**Multilingual Clickbait Detection– דוח סיכום**

**תקציר**

כותרות clickbait נפוצות כיום באינטרנט ומיועדות למשוך את תשומת לב המשתמשים ולעודד אותם ללחוץ על קישורים. גם כאשר אינם מעוניינים בכך. מטרתו של מחקר זה היא לפתח מערכת לזיהוי clickbait במספר שפות באמצעות מודלים מבוססי למידת מכונה, תוך התמקדות ביכולות רב-שפתיות multilingual ובין-שפתיות cross-lingual. לשם כך, נאספו מאגרי נתונים בשפות מגוונות (כגון עברית, אנגלית, ערבית, סינית, בנגלית ועוד), וכל אחד מהם תויג לפי תוכן clickbait או שאינו clickbait. בשלב הראשון, אומנו מודלים ייעודיים מבוססי BERT לכל שפה בנפרד כדי להשיג תוצאות השוואתיות. לאחר מכן, נבחנו תרחישים של זיהוי cross-lingual, בהם אומנו המודלים על שפה אחת ונבחנו על שפה אחרת, וכן תרחישים של זיהוי multilingual בהם אומנו המודלים על שילוב של מספר שפות ונבחנו על שפות שנמצאות באימון.

התוצאות מראות כי מודלי multilingual השיגו דיוק גבוה, במיוחד בשימוש ב-combinations של מספר שפות. מחקר זה מדגים את היתרונות בשימוש במודלי multilingual לזיהוי clickbait ומצביע על כיוונים לשיפור ביצועים במחקרי המשך.

התקציר מתמקד בעיקר במטרת המחקר, בשיטות העבודה, ובממצאים המרכזיים, ומספק מבט כולל על המחקר.

**הקדמה**

כותרות *clickbait* הפכו לחלק בלתי נפרד מנוף האינטרנט, ונעשה בהן שימוש נרחב במטרה למשוך את תשומת הלב של המשתמשים ולעודד אותם ללחוץ על קישורים. Clickbait מתייחס לכותרת המפתה את הקוראים לחפש מידע נוסף באמצעות קישור על ידי מניפולציה, הגזמה או חוסר יושר לגבי תוכן המאמר, בעוד שכותרת ללא *clickbait* מספקת מידע מפתח פשוט.[[1]](#endnote-1)

מחקרים בתחום זיהוי כותרות *clickbait* מתמקדים לרוב באנגלית, אך קיימים מאמצים מתקדמים לפתח מודלים רב-שפתיים שיכולים לזהות *clickbait* בשפות שונות ולבצע זיהוי בין-שפות. מחקר זה מתמקד בזיהוי כותרות *clickbait* במספר שפות, תוך שימוש במודלים של BERT המותאמים לשפות יעד שונות.

מחקרים קודמים חקרו מודלים של machine learning ו deep learningלאיתור clickbait. אומנו עד כה בכמה שפות. בשפה האנגלית במאמר "Clickbait Detection on Online News Headlines Using Naive Bayes and LSTM" מאת פביאן אריה ננדנה אדריאן וחבריו[[2]](#endnote-2) שבודק מודלי למידת מכונה לאיתור מודעות קליקבייט בשפה האנגלית רמת הדיוק שהתקבלה 0.98. מאגר הנתונים שלהם כלל [[3]](#endnote-3)32000 כותרות שונות מתויגות כclickbait או nonclickbait.

בשפה העברית בוצע מחקר ע"י טליה נתניה וחיה ליבסקינד עם מאגר נתונים הכולל 1315 כותרות מתויגות.

בשפה הטורקית בוצע מחקר לאיתור מודעות *clickbait*. מאגר הנתונים בשפה הטורקית כולל 20038 כותרות מתויגות[[4]](#endnote-4) והדיוק הוא 0.87 .

בשפה הגרמנית נמצא מאגר נתונים הכולל 69397 כותרות מתויגות[[5]](#endnote-5) (לחפש את המאמר גם הקודם)

השפה הרומנית גם נחקרה בהקשר של זיהוי *clickbait* במאמרם של ברוסקוטניו ד מ וחבריו [[6]](#endnote-6) בה בוצעו אימונים במודלי למידת מכונה, SVM , RandomForest וכן במודל למידה עמוקה RoBert, תוצאת הprecision המיטבי שהתקבל במחקר הוא 0.94 עבור מודל RoBert עם fine tunning. במאמר השתמשו במאגר הנתונים המכיל 8318 נתונים מתויגים.[[7]](#endnote-7)

בשפה הבנגלית התפרסם מאמרם של מהטב מ וחבריו[[8]](#endnote-8) המודלים השונים שנוסו הם LSTM, GRU ו - CNN. כאשר הדיוק המיטבי שהתקבל הוא 0.75, מאגר הנתונים כולל 15,406 נתונים מתויגים[[9]](#endnote-9).

בשפה האינדונזית המאמר שפורסם ע"י מוחמד נור פחרוזמאן וחבריו[[10]](#endnote-10) הביא למדד ROC-AUC של 0.89 אחוז. כאשר בוצע בו שימוש במודל למידה עמוקה IndoBERT. מאגר הנתונים עליו אומן המודל כולל 15000 כותרות מתויגות[[11]](#endnote-11).

בשפה הערבית בוצע מחקר לאיתור קליקבייט. במאמר שפורסם ע"י אלארבי ב. וחבריו[[12]](#endnote-12) המאגר שהשתמשו במחקר מכיל 15473[[13]](#endnote-13) רשומות מתויגות. עם מדד f1-score של 0.85.

בשפה הסינית במאמרם של ין וו וחבריו[[14]](#endnote-14) השתמשו במודל BERT לאימון מודל למידת מכונה כאשר הגיעו למדד recall מקסימלי 0.9. מאגר הנתונים במחקר כולל 16657 רשומות מתויגות. [[15]](#endnote-15)

המחקרים הקודמים, התמקדו בעיקר בפיתוח מודלי multilingual או cross-lingual אך בהיקף מצומצם. מחקר זה נבדל בכך שהוא בוחן את היכולת של מודלים לבצע זיהוי  *clickbait* בין-שפות, כלומר, לאמן את המודל על שפה אחת ולבדוק את ביצועיו בשפה אחרת. גישה זו מבוססת על אימון בעזרת מודלי multilingual כמו BERT במטרה לגלות שיפור משמעותי בדיוק הזיהוי גם בשפות שאינן בהכרח כלולות באימון המקורי.

**שיטות**

אימון לכל שפה לחוד עם מודל BERT מותאם:

בשלב הראשון, במטרה לבסס את הדיוק של מערכת הזיהוי בשפות שונות ולהתאים את המודל לרמת הדיוק של מחקרים קודמים, אומנו מודלים מבוססי BERT בהתאמה ספציפית לכל שפה. גישה זו מאפשרת להתייחס למאפיינים הייחודיים של כל שפה ולוודא שהמודלים מייצגים כראוי את נתוני ה-"clickbait" ו-"non-clickbait" בשפה הנבדקת. להלן שלבי האימון המפורטים:

איסוף וניקוי הנתונים לכל שפה:

לכל שפה נבחר מאגר נתונים נפרד שתויג לפי כותרות מסוג "clickbait" שסומן כ- 1 ו-"non-clickbait" שסומן כ- 0. מאגרי הנתונים נאספו ממקורות ציבוריים כגון Kaggle ו-GitHub, תוך הקפדה על פורמט מובנה הכולל עמודת כותרת (title) ועמודת תיוג (label) המגדירה האם הכותרת היא "clickbait" או "non-clickbait". המאגרים כוללים בין 8,000 ל-70,000 דוגמאות מתויגות בשפות: אנגלית, טורקית, גרמנית, רומנית, בנגלית, אינדונזית, עברית, ערבית, וסינית.

מבנה מאגרי הנתונים וההתאמות לכל שפה:

כל מאגר חולק לסט אימון וסט בדיקה ביחס של 80-20, תוך שמירה על יחס תיוג דומה בסט הבדיקה ובסט האימון. נתונים אלו הוזנו לפורמט CSV או Excel בהתאם לפורמט המקורי.

מודלי BERT עבור כל שפה:

לצורך התאמת המודלים לייחודיות הלשונית של כל שפה, נבחרו מודלי BERT ייעודיים, אשר כבר אומנו מראש על טקסטים מהשפה הנבדקת ומותאמים למאפיינים הדקדוקיים והתחביריים שלה. להלן דוגמאות למודלים שנבחרו:

- אנגלית: השתמשנו במודל BERT בסיסי עם התאמה לכותרות אנגלית. מאגר הנתונים באנגלית כלל 32,000 כותרות מתויגות.

- טורקית: המודל BERT-base-turkish-cased של חברת dbmdz אומן על מאגר בן 20,038 כותרות.

- גרמנית: אומן המודל bert-base-german-cased על מאגר נתונים גרמני רחב היקף, הכולל כ-69,397 כותרות.

- רומנית: המודל bert-base-romanian-cased-v1 הותאם על כ-8,318 כותרות מתויגות.

בשפה הבנגלית אומן מודל sagorsarker/bangla-bert-base. בשפה האינדונזית אומן indobenchmark/indobert-base-p1. בסינית bert-base-Chinese. בטורקית dbmdz/bert-base-turkish-cased ובעברית HeBERT.

פרוטוקול האימון: המטרה של כל אחד מהמודלים הייתה למקסם את דיוק הזיהוי של "clickbait" מול ""non-clickbait בשפה בה אומן המודל. להלן פרוטוקול האימון שהופעל:

- "אופטימיזציה": נעשה שימוש בoptimizer מסוג AdamW, שהתאים במיוחד למודל BERT בשל המורכבות הגבוהה שלו והרגולריזציה היעילה שהוא מספק.

- "loss function": לצורך מדידת שגיאות המודל, השתמשנו בפונקציית הפסד מסוג CrossEntropyLoss שמתאימה לבעיות סיווג בינאריות מסוג זה.

- "early stopping": האימון הוגבל ל-10 epochs כאשר התוצאות מתכנסות מבוצעת עצירה כדי למנוע overfitting.

- "learning rate scheduling ": בוצעה התאמה הדרגתית של קצב הלמידה תוך מעקב אחר מדדי דיוק על סט האימות. כל שפה נבחנה בנפרד ונמדדה במדדי Accuracy, Precision, Recall ו-F1 Score, כדי להבטיח ביצועים מקסימליים במודל הספציפי לכל שפה.

ניתוח הביצועים:

לאחר האימון, המודלים נבחנו על ה-test set של כל שפה, תוך חישוב מדדי הביצועים המרכזיים: Accuracy Precision, Recall ו-F1 Score, והשוואה לתוצאות שהתקבלו במחקרים קודמים.

**אימון cross-lingual.**

בשלב השני של המחקר, בחנו את היכולת של מודל cross-lingual לבצע זיהוי clickbait. בשיטה זו אומן מודל על מאגר נתונים בשפה אחת ונבדק על מאגר נתונים בשפה שונה, תוך שימוש במודל cross-lingual: bert-base-multilingual-cased . הבחירה במודל זה נעשתה בשל התאמתו לשפות מגוונות, מודל זה מאומן על שפות רבות כולל כל השפות שהורחבו לעיל. מה שמאפשר לו להרחיב את אימונו על מאגרי נתונים בשפות שאינן השפה המקורית.

**הכנה ואימון המודל**

1. **אימון על שפת מקור:** מודל ה- cross-lingual אומן תחילה על מאגר נתונים בשפה אחת (שפת מקור), כאשר נעשה שימוש באותם פרמטרים בסיסיים בהם השתמשנו באימון החד-שפתי. לדוגמה:
   * **אנגלית**: המודל אומן על כ-32,000 כותרות מתויגות באנגלית.
   * **טורקית**: בוצע אימון על כ-20,038 כותרות מתויגות בשפה זו.
   * **גרמנית**: המודל אומן על 69,397 כותרות בשפה הגרמנית.
   * **שפות נוספות**: אותן הגדרות אומצו גם למאגרים בשפות כגון רומנית, ערבית וסינית.
2. **תהליך אימון**: במהלך האימון, המודל עבר  **fine-tuning** בהתאמה לשפת המקור בלבד. כל נתוני שפת המקור שימשו לאימון של המודל במטרה למקסם את היכולת שלו לזהות  *clickbait*  במסגרת השפה הזו בלבד, תוך שימוש באופטימיזציה AdamW ו- gradual learning rate scheduling.

**הערכה על שפת יעד (Target Language)**

לאחר שלב האימון בשפת המקור, נבדק המודל על מאגר נתונים מתויג בשפה אחרת. לדוגמה, אם המודל אומן על מאגר באנגלית, הוא נבדק על כותרות בשפה העברית, הערבית ויתר השפות, וכך נבדקה היכולת של המודל לזהות  *clickbait* במגוון רחב של תרחישים בין-שפתיים.

**מדדי הביצועים**

כדי להעריך את איכות הזיהוי, חושבו מדדים סטנדרטיים לכל בדיקה בשפת יעד**Accuracy**,  **Precision**, **Recall** ו-**F1 Score** מטרת המדדים הללו היא להצביע על פערים אפשריים בין ביצועי המודל בשפת המקור לבין שפת היעד, ובאופן ספציפי על יכולתו הכללית של מודל BERT *multilingual* לבצע זיהוי  *clickbait* מדויק בשפה שונה מזו שעליה אומן.

**אימון *multilingual***

בשלב זה, התמקדנו בהערכת היכולת של מודל *multilingual* לזהות כותרות *clickbait* תוך שילוב של מספר שפות באימון אחד. גישה זו מאפשרת למודל ללמוד מאפיינים כלליים של *clickbait* על פני שפות שונות ולבחון אם השילוב של שפות נוספות באימון משפר את ביצועי המודל על שפת הבדיקה.

בשלב זה נרצה לבחון האם אימון המודל על מאגר נתונים הכולל שילוב של מספר שפות במקביל עשוי לשפר את יכולת הזיהוי של  *clickbait* בשפה ספציפית – זו שנמצאת ב - test. כך ניתן לבדוק האם שפות נוספות באימון מסייעות בזיהוי.

**אסטרטגיות אימון:**

1. **אימון על צמדי שפות**.  
   בשלב הראשון, המודל אומן על כל שילוב אפשרי של שתי שפות, ולאחר מכן נבדק על אחת מהן שנכללה גם באימון. לדוגמה, המודל אומן על אנגלית ובנגלית ונבדק על בנגלית.
2. **אימון על מספר רב של שפות (9 שפות) וtest בשפה העברית.**בשלב הבא, המודל אומן על כל תשע השפות יחד, תוך כדי כך ששפת הבדיקה עברית כלולה במאגר האימון. גישה זו נועדה לבדוק את יכולת הזיהוי של המודל כאשר הוא מקבל נתונים מגוונים בשפות שונות. והיכולת של השפות הנוספות להוות data augmentation עבור כותרות לזיהוי clickbait בעברית.
3. **החסרת שפה** – בדיקת שפות משמעותיות יותר לזיהוי clickbait בעברית  
   כדי לזהות את תרומת כל שפה לביצועי המודל, אומן המודל על 8 מתוך 9 השפות, כאשר בכל פעם הושמטה שפה אחת, והבדיקה נעשתה בעברית. תהליך זה נועד לבחון האם השמטת שפה מסוימת מהאימון משפיעה על הביצועים, ומה התרומה של כל שפה בנפרד לדיוק של המודל.

**פרוטוקול האימון**

בכל שלבי האימון ה-*multilingual*, נעשה שימוש בפרמטרים זהים לאלה ששימשו באימון החד-שפתי:

* **אופטימיזציה עם AdamW**: אופטימייזר זה נבחר בשל יעילותו עם מודלי BERT. **learning rate scheduling**: התאמות מדורגות של קצב הלמידה.:**Early Stopping** האימון הוגבל ל-10 epochs כאשר התוצאות מתכנסות מבוצעת עצירה כדי למנוע overfitting.

**הערכת הביצועים**

לאחר האימון בכל אסטרטגיה, חושבו מדדי ביצועים **Accuracy** , **Precision**, **Recall**, ו **F1 Score** על סט הבדיקה. מדדים אלו נועדו לספק מסקנות על ההשפעה של הוספת שפות שונות לאימון, ולזהות אילו שפות עשויות לתרום לדיוק הזיהוי של *clickbait* או לגרום לירידה בביצועים כאשר הן נכללות במודל.

שיטות אימון ה-*multilingual* אפשרו לנו להעריך כיצד שילוב שפות שונות משפיע על ביצועי המודל, והאם ניתן ליצור מודל כללי ועמיד יותר לפעולה רב-שפתית.

**תוצאות**

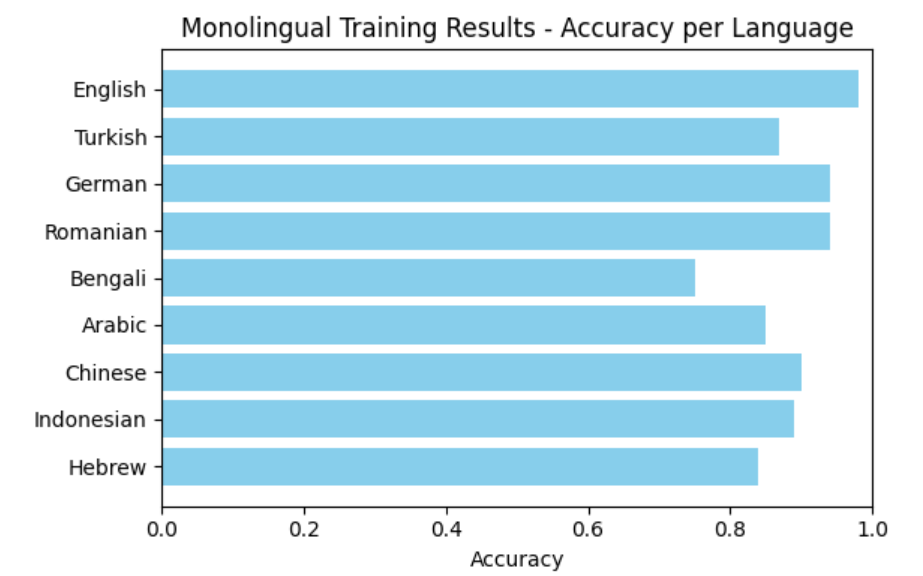
במהלך הניסויים, נבדקו הביצועים של המודל במגוון תרחישים על מנת לבחון את יכולתו לזהות  *clickbait* בשפות שונות ואת השפעת כל אסטרטגיית אימון על תוצאות המודל. כל תרחיש נותח לפי מדדי ביצועים שונים, בהם **Precision**, **Recall, Accuracy** ו**F1 Score**-.

**תוצאות אימון חד-שפתי**

מודלים שאומנו על כל שפה בנפרד השיגו דיוק גבוה, עם תוצאות הקרובות לאלו שנמצאו במחקרים קודמים. התוצאות המרכזיות היו כדלקמן:

* **אנגלית**: השיג דיוק של כ-0.98, בהשוואה לתוצאות במאמרים קודמים.
* **טורקית**: המודל הטורקי הגיע לדיוק של כ-0.92, שיפור ביחס לממצאים הקיימים.
* **עברית**: השיג ביצועים של כ-0.87 דיוק. זהה למחקרים קודמים.
* **רומנית, בנגלית, ערבית, סינית ואינדונזית**: המודלים השונים עבור השפות הללו השיגו רמות דיוק בטווח של 0.75 עד 0.94, עם ערכי Precision ו-Recall גבוהים בשפות בעלות מאגרים רחבי היקף, כמו גרמנית ורומנית

תמונה 1 – רמת דיוק עבור כל שפה בשימוש במודל BERT מתאים.



**תוצאות - אימון *cross-lingual***

בתרחיש אימון ה *cross-lingual* -אומן המודל על מאגר נתונים בשפה אחת ונבדק על מאגר נתונים בשפה אחרת. מטרת השלב הייתה לבחון את יכולת המודל לזהות *clickbait* בשפת יעד שאינה נכללת באימון המקורי. תוצאות האימון הראו ביצועים משתנים בהתאם למידת הקרבה הלשונית והתחבירית בין שפת המקור לשפת היעד.

**תוצאות מפתח באימון *cross-lingual***

1. **אימון באנגלית ובדיקה בעברית**  
   כאשר המודל אומן על מאגר הנתונים באנגלית ונבדק על מאגר הנתונים בעברית, הוא השיג:
   * **:Accuracy** כ-0.73
   * **:Precision** כ-0.70
   * **:Recall** כ-0.75
   * **F1 Score** : כ-0.72

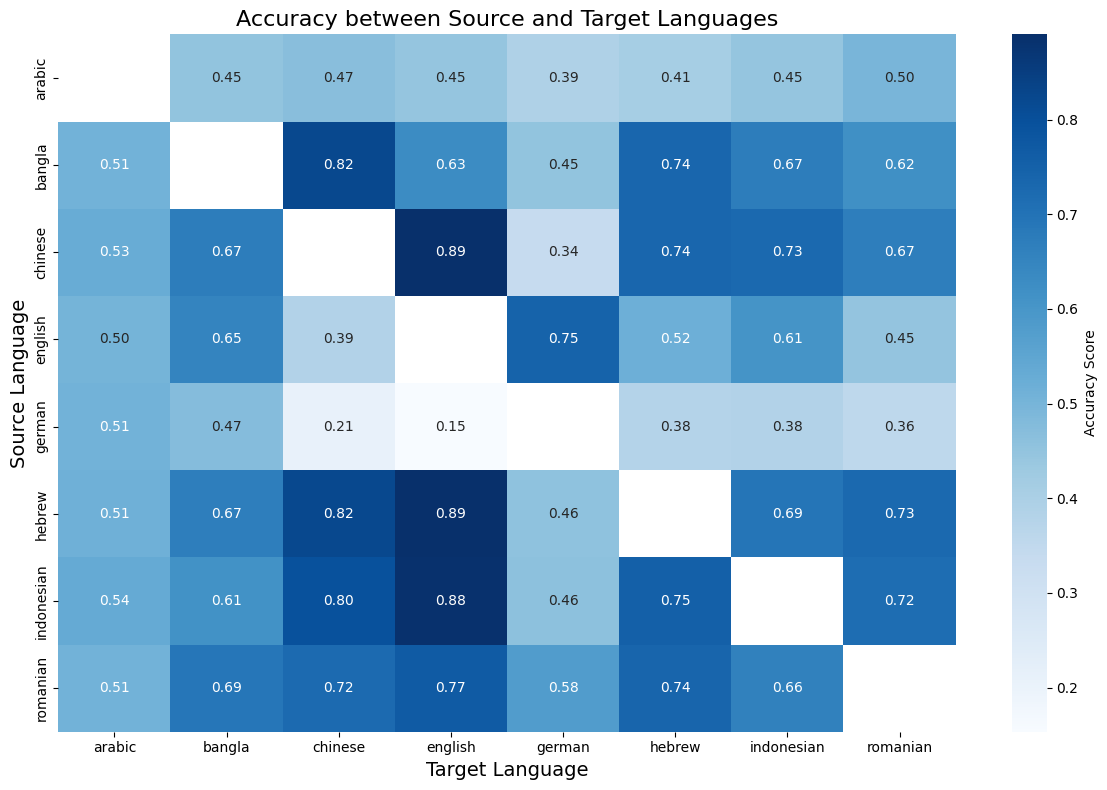
התוצאות מראות ביצועים מתונים בזיהוי כותרות  *clickbait* בעברית על בסיס מודל שאומן באנגלית בלבד. רמת הדיוק מצביעה על כך שהמודל הצליח לזהות את המאפיינים הכלליים של *clickbait*, אך לא תמיד הצליח לזהות מאפיינים ייחודיים לעברית.

1. **אימון בערבית ובדיקה באנגלית**  
   בתרחיש זה, המודל אומן על מאגר הנתונים בערבית ונבדק על מאגר הנתונים באנגלית. התוצאות היו כדלקמן:
   * **:Accuracy** כ-0.75
   * **:Precision** כ-0.72
   * **:Recall** כ-0.74
   * **:F1 Score** כ-0.73

בדומה לתוצאות הקודמות, הביצועים כאן היו טובים באופן חלקי, אך דיוק נמוך יותר בהשוואה למודלים שאומנו ובדקו באותה שפה. ייתכן שהשפה הערבית, בעלת מאפיינים תחביריים שונים משמעותית מהאנגלית, מקשה על המודל לשמר ביצועים גבוהים בזיהוי *clickbait* באנגלית.

**סיכום תוצאות אימון ה*cross-lingual***

נראה כי הביצועים הגבוהים ביותר הושגו כאשר שפת המקור ושפת היעד היו קרובות תחבירית, ואילו בזוגות שפות עם שונות גבוהה יותר, המודל התקשה להכליל את מאפייני *clickbait* בהצלחה.

תמונה 2- תוצאות Accuracyשל כל אימון מודלי ה-: cross-lingual

**תוצאות Multilingual**

בשלב ה-*Multilingual* המודל אומן על שילוב של שפות שונות, והביצועים הוערכו במספר אסטרטגיות במטרה לבדוק את השפעת השפות השונות על דיוק המודל.

**1 . אימון על צמדי שפות**

המודל אומן על קומבינציות של שתי שפות ונבדק בנפרד על כל אחת מהן, כדי לזהות שפות שמסייעות זו לזו בזיהוי  *.clickbait* השילוב של עברית וסינית הביא לשיפור ניכר בביצועים בשפה הסינית, והוביל לרמת דיוק של 0.968 – שיפור משמעותי בהשוואה לתוצאה המקורית, שהייתה 0.91. גם רומנית הושפעה לטובה כאשר שולבה עם גרמנית, והגיעה לדיוק של 0.8909 בהשוואה ל-0.88 באימון חד-שפתי​

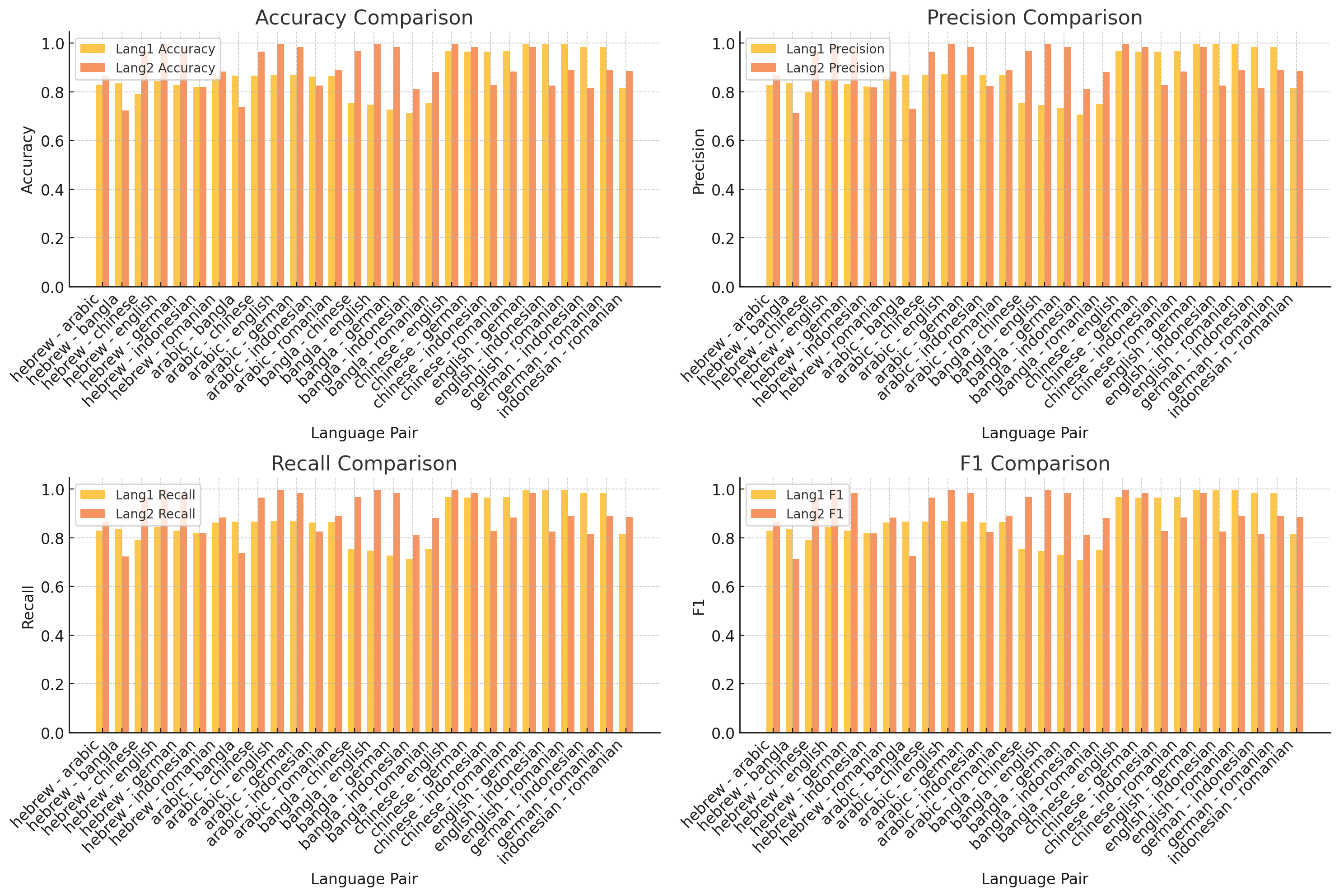
1. **אימון על כל השפות ובדיקה בעברית.**

לאחר מכן, אומן המודל על כל 9 השפות ובדיקת הביצועים נעשתה על מאגר נתונים בעברית, במטרה לבדוק את השפעת השפות הנוספות כ-*data augmentation* על הביצועים בעברית. התוצאה הראתה שהמודל השיג דיוק של כ-0.84 בעברית, והצליח להבחין בדפוסי  *clickbait* אף בשפה מורכבת כמו עברית. הדבר מעיד על יתרון באימון על מספר שפות בעת אימון על שפה עם מאגרי נתונים קטנים יותר​.

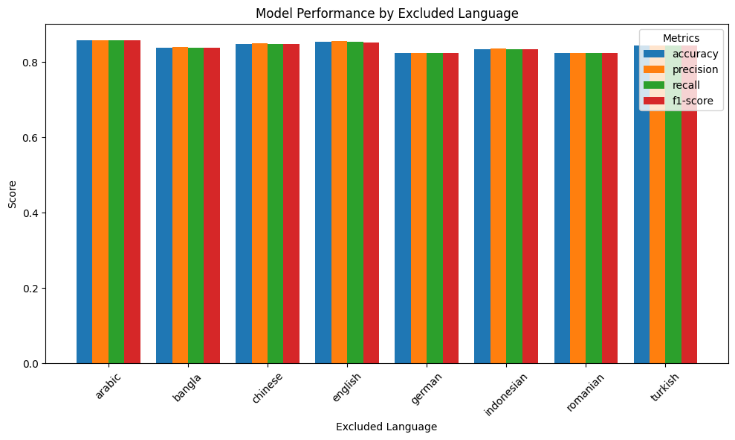
1. **החסרת שפה אחת בכל פעם**

במטרה לזהות את התרומה של כל שפה לביצועים של המודל בעברית, בכל ניסוי הושמטה שפה אחת מתוך תשע השפות, והמודל אומן על השפות הנותרות. לאחר מכן נבדק המודל על מאגר הנתונים בעברית, והתקבלו התוצאות הבאות:

* **השמטת גרמנית**: כאשר הושמטה השפה הגרמנית מהאימון, נרשמה ירידה של כ-4% בדיוק בביצועים על העברית, מה שמצביע על כך שגרמנית תורמת לשיפור הביצועים בעברית. ייתכן שהקרבה המבנית בין השפות והמאפיינים הסמנטיים מסייעים למודל להבחין טוב יותר בתבניות של *clickbait* .
* **השמטת רומנית**: גם במקרה של השמטת רומנית מהאימון, נרשמה ירידה של כ-3% בדיוק, ובנוסף ירידה בערכי ה-Recall, מה שמעיד על כך שרומנית מוסיפה למודל כיסוי טוב יותר לדפוסי  *clickbait* רחבים יותר.
* **השמטת סינית**: השמטה זו לא גרמה לירידה משמעותית בביצועים על העברית, אך בבדיקות המשולבות נמצא כי השילוב של סינית עם עברית שיפר את הדיוק בסינית
* **השמטת שפות אחרות (בנגלית, אינדונזית, טורקית)**: השמטת שפות אלו לא גרמה לירידה משמעותית בביצועי המודל, מה שמעיד כי תרומתן לזיהוי  *clickbait* בעברית הייתה נמוכה יחסית. סביר להניח שהשפות הללו כוללות דפוסים תחביריים וסמנטיים שונים מהעברית, ולכן פחות תורמות בזיהוי מבוסס שפה זו.

תמונה 3- תוצאות הבדיקות השונות במודל multilingual:

תמונה 4- תוצאות אימון על 8 שפות תוך החסרה של שפה אחת כל פעם:



בבדיקת זיהוי משפטים על מדגם ממאגר הנתונים קבלנו: (חלק מהמשפטים)

**משפטים שזיהה כ"non-clickbait" למרות שהינם "clickbait "false negative - .**

* תצפו לחיוך מושלם: כך תבחרו את הטיפול הנכון עבור החיוך שלכם
* ד"ר טטיאנה זייגר חושפת את הסוד שעומד מאחורי טיפולים אסתטיים מוצלחים
* גביע או צלחת? הטעות המביכה של השר אמסלם
* זה השעון החכם שכל גבר בישראל חיכה לו!
* סובלים מיתושים? יש לנו פתרון בשבילכם

**משפטים שזיהה כ" clickbait" למרות שהינם "" non-clickbait - false positive.**

* שי-לי ליפא, אביב משה ושרי אנסקי: 3 מתכוני חרוסת טעימים לפסח
* ראש הממשלה נתניהו החזיר את המנדט לנשיא
* "מצטער שלא הספקתי לעדכן אותך": יהודה פוליקר משיק שיר שכתב לו יהונתן גפן
* מיכאל בן דוד: 'כעסתי על אמא שלי שלא הגיעה לחתונה שלי'

**דיון**

תוצאות המחקר הנוכחי מדגישות את הפוטנציאל והאתגרים בשימוש במודלי multilingual ובמודלי *cross-lingual* לזיהוי *clickbait* בשפות שונות. במחקר זה, הוכחה היכולת של מודלי  *multilingual BERT* להשיג ביצועים טובים בזיהוי תבניות של כותרות, תוך גיוון השפות באימון, מה שמספק תרומה משמעותית למחקר בתחום עיבוד שפה טבעית ויישומי *clickbait* בשפות דלות משאבים.

**יתרונות המחקר ותרומתו**

אחד מהיבטי התרומה המרכזיים של המחקר הוא היכולת ליישם את המודלים גם בשפות עם מאגרי נתונים מצומצמים, כמו עברית, באמצעות שילוב שפות אחרות. שימוש בשפות כמו גרמנית ורומנית הביא לשיפור מדיד בביצועים על בדיקות בעברית, ומראה שניתן לנצל מאפיינים לשוניים משפות קרובות מבנית כדי להעשיר את יכולת הזיהוי. יתרון נוסף הוא היכולת לאמן את המודלים על שפות מרובות במקביל ולשפר את ביצועיהם במודל כללי אחד, מה שמציע חסכון בזמן ובמשאבים, ומרחיב את היישומים של המודל לעיבוד תוכן בשפות נוספות.

**אסטרטגיות נוספות ותיקוף התוצאות**

במחקר הנוכחי נבחנו גם שילובים נוספים שלא הוצגו, כגון אימון על שלוש שפות ובדיקה על עברית. שילוב זה נועד לבדוק האם ניתן להפיק תוצאות מדויקות אף יותר על עברית באמצעות הרחבת מספר השפות באימון. בנוסף, נעשתה בדיקה מעמיקה של תופעת ה-*overfitting* , שכן חלק מהשפות הציגו פערים בין דיוק גבוה באימון לדיוק נמוך בבדיקות. כדי לבדוק את ההשערה, אומן מודל  *TF-IDF* כמודל פשטני יותר על אותם מאגרי נתונים. התוצאות במודל זה היו נמוכות, מה שמראה כי ביצועי ה*BERT* היו טובים ולא נבעו מ- *overfitting* אלא ממאפייניו המתקדמים בזיהוי מבנים לשוניים מורכבים.

מבחינת זיהוי clickbait בעברית, המודל זיהה כותרות המעוררות רגשות אצל הקוראים, וסימן אותם כ- clickbait ( כ- 50% משגיאות (falsePositive) גם כשלא היו כאלו. כאשר לא היה תיאור של השיטה גם אז זיהה כ- clickbait. המודל התקשה לחזות כותרות המעוררות סקרנות וסימן את חלקן כ- non-clickbait (יותר מ- 50% משגיאות ה falseNegative ).כותרת או נושא השייכים ליישות לא מזוהים לא ידע לזהות כ- clickbait.

**מסקנות**

המחקר הנוכחי מדגים את הפוטנציאל של מודלי multilingual (multilingual BERT) בזיהוי כותרות clickbait במספר שפות, תוך התמקדות בשפות עם מאגרי נתונים דלילים. ביצועי המודל, במיוחד בשפות כמו עברית, הוכיחו ששילוב שפות קרובות באימון מסייע למודל לזהות דפוסי clickbait גם בשפות בהן היקף הנתונים מוגבל.

נוסף על כך, נמצא כי בשילובים מסוימים התקבלו תוצאות גבוהות במיוחד, כמו השילוב של סינית ועברית שהוביל לשיפור בביצועים של המודל בסינית. קשר זה עשוי לנבוע מדמיון בתחביר ובתבניות של רכיבי המשפט, ובהיותן שפות מסורתיות. כפי שמתואר במאמרם של מיהיל וחבריו[[16]](#endnote-16), שמציין את הקרבה בין עברית מקראית לסינית בהצבת אובייקטים בתחילת משפט. הדמיון במבני השפה בשתי השפות עשוי, אם כך, להעשיר את הלמידה המשותפת במודל ולסייע לזיהוי תבניות clickbait הדורשות הדגשת מידע, מאפיין נפוץ בסוגי הכותרות הללו. כמו כן יתכן וצורת האותיות השונה גם עוזרת ללמידה של השפות זו מזו.

המחקר תורם בכך שהוא מדגים את היתרונות והחסרונות של גישת cross-lingual ו-multilingual לזיהוי clickbait, ומספק בסיס לשילוב שפות נוספות באימון כדי להעשיר את יכולת המודל הכללית. עבודה עתידית יכולה לכלול אימון עם מודלים מתקדמים נוספים (כגון XLM-R) ויצירת מאגרי נתונים רחבים יותר בעברית, כדי לשפר את הביצועים ולמנוע מגבלות הנובעות מהיקפי נתונים מצומצמים.

1. [↑](#endnote-ref-1)
2. Natanya, T., & Liebeskind, C. (2023). Clickbait detection in Hebrew. Lodz Papers in Pragmatics, 19(2), 427-446.‏

   Adrian, F. H. N., Handradika, N. N., Prasojo, A. E., Gunawan, A. A. S., & Setiawan, K. E. (2024, February). Clickbait Detection on Online News Headlines Using Naive Bayes and LSTM. In 2024 IEEE International Conference on Artificial Intelligence and Mechatronics Systems (AIMS) (pp. 1-6). IEEE.‏ [↑](#endnote-ref-2)
3. <https://www.kaggle.com/datasets/amananandrai/clickbait-dataset?resource=download> [↑](#endnote-ref-3)
4. https://www.kaggle.comcodedatafan07clickbait-news-classification-using-sadedegelnotebook [↑](#endnote-ref-4)
5. https://github.com/youurt/german\_clickbaits\_tensorflow\_js/blob/main/01\_Datenaquise/ROHDATEN.csv [↑](#endnote-ref-5)
6. Broscoteanu, D. M., & Ionescu, R. T. (2023). A Novel Contrastive Learning Method for Clickbait Detection on RoCliCo: A Romanian Clickbait Corpus of News Articles. arXiv preprint arXiv:2310.06540.‏ [↑](#endnote-ref-6)
7. https://github.com/ralucaginga/ClickbaitSciTechRO/blob/main/dataset.xlsx [↑](#endnote-ref-7)
8. Mahtab, M. M., Haque, M., Hasan, M., & Sadeque, F. (2023). Banglabait: Semi-supervised adversarial approach for clickbait detection on bangla clickbait dataset. *arXiv preprint arXiv:2311.06204*.‏ [↑](#endnote-ref-8)
9. https://www.kaggle.com/datasets/motaharmahtab/banglabait-bangla-clickbait-dataset?resource=download [↑](#endnote-ref-9)
10. Fakhruzzaman, M. N., & Gunawan, S. W. (2021). Web-based Application for Detecting Indonesian Clickbait Headlines using IndoBERT. arXiv preprint arXiv:2102.10601.‏ [↑](#endnote-ref-10)
11. https://github.com/ruzcmc/ClickbaitIndo-textclassifier [↑](#endnote-ref-11)
12. Alharbi, B., Alhanaya, R., Alqarawi, D., & Alnejaidi, R. (2023, September). Baheta: Balanced and Unbalanced Dataset in Arabic Clickbait Detection Using a Deep Learning Model (LSTM). In International Conference on Innovation of Emerging Information and Communication Technology (pp. 1-12). Cham: Springer Nature Switzerland.‏ [↑](#endnote-ref-12)
13. https://github.com/RazanALhanaya/Clickbait-Fake-News- Dataset/blob/main/clickbaitNew%2BFakeNews%20(1).xlsx [↑](#endnote-ref-13)
14. Wu, Y., Cao, M., Zhang, Y., & Jiang, Y. (2023, May). Detecting Clickbait in Chinese Social Media by Prompt Learning. In 2023 26th International Conference on Computer Supported Cooperative Work in Design (CSCWD) (pp. 369-374). IEEE [↑](#endnote-ref-14)
15. <https://github.com/WeSeewy/Chinese-Clickbait/blob/main/data/news.txt>

    Myhill, J., & Xing, Z. (1993). The discourse functions of patient fronting: A comparative study of Biblical Hebrew and Chinese.‏ [↑](#endnote-ref-15)
16. [↑](#endnote-ref-16)