**המרכז האקדמי למשפט ולעסקים**

**הפקולטה למערכות מידע ומדעי המחשב**

**כיצד שימוש בכלים טכנולוגיים מבוססי בינה מלאכותית (AI) משפרים את הדיוק והיעילות בזיהוי וסינון דיסאינפורמציה ברשתות החברתיות?**

# הגשה מסכמת פרויקט מחקרי

דביר בורגר

ת.ז 209369016

**מנחה:** ד"ר סעיד עסלי

**שם הקורס:** מערכות מידע

**מס' הקורס:** 6172

ט"ז באב, תשפ"ה 10.08.2025

# אישור מנחה

# **תודות**

# **ברצוני להביע את תודתי והערכתי העמוקה למר סעיד עסלי, על ההנחיה המקצועית, התמיכה, הסבלנות והליווי הצמוד לאורך כל שלבי המחקר. תרומתו הרבה, הידע הרחב והעצות המועילות שסיפק, היוו גורם משמעותי בהצלחת העבודה.**

# **תוכן עניינים**

[הגשה מסכמת פרויקט מחקרי 1](#_Toc205586794)

[אישור מנחה 2](#_Toc205586795)

[**תודות** 3](#_Toc205586796)

[**תוכן עניינים** 4](#_Toc205586797)

[רשימת איורים, טבלאות וגרפים 5](#_Toc205586798)

[תקציר 6](#_Toc205586799)

[**Abstract (English Version)** 7](#_Toc205586800)

[מבוא 8](#_Toc205586801)

[רקע 8](#_Toc205586802)

[הגדרה ומהות בעיית המחקר 9](#_Toc205586803)

[רציונל המחקר 9](#_Toc205586804)

[חשיבות המחקר 9](#_Toc205586805)

[מטרות ויעדים של המחקר 10](#_Toc205586806)

[מטרות המחקר 10](#_Toc205586807)

[השערות המחקר 10](#_Toc205586808)

[היעדים של המחקר 11](#_Toc205586809)

[סקירה ספרותית 12](#_Toc205586810)

[שיטות 16](#_Toc205586811)

[תוצאות 28](#_Toc205586812)

[דיון 39](#_Toc205586813)

[סיכום ומסקנות 43](#_Toc205586814)

[ביבליוגרפיה 45](#_Toc205586815)

[פוסטר הפרויקט 46](#_Toc205586816)

# רשימת איורים, טבלאות וגרפים

1. איור מהלך המחקר – עמוד 16

2. גרף אורך ממוצע של טקסט – עמוד 17

3. גרף התפלגות רגשות לפי מקור נתונים – עמוד 18

4. גרף התפלגות רגשות לפי דירוג אמינות בPolitiFact – עמוד 18

5. גרף התפלגות תוויות עבור FakeNewsNet וPolitiFact – עמוד 19

6. גרף התפלגות רגשות לפי תווית עבור FakeNewsNet – עמוד 19

7. גרף התפלגות רגשות לפי תווית עבור PolitiFact – עמוד 19

8. גרף ניתוח 2-grams עבור FakeNewsNet – עמוד 20

9. גרף ניתוח 2-grams עבור PolitiFact – עמוד 20

10. Word Clouds עבור FakeNewsNet – עמוד 21

11. Word Clouds עבור PolitiFact – עמוד 21

12. טבלה גודל הדאטה לכל מקור – עמוד 22

13. גרף השוואת ייצוגי הטקסט השונים אל מול ארבעת המודלים הבודדים – עמוד 28

14. גרף תוצאות מדדי המודלים באמצעות ייצוג טקסט TF-IDF – עמוד 29

15. גרף השוואת המודלים הבודדים (Single) למודלים המשולבים (Stacking) – עמוד 30

16. גרף השוואת ביצועי מודלים עם וללא הפחתת ממדים (PCA) עבור ייצוגי טקסט שונים – עמוד 31

17. גרף סיכום תוצאות כלל המודלים מול כל ששת ייצוגי הטקסטים השונים – עמוד 32

18. גרף השפעת מאפיין הרגשות על ביצועי המודלים הבודדים במקור המידע FakeNewsNet – עמוד 33

19. גרף השפעת מאפיין הרגשות על ביצועי המודלים הבודדים במקור המידע PolitiFact – עמוד 34

20. גרף השפעת מאפיין הרגש על מודלי Stacking – עמוד 34

21. גרף השוואת רגשות לפי תגית דומיננטית אחת – עמוד 35

22. גרף השוואת רגשות לפי רגשות המקובצות לפי הקונפיגורציות השונות – עמוד 36

23. גרף תוצאות מודל SVM +XGBoost + LR לצורך ניתוח רגשות מעמיק לכל רגש – עמוד 36

24. גרף תוצאות מודל SVM +XGBoost + LR לצורך ניתוח רגשות מעמיק לכל קבוצת רגשות – עמוד 37

25. השפעת שגיאות כתיב ומאפייני רגש על ביצועי המודל לכל מקור – עמוד 38

# תקציר

המחקר עוסק בהתמודדות עם אתגר זיהוי מידע כוזב ברשתות החברתיות באמצעות כלים של עיבוד שפה טבעית (NLP) ולמידת מכונה (ML). הנתונים שנבחנו מקורם בשני מאגרי מידע מרכזיים: FakeNewsNet   
וPolitiFact, שעברו תהליכי סינון שפה, ניקוי טקסט, למיטיזציה ואיזון תוויות לצורך הכשרה מדויקת של המודלים.

הושוו מספר מודלים קלאסיים, ביניהם Support Vector Machine ‏(SVM), Random Forest, Logistic Regression ו־XGBoost, מול מודלים מבוססי Stacking, הכוללים שילובים שונים בין המודלים, לרבות Stacking דו־שכבתי (Meta-Ensemble). כל מודל נבחן תחת שלוש ייצוגי טקסט: TF-IDF, Word2Vec   
ו-FastText כאשר כל ייצוג טקסט נבחן בגרסה נוספת עם הפחתת ממדים באמצעות PCA.

הערכת הביצועים התבססה על ארבעה מדדים עיקריים: Accuracy, Precision, Recall ו־F1 Score. בנוסף, נבחנו השפעות שילוב רגשות ושגיאות כתיב על ביצועי המודלים.

מהממצאים עולה כי מודלים משולבים (Stacking) הציגו שיפור משמעותי במאגר FakeNewsNet, המאופיין בטקסטים קצרים ורגשיים, אך במאגר PolitiFact ההבדלים היו מזעריים והמודלים הבודדים השיגו תוצאות דומות.

השפעת הרגש נמצאה כתלוית הקשר, בFakeNewsNet נמצא שיפור מדדים עם שילוב רגשות, ובעיקר רגש Sadness, בעוד שבPolitiFact ההשפעה הייתה שולית. רגשות חיוביים הציגו את הביצועים הנמוכים ביותר בשני המאגרים, והוספת רגשות למודלים מורכבים (כגון Stacking עם Word2Vec) אף פגעה בביצועים, כנראה בשל ירידה בגודל תתי המדגמים ועלייה ברגישות להקשר. לכן, ניתוח הרגשות המעמיק התבסס על מודל פשוט ויציב יותר (SVM + Random Forest) עם ייצוג TF-IDF.

הוספת תכונת שגיאות כתיב לא תרמה באופן מובהק לשיפור הביצועים.

המחקר מדגיש את חשיבות ההתאמה בין שיטת הניתוח לאופי הטקסט, ומציע כיווני המשך כגון שילוב תכונות תחביריות, ניתוח הקשר רגש-כוונה, הרחבת המחקר לשפות נוספות, ושימוש במודלים מתקדמים ובדאטה בזמן אמת.

# **Abstract (English Version)**

This study addresses the challenge of detecting fake news on social media using Natural Language Processing (NLP) and Machine Learning (ML) techniques. The analyzed data originates from two major datasets: FakeNewsNet and PolitiFact, both of which underwent language filtering, text cleaning, lemmatization, and label balancing to ensure accurate model training.

Several classical models were compared, including Support Vector Machine (SVM), Random Forest, Logistic Regression, and XGBoost, alongside Stacking-based models that combine different classifiers, including a two-layer Stacking (Meta-Ensemble). Each model was evaluated using three text representation methods: TF-IDF, Word2Vec, and FastText, with additional versions using PCA for dimensionality reduction.

Model performance was assessed using four key metrics: Accuracy, Precision, Recall, and F1 Score. Furthermore, the impact of integrating emotion features and spelling error features into the models was examined.

Findings show that Stacking models yielded significant improvements on the FakeNewsNet dataset, which contains short and emotionally charged texts. However, on the PolitiFact dataset, the performance gap was marginal, and individual models achieved similar results.

The effect of emotion features was found to be context-dependent: in FakeNewsNet, incorporating emotions, particularly Sadness, improved model performance, while in PolitiFact, the impact was negligible. Positive emotions yielded the lowest performance scores across both datasets. Adding emotions to complex models (Stacking with Word2Vec) even degraded performance, likely due to smaller subgroup sizes and increased sensitivity to context. As a result, the deep emotion analysis was conducted using a simpler and more stable model (SVM + Random Forest) with TF-IDF representation.

Incorporating spelling error features did not lead to a significant improvement in model performance.

This study highlights the importance of aligning modeling approaches with the nature and style of the text, and suggests future research directions such as integrating syntactic features, analyzing emotion-intent relationships, expanding to additional languages, and applying advanced models to real-time data streams.

# מבוא

## רקע

מידע כוזב מוגדר כהפצת תוכן מטעה או שקרי, בין אם במכוון ובין אם בשגגה, תוך יצירת מצג שווא של עובדות. בעידן הדיגיטלי, ובעיקר ברשתות החברתיות, הפצת מידע כוזב הפכה לתופעה נרחבת ומשפיעה, שכן תכנים מסוג זה נוטים להתפשט במהירות רבה, לעיתים אף יותר מתכנים מבוססי אמת.

השפעותיו השליליות ניכרות במגוון תחומים קריטיים. ראשית, מידע כוזב עלול לעוות את תפיסת המציאות של הציבור, לעורר פחד, שנאה ובלבול, ולגרום להחלטות שגויות ברמה האישית והחברתית. בנוסף, הוא משמש ככלי לתעמולה פוליטית, מערער תהליכים דמוקרטיים, ומעמיק שסעים חברתיים. גם הכלכלה מושפעת, שכן מידע שגוי עשוי לערער את היציבות הפיננסית, לפגוע בעסקים ולגרום להפסדים משמעותיים.

למרות קיומן של מערכות לניטור ולסינון מידע כוזב, אתגר הזיהוי וההתמודדות עם תכנים אלו נותר משמעותי. מרבית הפתרונות הקיימים מתקשים לא רק להתמודד עם היקף ומהירות ההפצה של מידע ברשתות החברתיות, אלא גם בזיהוי מדויק של תוכן כוזב, במיוחד כאשר הוא מוסווה כתוכן אמין או מתבסס על מידע חלקי. מצב זה מדגיש את הצורך בפיתוח מנגנונים מתקדמים יותר, שיאפשרו סינון יעיל ומדויק של מידע כוזב בזמן אמת.

כתוצאה מכך, שאלת המחקר המרכזית היא: כיצד שימוש בכלים טכנולוגיים מבוססים AI משפרים את יעילות זיהוי וסינון מידע כוזב מכוון ברשתות החברתיות?

מחקר זה מתמקד בהבנת האתגרים הקיימים בזיהוי מידע כוזב מכוון (דיסאינפורמציה) ברשתות החברתיות, ובפיתוח פתרונות טכנולוגיים מתקדמים שיסייעו בהגברת אמינות המידע. בעזרת שימוש בטכניקות של למידת מכונה ועיבוד שפה טבעית (NLP), המחקר יבחן כיצד ניתן לשפר את היכולת של מודלים טכנולוגיים לנטר ולסנן מידע כוזב בצורה מדויקת יותר.

לאור זאת, השערת המחקר היא כי שילוב מודלים מבוססי למידת מכונה (Machine Learning) ועיבוד שפה טבעית (NLP) ישפר את רמת הדיוק בזיהוי וסינון דיסאינפורמציה ברשתות החברתיות, בהשוואה לשיטות קיימות. מודל זה צפוי להציע תוצאות מהירות, אמינות ואפקטיביות יותר, ובכך לצמצם את השפעת הדיסאינפורמציה ולהגביר את אמינות המידע במרחב הדיגיטלי.

## הגדרה ומהות בעיית המחקר

בעיית המחקר נוגעת לתופעת הפצת מידע כוזב מכוון (דיסאינפורמציה) ברשתות החברתיות, המהווה אתגר מרכזי בעידן הדיגיטלי. מידע זה מתפשט במהירות, עלול לעוות את תפיסת המציאות, לגרום לבלבול, לעורר פאניקה, לפגוע במוסדות דמוקרטיים ולהשפיע על דעת הקהל. חרף קיומן של מערכות קיימות לניטור מידע, עדיין קיימת מגבלה באיתור דיסאינפורמציה המוצגת באופן אמין, במיוחד בזמן אמת.

לאור זאת, המחקר מתמקד בפיתוח פתרון מבוסס למידת מכונה (Machine Learning) ועיבוד שפה טבעית (NLP), שמטרתו לשפר את רמת הדיוק בזיהוי וסינון מידע כוזב. הדגש יושם על בניית מודלים מסוגים שונים, בחינתם על בסיס מדדים כמותיים, והערכת התרומה של רכיבי ניתוח רגשות  
(Emotions Analysis) לשיפור יכולת הזיהוי. המחקר אינו מבצע השוואה למערכות מסחריות קיימות, אלא בוחן את יעילותן היחסית של קונפיגורציות שונות באמצעות דאטה אמיתי.

## רציונל המחקר

הרציונל לביצוע מחקר זה נשען על ההבנה כי הפצת דיסאינפורמציה ברשתות החברתיות היא תופעה רחבת היקף בעלת השפעות שליליות עמוקות על החברה, הכלכלה והדמוקרטיה. דיווחים כוזבים עלולים להתפשט במהירות רבה, ובמקרים רבים מצליחים לעקוף את מערכות הסינון הקיימות. הדבר יוצר צורך הולך וגובר לפיתוח מודלים מדויקים וגמישים שיידעו לזהות ולסנן מידע כוזב בזמן אמת, גם כאשר הוא עטוף במעטה של מהימנות.

מחקר זה שואף לגשר על הפער בין היכולת לזהות דיסאינפורמציה לבין קצב התפשטותה, באמצעות פיתוח מודל חדשני שמבוסס על טכניקות מתקדמות של למידת מכונה ועיבוד שפה טבעית. המודל שיפותח יתמקד בזיהוי דפוסים סמנטיים, לשוניים ורגשיים של מידע כוזב, תוך בחינה אמפירית של השפעת תוספת ניתוח רגשות על ביצועי המודל.

## חשיבות המחקר

חשיבות המחקר נעוצה בצורך הדחוף להתמודד עם תופעת הדיסאינפורמציה ברשתות החברתיות, שמהווה איום ממשי על אמון הציבור, יציבות חברתית, ותפקוד תקשורתי ודמוקרטי תקין. בעידן שבו תוכן שגוי מתפשט במהירות ויכול להשפיע על קהלים רחבים תוך שניות, עולה הצורך בכלים מדויקים ואפקטיביים שיידעו לאתר ולסנן מידע כוזב בזמן אמת.

בעוד שמרבית המערכות הקיימות מתקשות בזיהוי תכנים כוזבים המוסווים כתוכן לגיטימי, המחקר הנוכחי מבקש להציע פתרון מבוסס דאטה, הנשען על ייצוגים טקסטואליים מתקדמים, רגשות טקסטואליים (sentiment & emotion), וטכניקות של למידת מכונה. יישום המודל עשוי לשפר משמעותית את דיוק הסינון ולתרום לבניית סביבה דיגיטלית בטוחה יותר – הן מבחינה תקשורתית והן מבחינה חברתית ודמוקרטית.

# מטרות ויעדים של המחקר

## מטרות המחקר

מטרת מחקר זה היא לבחון את יעילותם של מודלים חישוביים מבוססי למידת מכונה ועיבוד שפה טבעית (NLP) בזיהוי וסינון דיסאינפורמציה ברשתות החברתיות. במסגרת המחקר פותחו והוערכו קונפיגורציות שונות של מודלים, תוך שילוב ייצוגים טקסטואליים מגוונים (כגון TF-IDF, Word2Vec ו(FastText- בשילוב תכונות רגש שהופקו באמצעות ניתוח רגשות.  
המחקר מתמקד בהערכת ביצועי הקונפיגורציות השונות במונחי דיוק, רגישות, ספציפיות ומידת ההתאמה הכללית ,(F1) במטרה לזהות את השילוב האופטימלי המוביל לתוצאות מיטביות בזיהוי מידע כוזב.  
בניגוד למחקרים המשווים למערכות קיימות, מחקר זה מתמקד בהשוואה פנימית בין גישות שונות לייצוג וניתוח הטקסט, תוך בחינת התרומה של ניתוח רגשות כמשתנה מחזק לשיפור ביצועי הסיווג.  
תוצרי המחקר עשויים לתרום לפיתוח פתרונות מדויקים ויעילים יותר להתמודדות עם הפצת דיסאינפורמציה, ולספק מסגרת אמפירית להמשך פיתוחם של כלים מבוססי NLP בתחום זה.

## השערות המחקר

1. יישום מודלים מבוססי למידת מכונה ועיבוד שפה טבעית יתרום לשיפור רמות הדיוק בזיהוי ובסינון של מידע כוזב לעומת ביצועים אקראיים או בסיסיים.

* ההשערה תיבחן באמצעות מדדי ביצוע סטנדרטיים כגון: Accuracy, Precision, Recall   
  ו-F1-Score

2. הביצועים של מודלים משולבים ,(Ensemble Learning) ובעיקר מודלים מסוג ,Stacking יהיו גבוהים באופן מובהק בהשוואה לאלגוריתמים יחידניים (Individual Classifiers) בזיהוי והבחנה בין מידע אמין למידע כוזב.

* ההשערה תיבחן באמצעות השוואת ביצועי מודלים בודדים לעומת מודלי Stacking

3. שילוב של טכניקות לייצוג טקסט יחד עם הפחתת ממדים (Dimensionality Reduction) באמצעות PCA יניב תוצאות מדויקות יותר, תוך הפחתת רעש ושיפור כללי של ביצועי המודל.

* ההשערה תיבחן באמצעות השוואת ביצועים בין קונפיגורציות הכוללות PCA לבין קונפיגורציות ללא הפחתת ממדים

4. שילוב של מאפייני רגש (Emotion Features) כחלק ממערך הקלט של המודל יתרום לשיפור ביצועיו, בהשוואה למודלים המבוססים על טקסט בלבד.

* ההשערה תיבחן באמצעות השוואת ביצועים בין קונפיגורציות הכוללות פיצ'רי רגשות לבין אלו שאינן כוללות אותם

5. שילוב של ממוצע שגיאות כתיב כפיצ'ר נוסף בייצוג הטקסט יתרום לשיפור ביצועי המודל, בהשוואה לשימוש בייצוג טקסט בלבד.

* ההשערה תיבחן באמצעות השוואת ביצועים בין קונפיגורציות הכוללות את פיצ'ר ממוצע שגיאות כתיב לבין קונפיגורציות ללא פיצ'ר זה

## היעדים של המחקר

1. הבנת הבעיה והקשר החברתי-דיגיטלי

יעד: לאפיין את תופעת הדיסאינפורמציה ברשתות החברתיות ולהבין את ההשלכות החברתיות, התקשורתיות והטכנולוגיות שלה.

מהלך העבודה: ביצוע סקירת ספרות עדכנית בתחום, ניתוח מגמות קיימות בהפצת מידע כוזב, והצגת הצורך במודלים חישוביים מבוססי NLP ולמידת מכונה.

2. פיתוח תשתית נתונים וייצוגים טקסטואליים

יעד: לאסוף ולעבד מאגרי מידע קיימים (כגון FakeNewsNet ו-PolitiFact), ולבצע עליהם פעולות עיבוד מוקדם (טוקניזציה, לימיטיזציה, הסרת מילים ריקות וכו').

מהלך העבודה: שילוב ייצוגים טקסטואליים שונים (TF-IDF, Word2Vec, FastText) והוספת מאפיינים לשוניים כגון רגשות, לצורך בניית תשתית להזנת המודלים.

3. הטמעת ניתוח רגשות כמשתנה תורם לזיהוי דיסאינפורמציה

יעד: לבדוק את תרומתו של ניתוח רגשות כמאפיין נוסף בזיהוי פוסטים כוזבים לעומת פוסטים אמינים.

מהלך העבודה: שילוב תכונות רגש (Emotion Features) כחלק מהקלט למודלים שונים, והשוואת ביצועים בין קונפיגורציות הכוללות רגשות לבין כאלה שלא.

4. בניית מודלים והשוואת קונפיגורציות סיווג שונות

יעד: לפתח מספר קונפיגורציות של מודלים מבוססי למידת מכונה (SVM, XGBoost, RF, Logistic Regression) בשילוב תכונות שונות, ולהשוות ביניהן לצורך זיהוי מידע כוזב.

מהלך העבודה: ביצוע Cross-Validation עם מדדים כמותיים (Accuracy, Precision, Recall, F1), והשוואה בין שילובים של ייצוגים טקסטואליים, רגשות והפחתת ממדים.

5. הערכת ביצועי הקונפיגורציות והסקת מסקנות

יעד: לזהות את השילובים האפקטיביים ביותר מתוך כלל הקונפיגורציות שנבדקו.

מהלך העבודה: ניתוח השוואתי של התוצאות, הדמיית הביצועים באמצעות גרפים וויזואליזציות, והסקת מסקנות ביחס לשיפור שחל באמצעות שילוב רכיבים כמו רגשות וPCA.

6. גיבוש המלצות ליישום עתידי

יעד: לנסח המלצות מחקריות ומעשיות להמשך פיתוח מודלים בתחום זיהוי דיסאינפורמציה.

מהלך העבודה: התבססות על תובנות המחקר לצורך תכנון עתידי של יישומים טכנולוגיים מבוססי NLP, ובחינת שילובם במערכות ניטור וסינון ברשתות החברתיות.

# סקירה ספרותית

מידע כוזב ברשתות החברתיות מתפשט במהירות ובנפח גדול יותר בהשוואה למידע אמין, ומשפיע לרעה על יחידים, קהילות, מוסדות חברתיים, מערכות פוליטיות, גופי בריאות וכלכלה.  
בזירה הפוליטית, דיסאינפורמציה משמשת ככלי תעמולה, משפיעה על תהליכי בחירות והצבעות, מערערת את האמון במוסדות הדמוקרטיים, ומחריפה קיטוב חברתי על ידי חיזוק שסעים אידיאולוגיים. בתחום הבריאות, במיוחד במגפות כמו COVID-19, מידע כוזב גרם להימנעות מחיסונים, לאימוץ טיפולים מסוכנים, ולהגברת חוסר האמון במערכת הבריאות, תופעה שהובילה לעלייה בתמותה. במהלך אסונות טבע, דיסאינפורמציה יכולה לעורר פאניקה ובלבול, לשבש מאמצי חילוץ והצלה, ולהוביל אנשים לקבל החלטות מסוכנות בהתבסס על מידע שגוי. מספר גורמים מרכזיים תורמים להפצת מידע כוזב. חרדה   
ואי-ודאות גורמות לאנשים להסתמך על מידע לא מבוסס, במיוחד בזמן משברים. אפקט תא ההדהוד מחזק דיסאינפורמציה בכך שמשתמשים נחשפים בעיקר לתוכן שמתאים לאמונותיהם, ללא בדיקת אמינות. אלגוריתמים של רשתות חברתיות מעדיפים תוכן מעורר אינטראקציה, מה שמוביל להפצה מהירה יותר של מידע כוזב בהשוואה למידע אמין. בנוסף, מניעים פוליטיים וכלכליים מעודדים הפצת דיסאינפורמציה ככלי להשפעה על דעת הקהל או להשגת רווחים. (Muhammed & Mathew, 2022)

כיום ניתן לסווג את השיטות לניטור וזיהוי דיסאינפורמציה לשתי גישות עיקריות: גישות מסורתיות וגישות מבוססות בינה מלאכותית (AI). הגישות המסורתיות כוללות בדיקת עובדות (Fact-Checking), במסגרתה משווים טענות למאגרי מידע מהימנים, זיהוי שמועות (Rumor Detection) שמתמקד בזיהוי שמועות בלתי מאומתות, לרוב ברשתות חברתיות, ניתוח עמדות (Stance Detection) שמעריך אם טקסט תומך, מתנגד או ניטרלי כלפי טענה מסוימת, וניתוח סנטימנט (Sentiment Analysis) המנתח את הרגש הכללי של טקסטים אך אינו קובע ישירות את אמינותם. בגישות מבוססות למידת מכונה ועמוקה, נעשה שימוש במודלים קלאסיים כמו SVM, Naïve Bayes, Random Forest ובמודלים מתקדמים יותר כמו רשתות עצביות (Neural Networks), כולל LSTM לרצפים ארוכים, CNN לזיהוי תבניות בטקסט, ו-GNN לניתוח התפשטות מידע כוזב ברשתות חברתיות. מנגנוני קשב (Attention Mechanism) מאפשרים למודלים להתמקד במילים קריטיות בטקסט, ומשפרים את ביצועי LSTM בזיהוי חדשות מזויפות. הטרנספורמרים (Transformers), בהם BERT, RoBERTa, GPT ו-T5, הם המתקדמים ביותר כיום ומשיגים דיוק גבוה על ידי הבנת הקשרים רחבים בטקסט. שיטות מתקדמות נוספות כוללות גישה רטורית (Rhetorical Approach), הבוחנת את מבנה המשפטים כדי להבחין בין חדשות אמיתיות למזויפות, וחיפוש והצלבת ראיות (Evidence-Based Detection), המצליבה תוכן עם מסמכים קיימים מבוסס RTE. למרות שכלים מסורתיים מסייעים בהתמודדות עם מידע כוזב, הם מוגבלים בדיוקם ואינם מזהים הקשרים סמנטיים עמוקים או מניפולציות בטקסט, ולכן נדרש מעבר למודלים מתקדמים יותר המבוססים על למידת עומק (Deep Learning). (Oshikawa, Qian, & Wang, 2020)

מידע כוזב הוא שם כולל הניתן לחלק לשלושה קטגוריות: מידע כוזב לא מכוון (Misinformation) הוא מידע שגוי שהופץ בטעות ללא כוונה להטעות, מידע כוזב מכוון (Disinformation) מתאר הפצה מכוונת של מידע כוזב במטרה להטעות, לייצר מניפולציה או לקדם אג'נדה פוליטית, כלכלית או חברתית. וחדשות מזויפות (Fake News) הוא סוג ספציפי של דיסאינפורמציה, המתיימר להיות דיווח חדשותי אך למעשה מבוסס על מידע שקרי או מסולף, לרוב עם מטרות אידיאולוגיות או מסחריות.  
הסיבה לאפיון סוגי המידע הכוזב היא כי קיימות שיטות שונות לזיהוי וסינון, בשל הבדלים במקור ההפצה, בכוונה ובמבנה התוכן. הפצת מידע כוזב לא מכוון דורש יותר בדיקות עובדתיות ובכך ניתן לזיהוי באמצעות בדיקת עובדות וטכניקות סיווג טקסט מסורתיות (Naïve Bayes, SVM). אך על מנת להתמודד עם מידע כוזב מכוון מאחר והוא מתוכנן להיראות אמין ולחמוק מכלי סינון פשוטים יש צורך בשימוש של AI מתקדם, רשתות עמוקות (BERT, RoBERTa) וניתוח דפוסי הפצה. חדשות מזויפות עלולות לכלול טקסט, תמונות וסרטונים מזויפים, ולכן נדרשת גישה רב-מודאלית (Multimodal AI) המשלבת ניתוח טקסט וניתוח חזותי. (Alghamdi, Lin, & Luo, 2023)

שלבי המחקר בתחום זיהוי מידע כוזב דומים זה לזה. תחילה, יש לייבא מערך נתונים שישמש כבסיס למחקר, עליו נבצע אימון ובדיקת מודלים, כאשר מערכי נתונים נפוצים בתחום כוללים LIAR, FakeNewsNet ו-Twitter15. לאחר מכן, יש לבצע עיבוד שפה טבעית (NLP) וייצוג טקסט, הכולל ניקוי נתונים, Stop Words - הסרת מילים נפוצות ,טוקניזציה - פירוק טקסט, סטמינג ולממטיזציה – קיצור מילים לצורת השורש. בשלב הבא, נייצג מילים וטקסטים באמצעות טכניקות שונות כמו Word2Vec, GloVe וBERT-. בנוסף, נבצע חילוץ מאפיינים מהטקסט כמו הפחתת ממדים, כדי למנוע Overfitting ולשפר את דיוק המודל. לאחר עיבוד הנתונים, נפתח וניישם מודלים מתקדמים לזיהוי דיסאינפורמציה. להלן מודלים עיקריים בתחום: רשתות קונבולוציה (CNN), רשתות נוירונים חוזרות (RNN) כגון LSTM ו-GRU, רשתות גרפים עצביות (GNN), רשתות גנרטיביות (GANs), מודלים מבוססי מנגנוני קשב (Attention Mechanisms), וכן מודלים מבוססי טרנספורמרים (Transformers). לבסוף, ננתח את ביצועי המודלים באמצעות מדדי הערכה כגון דיוק F1-score ,(Accuracy) ומדדים נוספים להערכת הביצועים וההתאמה של המודלים למשימה. (Mridha, Keya, Hamid, Monowar, & Rahman, 2021)

ממחקר שבוצע, אשר השווה בין מודלים קלאסיים של למידת מכונה לבין מודלים מבוססי למידה עמוקה (Deep Learning), נמצא כי למידת מכונה מסורתית הייתה מהירה יותר אך פחות מדויקת, בעוד שלמידה עמוקה השיגה דיוק גבוה יותר, אם כי דרשה כוח חישובי גבוה וזמן אימון ממושך. במחקר נעשה שימוש במאגרי הנתונים LIAR, FakeNewsNet ו-PHEME, ובוצעה השוואת ביצועים בין טכניקות שונות לסיווג מידע כוזב, בהן רגרסיה לוגיסטית, עצי החלטה, Random Forest ומודלים מבוססי Transformer. ההשוואה בחנה דיוק (Accuracy), זמן הרצה, רגישות (Recall) וערך F1 של המודלים השונים. מבין המודלים הקלאסיים, רגרסיה לוגיסטית הציגה את הביצועים הטובים ביותר, עם דיוק של עד 75%, אך כלל המודלים מבוססי למידה עמוקה הציגו תוצאות מרשימות יותר. המודל שהשיג את הדיוק המקסימלי הגבוה ביותר, 88%, היה BERT, אשר נמצא כמודל החזק ביותר לזיהוי מידע כוזב במחקר זה.  
BERT מבוסס על Transformer, שהוא סוג מתקדם של רשתות נוירונים עמוקות המתמקד בזיהוי קשרים בין מילים בטקסט באופן דו-כיווני. מודל זה מסוגל ללמוד יחסים סמנטיים בטקסט, אפילו במשפטים מורכבים או מבוססי הקשרים רחבים, ולכן השיג תוצאות טובות באופן משמעותי בהשוואה לשיטות אחרות. (Galli, Masciari, Moscato, & Sprli, 2022)

שילוב מודלים מבוססי למידה עמוקה הצליחו להשיג דיוק גבוהות יותר לעומת מודלים מסורתיים או נפרדים של למידה עמוקה. תוצאות המחקר הראו כי Logistic Regression (LR) השיג דיוק של 89.2% אך התקשה בזיהוי דפוסים סמנטיים מורכבים, Support Vector Machine (SVM) הציג דיוק של 91.6% אך היה רגיש להקשר, ואילו Naïve Bayes (NB) היה פחות מדויק (87.3%) והתאים בעיקר לניתוח טקסטים קצרים. מבין מודלי הלמידה העמוקה, CNN הגיע ל-93.5% אך התקשה בהבנת הקשרים רחבים, LSTM הציג שיפור עם דיוק של 94.5% בזכות יכולתו לנתח מבנים טקסטואליים לאורך זמן, ו-XGBoost שיפר את ביצועי המודלים הקודמים עם דיוק של 92.2%. המודל המוצע, MM-FND, ששילב בין CNN ו-LSTM, הציג את התוצאות הטובות ביותר עם דיוק של 96.3%, מה שמעיד על כך ששיטות היברידיות מספקות ביצועים טובים יותר לעומת מודלים מסורתיים או נפרדים של למידה עמוקה. המחקר ביצע שימוש במאגרי הנתונים LIAR, FakeNewsNet ו-PHEME. עם זאת, המאמר מציג כי שימוש ב-Transformer Models כמו BERT עשוי לשפר עוד יותר את הזיהוי והסינון של מידע כוזב במיוחד כאשר מדובר בטקסטים עם הקשרים סמנטיים מורכבים. אך הוא מדגיש כי שילוב כזה דורש משאבים רבים יותר, ולכן נכון לעכשיו, מודלים מבוססי CNN-LSTM הציעו איזון טוב יותר בין דיוק למהירות. (Soni, Sharma, & Sinha, 2025)

אף על פי זאת, מחקר אחר אשר ניסה למצוא פתרון מאוזן בין רמת דיוק גבוהה, יעילות חישובית והתאמה ליישום מעשי, הציג מודל היברידי חדש המשלב בין BERT ו-LSTM. החוקרים השתמשו במאגר נתונים ייעודי שנאסף ממקורות שונים ברשתות חברתיות באמצעות סורק רשת (Web Crawler) מותאם אישית. המחקר השווה את BERT-LSTM למודלים קיימים, כולל Logistic Regression, CNN, LSTM, GNN, GPT-3 ו-T5, לצד מודלים מסורתיים נוספים. תוצאות המחקר הצביעו על כך שמודלים מסורתיים אינם מספקים דיוק מספק, כאשר רמת הדיוק שלהם נעה בין 83% ל-84%, בעוד שמודלים מבוססי למידה עמוקה הציגו דיוק של מעל 90%. המודל המוצע BERT-LSTM השיג דיוק של 93.51%, בעוד ש-GPT-3 ו-T5 הציגו דיוק מעט גבוה יותר, כאשר BERT-Large הוביל עם 94.10%. למרות זאת, ההמלצה המרכזית של המחקר היא להשתמש במודל BERT-LSTM, שכן הוא מציע איזון אופטימלי בין ביצועים לדרישות חישוביות, מבלי להצריך משאבי עיבוד משמעותיים יותר. (Wang, Wang, & Yu, 2025)

CMAFD הינו מודל חדשני לזיהוי דיסאינפורמציה ברשתות החברתיות, המבוסס על למידה עמוקה וניתוח הקשרים סמנטיים בין גוף החדשות, כותרות המאמרים ותגובות המשתמשים. המודל משלב BERT עם BiGRU ליצירת ייצוגים סמנטיים מעמיקים ומיישם מנגנון תשומת לב רב-מודאלי (Cross-Modality Attention) המאפשר למערכת להבין קשרים בין סוגי טקסט שונים. כמו כן, הוא כולל מקודד דפוסי פרסום, הבוחן את היסטוריית השיתופים והמעורבות החברתית של משתמשים כדי לשפר את רמת הדיוק. המחקר השווה את המודל למספר שיטות קיימות ומצא כי CMAFD השיג דיוק של 95.83% ב-PolitiFact ו-93.51% ב-GossipCop, מה שהיווה שיפור משמעותי לעומת מודלים קיימים. מסקנות המחקר מצביעות על כך ששימוש בשיטות רב-מודאליות המשלבות טקסט, כותרות ותגובות, לצד ניתוח דפוסי משתמשים, מוביל לזיהוי אמין ומדויק יותר של דיסאינפורמציה. (Alghamdi, Lin, & Luo, 2024)

חוקרים שילבו מדדי סנטימנט של VADER כמאפיין נוסף למודל למידת העומק BERT + CNN. התוצאות הראו שיפור בביצועים של המודל כאשר ניתוח הסנטימנט שולב, מה שמעיד על כך שלרגשות בטקסט יש תפקיד בזיהוי דיסאינפורמציה. המחקר ביסס את המודל שלו על ציוצים בטוויטר מתקופת הקורונה, תוך שימוש בפילטרים של האשטאגים, ניתוח סנטימנט ותיוג מבוסס מקורות בדיקת עובדות. בסיס הנתונים אפשר לאמן ולבדוק את המודל על מידע אמיתי מרשת חברתית, אך מגבלותיו כוללות התמקדות בנושא יחיד והסתמכות על מקורות בדיקת עובדות חיצוניים. (Maathuis & Godschalk, 2023)

מחקר נוסף אשר חקר את שילוב המודלים CNN וLSTM פיתח אלגוריתם חדשני לזיהוי תוכן מטעה, BER-DLEDCD, המבוסס על שילוב המודלים, יחד עם אופטימיזציה של פרמטרים באמצעות אלגוריתם BER. המחקר השתמש במאגרי הנתונים PolitiFact ו-BuzzFeed לצורך אימון והערכת המודל לזיהוי מידע כוזב. נעשה שימוש ברשתות הנוירונים CNN-LSTM, כאשר ה-CNN מזהה דפוסי טקסט ברמה המיקרו-מבנית, בעוד ה-LSTM מנתחת קשרים טקסטואליים ארוכי טווח, מה שמאפשר למודל להבין טוב יותר הקשרים סמנטיים מורכבים. לבסוף, מתבצעת אופטימיזציה מבוססת BER, המתבססת על חישובי אל-בירוני לשיפור בחירת ההיפר-פרמטרים, מה שמעלה את דיוק הזיהוי ומפחית שגיאות בסינון המידע. תוצאות המחקר הראו כי המודל הציג ביצועים טובים יותר בהשוואה למודלים קיימים, עם דיוק של 94% בזיהוי חדשות כוזבות, תוך השגת precision של 94.83% ו-F1-score של 94.30%. שילוב רשתות CNN ו-LSTM תרם להבנת דפוסי טקסט עמוקים, שיפור יכולת ההבחנה בין מידע אמין לכוזב, והגברת דיוק הסינון. בנוסף, אופטימיזציית BER תרמה לשיפור היעילות החישובית של המערכת, הפחיתה את שיעור ההתראות השגויות, ואפשרה זיהוי מהיר ומדויק יותר של מידע ברשתות החברתיות.  
(Jain, Neelakandan, Vidyarthi, Mishra, & Alkhayyat, 2025)

כלל המחקרים שבוצעו בנושא אשר השתמשו בשיטות חדשניות המבוססות למידה עמוקה וNLP הצדיקו את השערת המחקר שלנו, ניתן לראות כי הם הצליחו להביא לשיפור בדיוק ובמהירות זיהוי וסינון מידע כוזב מכוון ברשתות החברתיות. השיפור נבע מהיעילות של מודלים של למידה עמוקה שהצליחו לזהות תבניות מורכבות בטקסטים, להתמודד עם נתונים רועשים ולקבל החלטות מדויקות יותר על סמך למידה עצמאית, ללא צורך בהגדרה ידנית של תכונות.

# שיטות

מהלך המחקר:

1. בסיס הנתונים - נבחרו שני מאגרי מידע שונים:

FakeNewsNet 1.1 – מאגר הנתונים מכיל ארבעה קבצים מסוג CSV:   
polisitfact\_real, politifact\_fake, gossipcoc\_real, gossipcop\_fake  
הקבצים מסווגים למקורות השונים שמהם נלקחו והאם ההצהרות בתוכן אמיתיות או שקריות.  
המאגר מכיל מגוון הצהרות, מתוכם אלפי הצהרות אשר נאספו מהרשתות החברתיות מתוך פוסטים ותגובות. כלל ההצהרות מסווגות לפי תוויות אמינות של אמת או שקר.  
גודל בסיס הנתונים: 23,196  
((Shu, Mahudeswaran, Wang, Lee, & Liu, 2018

PolitiFact 1.2 – מאגר נתונים זה מכיל קובץ מסוג JSON אשר מכיל אלפי הצהרות פוליטיות המדורגות לפי דירוג אמינות (כגון "נכון", "חלקית נכון", "לא נכון").   
גודל בסיס הנתונים: 21,152  
 PolitiFact (Misra, 2020)

1.3 שימוש במאגרי המידע השונים נבחרו על מנת לכסות מגוון רחב של נושאים ומקורות מידע, לספק איכות נתונים ומדדי אמינות ברורים, והסתמכות על מקורות מהימנים. השניים יחד מאפשרים ניתוח מעמיק של דיסאינפורמציה ברשתות חברתיות ומסייעים בפיתוח מודלים מדויקים ומאוזנים.

2. עיבוד מקדים וניתוח נתונים ראשוני (EDA):

2.1 קריאת נתונים – הצגת שורות ראשונות מכל קובץ על מנת להבין את מבנה בסיס הנתונים, אילו עמודות קיימות, מה הוא סוג העמודה וגודל הדאטה.

2.2 איחוד נתונים – קבצי המקור בFakeNewsNet אוחדו לדאטהסט אחיד תוך הוספת עמודת מקור ותווית בינארית. עבור מאגר PolitiFact נבחרו רק העמודות הרלוונטיות (statement, verdict), והומרו לפורמט CSV על מנת לשמור על עקביות.

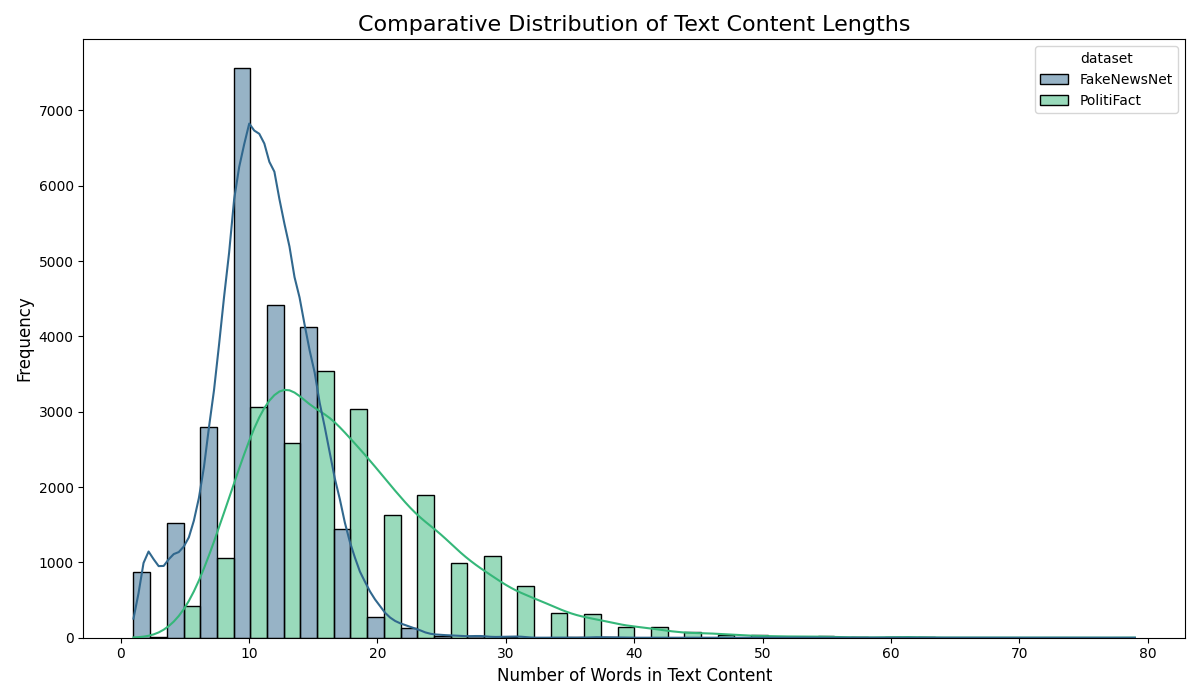
2.3 המרת תוויות - עמודת verdict בדאטהסט PolitiFact הומרה לעמודת label בינארית: תווית true קודדה כ-1 עבור הדירוגים 'true' ו-'mostly-true' ותווית false כ-0 אשר הכילה את כל שאר הדירוגים. נעשה לצורך התאמה למודל סיווג דיכוטומי ולשמירה על אחידות בין שני המקורות.

2.4 אחידות נתונים – כל הטקסט נשמר בעמודת text\_content

2.5 בדיקת שפה - בוצעה בדיקה האם הטקסטים כתובים באנגלית (באמצעות langdetect).   
 השורות הלא אנגליות דווחו אך לא הוסרו בשלב זה לצורך שקיפות.

* FakeNewsNet: 1,484 שורות (מתוך 23196)
* PolitiFact: 139 שורות (מתוך 21152)

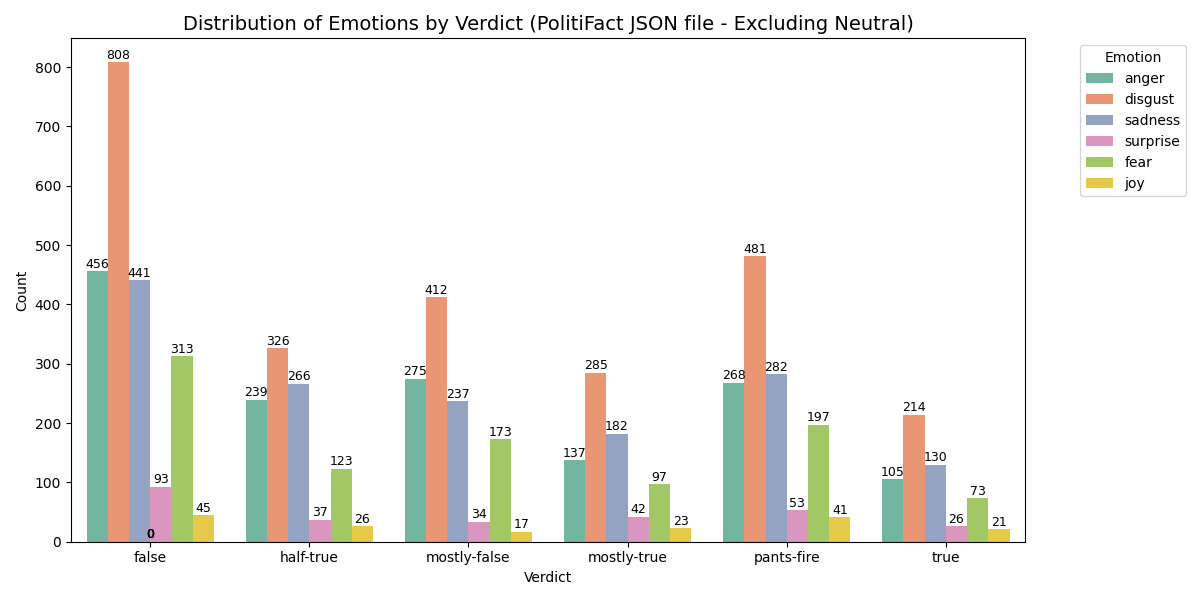
2.6 בדיקת אנומליות - חושב אורך ממוצע של טקסט, סטיית תקן, וזוהו טקסטים אנומליים בעלי אורך חריג (יותר מ-3 סטיות תקן מהממוצע).



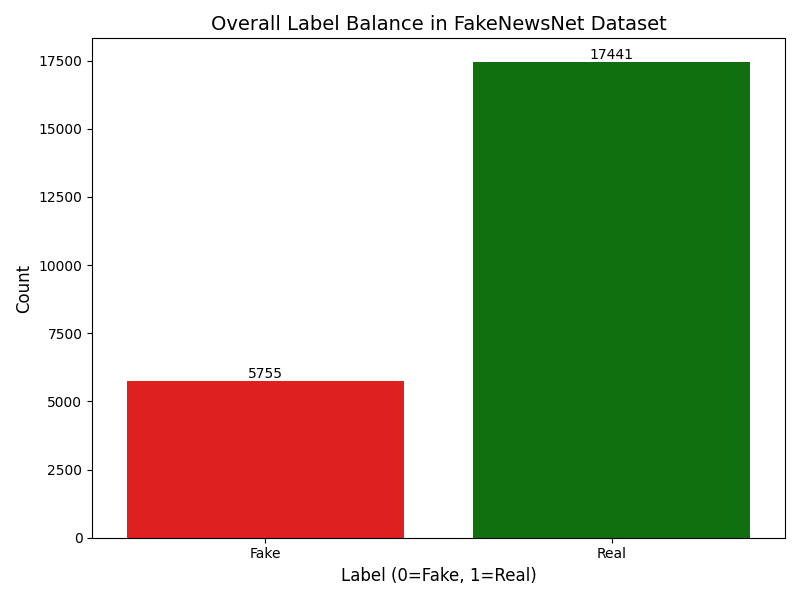
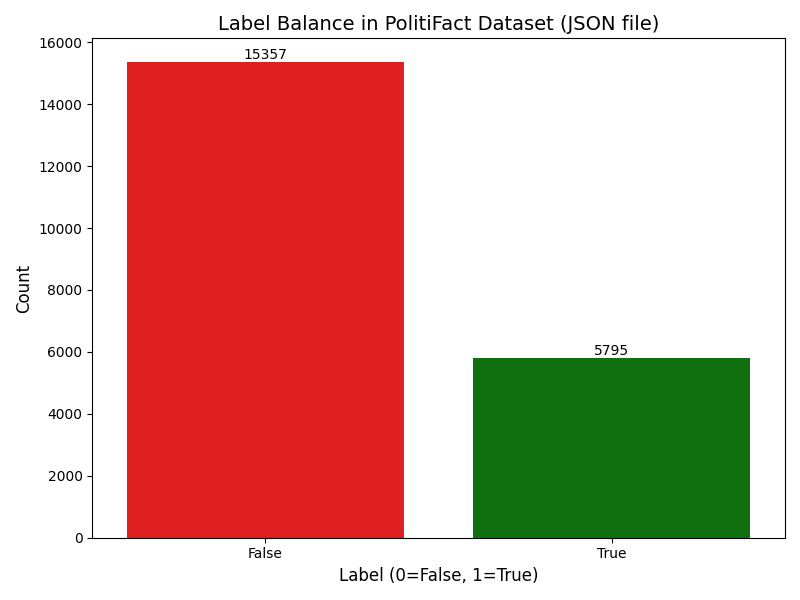
2.7 בדיקת שגיאות כתיב - עבור כל טקסט חושבה כמות המילים השגויות לפי ספריית pyspellchecker.

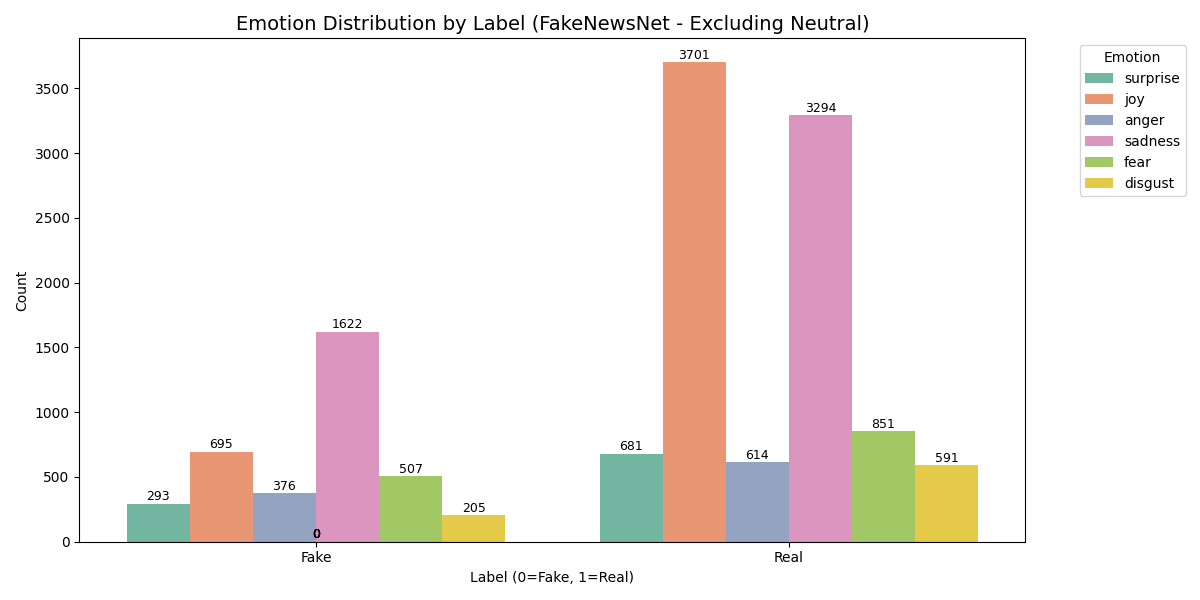
* FakeNewsNet – ממוצע שגיאות 1.89
* PolitiFact – ממוצע שגיאות 2.85

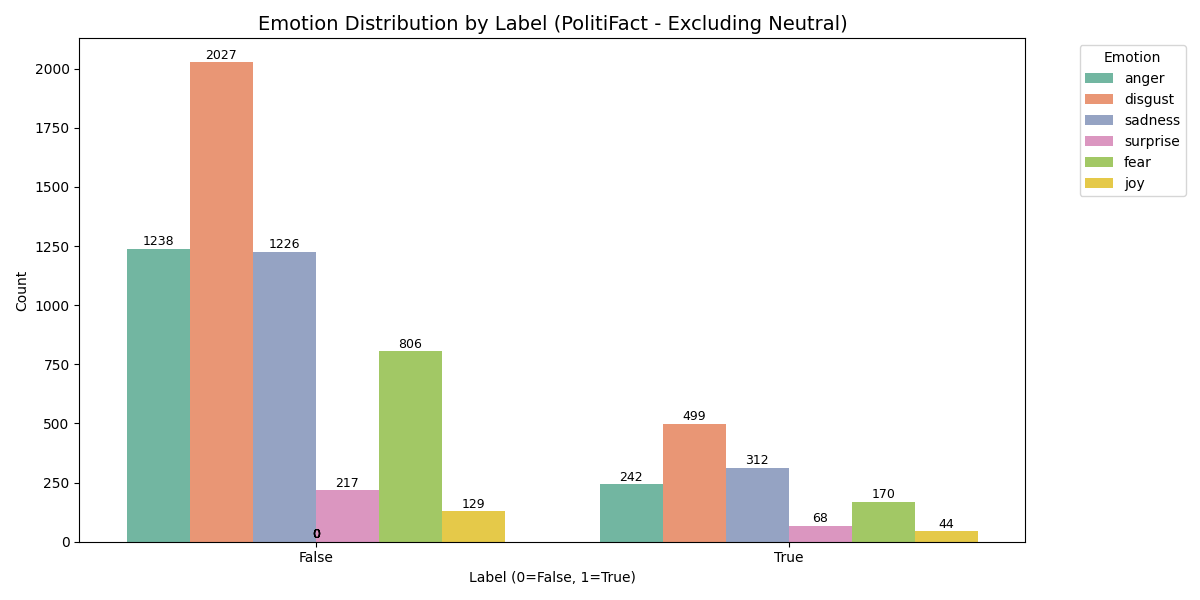
2.8 ניתוח רגשות - לכל טקסט בוצע ניתוח רגש באמצעות מודל טרנספורמר   
("j-hartmann/emotion-english-distilroberta-base"). עבור כל טקסט נשמר הרגש הדומיננטי בעמודת emotion.



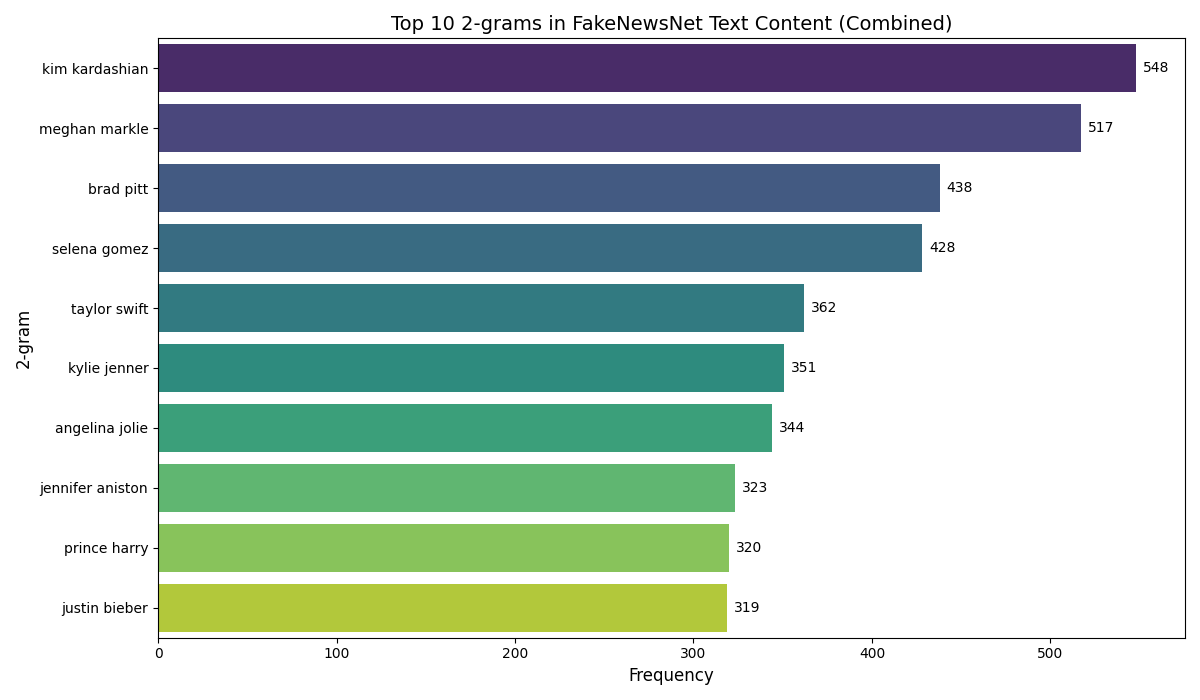
2.9 ניתוח גרפי וסטטיסטי - בוצעו ניתוחים ויזואליים וסטטיסטיים כגון:

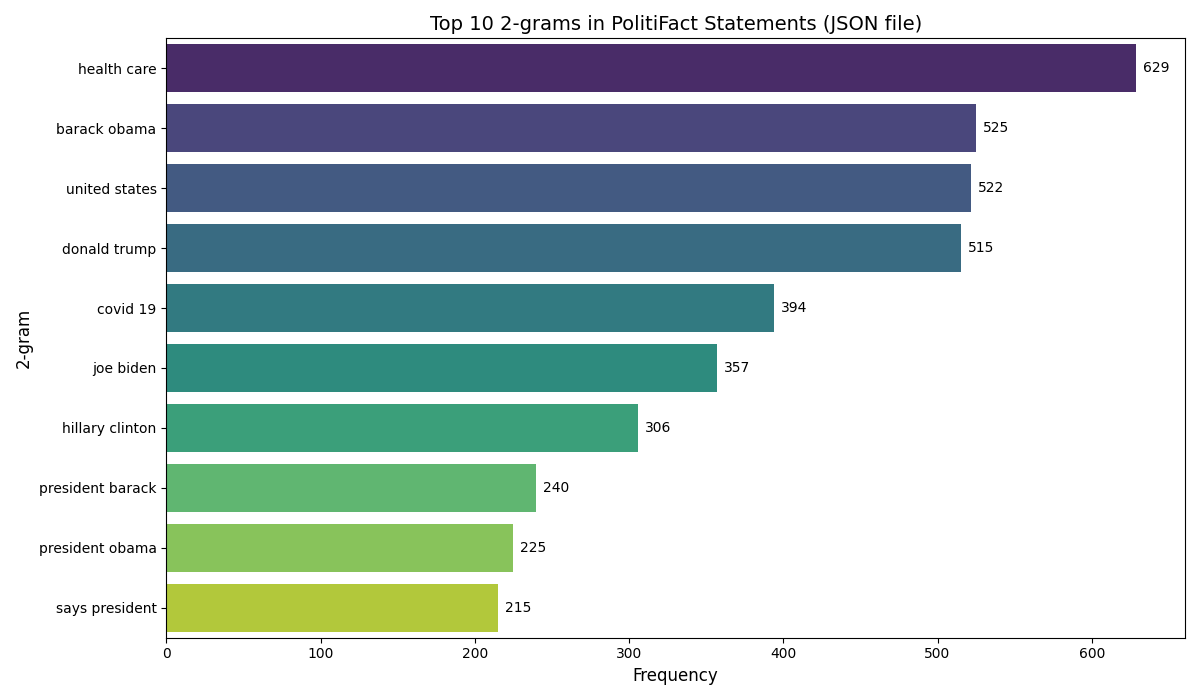
* התפלגות תוויות (Fake/Real) עבור כל מקור.
* התפלגות רגשות לפי תווית – הוצגו רגשות דומיננטיים בטקסטים אמתיים וכוזבים.



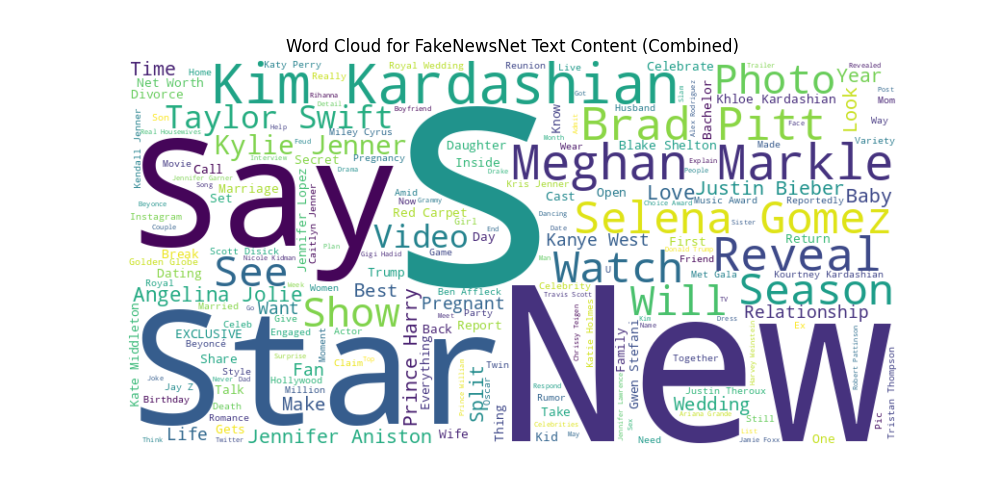


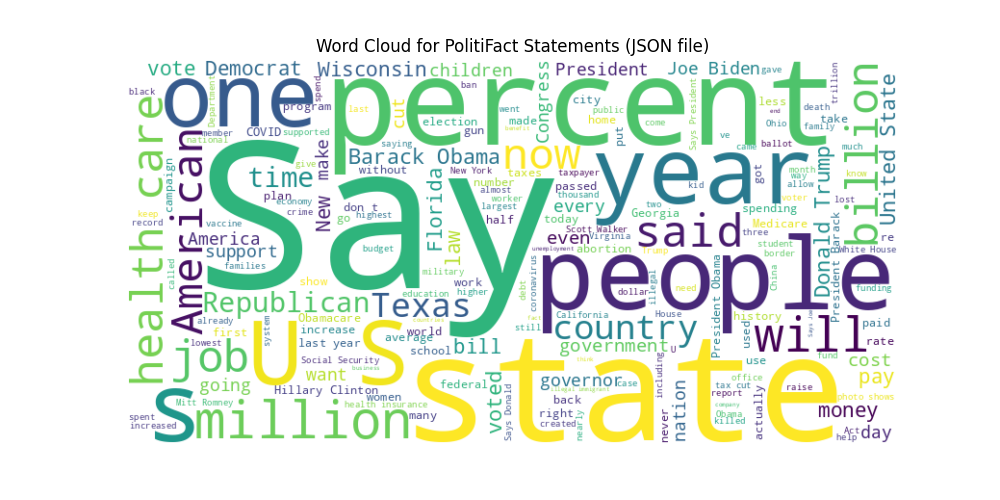
* ניתוח -grams 2 – זוהו הביטויים הדו-מילתיים הנפוצים ביותר בכל דאטהסט.





* Word Clouds – מיפוי חזותי של מילים נפוצות





2.10 שמירה וסטנדרטיזציה - שני הדאטהסטים נשמרו כקבצי CSV אחידים במבנה העמודות:  
['text\_content', 'text\_length', 'emotion', 'misspellings', 'label', 'source']  
בכך נוצר בסיס אחיד לניתוח והשוואה בין מקורות המידע.

3. הכנת הנתונים – עיבוד נתונים:

3.1 טעינת מאגרי המידע - הקבצים הנטענים הם גרסאות שעברו עיבוד מקדים במסגרת הEDA, ומכילים טקסטים ותוויות.

3.2 טיפול בערכים חסרים - הוסרו שורות המכילות ערכים חסרים בעמודות המרכזיות.

3.3 בדיקת שפה וסינון טקסטים לא באנגלית - באמצעות langdetect זוהתה שפת הטקסט, והוסרו שורות שאינן בשפה האנגלית.

3.4 ניקוי טקסט - הפיכת כל הטקסט לאותיות קטנות (lowercase), הסרת רווחים מיותרים,  
תווים מיוחדים למעט גרש וסימני פיסוק.

3.5 טוקניזציה ולממטיזציה – בוצע פיצול הטקסטים למילים (tokenization), הסרה של מילים חסרות משמעות (Stop Words) בשפה האנגלית, ולבסוף המרת המילים לצורת השורש הדקדוקית שלהן (lemmatization) באמצעות WordNetLemmatizer.

3.6 שימור טקסט מנוקה - שמירת התוצאה בעמודה חדשה בשם clean\_text אשר שימשה את המודלים בשלב הלמידה.

3.7 וידוא תוויות וסינון תוויות חריגות – בדיקה שכל התוויות בינאריות (0 או 1)

|  |  |  |
| --- | --- | --- |
| גודל הדאטה | FakeNewNet | PolitiFact |
| מקור | 23,196 | 21,152 |
| לאחר עיבוד נתונים (ללא איזון נתונים) | 21,695 | 21,014 |
| חלוקה לתוויות לפני איזון | 0 – 5,277 1 – 16,418 | 0 – 15,245 1 – 5,769 |
| חלוקה לתוויות לאחר איזון | 0 – 5,277 1 – 5,277 | 0 – 5,769 1 – 5,769 |

3.8 איזון הדאטהסט - עבור כל אחד מהמקורות, בוצע תת-דגימה (undersampling) על הקטגוריה הדומיננטית כך שכל אחת מהתוויות תהיה מיוצגת באופן שווה. לאחר מכן הדאטהסט עבר ערבוב (shuffle) להבטחת אקראיות.

3.9 חלוקה הנתונים לסטים: בוצע פיצול הנתונים לקבוצת אימון (80%) וקבוצת בדיקה (20%)

3.10 שמירת הקבצים המעובדים - שני קבצים חדשים נוצרו עבור FakeNewsNet ו PolitiFact-עם טקסט מנוקה ומאוזן, כולל תווית בינארית לצורך בניית המודלים.

אלגוריתמים ומודלים:

במחקר זה נעשה שימוש במספר אלגוריתמים מבוססי למידת מכונה לסיווג טקסטים כמידע אמין או כוזב, תוך השוואת ביצועיהם על פני ייצוגים שונים של הטקסט ובאמצעות שילובים היררכיים (Stacking) בין המודלים.

4.1 מודלים בודדים:

* Support Vector Machine (SVM):  
  אלגוריתם השואף למצוא את ההיפר-מישור המפריד האופטימלי בין שתי מחלקות, תוך שימוש בפונקציית גרעין (kernel) להפרדה לא ליניארית.

האלגוריתם יעיל במיוחד במרחבים בעלי מימד גבוה ומתמודד היטב עם בעיות של הפרדת גבולות מורכבת.

* Random Forest (RF):  
  יער של עצי החלטה הנבנים על תת-מדגמים אקראיים של הדאטה ושל מאפיינים. האלגוריתם מצביע לפי רוב קולות העצים (Bagging), ומשפר יציבות ודיוק של סיווגים.
* Logistic Regression (LR):  
  מודל ליניארי לסיווג בינארי, המחשב את ההסתברות של דוגמה להשתייך למחלקה חיובית באמצעות פונקציית סיגמואיד. נחשב למודל פשוט אך אפקטיבי להבנת תרומת כל פיצ'ר.
* Extreme Gradient Boosting (XGBoost):  
  שיטת Boosting משוכללת הבונה עצים בסדרה, כאשר כל עץ לומד את השגיאות של העץ הקודם. XGBoost כולל מנגנוני regularization ויעילות חישובית גבוהה במיוחד.

4.2 שילובי Stacking  
שילובי המודלים במסגרת רכיבי ה-Stacking נבחרו מתוך כוונה לאזן בין מסווגים ליניאריים (כדוגמת SVM ו-Logistic Regression) למסווגים מבוססי עצים (Random Forest ו-XGBoost), במטרה להפיק יתרון מהשונות במבנה ובאופן קבלת ההחלטות של כל מודל. השילוב של מודלים בעלי מנגנוני למידה שונים מאפשר למודל ההרכבה לקלוט תבניות מורכבות יותר ולשפר את ביצועי הסיווג, תוך הפחתת הטיה והגדלת הגיוון (diversity) בתוך ההרכב.

נבחנו ההרכבים הבאים:

* SVM + Random Forest : 1 Stacking
* SVM + Random Forest + Logistic Regression : 2 Stacking
* SVM + XGBoost + Logistic Regression :3 Stacking

4.3 Meta-Ensemble (Stacking דו-שכבתי)  
שכבת המטא מאפשרת גמישות בשקלול התחזיות, תוך ניצול היתרונות היחסיים של כל אחד מהמודלים הבסיסיים, ומובילה לרוב לשיפור מדדי הדיוק והיציבות של המודל הכולל. שימוש ב XGBoost כשכבת Meta נבחר בזכות ביצועיו הגבוהים ויכולתו להתמודד עם תבניות מורכבות תוך הפחתת סיכון   
ל- overfitting.  
נבחן ההרכב הבא:

SVM, Logistic Regression, Random Forest :Base Layer

XGBoost :Meta Layer

4.4 ייצוגי טקסט ומאפיינים נוספים

4.4.1 ייצוגי טקסט

כל מודל נבחן תחת שישה ייצוגים שונים של הטקסט:

* TF-IDF: ייצוג מספרי של תדירות מילים, תוך הענשת מילים שכיחות מדי
* TF-IDF + PCA: הפחתת ממדיות (Dimensionality Reduction) על ידי ניתוח רכיבים עיקריים
* Word2Vec: ייצוג מילים כוקטורים רציפים לפי הקשרן הסמנטי
* Word2Vec + PCA: יישום הפחתת ממדים על הווקטורים שהתקבלו
* FastText: ייצוג המילה לפי תתי-מחרוזות (subword units), יתרון עבור מילים חדשות או נדירות
* FastText + PCA: בוצעה הפחתת ממד לצמצום מורכבות
* ייצוגי Word2Vec וFastText אומנו בשיטת CBOW (Continuous Bag of Words), שהיא ברירת המחדל במודל gensim, כאשר כל מילה נלמדה על בסיס ההקשר הסמנטי שלה במשפטים שנלקחו מהדאטה.

**4.2.2 נרמול נתונים:**  
עבור ייצוגי Word2Vec ו FastTextבוצע נרמול הווקטורים באמצעות StandardScaler מאחר שווקטורים אלו אינם מנורמלים באופן טבעי לעומת ייצוג TF-IDF שכן הוא כבר מנורמל מראש  
 (L2 normalization) כחלק ממנגנון החישוב של TfidfVectorizer.

4.2.3 מאפיינים נוספים שנבדקו:

* רגשות (Emotions): נבדקה השפעת שילובו של מאפיין רגשי שהופק מראש באמצעות מודל Transformer ייעודי (j-hartmann/emotion-english-distilroberta-base). לכל טקסט שויכה תגית רגש דומיננטי, אשר קודדה באמצעות One-Hot Encoding והתווספה לוקטור ייצוג הטקסט   
  (TF-IDF, Word2Vec או FastText). המודלים הוערכו בשתי גרסאות – עם ובלי המאפיין לצורך בחינת תרומתו למדדי הביצוע
* שגיאות כתיב (Spelling Errors): נבדקה השפעתו של מאפיין כמותי המייצג את ממוצע שגיאות הכתיב בטקסט, שחושב באמצעות ספריית pyspellchecker. מאפיין זה הוזן למודל כערך מספרי רציף והתווסף לוקטור הקלט של ייצוג הטקסט, ללא המרת קידוד קטגורית. המודלים הורצו בשתי גרסאות – עם ובלי המאפיין – לצורך הערכת תרומתו לדיוק הסיווג.

5. שיטות סטטיסטיות לבדיקת טיב התוצאות:

5.1 מדדי הביצוע:

לצורך הערכת הביצועים של המודלים, נעשה שימוש בארבעה מדדי ביצוע נפוצים בתחום הסיווג הבינארי:

* Accuracy (דיוק) – אחוז התחזיות הנכונות מתוך כלל הדגימות.
* Precision (דיוק חיובי) - מתוך כל הדגימות שסווגו כ"חיוביות", כמה מהן באמת חיוביות.
* Recall (רגישות) - מתוך כל הדגימות שבאמת חיוביות, כמה מהן זוהו כראוי.
* F1-Score - ממוצע הרמוני בין Precision ל-Recall.

5.2 שיטות להערכת מודלים:

לצורך בחינת ביצועי המודלים שפותחו במסגרת המחקר, נעשה שימוש בשתי שיטות הערכה עיקריות:

* חלוקה ל־Train/Test:

בשלב הראשוני, כלל המודלים הן המודלים הבודדים (SVM, Random Forest, Logistic Regression, XGBoost) והן מודלי הStacking הוערכו באמצעות פיצול הנתונים  
לקבוצת אימון (80%) וקבוצת בדיקה (20%). שיטה זו אפשרה לבצע הערכה ראשונית מהירה של ביצועי כל מודל.

* K-Fold Cross-Validation

על מנת לשפר את מהימנות התוצאות ולהפחית את ההשפעה של חלוקה מקרית של הדאטה, בוצעה גם הערכה באמצעות תיקוף צולב (Cross-Validation) עם k=3 קפלים. במסגרת שיטה זו, כל מודל אומן והוערך שלוש פעמים על תתי-מדגמים שונים, והמדדים חושבו כממוצע של כלל הקפלים.

שילוב שתי שיטות ההערכה מאפשר לבחון הן את הביצועים הראשוניים של המודלים על חלוקה אחת של הנתונים, והן את יציבותם לאורך פיצולים שונים – ובכך לתרום לתוקף החיצוני של הממצאים.

5.3 השוואה בין קונפיגורציות ייצוג טקסט:

כל מודל נבחן תחת שש קונפיגורציות של ייצוג טקסט:

* TF-IDF
* TF-IDF עם הפחתת מימדים באמצעות PCA
* Word2Vec
* Word2Vec עם הפחתת מימדים באמצעות PCA
* FastText
* FastText עם הפחתת מימדים באמצעות PCA

5.4 השוואת השפעת רגשות כמאפיין נוסף:

בנוסף להערכות הבסיסיות של המודלים, נבחנה תרומתו של מאפיין רגשי (Emotion) כמשתנה קלט נוסף. לצורך כך, בוצע חיזוי רגש דומיננטי עבור כל טקסט באמצעות מודל Transformer מסוג j-hartmann/emotion-english-distilroberta-base. כל טקסט קיבל תגית רגש מתוך שבע קטגוריות, והתגית קודדה כמאפיין נוסף ונוספה לוקטור הקלט.

בשלב הראשון, כלל המודלים שנבדקו הוערכו הן בגרסה הכוללת את מאפיין הרגש והן בגרסה שאינה כוללת אותו, במטרה לבחון את התרומה הכללית של רכיב זה לשיפור הביצועים.

בשלב נוסף, בוצע ניתוח ממוקד שכלל פיצול של רגשות לקבוצות מרכזיות – רגשות חיוביים, רגשות שליליים, וכלל הרגשות ללא רגש ניטרלי – לצורך הערכת התרומה היחסית של כל קבוצת רגשות. ניתוח זה יושם על מודל נבחר אחד: מודל Stacking המשלב את האלגוריתמים SVM ו-Random Forest עם ייצוג טקסט מסוג TF-IDF, אשר נמצא כאחד הביצועיים ביותר בניסויים הקודמים.

לצורך העמקה נוספת, בוצעה בדיקה מקבילה של פיצול רגשות גם ביחס לעמודת ה־verdict המקורית במאגר הנתונים PolitiFact, שבה מופיעים דירוגים ברמת פירוט גבוהה יותר (למשל: true, half-true, false, pants-fire וכו’). מטרה זו נועדה לזהות קשרים אפשריים בין סוגי רגשות לתחושת מהימנות כפי שסווגה במקור על ידי בודקים אנושיים.

5.5 השוואה בין מודלים – ניתוח סופי והשוואתי:  
תוצאות ההרצה של כל קונפיגורציה נשמרו בפורמטים CSV ו- JSONלצורך ניתוח שיטתי והשוואתי. בנוסף, תוכננה הצגה גרפית של הביצועים לצורך זיהוי חזותי של מגמות והבדלים בין האלגוריתמים והקונפיגורציות השונות. השוואה זו נועדה לזהות את היתרונות היחסיים של כל מודל בהתאם לייצוג הטקסט, שימוש במאפיינים נוספים (כגון רגשות), וסוג האלגוריתם.

6. מרכיבי התוכנה והממשק:

6.1 שפת תוכנה: Python

6.2 סביבת עבודה: PyCharm, Google Colab

6.3 מערכת הפעלה: Windows 10

6.4 ספריות:

6.4.1 ספריות לעיבוד נתונים וטבלאות:

* pandas – קריאה, עיבוד וניהול של מבני נתונים (DataFrames)
* numpy – חישובים מתמטיים ווקטוריים

6.4.2 ספריות לניתוח ויזואלי וסטטיסטי:

* matplotlib.pyplot – יצירת גרפים ויזואליים
* seaborn – גרפים סטטיסטיים, Heatmaps ו-Boxplots
* tabulate – הצגת טבלאות בפורמט טקסטואלי מסודר

6.4.3 ספריות לעיבוד שפה טבעית (NLP):

* nltk – עיבוד טקסט, טוקניזציה, לימיטיזציה, סטופ-וורדס, ניתוח רגשות בסיסי
* nltk.tokenize.word\_tokenize
* nltk.corpus.stopwords
* nltk.corpus.words
* nltk.stem.WordNetLemmatizer
* nltk.sentiment.SentimentIntensityAnalyzer
* langdetect – זיהוי שפת הטקסטים
* re – ביטויים רגולריים לניקוי טקסט
* WordCloud – יצירת ענני מילים
* spellchecker.SpellChecker – בדיקת שגיאות כתיב
* transformers.pipeline – ניתוח רגשות באמצעות מודל Transformer

6.4.4 למידת מכונה – עיבוד וייצוג טקסטים:

* sklearn.feature\_extraction.text.TfidfVectorizer – ייצוג טקסט ב-TF-IDF
* sklearn.feature\_extraction.text.CountVectorizer – N-grams
* gensim.models.Word2Vec – ייצוג מילים באמצעות Word2Vec
* gensim.models.FastText – ייצוג מילים באמצעות FastText

6.4.5 למידת מכונה – מודלים ואלגוריתמים:

* sklearn.model\_selection.train\_test\_split – חלוקת הדאטה ל-Train/Test
* sklearn.model\_selection.StratifiedKFold – Cross Validation
* sklearn.decomposition.PCA – הפחתת ממדים
* sklearn.ensemble.RandomForestClassifier – יערות החלטה
* sklearn.linear\_model.LogisticRegression – רגרסיה לוגיסטית
* sklearn.svm.SVC – מכונת וקטורים תומכת (SVM)
* sklearn.ensemble.StackingClassifier – Stacking של מודלים
* xgboost.XGBClassifier – מודל Gradient Boosting מבוסס XGBoost
* sklearn.preprocessing.StandardScaler – סטנדרטיזציה של נתונים

6.4.6 מדדי ביצוע והערכה:

* sklearn.metrics:
  + accuracy\_score
  + precision\_score
  + recall\_score
  + f1\_score

6.4.7 אחר:

* os – ניהול נתיבי קבצים
* json – קריאה וכתיבה של קבצי JSON
* sys – פונקציות מערכתיות (טיפול בשגיאות)
* files – from google colab

# תוצאות

השערות המחקר ותוצאותיהם:

1. השערת מחקר 1:

1.1 יישום מודלים מבוססי למידת מכונה ועיבוד שפה טבעית יתרום לשיפור רמות הדיוק בזיהוי ובסינון של מידע כוזב לעומת ביצועים אקראיים או בסיסיים.

1.2 אופן הבחינה:

הוערכו ביצועי ארבעה אלגוריתמים בודדים (SVM, Random Forest, Logistic Regression, XGBoost) על שני מקורות מידע (FakeNewsNet, PolitiFact), תחת שלושה ייצוגי טקסט שונים:

TF-IDF, Word2Vec, FastText

1.3 מדדים שנמדדו:

Accuracy, Precision, Recall ו־F1-Score (ממוצעים משוקללים)

1.4 תוצאות עיקריות:

1.4.1 השוואת ייצוגי הטקסט השונים אל מול ארבעת המודלים הבודדים:

הגרף מציג את ממוצעי מדד F1 עבור ארבעה מודלים (Logistic Regression, Random Forest, SVM XGBoost) על פני שלושה ייצוגי טקסט שונים: TF-IDF, Word2Vec ו-FastText, עבור מקורות הנתונים PolitiFact וFakeNewsNet-. ניתן לראות כי עבור PolitiFact, הערכים גבוהים באופן יחסי (בסביבות 60%), בעוד שבFakeNewsNet המודלים מציגים ערכים נמוכים יותר, עם שונות גדולה יותר בין הייצוגים והמודלים.

1.4.2 תוצאות מדדי המודלים באמצעות ייצוג טקסט TF-IDF:

הגרף מציג את תוצאות ארבעת המדדים (Accuracy, Precision, Recall, F1-Score) עבור כל אחד מארבעת המודלים, בשני מקורות מידע, תוך שימוש בייצוג טקסט TF-IDF. עבור PolitiFact, כל המדדים נעים סביב טווח של 0.65–0.6 בכל המודלים. עבור FakeNewsNet, קיימת ירידה ניכרת ברוב המדדים, כאשר הערכים נעים בטווח של כ0.2–0.3, בהתאם למודל ולמדד.

2. השערת מחקר 2:

2.1 הביצועים של מודלים משולבים ,(Ensemble Learning) ובעיקר מודלים מסוג ,Stacking יהיו גבוהים באופן מובהק בהשוואה לאלגוריתמים יחידניים (Individual Classifiers) בזיהוי והבחנה בין מידע אמין למידע כוזב.

2.2 **אופן הבחינה:**  
השוואת תוצאות בין מודלים בודדים לבין מודלים משולבים ,(Stacking) תוך בחינת ארבעה הרכבים שונים:

* SVM + Random Forest
* SVM\_Random Forest\_Logistic Regression
* SVM\_XGBoost\_Logistic Regression
* SVM + Logistic Regression + Random Forest + שכבת Meta: XGBoost

2.3 תוצאות עיקריות:

2.3.1 השוואת המודלים הבודדים (Single) למודלים המשולבים (Stacking):

הגרף מציג את ממוצע F1 עבור שמונה תצורות של מודלים: ארבעה מודלים בודדים (XGBoost, SVM, Random Forest, Logistic Regression) וארבעה מודלים משולבים (Stacking) הכוללים שילובים שונים של אותם אלגוריתמים, מבוסס ייצוג טקסט TF-IDF משני מקורות המידע.

ניתן לראות כי חל שיפור ניכר בין המודלים הבודדים למודלים המשולבים בFakeNewsNet אך ללא שינוי משמעותי בPolitiFact.  
ניתן לראות בגרף ערכי המדד עבור המודלים הבודדים בFakeNewsNet נעים בין 0.17-0.32 לעומת המודלים המשולבים אשר נעים בין 0.73-0.78, לעומת זאת, בPolitiFact ערכי המדד הן במודלים הבודדים והן במודלים המשולבים נעים בטווח שבין 0.6-0.65.

3. השערת מחקר 3:

3.1 שילוב של טכניקות לייצוג טקסט יחד עם הפחתת ממדים (PCA) יניב תוצאות מדויקות יותר, תוך הפחתת רעש ושיפור כללי של ביצועי המודל.

3.2 אופן הבחינה:

בוצעה השוואה בין כל ייצוג רגיל (TF-IDF, Word2Vec, FastText) לבין גרסתו המצומצמת לאחר החלת PCA.

3.3 תוצאות עיקריות:

3.3.1 השוואת ביצועי מודלים עם וללא הפחתת ממדים (PCA) עבור ייצוגי טקסט שונים:

הגרף מציג את תוצאות ההשוואה בין ביצועי המודלים השונים שנבדקו בשני מקורות המידע (FakeNewsNet וPolitiFact), כאשר נעשה שימוש בייצוגי טקסט שונים (TF-IDF, Word2Vec, FastText), הן בגרסתם המקורית והן לאחר החלת הפחתת ממדים באמצעות PCA. לצורך ההשוואה חושבו ממוצעי Accuracy, Precision, Recall וF1-Score עבור כל קונפיגורציה. הגרף מאפשר לבחון האם החלת הפחתת ממדים תרמה לשיפור ביצועים.

3.3.2 סיכום תוצאות כלל המודלים מול כל ששת ייצוגי הטקסטים השונים

הגרף מציג את תוצאות ההשוואה על פי מדד F1-Scrore של כלל המודלים – בודדים ומודלי Stacking עבור כל אחד מהייצוגי טקסטים השונים.

4. השערת מחקר 4:

4.1 שילוב של מאפייני רגש (Emotion Features) כחלק ממערך הקלט של המודל יתרום לשיפור ביצועיו, בהשוואה למודלים המבוססים על טקסט בלבד.

4.2 אופן הבחינה:

לשם בחינת ההשערה, הוספו פיצ’רי רגשות (שהופקו מראש על ידי מודל Transformer) לכלל הקונפיגורציות שנבדקו בתור עמודת קלט נוספת למודל באמצעות One-Hot Encoding והושוו התוצאות למדדים שהושגו באותן קונפיגורציות ללא מאפיין זה. ההשוואה כללה מודלים בודדים (SVM, RF, LR, XGBoost) וכן מודלי Stacking בהרכבים שונים.

4.3 ניתוח מעמיק במודל הנבחר:

ניתוח מעמיק בוצע על המודל Stacking: SVM + Random Forest תחת ייצוג טקסט מסוג TF-IDF, בוצעה עליו בדיקה מעמיקה של השפעת הרגש באמצעות קונפיגורציות שונות של רגשות:

* רגשות לפי תגית דומיננטית אחת
* רגשות חיוביים בלבד (joy, surprise)
* רגשות שליליים בלבד (sadness, anger, disgust, fear)
* כלל הרגשות למעט ניטרלי
* כלל הרגשות כולל ניטרלי

4.4 מדדים שנמדדו: Accuracy, Precision, Recall, F1-Score.

4.5 תוצאות עיקריות:

4.5.1 השפעת מאפיין הרגשות על ביצועי המודלים הבודדים במקור המידע FakeNewsNet:

גרף 4.5.1 מציג את השפעת הוספת מאפייני רגשות על ביצועי F1-Score של המודלים הבודדים (SVM, RF, XGBoost Logistic) במקור המידע FakeNewsNet עבור ששת הקונפיגורציות השונות של ייצוג הטקסט שנבחנו. הגרף מייצג את הנתונים שהתקבלו עם (yes) או בלי (no) פיצ'ר הרגשות.

4.5.2 השפעת מאפיין הרגשות על ביצועי המודלים הבודדים במקור המידע PolitiFact:

גרף 4.5.2 מציג את השפעת הוספת מאפייני רגשות על ביצועי F1-Score של המודלים הבודדים (SVM, RF, XGBoost Logistic) במקור המידע PolitiFact עבור ששת הקונפיגורציות השונות של ייצוג הטקסט שנבחנו. הגרף מייצג את הנתונים שהתקבלו עם (yes) או בלי (no) פיצ'ר הרגשות.

4.5.3 השפעת מאפיין הרגש על מודלי Stacking :

גרף 4.5.3 מציג את ביצועי F1-Score על כלל מודלי הStacking עם וללא שילוב רגשות, תוך שימוש בייצוג טקסט TF-IDF. מהגרף עולה כי בכל המקרים שילוב רגשות אינו תרם לשינוי משמעותי במדד הביצוע, הן בFakeNewsNet והן בPolitiFact.

4.5.4 ניתוח רגשות מעמיק במודל SVM ּ + Random Forest באמצעות ייצוג טקסט TF-IDF:

4.5.4.1 השוואת רגשות לפי תגית דומיננטית אחת

גרף 1.4.5.4 מציג את ההתפלגות הממוצעת של תוויות הרגש הדומיננטיות שנחזו בטקסטים בכל מקור מידע ((Joy, Surprise, Neutral, Sadness, Anger, Disgust, Fear. ניתן לראות כי טקסטים ממקור FakeNewsNet נוטים להציג רגשות עזים יותר, עם דגש על רגש sadness, בעוד שבPolitiFact רוב הרגשות נמצאים סביב ערכים נמוכים יותר.

4.5.4.2 השוואת רגשות לפי רגשות המקובצות לפי הקונפיגורציות השונות:

גרף 4.5.4.2 משווה את הערכים הממוצעים של רגשות לפי קטגוריות: כל הרגשות, רגשות ללא ערך ניטרלי, רגשות שליליים (Sadness, Anger, Disgust, Fear) ורגשות חיוביים (Joy, Surprise). נמצא כי בFakeNewsNet ערכי הרגשות גבוהים יותר בכל הקטגוריות בהשוואה לPolitiFact.

4.5.4.3 תוצאות מודל SVM +XGBoost + LR לצורך ניתוח רגשות מעמיק

שני הגרפים ב4.5.4.3 מציגים את תוצאות המודל SVM +XGBoost + LR בניתוח רגשות מעמיק עבור כל רגש בנפרד ועבור רגשות המקובצים בקטגוריה של רגשות חיוביים, רגשות שליליים וכלל הרגשות ללא רגש ניטרלי.

5. השערה מספר 5:

5.1 שילוב של ממוצע שגיאות כתיב כפיצ'ר נוסף בייצוג הטקסט יתרום לשיפור ביצועי המודל, בהשוואה לשימוש בייצוג טקסט בלבד.

5.2 אופן הבחינה:

ההשערה נבדקה באמצעות השוואת ביצועי המודל (Stacking: SVM + Random Forest) באמצעות ייצוג TF-IDF בשלוש קונפיגורציות שונות:

* ייצוג TF-IDF בלבד
* ייצוג TF-IDF בתוספת ממוצע שגיאות כתיב
* ייצוג TF-IDF בתוספת ממוצע שגיאות כתיב ופיצ’ר רגשות

5.3 מדדים שנמדדו:

Accuracy, Precision, Recall, F1-Score

5.4 תוצאות עיקריות:

5.4.1 השפעת שגיאות כתיב ומאפייני רגש על ביצועי המודל לכל מקור

הגרף מציג את תוצאות הביצועים של מודל SVM + Random Forest על שלושת הקונפיגורציות השונות עבור כל אחד ממקורות הנתונים PolitiFact וFakeNewsNet:

* שימוש ב-TF-IDF בלבד
* TF-IDF בשילוב תכונת שגיאות כתיב
* TF-IDF בשילוב תכונת שגיאות כתיב ורגשות

בשני המקורות נתונים השונות בין הקונפיגורציות הייתה מתונה בכל ארבעת המדדים: F1, Recall, Precision, Accuracy.

# דיון

1. סקירת השערות המחקר והממצאים המרכזיים:

1.1 מטרת מחקר זה הייתה לבחון את תרומתם של יישומי למידת מכונה וטכניקות עיבוד שפה טבעית (NLP) לזיהוי וסינון מידע כוזב ברשתות החברתיות, תוך ניתוח של מספר פרמטרים: סוג האלגוריתם, ייצוג הטקסט, הפחתת ממדים, שילוב רגשות ושגיאות כתיב. חמש השערות הוצגו ונבחנו על פני שני מאגרי נתונים: FakeNewsNet ו-PolitiFact.

1.2. השערה 1 – אימות חלקי  
במאגר הנתונים PolitiFact אכן התקבלו תוצאות משמעותיות של המודלים הבודדים (מדדי F1 סביב 60%), המעידות על יכולת סיווג העולה בבירור על קו בסיס אקראי.  
עם זאת, במאגר FakeNewsNet, מרבית המודלים הבודדים הציגו מדדים נמוכים למדי (מתחת ל־50%), דבר שמעיד על קושי לסווג את הטקסטים בדאטהסט זה.  
הבדל זה ניתן להסבר על רקע מאפייני הדאטה: בFakeNewsNet הטקסטים לרוב קצרים, רגשיים, ולעיתים לא תקניים לשונית – מה שמייצר קושי למודלים פשוטים להבחין בין מידע אמין לכוזב, שכן גם טקסטים אמיתיים מכילים רמות רגש גבוהות וסגנון דרמטי. לעומת זאת, ב־PolitiFact מופיעים ציטוטים מנומקים של פוליטיקאים, בשפה תקנית ופורמלית, שבה קיימת הבחנה ברורה יותר בין ניסוח אמין לניסוח מטעה. לכן, המודלים מצליחים טוב יותר בזיהוי דיסאינפורמציה כאשר הסגנון הלשוני תורם להבחנה סגנונית בין טקסטים אמיתיים לכוזבים.  
ממצא זה מדגיש את החשיבות של איכות וסגנון הטקסט בזיהוי דיסאינפורמציה, ואת מגבלותיהם של מודלים פשוטים כאשר אין הבחנה לשונית ברורה בין מחלקות.

1.3. השערה 2 – אימות חלקי  
ההשערה בדבר השפעה של סוג האלגוריתם על הביצועים אומתה באופן חלקי. אכן ניכר הבדל בין מודלים בודדים לבין מודלי Stacking, בעיקר עבור מאגר FakeNewsNet, שבו המודלים המשולבים הציגו שיפור משמעותי במדדי F1 (טווח של 0.73–0.78) לעומת המודלים הבודדים (0.17–0.32).  
עם זאת, במאגר PolitiFact ההבדלים היו מצומצמים בהרבה: גם המודלים הבודדים וגם המשולבים הציגו ערכים בטווחים קרובים (0.65–0.6), ללא שיפור מובהק. ייתכן כי אופיו של הדאטה הכולל ניסוחים פורמליים ועקביים אפשר למודלים הבודדים להפיק תוצאות טובות יחסית גם ללא שילוב מודלים.  
ממצא זה מדגיש את התרומה של מודלי Stacking בעיקר כאשר מדובר בטקסטים פחות מובנים או בעלי סגנון מגוון, כפי שקיים בFakeNewsNet, ומחזק את החשיבות של התאמת השיטה לאופי הנתונים.

1.4. השערה 3 – אימות חלקי  
לגבי תרומת הפחתת ממדים (PCA), התוצאות הצביעו על שיפור נקודתי בלבד. אמנם בחלק מהקונפיגורציות נרשם שיפור קל במדדים, אך לא היה מדובר בעלייה עקבית או מובהקת בכלל הדגימות. היעדר שיפור לאחר הפחתת ממדים מצביע על כך שהייצוגים המקוריים הכילו את רוב המידע המבחין הנדרש, וצמצום הממדים לא תרם לביצועים.

1.5. השערה 4 – אימות חלקי  
שילוב מאפייני רגשות במערך הקלט הציג השפעה שונה בין המודלים והמקורות. במודלים הבודדים עבור FakeNewsNet נצפתה עלייה מובהקת במדדי F1 לאחר שילוב הרגשות. לעומת זאת, עבור מקור הנתונים PolitiFact, השיפור היה חלקי ולעיתים אף זניח. במודלי Stacking, הוספת מאפייני הרגש לא תרמה לשיפור נוסף, ובחלק מהשילובים אף נרשמה ירידה קלה במדדי F1. ייתכן כי במודלים מורכבים אלו נוצרת רוויה או חפיפה מידעית בין מאפייני הרגש לבין ייצוגי הטקסט, ולכן לא מתקבל ערך מוסף. ממצא זה מדגיש כי תרומתם של מאפייני רגש תלויה לא רק ברמת האקספרסיביות של הטקסטים, אלא גם במורכבות הארכיטקטורה של המודל.

1.5.2 בחירת מודל SVM + Random Forest עם ייצוג טקסט TF-IDF לניתוח רגשות עמוק:   
המודל המשולב שהציג את התוצאות הגבוהות ביותר בכל המחקר היה Stacking: SVM + XGBoost + Logistic Regression, תוך שימוש בייצוג טקסט מסוג Word2Vec, עם ערך F1 מרשים של 0.846 ללא שילוב רגשות.  
ניסיון להוסיף ניתוח רגשות מעמיק למודל זה הוביל לירידה חדה בביצועים, עם ערכי F1 שנעו בין 0.6 ל0.67. הסיבה לכך היא ככל הנראה ירידה משמעותית בגודל הדאטה מאחר שחלוקת הדוגמאות לפי קטגוריות רגש יצרה תתי מדגמים קטנים מאוד.  
בנוסף, קיימת סבירות ששיטת ייצוג הטקסט Word2Vec דורשת הקשר רחב, ולכן פחות מתאימה לעבודה עם קבוצות טקסט קטנות, דבר שעשוי לפגוע בביצועי המודל.  
לפיכך, הוחלט לבסס את ניתוח הרגשות המעמיק על מודל פשוט יותר: Stacking: SVM + Random Forest תוך שימוש בTF-IDF, המתאים יותר לטקסטים קצרים, מספק ייצוג מנורמל כברירת מחדל, והציג את הביצועים הגבוהים ביותר במסגרת מודל זה.

1.5.3 שילוב רגשות כמאפיין קלט במודל הסיווג הביא לשיפורים מתונים במדדי הביצוע, בעיקר במאגר FakeNewsNet. מהניתוח עולה כי רוב הרגשות תרמו באופן דומה, אך רגש Sadness הציג את הביצועים הגבוהים ביותר (F1 ≈ 0.71), והיה היחיד שהציג פער מובהק בהשוואה לאחרים.  
לעומתו, רגשות חיוביים (Joy, Surprise) הציגו את המדדים הנמוכים ביותר בשני המאגרים, מה שמעיד על תרומה פחותה של רגשות אלו בזיהוי אמינות הטקסט.  
במאגר PolitiFact, לא נצפה שינוי משמעותי בין הרגשות השונים, וההשפעה הכוללת של מאפייני הרגש הייתה מוגבלת.  
השפעת הרגש אינה אחידה בין המאגרים, היא תלויה במבנה ובאופי של הטקסטים: רגשות תורמים יותר כשיש להם ביטוי ברור בטקסט כמו במאגר FakeNewsNet שם התוכן רגשני ובעל אופי חברתי יותר מאשר בPolitiFact המתמקד בהצהרות פוליטיות רשמיות. כאשר רגשות שליליים, ובעיקר עצב, בולטים יותר בתרומתם.

1.6. השערה 5 – לא אוששה

הוספת תכונת שגיאות כתיב לא תרמה לשיפור מובהק בביצועי המודל. הפערים בין הקונפיגורציות עם וללא תכונה זו היו מתונים ולא קונסיסטנטיים. ייתכן כי מאפיין זה פחות אפקטיבי בזיהוי דיסאינפורמציה, או שמא נדרש מנגנון חכם יותר לזיהוי שגיאות תחביריות עמוקות מאשר בדיקת איות שטחית.

2. תרומת המחקר

2.1 תרומה תיאורטית  
מבחינה תיאורטית, המחקר תורם להבנה עמוקה יותר של האופן שבו מאפייני טקסט, רגשות ושגיאות כתיב משפיעים על ביצוע מודלים בתחום זיהוי דיסאינפורמציה. נמצא כי מודלי Stacking עשויים להציע יתרון בביצועים, בעיקר בדאטהסטים המאופיינים בטקסטים מגוונים או פחות מובנים, אך השיפור אינו עקבי בכל המקרים. כמו כן, המחקר מצביע על הבדלים מובהקים בין מקורות מידע שונים, מה שמחדד את הצורך בהתאמת מודל לסוג הדאטה.

2.2 תרומה יישומית  
המחקר מספק בסיס יישומי להטמעה של מערכות סינון מתקדמות בפלטפורמות מידע. שילוב של מודלים משולבים עם טכניקות ייצוג טקסט כמו TF-IDF, ולעיתים גם רגשות או PCA, עשוי לשפר את ביצועי סיווג הדיסאינפורמציה - במיוחד כאשר מדובר ברשתות עם טקסטים אמוציונליים או לא פורמליים, כפי שנצפה במאגר FakeNewsNet.

3. מגבלות מתודולוגיות

3.1 שימוש בשני מאגרים בלבד:  
למרות היתרון שבשונות בין FakeNewsNet ו-PolitiFact, השימוש רק בשני מקורות עלול להגביל את היכולת להכליל את הממצאים למקורות אחרים של דיסאינפורמציה.

3.2 איזון מלאכותי של תוויות :  
בוצעה תת-דגימה ליצירת איזון מלאכותי בין אמת/שקר, מה שעלול לפגוע בייצוגיות של המציאות שבה דיסאינפורמציה מופיעה לרוב בשכיחות נמוכה יותר או גבוהה יותר.

3.3 המרת תוויות במאגר PolitiFact לפורמט בינארי:  
איחוד קטגוריות רבות תחת תווית "שקר" עלול לטשטש הבדלים סמנטיים חשובים בין רמות שונות של חוסר אמינות (כגון “half-true” לעומת “pants on fire”).

3.4 שימוש בתג הרגש הדומיננטי בלבד:  
שמירת רגש אחד (הדומיננטי) בלבד מתוך שבע קטגוריות עלולה לאבד מידע רגשי עשיר שקיים בטקסט.

3.5 ניתוח רגשות אוטומטי מבוסס Transformer:  
תיוג רגשי אוטומטי עשוי להיות מושפע משגיאות או הטיות של המודל, במיוחד בטקסטים קצרים או עמומים.

3.6 גודל נמוך של K בK-Fold:  
נעשה שימוש ב3 קפלים (k=3) בK-Fold Cross Validation מספר זה נחשב נמוך יחסית, ועלול לפגוע במהימנות ההערכה.

4. השלכות המחקר והצעות למחקרי המשך

4.1 שימוש ברגשות מותאמים הקשר - ניתוח רגשות לפי סגנון, כוונה או מסר ולא כקטגוריות גנריות עשוי לחשוף הבדלים בין רגש טבעי לרגש מניפולטיבי, במיוחד ברשתות חברתיות.

4.2 העמקה בטקסטים ארוכים לעומת קצרים - מומלץ לבחון האם אפקטיביות האלגוריתמים משתנה בהתאם לאורך הטקסט, שכן ייתכן שקיים הבדל בין טקסטים מורכבים (כתבות) לבין קצרים (פוסטים/כותרות).

4.3 בחינת רשתות עצביות מתקדמות - שילוב של מודלים מתקדמים כמו BERT, RoBERTa או GPT עשוי לשפר את הביצועים באופן משמעותי, לאור יכולתם להבין הקשרים סמנטיים עמוקים ולהתמודד עם שפה טבעית מורכבת.

4.4 פיתוח מדדים מתקדמים לשגיאות ניסוח - מעבר למדידת שגיאות כתיב פשוטות, ניתן לבחון כלים מבוססי Language Models המזהים ניסוחים לא טבעיים, חוסר עקביות תחבירית, או שימוש יתר בביטויים דרמטיים.

4.5 שילוב ניתוח תחבירי עמוק (Syntactic Analysis) - מוצע לשלב במודלים תכונות תחביריות כמו תפקידי מילים (POS tagging), יחסי תלות תחביריים (Dependency Parsing) או מבני עץ תחבירי, כדי לבדוק האם מבנה המשפטים שונה באופן מובהק בין טקסטים אמינים לכוזבים. לדוגמה, ייתכן וטקסטים מזויפים כוללים פחות פסוקים מורכבים או דפוסי ניסוח תחביריים חריגים, דבר שעשוי לשמש כמדד נוסף לשיפור הסיווג.

4.6 שימוש בדאטה בזמן אמת - לצורך בדיקת המודלים במצבים "חיים", כפי שנדרשים ברשתות חברתיות דינמיות, מומלץ ליישם את המודלים על זרם מידע חדש (Streaming) ולבחון את עמידותם לאורך זמן.

4.7 הרחבת המחקר לשפות נוספות (כגון עברית) - מאחר שרוב המחקר בתחום זיהוי דיסאינפורמציה נעשה באנגלית, מומלץ לבחון את ביצועי המודלים גם בשפות אחרות, ובעיקר שפות פחות נחקרות כמו עברית. הרחבה זו תאפשר לבדוק את יכולת ההכללה של המודלים, לחשוף אתגרי שפה ייחודיים (כגון נטייה לצורת פועל, חוסר ניקוד, שימוש בארמית או סלנג), ולבחון האם קיימים הבדלים בדפוסי שפה בין מידע אמין למידע כוזב בשפות שונות. יישום זה ידרוש התאמה של כלים קיימים (tokenizers, וקטוריזציה, ניתוח רגשות) לשפה העברית או שימוש במודלים רב לשוניים כגון mBERT.

4.8 תכונות מבוססות התנהגות משתמשים - להרחיב את המחקר כך שיכלול מידע הקשרי או מבני (כגון דפוסי הפצה, תזמון פרסום, ומאפייני משתמש), ולא להסתמך רק על מאפיינים לשוניים.

# סיכום ומסקנות

1. סיכום תהליך העבודה

המחקר הנוכחי נבנה בשלבים מדורגים, החל מבחירת הנושא, דרך סקירת ספרות מקיפה ועד לבניית מודלים, ניתוח תוצאות והסקת מסקנות. בשלב הראשוני בוצעה סקירה תיאורטית ומעמיקה של תחום זיהוי דיסאינפורמציה, הוגדרו מטרות המחקר והושלמה בניית מסגרת תיאורטית ברורה שכללה את שאלת המחקר, ההשערות והמתודולוגיה המוצעת. לאחר מכן נבחרו שני מאגרי מידע מהימנים FakeNewsNet ו-PolitiFact לצורך ניתוח השוואתי בין מקורות מידע שונים, תוך שמירה על גיוון במבנה התוכן והמקורות.

בהמשך, בוצע תהליך של עיבוד מקדים (EDA), שכלל ניקוי טקסטים, ניתוח רגשות, בדיקת שגיאות כתיב, סינון שפה, ולבסוף איזון של התוויות. לאחר עיבוד הנתונים, יושמו שיטות שונות לייצוג טקסט, כולל TF-IDF, Word2Vec ו-FastText כל אחת גם בגרסה מופחתת ממדים באמצעות PCA. על בסיס ייצוגים אלו נבנו והוערכו מודלים שונים הן מודלים בודדים (SVM, Random Forest, Logistic Regression, XGBoost) והן מודלים משולבים (Stacking), תוך שימוש במדדי דיוק, רגישות, Precision וF1-Score.

2. מסקנות והישגים

2.1 שימוש באלגוריתמים קלאסיים תרם לביצועים טובים יותר מBaseline פשוט, אך במודלים הבודדים בלבד (SVM, XGBoost, Logistic Regression) – הביצועים בFakeNewsNet היו מוגבלים ונמוכים מ50% במדדי F1, ככל הנראה בשל מורכבות גבוהה של הטקסטים או חוסר יציבות בייצוגים.

2.2 מודלי Stacking שיפרו משמעותית את הביצועים בFakeNewsNet בשל אופיו הרועש והבלתי עקבי של הטקסטים, בו מודל יחיד מתקשה ללמוד תבניות. השילוב בין מודלים מאפשר תפיסה רחבה ומאוזנת של הדאטה. לעומת זאת, בPolitiFact הטקסטים מסודרים וברורים יותר, ולכן היתרון של מודלים משולבים מתמתן.

2.3 מאפייני רגשות שיפרו מודלים בודדים בלבד ב FakeNewsNetייתכן כי בשל טקסטים עשירים רגשית, הוספת רגשות תרמה משמעותית למודלים פשוטים. אך במודלי Stacking לא נרשמה תרומה מובהקת, ולעיתים אף נרשמה ירידה קלה.   
ככל שהמודל מורכב ועשיר יותר כך פוחת הערך של הוספת מאפיינים נוספים, במיוחד כאשר קיימת חפיפה מידעית בין פיצ'רים (כמו רגשות) לוקטור הקלט שכבר כולל סמנטיקה ותחביר.

2.4 מאפייני שגיאות כתיב לא שיפרו את המודלים, כנראה בשל מגבלות במדד הפשטני (ממוצע שגיאות פר טקסט) שאינו מייצג תחביר או רמת ניסוח באופן אמין.

2.5 המודל הבולט ביותר במחקר היה Stacking המשלב SVM + XGBoost + Logistic Regression עם ייצוג Word2Vec, שהשיג את התוצאה הגבוהה ביותר במדד F1 (0.84) – ללא שימוש ברגשות. עם זאת, כאשר נעשה ניסיון לשלב רגשות בניתוח מעמיק על מודל זה, חלה ירידה משמעותית בביצועים, ככל הנראה עקב ירידה בכמות הנתונים לכל תת-רגש, יחד עם מגבלות של ייצוג Word2Vec בהקשרים מצומצמים.

3. מגבלות בתהליך העבודה

* מגבלות מחשוב: עקב העדר כרטיס גרפי ייעודי, מודלים מורכבים שכללו גם ניתוח רגשות היו בעלי זמני ריצה ארוכים בהרצה בסביבת Pycharm. לכן הוחלט לעבור להרצה חלקית בסביבת Google Colab.
* פערים מבניים בין הדאטהסטים: קיים הבדל מובהק באופי הטקסטים, אורך המשפטים ורמות הרגש בין FakeNewsNet לPolitiFact, מה שהקשה על אחידות בביצועי המודלים ודרש ניתוחים נפרדים לכל מקור.

4. כלים שנרכשו במהלך העבודה:

* חשיבות בניית תשתית תיאורטית מבוססת לפני כניסה לשלבים טכניים
* ההבנה כי עיבוד מקדים איכותי תורם לא פחות מהאלגוריתם עצמו
* זיהוי תרומתם היחסית של אלגוריתמים שונים, כולל הייחודיות של כל מודל
* בחינה שיטתית של כל רכיב במודל (ייצוג טקסט, רגשות, הפחתת ממדים וכו') מאפשרת הבנה מעמיקה של תהליך הסיווג והשלכותיו

5. פערים והצעות לשיפור במחקר:

5.1 **אי הצלחת מאפיין שגיאות כתיב לתרום לביצועים**שילוב מדדים מתקדמים לזיהוי איכות ניסוח מעבר לזיהוי שגיאות כתיב בסיסיות, ניתן היה לשלב כלי שפה מתקדמים (כגון Language Models) שמזהים שפה לא טבעית, תחביר לקוי, או סגנון ניסוח המאפיין דיסאינפורמציה.

5.2 אי הכללת ניתוח טקסטים מרובי רגשות בו-זמנית  
תווית הרגש נבחרה כרכיב בודד לכל טקסט (או לפי קבוצות רגשות), אך לא בוצע ניתוח של מקרים בהם קיימים מספר רגשות יחד, או של עוצמת רגשות (intensity).

5.3 אי השפעת מאפיין רגשות כתורם מובהק לביצועים  
בדיקת השפעת ניתוח תחבירי עמוק – שילוב תכונות תחביריות כמו תפקידי מילים (POS tagging), יחסי תלות תחביריים (Dependency Parsing) או מבנים תחביריים מורכבים (Constituency Trees), עשוי לחשוף דפוסים לשוניים ייחודיים למידע כוזב, כגון שימוש יתר במבנים פשוטים, סדר תחבירי חריג או תבניות ניסוח דרמטיות.

5.4 בסיס נתונים מוגבל  
המחקר הסתמך על FakeNewsNet וPolitiFact בלבד. שילוב מקורות נוספים (כמו LIAR, ISOT או Reddit) היה עשוי להרחיב את תוקף המסקנות.

5.5 מקור המידע PolitiFact ללא שינויים מובהקים במחקר  
כיוון ש-PolitiFact מכיל טקסטים פחות רגשיים, ייתכן שמודלים רגישים לרגש אינם אפקטיביים באותו האופן. לכן, יש לשקול התאמת תהליך החילוץ או בחירה בפיצ'רים חלופיים שיתאימו לסוגי טקסט פחות אקספרסיביים.

# ביבליוגרפיה

Alghamdi, J., Lin, Y., & Luo, S. (2023). A comprehensive survey on machine learning approaches for fake news detection.

Alghamdi, J., Lin, Y., & Luo, S. (2024). Unveiling the hidden patterns: A novel semantic deep learning approach to fake news detection on social media.

Galli, A., Masciari, E., Moscato, V., & Sprli, G. (2022). A comprehensive Benchmark for fake news detection.

Jain, D. K., Neelakandan, S., Vidyarthi, A., Mishra, A., & Alkhayyat, A. (2025). A knowledge Aware NLP-Driven conversational model to detect deceptive contents on social media posts.

Maathuis, C., & Godschalk, R. (2023). Social Media Manipulation Deep Learning based Disinformation Detection.

**(Misra, R.** (2020). PolitiFact Fact-Check Dataset [Data set]. Kaggle.) https://www.kaggle.com/datasets/rmisra/politifact-fact-check-dataset

Mridha, M. F., Keya, A. J., Hamid, M. A., Monowar, M. M., & Rahman, M. S. (2021). A comprehensive review on fake news detection with deep learning.

Muhammed, T' S', & Mathew, S' K'. (2022). The disaster of misinformation: a review of research in social media. International journal of data science and analytics.

Oshikawa, R., Qian, J., & Wang, W. Y. (2020). A survey on natural language processing for fake news detection.

**Shu, K., Mahudeswaran, D., Wang, S., Lee, D., & Liu, H. (2018).** FakeNewsNet: A data repository with news content, social context and dynamic information for studying fake news on social media [Preprint]. arXiv. HYPERLINK "https://arxiv.org/abs/1809.01286" \t "\_new" https://arxiv.org/abs/1809.01286

Soni, H. K., Sharma, S., & Sinha, G. R. (2025). Text and Social Media Analytics for Fake News and Hate Speech Detection.‏.

Wang, J., Wang, X., & Yu, A. (2025). Tackling misinformation in mobile social networks a BERT-LSTM approach for enhancing digital literacy.

# פוסטר הפרויקט

