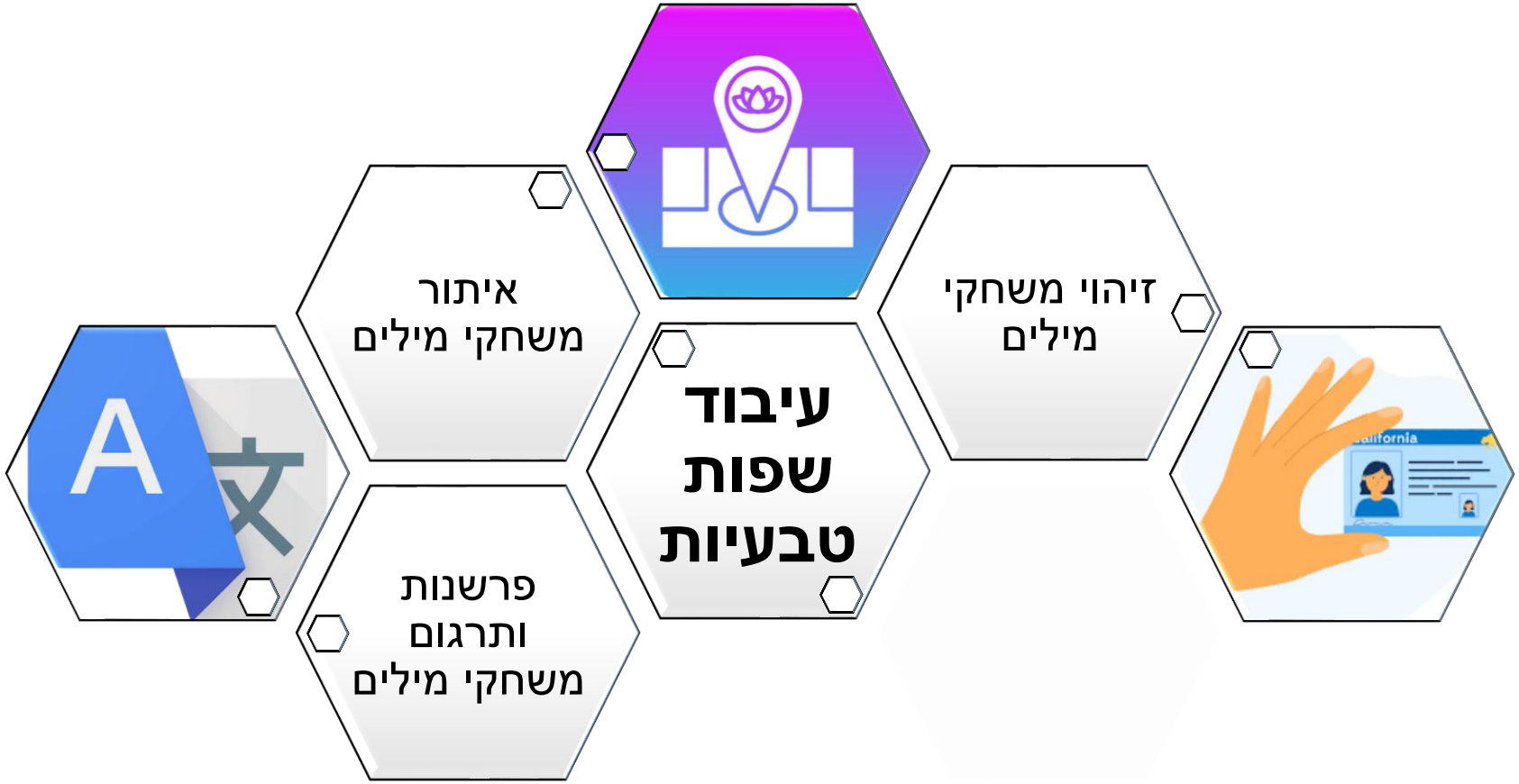
Exploring Humor in Natural Language Processing: A Comprehensive Review of JOKER Tasks at CLEF Symposium 2023

השם : הדיל חמודי

הקדמה - הומור

דו-משמעות	• משחק מילים שבו ביטוי או משפט מכיל שתי משמעויות שונות, אחת מהן לרוב תמימה או מילולית, והשנייה לעיתים קרובות בעלת אופי שובב או הומוריסטי.
משחק מילים קלאסי	• צורת משחק מילים קלאסית שנשענת על ניצול המשמעויות המרובות של מילה אחת או על שימוש במילים שנשמעות דומות אך בעלות משמעויות שונות.
אירוניה	• סוג של משחק מילים שבו משתמשים במילים כדי להעביר משמעות שהיא הפוכה למשמעות המילולית שלהן. אירוניה לרוב תלויה בהקשר ובטון הדיבור של הדובר.
ספונריזם (ביטוי משוכל)	• סוג של משחק מילים שבו מחליפים את הצלילים או האותיות הראשונות של מילים כדי ליצור אפקט הומוריסטי.

מושגים בסיסים שקשורים למאמר



מאגר הנתונים JOKER

הוא אוסף רחב של תוכן

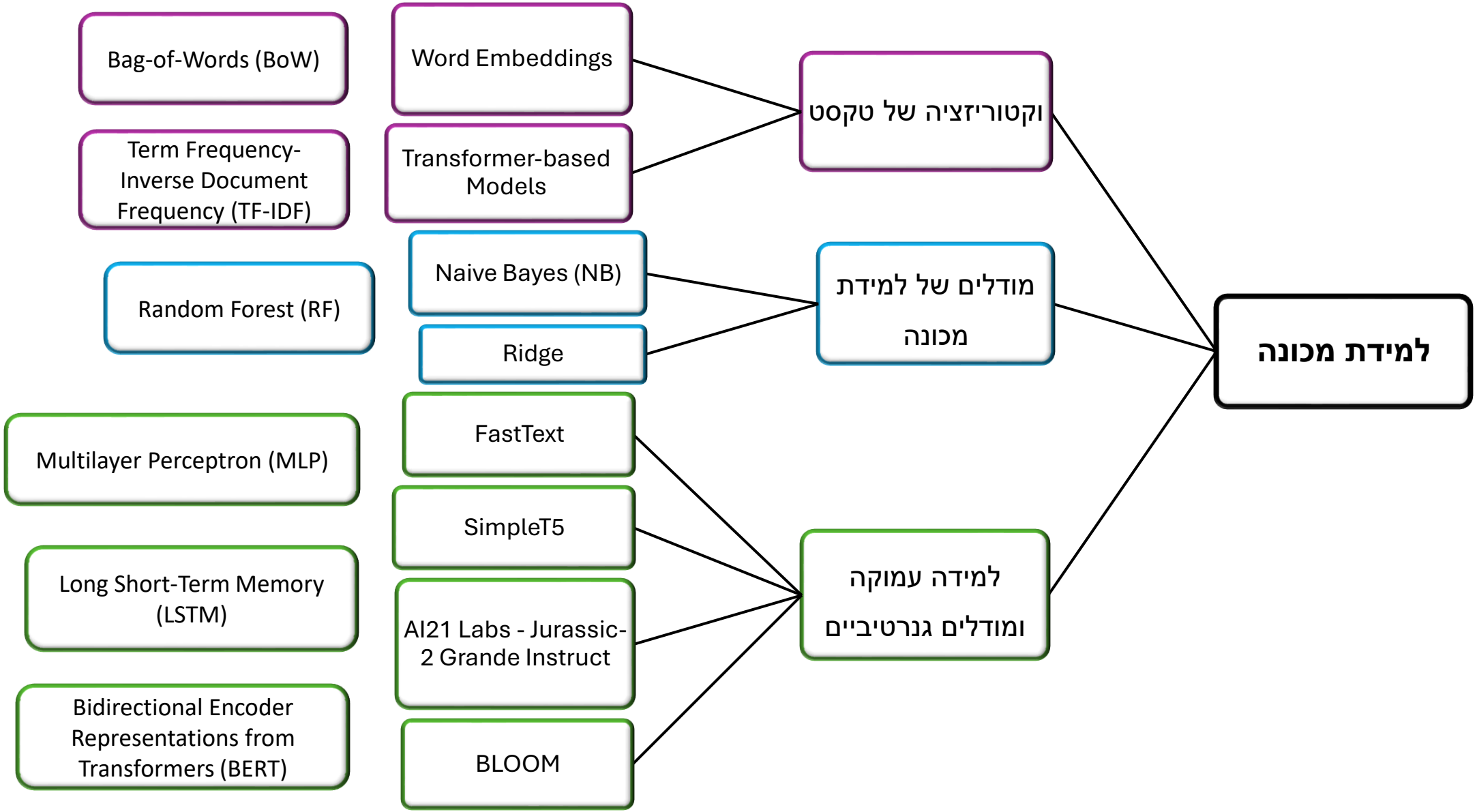
הקשור להומור, כולל בדיחות

וקללות משפות ותרבויות שונות.

מאגר זה משמש מקור יקר ערך לאימון

והערכת מודלים של עיבוד שפה טבעית.

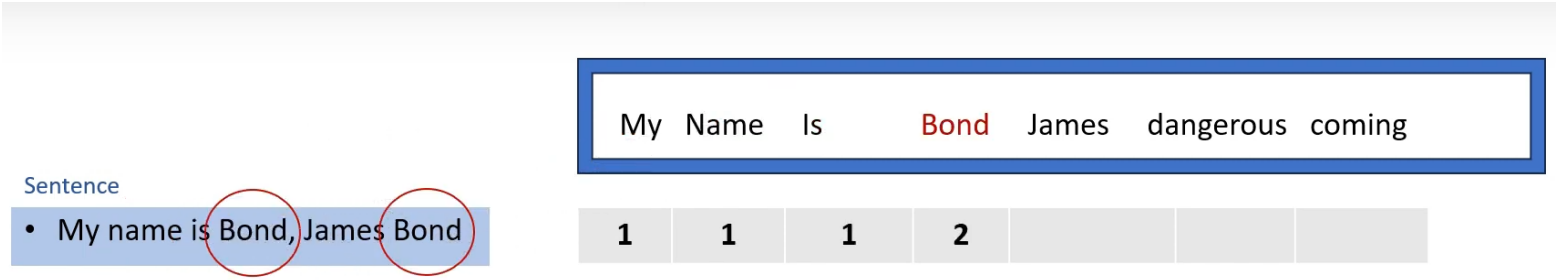
שימוש במודלים וטכניקות קיימות

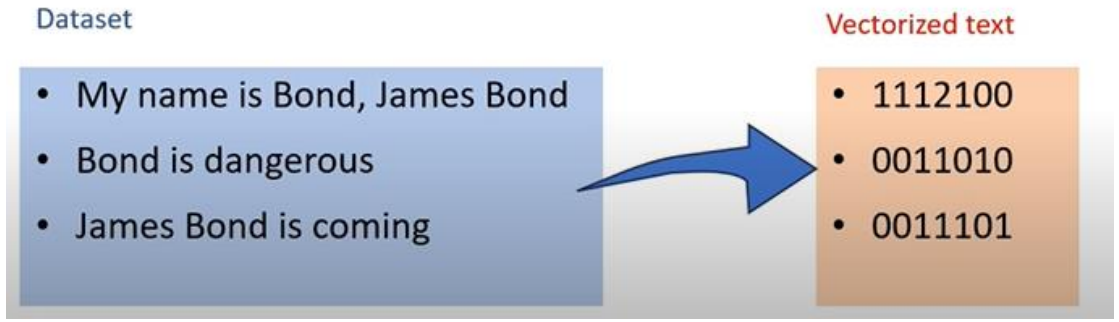


וקטוריזציה של טקסט

וקטוריזציה של טקסט - Bag-of-Words

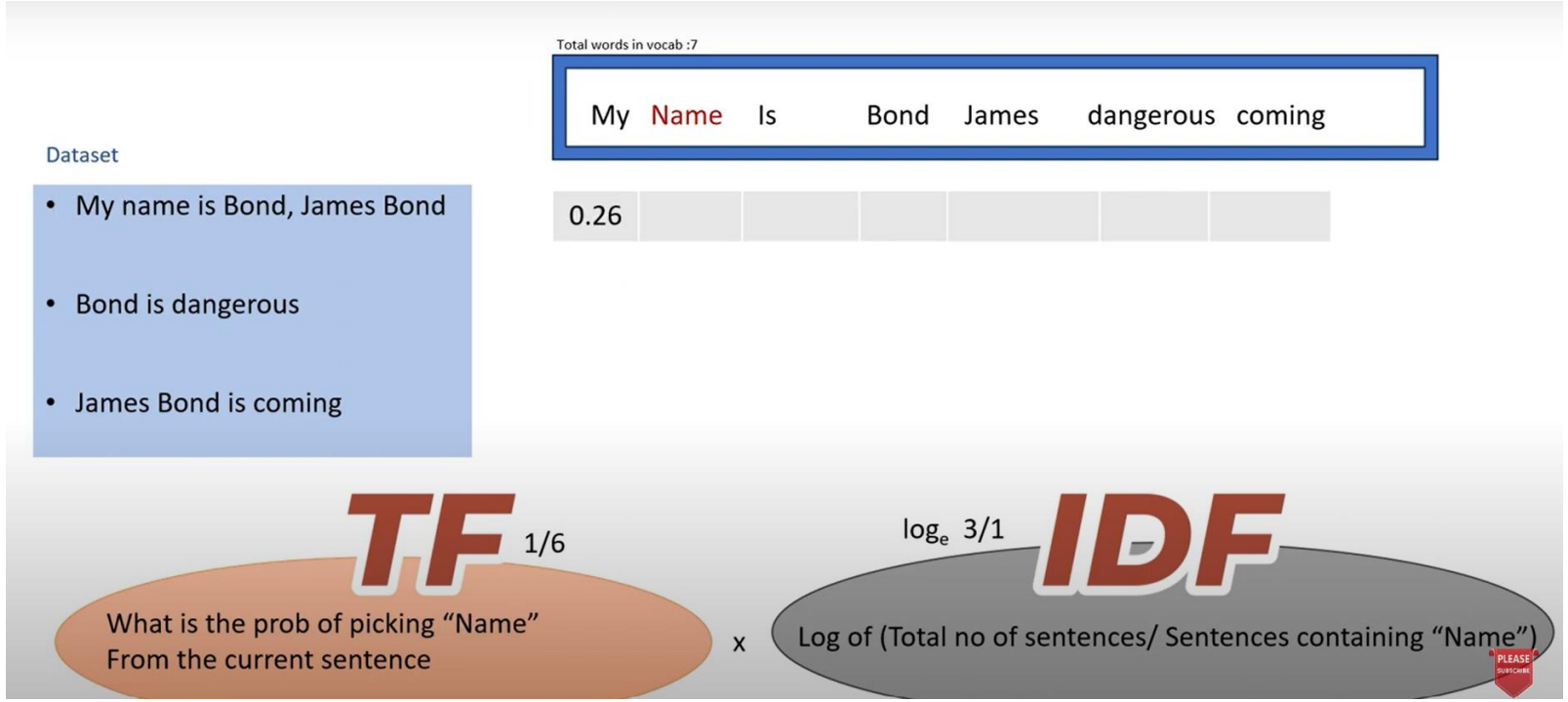
1. BoW : Bag-of-Words (BoW) מייצג מסמך כקולקציה של מילים, תוך התעלמות מהדקדוק וסדר המילים. הוא יוצר אוצר מילים של מילים ייחודיות ומקצה ערך בינארי או ערך מבוסס תדירות לכל מילה, המצביע על נוכחותה או התרחשותה במסמך.





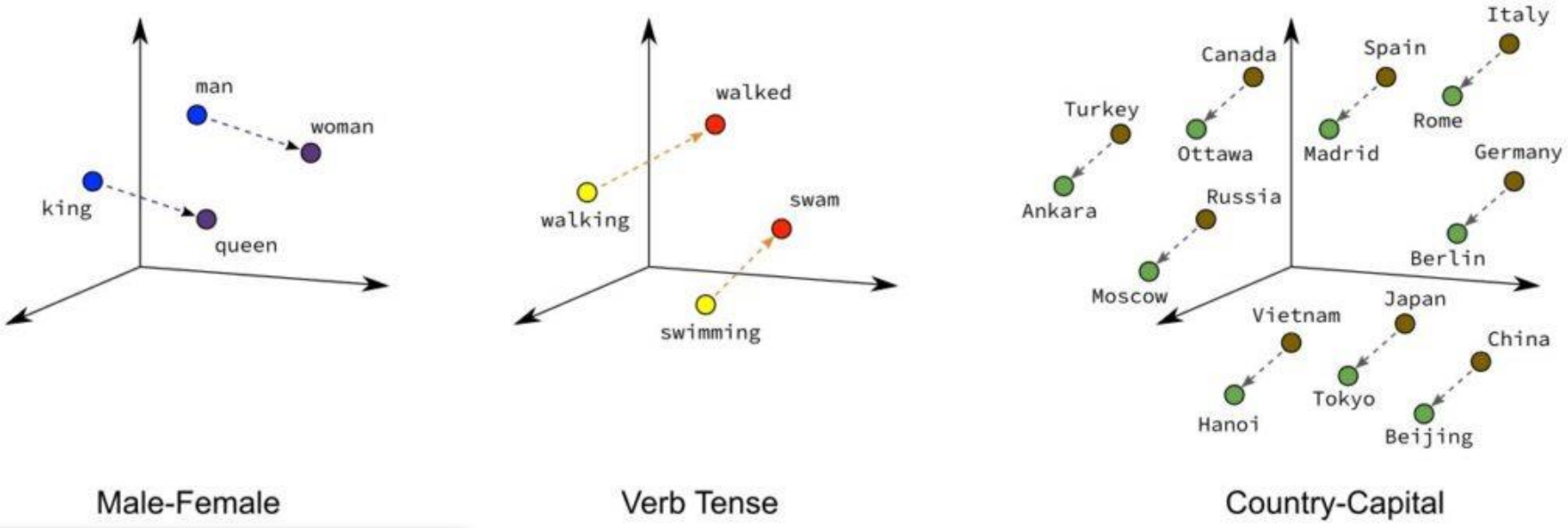
וקטוריזציה של טקסט - TF-IDF

2 . TF-IDF (תדירות מילה-תדירות מסמך הפוכה): TF-IDF היא טכניקה המשמשת להערכת חשיבות של מילה בתוך מסמך מסוים ביחס לכלל המסמכים . היא מקצה משקל גבוה יותר למילים המיוחדות למסמך מסוים.



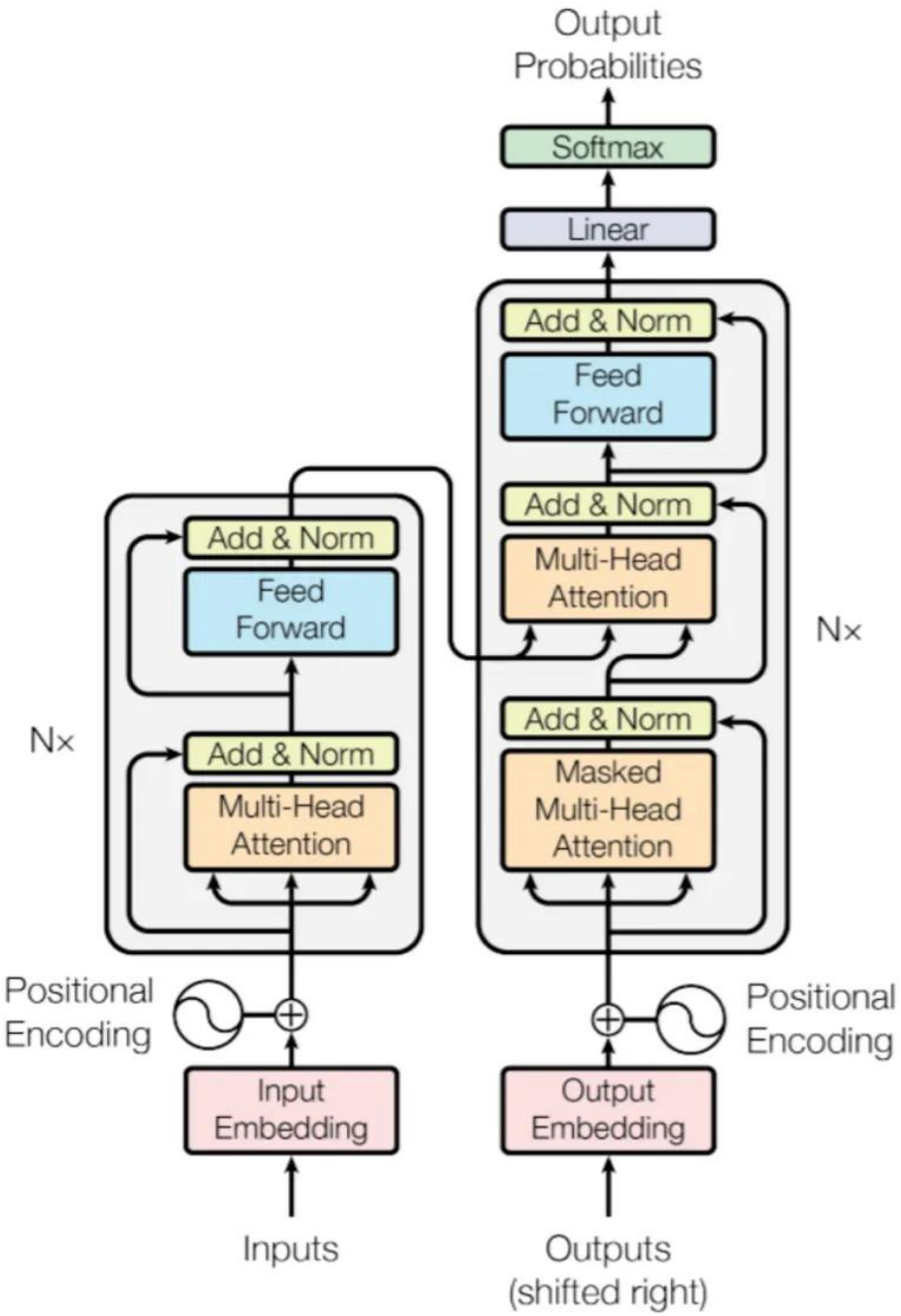
וקטוריזציה של טקסט - Word Embeddings

3 . Word Embeddings : היא דרך לייצג מילים כווקטורים, שהם רשימות של מספרים, שמבטאים את המשמעות שלהן ואת היחסים ביניהן. שיטה זו מאפשרת למילים עם משמעות דומה להיות מיוצגות במספרים קרובים זה לזה.
אלגוריתמים נפוצים ליצירת ייצוגים כאלה כוללים את Word2Vec ו-FastText. את ה- Word Embeddings ניתן לאמן מראש על בסיס כמויות גדולות של טקסט, או להתאים אותן במיוחד למשימה מסוימת כדי לשפר ביצועים בהקשר מסוים.



King – man + woman = Queen

וקטוריזציה של טקסט – transformer-based Models

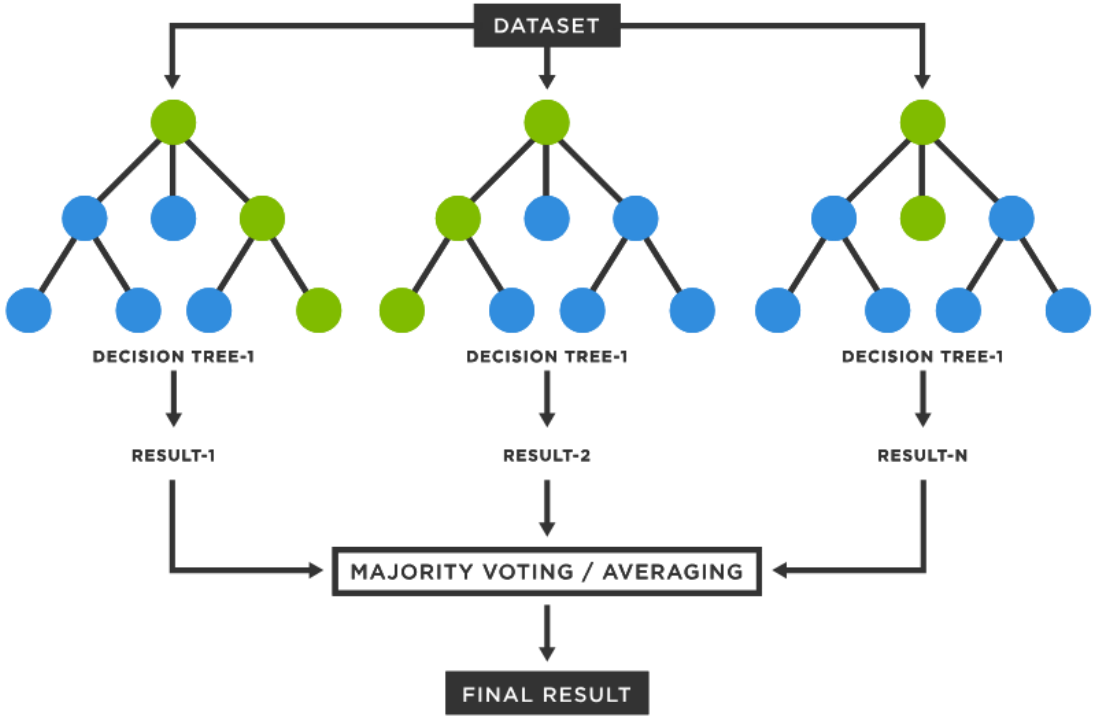


4 . מודלים מבוססי טרנספורמר, כמו BERT ו-GPT, שינו את הדרך בה מייצגים טקסט. הם משתמשים במנגנון שנקרא "תשומת לב" (Attention) כדי להתייחס לכל חלקי הטקסט בקלט, מה שמאפשר להם להבין טוב יותר את ההקשר של מילים ומשפטים. בזכות זה, המודלים יכולים ליצור ייצוגים מדויקים יותר של המשמעות של מילים בתוך המשפטים והקשרים שבהם הן מופיעות.

מודלים של למידת מכונה

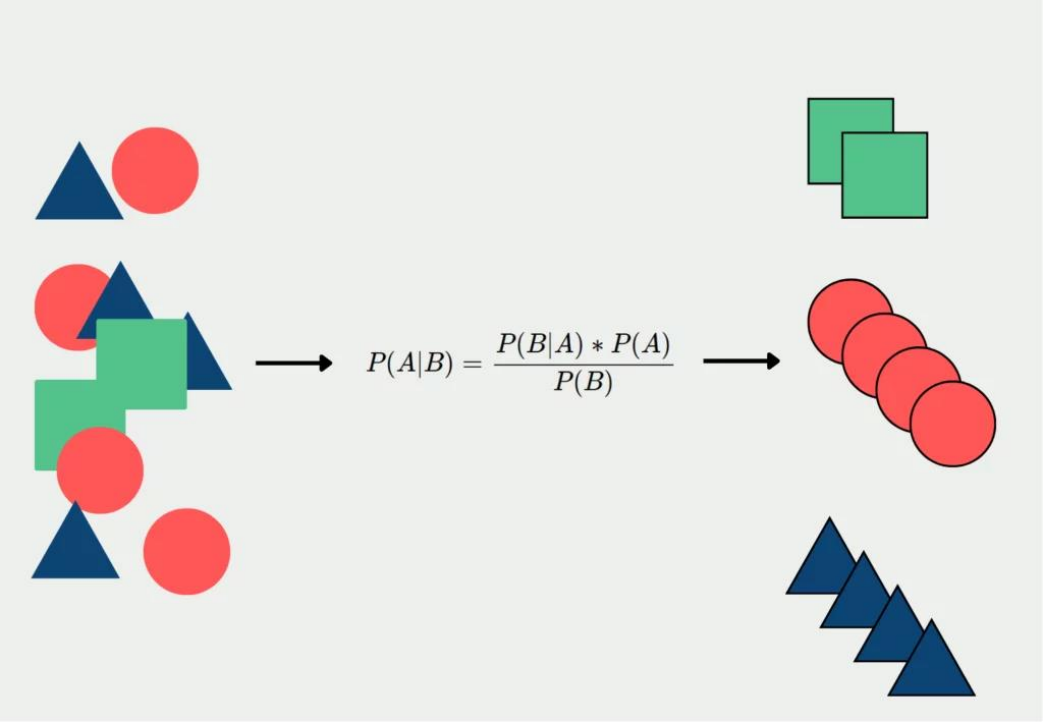
מודלים של למידת מכונה - Random Forest

- Random Forest (RF) : מודל ה-Random Forest הוא שיטת למידת חבורה (ensemble learning) שמשלבת מספר עצי החלטה לצורך ביצוע תחזיות. הוא פועל על ידי בניית מספר רב של עצי החלטה ואגירת הפלטים שלהם כדי לקבוע את התחזית הסופית. גישה זו משפרת את הדיוק של המודל ומפחיתה את הסיכון ל-overfitting על ידי ניצול מגוון של עצים מרובים.

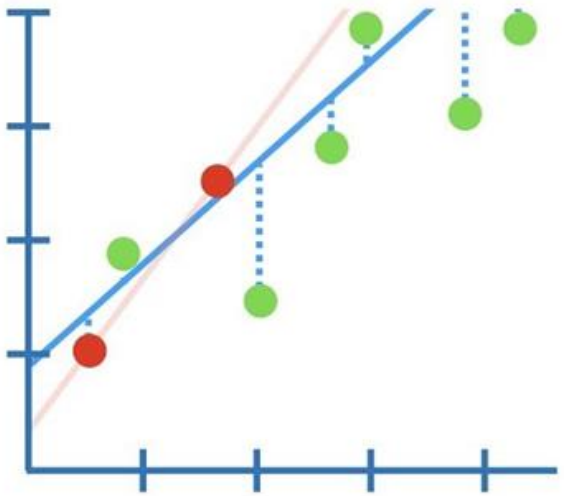


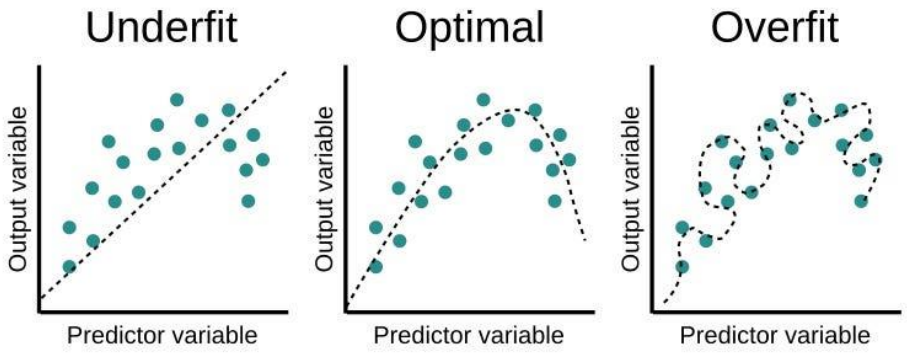
מודלים של למידת מכונה - Naive Bayes

- Naive Bayes (NB) : מודל ה-Naive Bayes הוא מסווג **הסתברותי** פשוט אך חזק. הוא מניח שהמאפיינים לא תלויים זה בזה ומחשב את הסבירות של קבוצה בהתבסס על הנתונים הקלטיים. על אף ההנחה של עצמאות בין המאפיינים, ה-Naive Bayes לעיתים קרובות מציג ביצועים טובים ויעיל חישובית במשימות סיווג טקסט.



מודלים של למידת מכונה - Ridge



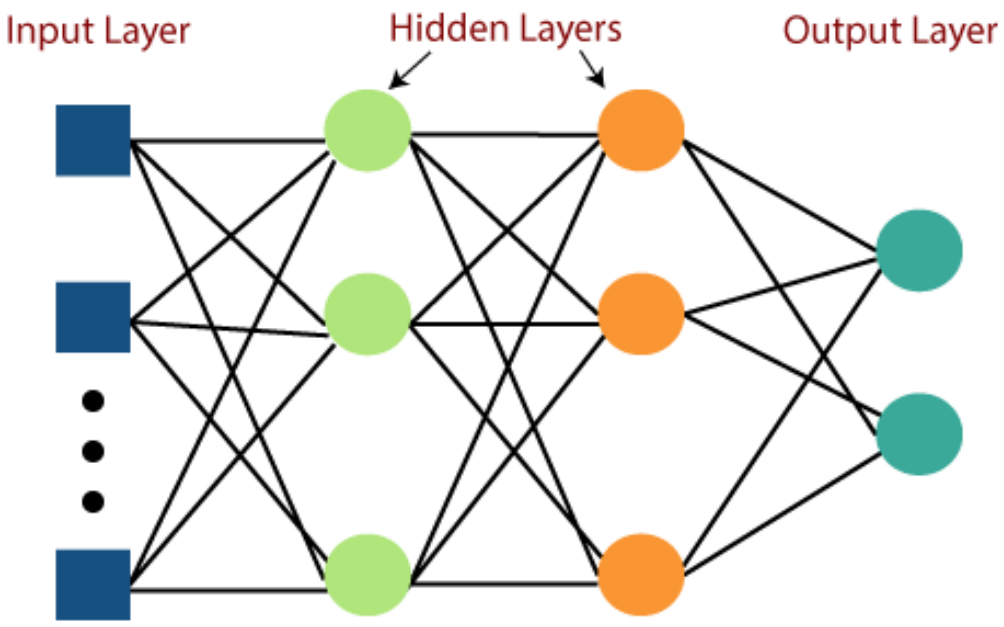


- Ridge: מודל Ridge הוא שיטת רגרסיה ליניארית שמשתמשת ברגולריזציה כדי למנוע בעיות של overfitting, במיוחד במצבים שבהם יש תלות בין המשתנים . המודל מוסיף מעין "קנס" למקדמים, שמקטין אותם לכיוון אפס, אך לא מבטל אותם לחלוטין. זה עוזר למנוע מורכבות יתר של המודל ולשפר את היכולת שלו לפעול טוב יותר על נתונים חדשים.

למידה עמוקה ומודלים גנרטיביים

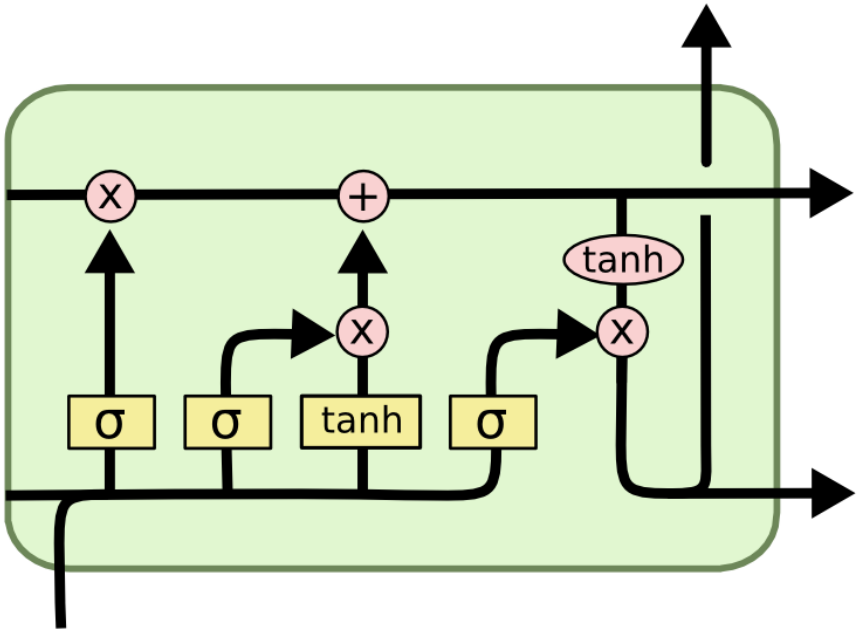
למידה עמוקה ומודלים גנרטיביים - MLP

- Multilayer Perceptron (MLP) : מודל ה-MLP הוא סוג של רשת עצבית מלאכותית המורכבת מכמה שכבות של צמתים מחוברים, או נוירונים. זהו מודל רשת עצבית שבו המידע זורם בכיוון אחד, משכבת הקלט דרך השכבות הסמויות לשכבת הפלט. ה-MLP מסוגל ללמוד תבניות מורכבות ויחסים לא ליניאריים, מה שהופך אותו לראוי למגוון רחב של משימות סיווג ורגרסיה.



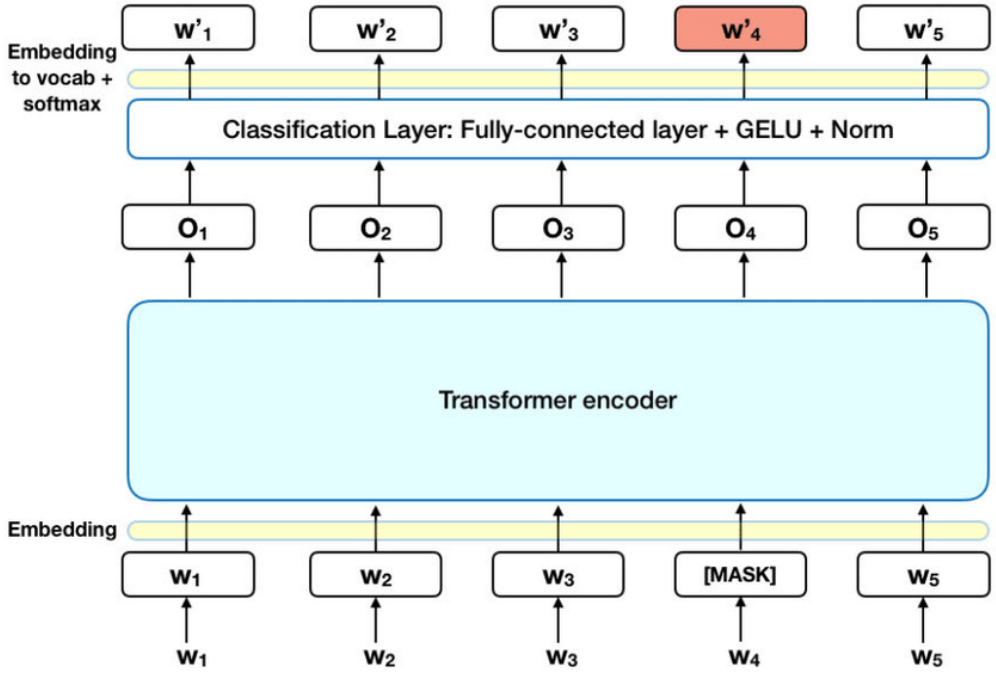
למידה עמוקה ומודלים גנרטיביים - LSTM

LSTM: Long Short-Term Memory (LSTM) הוא סוג של רשת עצבית המיועדת להתמודד עם קשרים לטווח ארוך בנתונים רציפים. בניגוד לרשתות חוזרות רגילות, LSTM משתמש במנגנונים שמאפשרים לו לזכור או לשכוח מידע באופן חכם. תכונה זו הופכת אותו למועיל בעיבוד שפה טבעית ובניתוח נתונים רציפים.



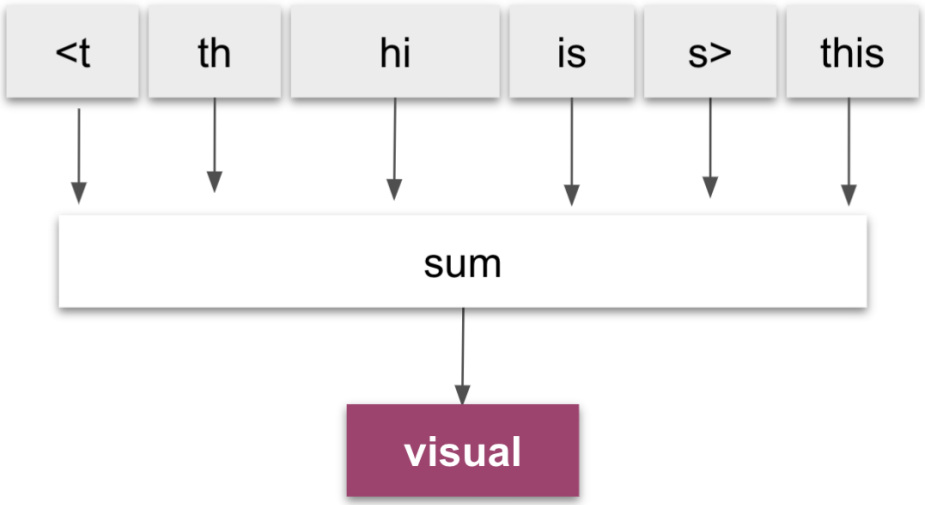
למידה עמוקה ומודלים גנרטיביים - BERT

- Bidirectional Encoder Representations from Transformers (BERT): מודל BERT הוא מודל לייצוג שפה המבוסס על ארכיטקטורת הטרנספורמר. הוא משתמש בשיטת אימון דו-כיוונית, מה שמאפשר לו להבין את ההקשר של מילים הן מהמילים לפני והן מהמילים אחרי. בזכות זה, BERT מציג ביצועים מרשימים במשימות עיבוד שפה טבעית, כמו סיווג משפטים, זיהוי ישויות ושאלות-תשובות.



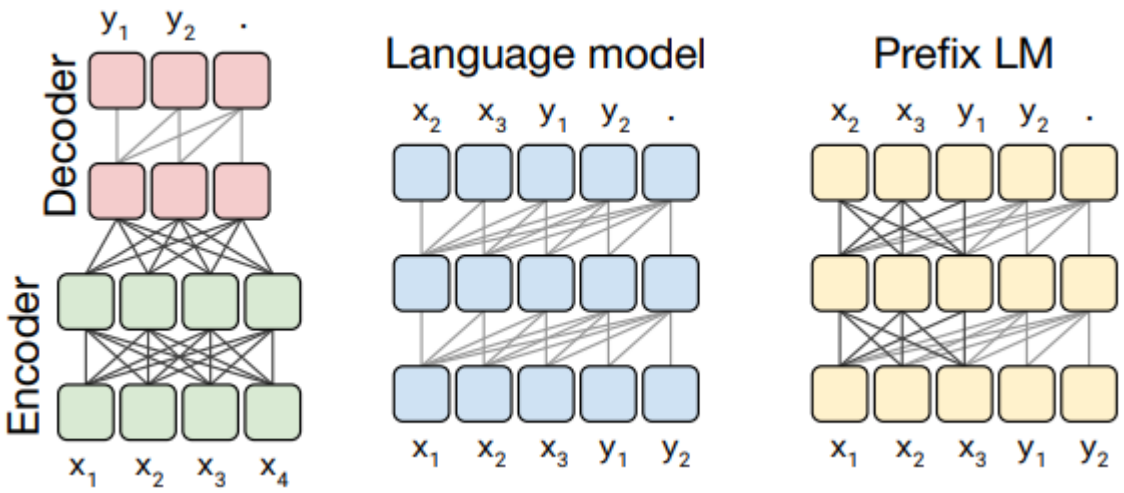
למידה עמוקה ומודלים גנרטיביים - FastText

- FastText הוא מודל סיווג טקסט המשתמש ב-word embeddings כדי לייצג טקסט. הוא מפרק מילים לתת-מילים, מאפשר טיפול במילים שאינן באוצר המילים, וידוע ביעילותו ודיוקו בסיווג טקסטים גדולים למשימות כמו ניתוח רגשות וקיטלוג נושאים.



למידה עמוקה ומודלים גנרטיביים - SimpleT5

- SimpleT5 הוא מודל מבוסס Transformers שמאפשר לאמן בקלות מודלי T5 בעזרת כמה שורות קוד. מודלים אלו רב-תכליתיים ומשמשים למשימות NLP שונות כגון סיכום, שאלות ותשובות, יצירת שאלות, תרגום ויצירת טקסט.



למידה עמוקה ומודלים גנרטיביים - Jurassic-2

- Jurassic-2 Grande Instruct הוא מודל שפה אוטו-רגרסיבי מבוסס טרנספורמר מבית AI21 Labs, המיועד לביצוע יעיל של בקשות zero-shot ו few-shot ללא צורך בדוגמאות. המודל מציע שיפורים על פני GPT-3, כולל גודל אוצר מילים ומבנה הרשת, ומאפשר אינטראקציה טבעית עם מודלים גדולים להשגת תוצאות אופטימליות.

למידה עמוקה ומודלים גנרטיביים - BLOOM

- BLOOM (BigScience Large Open-science Open-access Multilingual Language Model) : הוא מודל שפה אוטו-רגרסיבי מבוסס טרנספורמר עם decoder בלבד, שאומן על 366 מיליארד טוקנים מ-46 שפות טבעיות ו-13 שפות תכנות. המודל נבנה על מאגר נתונים של 1.6 טרה-בייט ומאפשר הבנה רחבה של הקשרים לשוניים ותכנותיים.

a BigScience initiative

BLOOM

176B params · 59 languages · Open-access

ארכיטקטורת Transformer

- Transformer המורכב רק ממפענח
- עיבוד טוקן אחד בכל פעם
- ממנגנוני תשומת לב עצמית ו feedforward

אימון לינארי

- המודל למד דפוסים וייצוגים של טקסט שעובד מראש
- מערך נתונים עצום של 366 מיליארד

יכולות מגוונות

- יכול לטפל במשימות בשפות שונות ואף בקוד
- יכול להכליל בין השפות ולתפוס ניואנסים לשוניים

יצירת טקסט אוטו רגרסיבית

- יוצר טקסט על ידי חיזוי מילה אחת בכל פעם, תוך שימוש במילים שהופקו קודם לכן כהקשר
- few-shot \ zero-shot

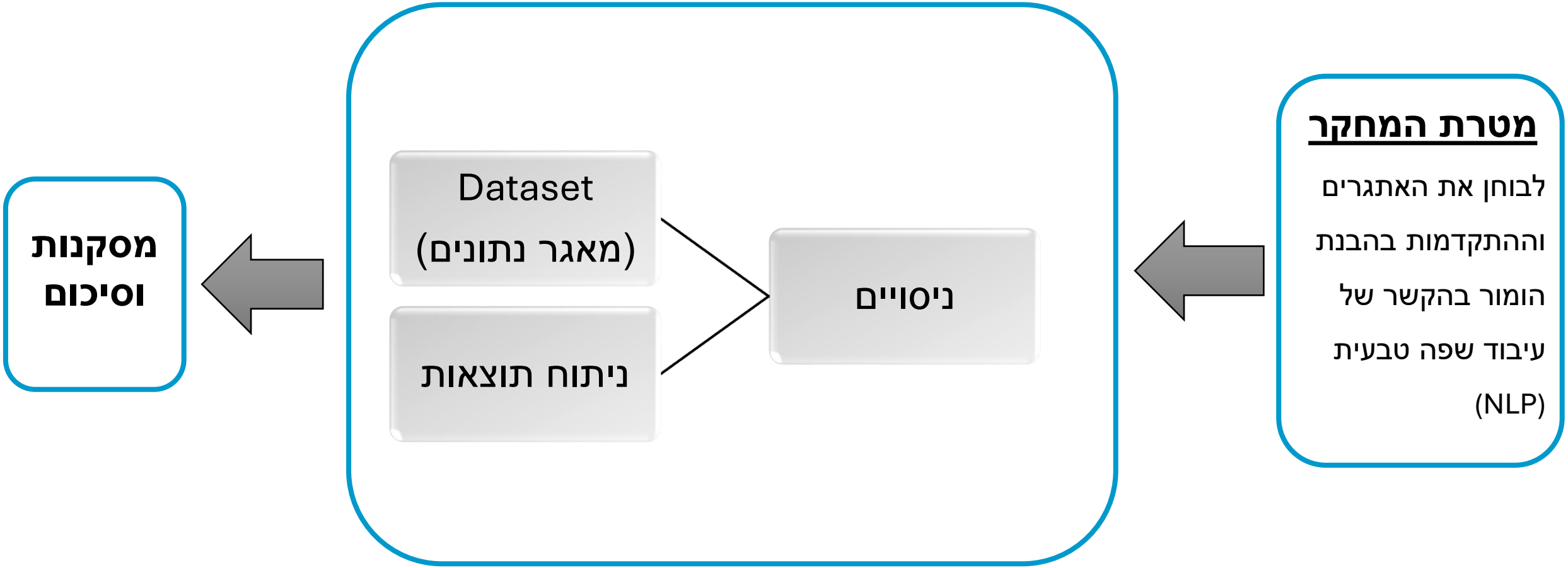
שימושים

- בעיבוד שפה טבעית (NLP) כגון תרגום, יצירת טקסט, סיכום, מענה על שאלות
- ומשימות בשפות תכנות כמו יצירת קוד

open access ייצוג פתוח

- BLOOM הוא במודל של גישה פתוחה. פתיחות זו מאפשרת לחוקרים ולמפתחים להשתמש ולשפר את BLOOM עבור יישומים ספציפיים

שלבי המחקר


מטרת המחקר
לבוחן את האתגרים
וההתקדמות בהבנת
הומור בהקשר של
עיבוד שפה טבעית
(NLP)
ניסויים
Dataset
(מאגר נתונים)
ניתוח תוצאות
מסקנות
וסיכום

מאגר נתונים 1

- האוסף במחקר כולל מעל 2,000 דוגמאות מתורגמות של משחקי מילים ממקורות שונים ,בעיקר באנגלית ובצרפתית .כל דוגמה מוינה וסווגה לפי סוגים שונים ,עם הערות שמסבירות את משמעות המילים או דרך הבנייה שלהן.
- במחקר הסתמכו על מסד נתונים מתויג מראש , הכולל שלושה עמודות : מזהה ,טקסט ויעד חיזוי שמשתנה בהתאם למשימה.
- בגילוי משחקי מילים ,היעד הוא אינדיקטור בינארי (כן/לא) . בזיהוי מיקום ,היעד הוא המילה שיוצרת את המשחק .בתרגום ,היעד הוא תרגום המשפט לצרפתית כדי להעריך שימור משחק המילים.

ID	PUN	PREDICTED TARGET		
		גילוי	איתור	תרגום

מאגר נתונים 2

Table 1
Task 1 and Task 2 dataset statistics

Language	Task 1		Task 2	
	Train	Test	Train	Test
English	5,292	3.183	2,315	1,205
French	3,999	12,873	2,000	4,655
Spanish	1,994	2,241	876	960

הפרטים של הנתונים ששימשו למשימה 1 ולמשימה 2 מוצגים בטבלה 1. מהטבלה, ברור כי מערכי הנתונים מכילים מספר לא שווה של דוגמאות עבור כל קטגוריית שפה. במשימה 1 יש הבדל גדול בין מספר הדגימות השליליות לדגימות החיוביות, כאשר יש יותר דגימות שליליות מאשר חיוביות.

Table 2
Task 1 dataset statistics

Language	Train		Test	
	Postive	Negative	Positive	Negative
English	3,085	2,207	809	2,374
French	1,998	2,001	5,308	7,565
Spanish	855	1,139	952	1,289

חוסר האיזון בנתונים בולט במיוחד במשימה 1, שבה יש הרבה יותר דגימות שליליות מאשר חיוביות. פיזור לא אחיד זה עלול לגרום לבעיות של הטיה ולהגביל את הדיוק של המודלים. עקב חוסר האיזון הזה, עשוי להיווצר קושי בחיזוי נכון של הקבוצה החיובית הקטנה יותר, מה שעלול להשפיע על הביצועים הכוללים ועל אמינות המודל.

מושגים חשובים ב- ML

		Actual Values	
		Positive	Negative
Predicted Values	Positive	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

Accuracy - (דיוק)

המדד הפשוט ביותר הוא דיוק שהוא היחס בין הסיווגים הנכונים לבין סך כל הסיווגים.

$$Accuracy = \frac{TN + TP}{TN + FN + TP + FP}$$

Recall -(רגישות)

היא הפרופורציה של דוגמאות חיוביות שהמודל זיהה מכל הדוגמאות החיוביות במציאות.

$$Recall = \frac{TP}{TP + FN}$$

Precision

הוא היחס של תצפיות חיוביות שהמודל זיהה נכונה מכל התצפיות שהמודל זיהה שהם חיוביות (בצדק או שלא בצדק).

$$Precision = \frac{TP}{TP + FP}$$

F1 score

מדד **F1** עושה ממוצע הרמוני של ה- Precision וה- Recall, ובכך לוקח בחשבון את השגיאות משני הסוגים.

$$F_1 = \frac{2}{\frac{1}{Recall} + \frac{1}{Precision}} = 2 \times \frac{(Precision \times Recall) \times 1}{(Precision \times Recall) \times \frac{Precision + Recall}{Precision \times Recall}} = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

F1 score	Interpretation
> 0.9	Very good
0.8 - 0.9	Good
0.5 - 0.8	OK
< 0.5	Not good

ניתוח התוצאות 1

טבלאות 3 ו-5 מסכמות את הביצועים של מודלי הלמידה וה-NLP, ומציגות מדדים כמו דיוק (Accuracy), F1, רגישות (Recall) ודיוק חיובי (Precision) שהושגו במהלך האימון.

התוצאות מראות שהמודלים אומנו היטב על הנתונים והפגינו ביצועים טובים על קבוצת האימות, עם דיוק גבוה, ערכי F1 מאוזנים, ורגישות ודיוק חיובי טובים. המשמעות היא שהמודלים למדו את דפוסי הנתונים בצורה יעילה והניבו תוצאות מבטיחות.

Table 3
Accuracy, Precision, Recall and F1-Score on the Training Data-set of Task 1 (1.1).

Model	Precision	Recall	F1-Score	Accuracy
Jurassic-2	0.51	0.07	0.14	0.41
BLOOM	0.58	0.05	0.01	0.41
FastText	0.72	0.84	0.78	0.72
RF-TFIDF	**0.99**	**0.99**	**0.99**	**0.99**
ST5	0.74	0.92	0.86	0.77
TFidfRidge	0.87	0.97	0.92	0.90

Table 5
Accuracy score on Train Data set of Task 1 (2.1).

Model	Accuracy
Ai21	0.42
BLOOM	0.36
ST5	**0.85**

ניתוח התוצאות 2

Table 4
Accuracy, Precision, Recall and F1-Score on the Test Data-set Task1 (1.1).

Model	Precision	Recall	F1-Score	Accuracy
Jurassic-2	0.27	0.09	0.019	0.74
BLOOM	**0.30**	0.03	0.07	**0.74**
FastText	0.25	0.80	0.39	0.35
RF-TFIDF	0.25	0.83	0.39	0.34
ST5	0.26	**0.93**	**0.41**	0.34
TFidfRidge	0.26	0.93	0.41	0.34

עם זאת, תוצאות הבדיקה מצביעות על ביצועים ירודים. סביר להניח שהדבר נובע ממערך נתונים בלתי מאוזן מאוד, מה שמקשה על המודל להכליל ולחזות במדויק את הקבוצות הפחות נפוצות.

חשוב להתמודד עם חוסר האיזון בנתונים כדי לשפר את ביצועי המודל על נתונים חדשים. טכניקות כמו דגימה מוגברת (oversampling), מתן משקל שונה לקבוצות או שימוש באלגוריתמים ייעודיים יכולות לסייע למודל להתמודד טוב יותר עם נתונים בלתי מאוזנים, ולהפוך את התחזיות שלו לאמינות וחזקות יותר.

Table 6
Accuracy score on Test Data set of Task 1 (2.1).

Model	Accuracy
Ai21	0.43
BLOOM	0.46
ST5	**0.80**

מסקנות וסיכום

פרויקט JOKER השיג התקדמות משמעותית בהבנת השפה היצירתית, בעיקר בתחום ההומור ומשחקי המילים. מאגר הנתונים JOKER סיפק תובנות חשובות לגבי סוגים שונים של משחקי מילים. המודל שפיתחו במחקר הציג ביצועים חזקים בחיזוי הומור בשפות שונות, תוך התחשבות בתכונות לשוניות ייחודיות כמו הומופונים ומשמעויות כפולות.

ביקורת

1. אי התייחסות להבדלים בין המודלים השונים ולגורמים להבדלים אלו בניתוח התוצאות מהווה חיסרון משמעותי. הבנת ההבדלים בין המודלים ומקורם יכולה לתרום לשיפור מחקרים עתידיים, לאפשר דיוק גבוה יותר בהסקת מסקנות ולשפר את איכות המחקר
2. מגוון השפות שנכלל במחקר היה מצומצם מדי לדעתי, והתמקד באנגלית, צרפתית וספרדית בלבד. הרחבת מגוון השפות יכולה להעשיר את המחקר ולאפשר בחינה רחבה ומעמיקה יותר של התופעות הנחקרות, במיוחד כשמדובר בתופעות לשוניות שעשויות להשתנות משמעותית בין שפות שונות.
3. לא היה שיתוף פעולה עם חוקרי שפה שמבינים את המורכבויות וההבדלים בין שפות, כמו גם את התופעות המשותפות להן, יכול לתרום משמעותית לניתוח מעמיק ומדויק יותר של הנתונים. הבנה מעמיקה של המאפיינים הייחודיים לכל שפה והשוואה בין תבניות לשוניות יכולה להעשיר את ממצאי המחקר ולהוביל לתובנות חדשות ואיכותיות יותר.

Thank you!

GRACIAS

Merci!