג הזה ונחזיר ציון תמיכה = ממוצע תיוגי העמדה שלהם.
לאחר הרצת השוואה בין מודלים ואלגוריתמים שונים עבור שתי המשימות, נשכלל את התוצאות לפי פוסטים עבור האלגוריתמים עם הביצועים האופטימליים, ונריץ שתי אנליזות על התוצאות כדי להדגים את השימושיות שיכולה להיות לכלי כזה:

1. עבור הדאטה סט הכללי, נבדוק התפלגות נושאים עבור 3 הנושאים שבחרנו לפי מפלגות.
2. עבור הדאטה סט "טראמפ", נבדוק התפלגות עמדות לפי מפלגות.
3. עבור הדאטה סט "טראמפ", נבדוק התפלגות עמדות לפי חודשים.

5 תוצאות

5.1 המשימה הראשונה: Topic Analysis

עבור משימה זו בחנו ממוצע של שני מדדים: Accuracy ו Accuracy עבור כל המשפטים שלא תוייגו "אחר" ידנית.
המדד העיקרי הוא Accuracy כללי, אך מכיוון שיש הרבה יותר משפטים שמתוייגים "אחר" מאשר משפטים שמתוייגים עבור כל אחד מהנושאים, המדד השני יכול לתת תובנות לגבי כמה מהנושאים באמת תוייגו נכון.
עבור המודל השלישי (הרצה בשני שלבים) נבחר שני אלגוריתמים להרצה כאלגוריתם שלב 1, על פי הדיוק (Accuracy) שלהם עבור משפטים שתוייגו "אחר". ננרמל את הציון הזה במספר האפסים שתוייגו סה"כ כדי למנוע מצב שתיוג של כל המשפטים כנושא "אחר" יקבל ציון גבוה.

SVD	Random SVD	PCA	NMF	NMF with TF-IDF	LDA	
0.630	0.536	**0.768**	0.516	0.521	-	דיוק עבור שלב 1 = num accurate zeros / num zeros

טבלה 2: תוצאות עבור שלב 1 של המודל של הרצה בשני שלבים, כאשר הציון נקבע על פי הדיוק עבור משפטים שתוייגו "אחר" (0), מנורמל ע"י מספר המשפטים שתוייגו "אחר".

מטבלה 2 ניתן לראות ש PCA ו SVD מקבלים את התוצאות האופטימליות, לכן ניקח אותם להיות שני אלגוריתמים אפשריים להרצה כאלגוריתם שלב 1 במודל 3.
עבור כל אלגוריתם וכל מודל נמצע את התוצאות על 10 הרצות מכיוון שלחלק מהאלגוריתמים יש בסיס הסתברותי.

SVD	Random SVD	PCA	NMF	NMF with TF-IDF	LDA	
0.537	0.559	0.434	0.580	0.587	0.332	ציון עבור חלוקה ל 4 נושאים
0.420	0.639	0.469	**0.666**	**0.670**	-	ציון עבור חלוקה ל 3 נושאים + נושא "אחר"
						הרצה בשני שלבים
0.459	0.545	0.586	0.539	0.543	-	ציון עבור שלב 2 עם אלגו 1 = SVD
0.412	0.614	0.562	**0.643**	**0.656**	-	ציון עבור שלב 2 עם אלגו 1 = PCA

טבלה 3 : תוצאות עבור משימת Topic Analysis אחרי מיצוע על 10 הרצות עבור המודלים והאלגוריתמים שבחרנו.

בטבלה 3 ניתן לראות ש NMF ו NMF with TF-IDF vectorization נותנים את התוצאות הכי טובות במודלים של חלוקה ל 3 נושאים + נושא "אחר" והרצה בשני שלבים עם אלגו1 = PCA. NMF with TF-IDF vectorization מקבל תוצאות קצת יותר טובות בשני המודלים. במודל של חלוקה ל 3 נושאים + נושא "אחר" מקבלים תוצאות מעט יותר טובות, אך אם נסתכל על שני המדדים בנפרד (טבלה 4) נראה שבמודל זה נקבל הבדלים גדולים ביניהם, בעוד שמודל שני השלבים מקבל תוצאות מאוד מאוזנות.

NMF			NMF with TF-IDF			
Average	Accuracy	Accuracy for main topics	Average	Accuracy	Accuracy for main topics	
0.666	0.568	0.764	0.670	0.569	0.771	חלוקה ל 3 נושאים + נושא "אחר"
0.643	0.642	0.643	0.656	0.663	0.650	הרצה בשני שלבים עם אלגו1 = PCA

טבלה 4 : פירוט הרכב התוצאות עבור NMF ו NMF with TF-IDF vectorization במודלים של חלוקה ל 3 נושאים + נושא "אחר" והרצה בשני שלבים עם PCA.

לסיכום, כל השילובים הנ"ל נותנים תוצאות דומות, כאשר NMF with TF-IDF vectorization מעט יותר טוב, ושני המודלים נבדלים ברמת האיזון בין המדדים.

נתבונן בנוסף במילים המתארות את הנושאים שהתקבלו בהרצה בשני שלבים עם PCA ו NMF with TF-IDF vectorization וחלוקה ל 3 נושאים + "אחר" עם NMF with TF-IDF vectorization (איור 5). ניתן לראות שעבור שתי האופציות, הנושאים ממוקדים, מחולקים היטב, והמילים המתארות כל נושא בכל אחת מהאופציות דומות.

הרצה בשני שלבים עם PCA ו NMF with TF-IDF vectorization :
['חוק', 'לאום', 'מדינת', 'יסוד', 'יהודי', 'עבר', 'מדינה', 'ביטול', 'קמיניץ', 'הצעת']
['טראמפ', 'נשיא', 'ארהב', 'דונלד', 'ירושלים', 'נבחר', 'אמת', 'מברך', 'ברית', 'ממשל']
['נתניהו', 'נימין', 'ממשלה', 'ראשות', 'ליכוד', 'בנימין', 'ממשלת', 'לאומית', 'מחל', 'מדינה']

חלוקה ל 3 נושאים + "אחר" עם NMF with TF-IDF vectorization :
['חוק', 'לאום', 'מדינת', 'יהודי', 'יסוד', 'מדינה', 'עבר', 'ביטול', 'קמיניץ', 'הצעת']
['טראמפ', 'נשיא', 'ארהב', 'דונלד', 'ירושלים', 'אמת', 'נבחר', 'מברך', 'ברית', 'ממשל']
['נתניהו', 'ממשלה', 'נימין', 'ראשות', 'ליכוד', 'בנימין', 'ממשלת', 'מדינת', 'מדינה', 'לאומית']

איור 5 : 10 המילים המשמעותיות ביותר לכל נושא, בהרצה בשני שלבים עם PCA ו NMF with TF-IDF vectorization (למעלה) וחלוקה ל 3 נושאים + "אחר" עם NMF with TF-IDF vectorization (למטה).

5.2 המשימה השנייה: Sentiment Analysis

עבור משימה זו ניתחנו שלושה דאטה סטס, כל אחד מכיל משפטים מנושא אחד. בחנו את התוצאות של אלגוריתמי התיוג (ל - 3 מחלקות) על פי מדד Accuracy. נריץ את האלגוריתמים השונים על 10 הרכבים שונים של קבוצות אימון וטסט (CV עם 10 folds) ונמצע את הציונים כדי להימנע מחלוקה ספציפית שתעזור/ תפגע בביצועים.

Naive Bayes	Logistic Regression		Random Forest Classifier		Bert	
	Balanced	Unbalanced	Balanced	Unbalanced		
0.516	0.583	0.551	0.548	0.545	**0.606**	דאטה סט טראמפ
0.600	0.629	0.606	0.611	0.627	**0.659**	דאטה סט נתניהו
0.535	**0.576**	0.547	0.500	0.508	**0.573**	דאטה סט חוק הלאום

טבלה 5: תוצאות עבור משימת Sentiment Analysis אחרי מיצוע על 10 הרכבים שונים של קבוצות אימון וטסט עבור האלגוריתמים שבחרנו.

בטבלה 5 ניתן לראות ש Bert נותן לנו את התוצאות האופטימליות, כאשר עבור דאטה סט "חוק הלאום" יש מעט יתרון עליו Logistic Regression עם משקולות מאוזנות נותן תוצאה דומה.

חוק הלאום			נתניהו			טראמפ			סה"כ	
-1	0	1	-1	0	1	-1	0	1		
142	103	103	240	130	72	56	104	154	1104	מספר משפטים (לפי תיוג דעה)

טבלה 6: הרכב הדאטה סטס של משימת Sentiment Analysis.

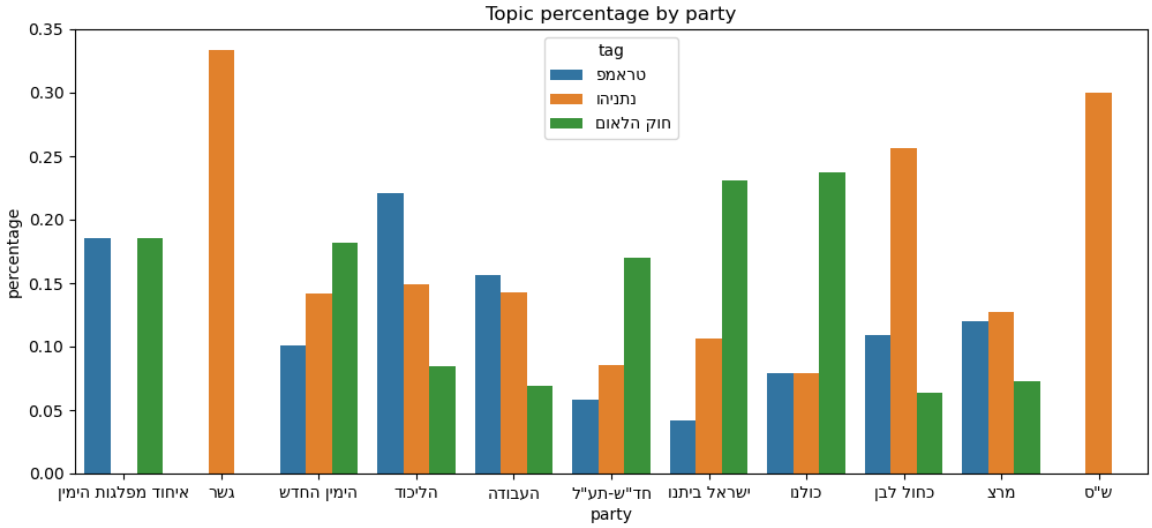
נשים לב שהתוצאות נמוכות משמעותית עבור דאטה סט "חוק הלאום" בכל האלגוריתמים. נסתכל שוב בהרכב של הדאטה סטס (טבלה 6). ניתן לראות שהדאטה סט "חוק הלאום" הוא הכי מאוזן מבין השלושה. אם נסתכל (טבלה 7) על הפרדיקציות של האלגוריתמים (לדוגמא, Bert) עבור הסטים האחרים (למשל, "טראמפ"), ניתן לראות שהפרדיקטור חוזה את רוב התיוגים להיות התיוג שמופיע הכי הרבה בסט, ומעט מאוד עבור התיוג שמופיע הכי מעט בסט. זה ככל הנראה מאפשר לנו לקבל ציון יחסית גבוה בסטים האלו לעומת סט מאוזן.

Predicted -1	Predicted 0	Predicted 1	
1	0	3	True 1
3	4	6	True 0
2	0	12	True -1

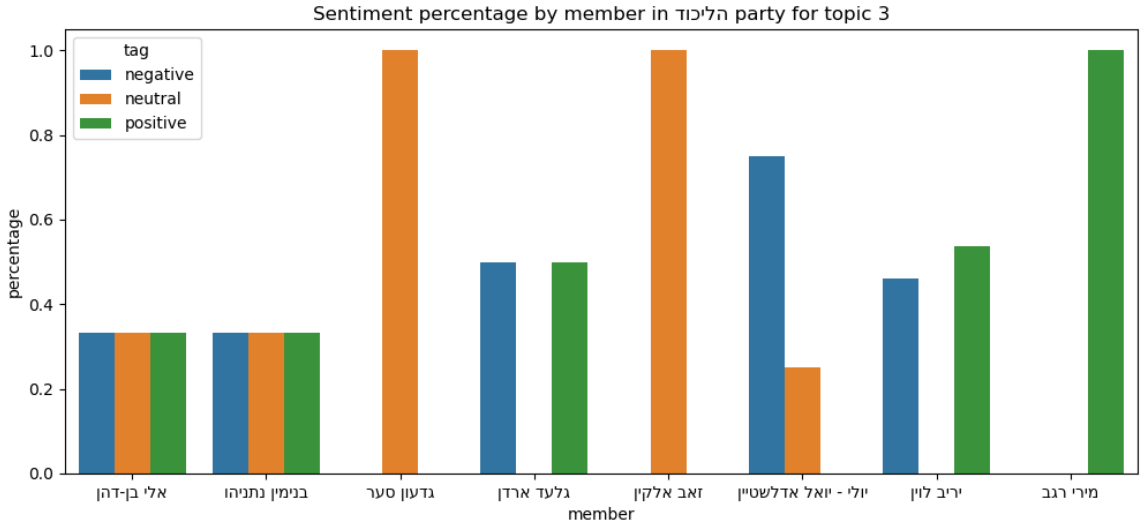
טבלה 7: Confusion matrix עבור אלגוריתם Bert בהרצה יחידה על דאטה סט "טראמפ".

5.3 שקלול לפי פוסט

נשכלל את התוצאות לפי פוסטים עבור מודל שני השלבים עם PCA ו NMF with TF-IDF vectorization עבור זיהוי הנושאים (התאמת הנושא לתיוג בוצעה ידנית), ואלגוריתם Balanced Logistic Regression עבור זיהוי עמדות. נזכור שרצינו לבצע את האנליזות הבאות כדי להדגים הסקת מסקנות מעיבוד כמו זה שביצענו. עם זאת, מכיוון שהציונים שהתקבלו עבור האלגוריתמים והמודלים שבדקנו (אפילו האופטימליים שנשתמש בהם כעת) יחסית נמוכים, יש לקחת את תוצאות האנליזות האלו בעירבון מוגבל.



איור 6: התפלגות נושאים עבור 3 הנושאים שבחרנו לפי מפלגות בדאטה סט הכללי, על פי סיווג של מודל שני השלבים עם PCA ו NMF with TF-IDF vectorization (התאמת הנושא לתיוג בוצעה ידנית).



איור 7: התפלגות עמדות לפי חברי כנסת במפלגת הליכוד בדאטה סט "חוק הלאום", על פי סיווג של אלגוריתם Balanced Logistic Regression.

6 מסקנות

בפרוייקט זה בנינו כלי למציאת הנושאים בפוסטים של חברי כנסת בפייסבוק והעמדות שהם מבטאים לגביהם.
הבנייה התחלקה לשתי משימות:

1. **Topic Analysis**: אילו נושאים ניתן למצוא באוסף כללי של פוסטים?
2. **Sentiment Analysis**: עבור פוסטים מנושא מסויים, אילו עמדות מבטאים בהם?

הפרקטיקה שהתמקדנו בה הייתה Aspect-based sentiment analysis (ABSA), כלומר ראשית למצוא נושאים שמופיעים בפוסט ואז לנתח את מידת התמיכה של הכותב בכל אחד מהנושאים בנפרד. חילקנו את הפוסטים למשפטים כדי לנתח עמדות עבור נושא מסויים בכל פעם.
עבור כל משימה השוונו בין מספר אלגוריתמים ומודלים. במשימה הראשונה קיבלנו ש NMF with TF-IDF vectorization נותן את התוצאות הכי טובות עם המודלים של חלוקה ל 3 נושאים + נושא "אחר" והרצה בשני שלבים עם אלגו PCA = 1, על פי ממוצע של מדד Accuracy ומדד Accuracy על שלושת הנושאים העיקריים. ראינו שהמודל של חלוקה ל 3 נושאים + נושא "אחר" נותן תוצאות טובות יותר באופן משמעותי עבור המדד השני, בעוד שהרצה בשני שלבים עם אלגו PCA = 1 נותנת תוצאות יותר מאוזנות בין שני המדדים. בסה"כ האלגוריתם (עם שני המודלים) זיהה היטב את הנושאים מבחינת המילים המשמעותיות עבורם, כך שניתן להשתמש בו גם כדי לזהות נושאים באוסף כללי של פוסטים. במשימה השנייה, ראינו ש Bert נותן לנו את התוצאות האופטימליות. עם זאת, התוצאות עבור סטים לא מאוזנים ("טראמפ", "נתניהו") גבוהות משמעותית מאלו של סטים מאוזנים ("חוק הלאום"), ובסטים לא מאוזנים כל האלגוריתמים חוזים את רוב התיוגים להיות התיוג שמופיע הכי הרבה בסט, ומעט מאוד עבור התיוג שמופיע הכי מעט בסט. זה ככל הנראה מאפשר לנו לקבל ציון יחסית גבוה בסטים האלו לעומת סט מאוזן, אך לא מבטא למידה אמיתית של הסנטימנט. בסוף התהליך ביצענו מספר אנליזות על הפרדיקציות כדי להדגים הסקת מסקנות מעיבוד כמו זה שביצענו, אך מכיוון שאחוזי ההצלחה בשתי המשימות לא גבוהים במיוחד, יש לקחת את התוצאות בעירבון מוגבל.
לסיכום, השאלה שעומדת במרכז הפרוייקט היא שאלה מורכבת. היכולת לענות עליה נשענת על היכולת להגדיר מהו נושא ומהן עמדות של תמיכה והתנגדות בהקשר לנושא, ואלו הגדרות שמורכב לתת, גם עבור כל נושא בנפרד. בעקבות כך, אפילו תיוג ידני של עמדות הוא מסובך, ומתבסס לעיתים קרובות על ההקשר שבו המשפט נאמר, ולא רק על הטקסט במשפט עצמו. זיהוי הנושאים מתבצע עם אחוזי הצלחה סבירים (לפחות עם מדד ה Accuracy על שלושת הנושאים העיקריים), אך האלגוריתמים לזיהוי העמדות נותרו עם ביצועים נמוכים יחסית עבור סטים מאוזנים, ולא מראים למידה אמיתית של הסנטימנט. ישנה עבודה רבה להמשך.

7 עבודה עתידית

ישנה עבודה רבה שניתן לעשות על מנת להרחיב את הפרוייקט ולשפר את התוצאות:

ניקיון וחלוקה יותר טובה למורפמות: ישנן בעיות רבות עם שמות, מילים שנחתכות, מילים שלא מחולקות למורפמות כראוי ועוד.

תיוג יותר דאטה: עבור זיהוי עמדה, תיוג דאטה נוסף הוא הכרחי, גם על מנת לשפר את הכלים בהם השתמשנו, אך גם על מנת לאפשר שימוש בכלים של למידה עמוקה, שהוכחו שימושיים בעבודות קודמות באנליזה של טקסט בעברית [7].

הכנסת רכיב סקוונטיאלי לזיהוי עמדה: תיוג משפטים מנושא מסויים לפי עמדות היא משימה קשה אפילו באופן ידני, מכיוון שלעיתים קרובות העמדה במשפט מנוסחת באופן שברור רק מתוך ההקשר של המשפטים מסביב. שימוש באלגוריתמים שמסוגלים ללמוד הקשר יכול לשפר משמעותית את היכולת לזהות עמדות.

למידת עמדות לא מפוקחת: קו מרכזי בפרוייקט היה הרצון לאפשר ניתוח מהיר של פוסטים של חברי כנסת, עם תיוג ידני מינימלי. עם זאת, ניתוח עמדות מתבצע לפי נושא, לכן לא ניתן לאמן מראש מסווג עבור נושאים חדשים ויש לתייג קבוצת אימון עבור כל נושא שנרצה לנתח. אפשרות נוספת היא למידת עמדות לא מפוקחת, למשל כזו המתבססת על ייצוג המשפטים על ידי Bert וקלאסטרינג עם אלגוריתם כלשהו. בניסויים שערכנו עם הדאטה הקיים התוצאות יצאו נמוכות מאוד (ולא נכנסו לסיכום הזה), אך יתכן שניתן לשפר אותן בעבודה עתידית.

בניית מסווגים עבור פוסטים חדשים: כיוון נוסף שניתן לפתח הוא האפשרות לסווג פוסטים חדשים על פי ניתוח שנערך על אוסף פוסטים, על פי הנושאים שנותחו באוסף המקורי.

מקורות

[1] אתר כיכר המדינה, הסדנא לידע ציבורי https://kikar.org/

[2] More, Amir, et al. "Joint transition-based models for morpho-syntactic parsing: Parsing strategies for MRLs and a case study from modern Hebrew." Transactions of the Association for Computational Linguistics 7 (2019): 33-48.

[3] Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805 (2018).

[4] "Topic Modeling with LSA, PLSA, LDA & lda2Vec", Medium, Joyce Xu, May 25, 2018, https://medium.com/nanonets/topic-modeling-with-lsa-psla-lda-and-lda2vec-555ff65b0b05

[5] WordNet for Hebrew, Computational Linguistics Group, Department of Computer Science, University of Haifa, http://cl.haifa.ac.il/projects/mwn/index.shtml

[6] Mughaz, Dror, Tzeviya Fuchs, and Dan Bouhnik. "Automatic opinion extraction from short Hebrew texts using machine learning techniques." Computación y Sistemas 22.4 (2018).

[7] Amram, Adam, Anat Ben David, and Reut Tsarfaty. "Representations and Architectures in Neural Sentiment Analysis for Morphologically Rich Languages: A Case Study from Modern Hebrew." Proceedings of the 27th International Conference on Computational Linguistics. 2018.