גירסה 23.8
16/01/2022

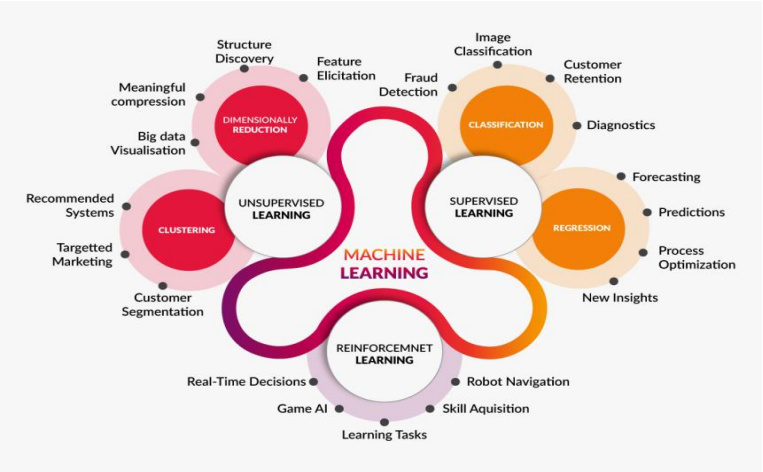
השוואה וניתוח אלגוריתמים ללימוד מכונה כדי לסווג את "همزتي الوصل والقطع"בשפה הערבית

פתרון הבעיה של שימוש שגוי ב- "همزتي الوصل والقطع" בטקסטים בשפה הערבית

קוד פרויקט: לדוגמא 21220570

דוח התקדמות - פרויקט גמר תשפ"ב

מנחה אקדמי : דר' יורם סגל



מגיש

שם סטודנט : 025879123 נזאר מערוף

חתימת מנחים

תוכן עניינים

1 תקציר 2
1.1 תקציר בעברית 2
1.2 תקציר באנגלית 3
1.3 רשימת מילות מפתח 4
2 טבלאות ומונחים 4
2.1 טבלת שרטוטים 4
2.2 טבלת איורים 4
2.3 מילון מושגים 5
3 מבוא 7
3.1 הגדרת הבעיה 10
3.2 האתגר הטכנולוגי 10
4 דרכי פתרון הבעיה 11
5 תוצר מצופה מהפרויקט 12
6 תיאור רעיון דומה שיכול להוות השראה 13
7 סיכונים, אי וודאות ואילוצי הפרויקט 13
8 עבודות בנושא 14
9 מה בוצע עד כה הפרויקט 15
10 תכנון 22
10.1 טבלת אבני דרך 22
10.2 טבלת משימות 23
10.3 סכמת בלוקים 24
..... 24
10.4 גאנט 25
11 מקורות קריאה 25
מכתב חוות דעת עבור חשיבות ותרומה של פרויקט לשפה וספרות הערבית 27

1 תקציר

1.1 תקציר בעברית

השפה הערבית נחשבת לאורגניזם חי גדל ומתפתח באמצעות תרגול ויישום נכון של כל ספרותיה וענפיו התחביריים, המורפולוגיים, הסמנטיים והלקסיקליים. בפרויקט זה, נסקור את תרומת הטכנולוגיה לפיתוח השפה הערבית, במיוחד הכתיבה הנכונה של "همزتي الوصل والقطع" . פרויקט זה נועד לבנות מודל מסווג חכם לסיווג מילים ערביות המתחילות באות "أ" ו- "همزتي الوصل والقطع" [1] באמצעות שימוש בטכניקות בינה מלאכותית ובכמה אלגוריתמים של למידת מכונה כדי לקבוע קריטריונים מדויקים ונכונים בכתיבת "همزتي الوصل والقطع" כתוצאה מכך, הטכנולוגיה תהיה מותאמת לתרום לשירות השפה הערבית נכון. בפרויקט זה הסתמכתי על כמה ספרים בשפה וספרות הערבית [2] כדי למצוא ולבנות מאגר מילים בערבית המתחילות " بالهمزة " [3].המאגר הכיל 477 מילים המתחילות " بالهمزة " ונשאר לסווגן ל- "همزتي الوصل والقطع" על פי הכללים הדקדוקיים הננקטים בתהליך זה ,לכן שלחתי למומחים .

המאגר הועבר ונבדק על ידי שישה מומחים בתחביר בדרגים אקדמיים שונים ,באיור(1) למטה רואים אחד מהמומחים עם מכתב על חשיבות פרויקט זה לשפה הערבית. המספר הכולל של המילים המסווגות הגיע ל-477 מילים, ולאחר עיבוד ואי הכללה של המילים החוזרות, 101 מילים, השגתי 376 [4] מילים תקפות ליישום במודל המסווגן, ובהתבסס על גודל וסוג הנתונים שאספתי ומנגנון הסיווג לאחר מכן, הוחלו אלגוריתמי סיווג שמתאימים לדגימה שנאספה, כגון אלגוריתמי התמיכה הוויקטורית (VSM), אלגוריתמי(NB) של Naif Biz ואלגוריתמים Nearest neighbor (KNN) על ידי שימוש בשפת Python ובספריית sk-learn. לאחר הרצת אלגוריתמי הסיווג שבחרתי להשתמש בהם וקבלת התוצאות הסופיות, אחליט איזה אלגוריתם הוא הטוב ביותר לתהליך תיקון כתיבת "همزة الوصل وهمزة القطع" במאמרים בשפה הערבית .

איור 1- מכתב מדר' שחאדה הארון אחד המומחים והגאונים בשפה וספרות ערבית, מאשר את חשיבות ותרומת הפרויקט לשפה הערבית .

1.2 תקציר באנגלית

The Arabic language is considered a living organism that grows and develops through the correct practice and application of all its literature and its syntactic, morphological, semantic, and lexical branches. In this research study, we review the contribution of technology to the development of Arabic language, especially the correct writing of **Hamzat Alwasl and Hamzat Alqatae** "همزتي الوصل والقطع".
This project aims to build a smart Classifier model to classify Arabic words beginning with the letter alif and their **hamza** " الهمزة". into conjunctive and disjunctive through the using of artificial intelligence techniques in general and machine learning algorithms to establish accurate and correct criteria in writing the conjunctive and disjunctive hamza correctly.
Consequently, technology would be adapted to contribute to the service of Arabic language.
In this project I relied on several books in Arabic language and literature to find and build a database of Arabic words starting with "بالهمزة".
The database contains 477 words starting with "بالهمزة" and remains to be classified as "همزتي الوصل والقطع" according to the grammatical rules adopted in this process, so I sent to the experts.
The database was transferred and examined by six syntax experts at various academic levels , Figure (1) above shows one of the experts with a letter about the importance of this project to the Arabic language. The total number of classified words reached 477 words, and after processing and excluding the repeated words, 101 words, I obtained 376 valid words to be applied to the Classifier model, And based on the size and type of data I collected and the subsequent classification mechanism, classification algorithms appropriate to the collected sample were applied, such as Victor Support algorithms (VSM), Naif Biz algorithms (NB) and Nearest neighbor algorithms (KNN) using Python and sk-learn library. After running the classification algorithms, I have chosen to use and getting the results, I will decide which algorithm is best for the process of correcting writing " همزة الوصل وهمزة القطع" in articles in Arabic.

1.3 רשימת מילות מפתח

אינטליגנטי מלאכותי, אלגוריתמי למידת מכונה, "الهمزة", "الوصل", "القطع", אלגוריתמי סיווג, שפה ערבית , Correlation, עקומת ROC .

2 טבלאות ומונחים

2.1 טבלת שרטוטים

טבלה 1- טבלת קורלציה שמראה קשר בין עמודות הנתונים 10
טבלה 2- מתארת חלק מטבלת הנתונים הכללית שנתתי לה שם (ARABIC_WORD.CSV). 16
טבלה 3- מסביר את הכללים המשמשים לקביעת מיקומי ה-"همزتي الوصل والقطع" 18
טבלה 4- קידוד נתונים בקובץ הנתונים (DATA_ENCODING.CSV) 18

2.2 טבלת איורים

איור 1- מכתב מדר' שחאדה הארון אחד המומחים והגאונים בשפה וספרות ערבית, מאשר את חשיבות ותרומת הפרויקט לשפה הערבית 2
איור 2- הסבר על קורלציה 6
איור 3- מציג את המנגנון של סיווג טקסט באמצעות אלגוריתמים לסיווג נתונים. 8
איור 4- מציג את ששת השלבים של מודול אלגוריתמים סיווג הנתונים. 11
איור 5- מציג כלים של חומרה והשוואה ביניהם 12
איור 6- מציג את ששת השלבים של מודול אלגוריתמים סיווג הנתונים. 15
איור 7- ביצוע בדיקה באמצעות שפת פיתון לבדיקה שיש רצף בנתונים ללא מחסור. 17
איור 8- ייצוג נתונים באמצעות עורך VSCODE 19
איור 9- ייצוג נתונים באמצעות עורך VSCODE 19
איור 10- בניית מודלים לסיווג באמצעות העורך (VSCODE) 20
איור 11- סכמת בלוקים , חלוקת שלבי יישום מודלים של סיווג נתונים 24
איור 12- תרשים גאנט למשימות של הפרויקט 0
איור 13- מכתב חוות דעת 1

2.3 מילון מושגים

1- **הבינה המלאכותית:** מדע הבינה המלאכותית (Intelligence Artificial) הוא אחד מענפי מדעי המחשב, ואחד מעמודי התווך של תעשיית הטכנולוגיה בעידן המודרני שלנו, המכונה בקיצור AI. ניתן להגדיר את מדע הבינה המלאכותית כיכולת של מכונות ומחשבים לביצוע משימות המחקות במידה רבה את מה שהמוח עושה.

2- **בינה מלאכותית מוגבלת:** זה אחד מהסוגים שיכולים לבצע משימות ספציפיות וברורות כמו יישומי רכב בנהיגה עצמית, תוכנות לזיהוי דיבור, תמונות או שחמט, וסוג זה של בינה מלאכותית הוא הנפוץ ביותר.

3- **בינה מלאכותית כללית:** זהו אחד מהטיפוסים שיש להם יכולות חשיבה דומות ליכולת האנושית, שכן הוא גורם למכונה להיות מסוגלת לחשוב בעצמה ודומה מאוד לחשיבה האנושית, למעשה אין יישומים מעשיים לסוג זה, אלא רק מחקרים שצריכים מאמץ רב כדי להפוך אותם למציאות. שיטת רשתות עצבים היא אחד המודלים של בינה כללית מלאכותית, שכן היא עוסקת בייצור מערכת של רשתות עצביות למכונה הדומה לאלו שהיו אינו כלול בתודעה האנושית.

4- **בינה מלאכותית בלתי מוגבל:** בינה מלאכותית בלתי מוגבלת היא מהסוג שעלול לעלות על רמת האינטליגנציה האנושית, ויכולתה לבצע משימות בצורה שיכולה להיות טובה יותר מהיכולת של בני אדם מומחים בעלי ידע, ולסוג זה מאפיינים הכרחיים רבים, כגון: למידה, תכנון אוטומטי, היכולת לתקשר ולקבל את ההחלטה המתאימה, אבל המושג של בינה-על מלאכותית נחשב למושג היפותטי שלא קיים בזמננו.

5- **למידת מכונה:** למידת מכונה היא ענף של בינה מלאכותית, הכולל תכנון ופיתוח של אלגוריתמים וטכניקות המאפשרות למחשבים להיות בעלי תכונות "למידה". באופן כללי, הלמידה מתחלקת לשתי רמות: אינדוקטיבית ודדוקטיבית, כאשר הגישה הדדוקטיבית מוציאה כללים ושיפוטים כלליים מביג דאטה.

6- **אלגוריתמים לסיווג נתונים:** ישנם מספר אלגוריתמים לסיווג נתונים המתאימים ליישומי כריית נתונים בטקסט בקלות לאחר עיבודם, והם גם קלים לאימון, אם עם כמויות גדולות או קטנות של נתונים מסופקים. להלן נסקור את האלגוריתמים המפורסמים ביותר של למידת מכונה.

7- **אלגוריתם מכונות וקטור-תמיכה SVM:** אלגוריתם זה מכונה בקיצור (SVM) אלגוריתם תחת למידת מכונה המסתמך על מערך נתונים עם תוצאות ידועות מראש (ערכת אימון) באימון האלגוריתם כך שיוכל לנתח ולסווג כל סט חדש של נתונים או לקבוע את הנטיות שלו. , אלגוריתם זה פותח על ידי שני המדענים ולדימיר פאבניק ואלכס שרבוניקס ב-1963, לאחר מכן פותח על ידי קורינה קורץ ופאבניק ב-1993 ופורסם ב-1995.

8- **אלגוריתם נאיבי של בייס (Bayes Naive)**: אלגוריתם ה- Knife Base (NB) נחשב לאחד האלגוריתמים של למידת מכונה, והוא תלוי בכללי ההסתברות המותנית שגיבש המדען תומאס ביס היכן הוא מחשב את ההסתברות באמצעות מספר איטרציות הערכים והאיטרציות והשילובים של ערכים בנתונים הידועים מראש בתוצאות (נתוני אימון). מפרצי סכין כקבוצה של מסווגים הסתברותיים פשוטים המבוססים על ההנחה הכללית שכל התכונות אינן תלויות זו בזו בהתאם למחלקה הספציפית, וכן עבור הקלות והמהירות של היישום של מסווג זה, הוא נחשב לקו הבסיס בסיווג טקסטים ונחשב יעיל בתחומים רבים, אם כי ישנם מספר מסווגים אחרים בעלי דיוק גבוה יותר כמו מודל SVM, שבו המודל של Bayes Naive מפיץ את טקסטים לכל מחלקה באמצעות מודל הסתברותי עם הנחות בלתי תלויות, שיטה זו פופולרית מאוד בתחום סיווג הטקסט, שכן המסווג הבינארי הוא אחת השיטות הידועות ביותר של המודל Bayes Naive שהשתמש בייצוג רדיאלי דו-ערכי של טקסטים.

9- **אלגוריתם השכן הקרוב : K-Nearest Neighbor**: אלגוריתם KNN יכול לשמש כמסווג פשוט ויעיל לסיווג טקסטים. למסווג KNN יש שני חסרונות: המורכבות החישובית אם הדגימות דומות, והביצועים שלו מושפעים בקלות אם דגימות האימון אינדיבידואליות. ניתן להפחית את המורכבות של ה-KNN על ידי שימוש בשלוש שיטות: או על ידי הגבלת ממדי הווקטור המיוצג על ידי הטקסט, על ידי הגבלת כמות דגימות האימון, או על ידי הגבלת מציאת השכן הקרוב, כלומר הערך של k . KNN משתמש בסיווג טקסט על ידי חישוב המרחק בין הטקסט לכל הטקסטים במערך הנתונים של ההדרכה תוך שימוש במדד של

הבדל או דמיון ביניהם, ולאחר מכן מציאת ה-K הקרובה ביותר מבין כל טקסטי ההדרכה ובחירה בשיעור הטקסט לזה עם המספר הגדול ביותר של טקסטים בשכנים הקרובים ביותר הם סקריפטים, וכמו אלגוריתמים אחרים, הם שופרו ביותר מדרך אחת.

10- **עקומת ROC** : היא גרף המציג את הביצועים של מסווג דו-ערכי ,לאור סף ההחלטה שנקבע לו. העקומה נוצרת על ידי התוויית שיעור החיוביים האמיתיים (TPR) מול שיעור החיוביים הכוזבים (FPR) תחת ספי קבלה שונים. שיעור החיוביים האמיתיים ידוע גם כרגישות או כיסוי בלמידת מכונה .שיעור החיוביים הכוזבים ידוע גם כדלף וניתן לחשב אותו כ־1 פחות הסגוליות .עקומת ROC היא, אם כן, הרגישות כפונקציה של הדלף. באופן כללי, אם התפלגות ההסתברות ידועה הן לפגיעה ,Hit) חיובי אמיתי (והן לאזעקת שווא (חיובי כוזב), ניתן לייצר את עקומת ROC על ידי התוויית פונקציית הסתברות הפגיעות (שיעור חיוביים אמיתיים) בציר ה־ yלעומת פונקציית ההסתברות לאזעקת שווא (שיעור חיוביים כוזבים) בציר ה־.x

ניתוח ROC מספק כלים לבחירת מבחן (או מסווג) אופטימלי, בטרם מתחשבים בהקשר העלות של כל אחת מסוגי הטעויות או בהתפלגות הפרטים בין הקבוצות. ניתוח ROC קשור באופן ישיר וטבעי לניתוח עלות/תועלת של קבלת החלטות.

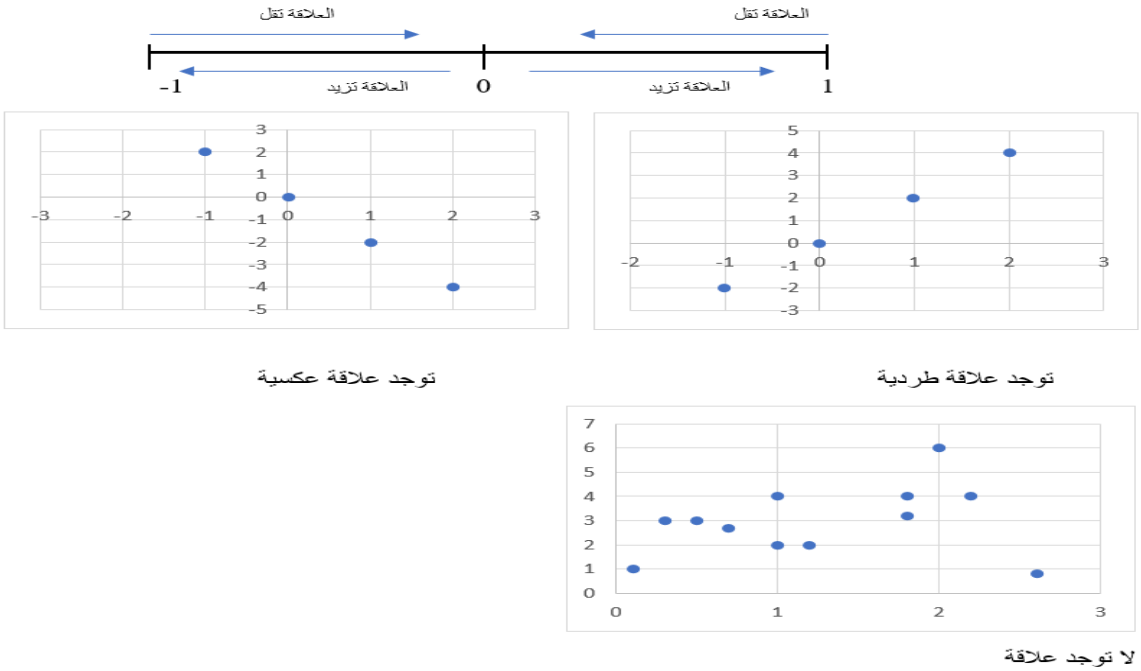
11- **מתאם(Correlation)**: זה הוא מבחן למדידת מידת הקשר בין שני משתנים. כלומר, בדיקה זו נותנת לנו ערך המעיד על נוכחות או היעדר מתאם בין שני משתנים, כלומר, אם יש שינוי בערכים של אחד המשתנים, הוא ילווה בעלייה או ירידה ב- הערכים של המשתנה האחר. ערך זה נקרא מקדם המתאם.

חשוב להבין שמקדם המתאם משנה את מידת הקשר בין שני משתנים מבחינה סטטיסטית ולא סיבתית לפי איור(1).

נימוק לשימוש במבחן זה:

אנו משתמשים במבחן מקדם המתאם של פירסון (קורלציה) כדי לחשב את הקשר בין שני משתנים מסוג מספרי רציף, בהנחה שהקשר בין שני המשתנים הוא ליניארי. אבל הנחה זו לא אומרת שכל הנקודות חייבות להיות ממוקמות על קו ישר, אלא שצורת הדיפוזיה מצביעה על כך שיש נטייה לנוכחות של סגירה ליניארית, לא עקומה. השערותיו הן כדלקמן:

- השערה אפס (H0) - אין קשר בין שני המשתנים במדגם הנבדק.
- השערה חלופית (H1) - קיים קשר בין שני המשתנים במדגם הנבדק.



איור 2- הסבר על קורלציה .

3 מבוא

ראשית: מושג הבינה המלאכותית:

מדע הבינה המלאכותית (Intelligence Artificial) הוא אחד מענפי מדעי המחשב, ואחד מעמודי התווך של תעשיית הטכנולוגיה בעידן המודרני שלנו, המכונה בקיצור AI. ניתן להגדיר את מדע הבינה המלאכותית כיכולת של מכונות ומחשבים לביצוע משימות המחקות במידה רבה את מה שהמוח עושה. האדם, המתאפיין באינטליגנציה, וניתן לסכם את המשימות הללו ביכולת לחשוב או ללמוד מניסיונותיו הקודמים, כך שאנו יכולים לומר כי בינה מלאכותית מכוונת להגיע למערכות שמתנהגות, לומדות ומבינות בזמן שהן פועלות, לומדות ומבינות בני אדם כפי שיש להן תכונה של אינטליגנציה [5].

סוגי בינה מלאכותית:

על פי יכולותיה, ניתן לחלק את הבינה המלאכותית לשלושה סוגים [5]. כדלקמן:

- **בינה מלאכותית מוגבלת:**

זה אחד מהסוגים שיכולים לבצע משימות ספציפיות וברורות כמו יישומי רכב בנהיגה עצמית, תוכנות לזיהוי דיבור, תמונות או שחמט, וסוג זה של בינה מלאכותית הוא הנפוץ ביותר.

- **בינה מלאכותית כללית:**

זהו אחד מהטיפוסים שיש להם יכולות חשיבה דומות ליכולת האנושית, שכן הוא גורם למכונה להיות מסוגלת לחשוב בעצמה ודומה מאוד לחשיבה האנושית, למעשה אין יישומים מעשיים לסוג זה, אלא רק מחקרים שצריכים מאמץ רב כדי להפוך אותם למציאות. שיטת רשתות עצבים היא אחד המודלים של בינה כללית מלאכותית, שכן היא עוסקת בייצור מערכת של רשתות עצביות למכונה הדומה לאלו שהיו אינו כלול בתודעה האנושית [6]

- **בינה מלאכותית בלתי מוגבל:**

בינה מלאכותית בלתי מוגבלת היא מהסוג שעלול לעלות על רמת האינטליגנציה האנושית, ויכולתה לבצע משימות בצורה שיכולה להיות טובה יותר מהיכולת של בני אדם מומחים בעלי ידע, ולסוג זה מאפיינים הכרחיים רבים, כגון: למידה, תכנון אוטומטי, היכולת לתקשר ולקבל את ההחלטה המתאימה, אבל המושג של בינה-על מלאכותית נחשב למושג היפותטי שלא קיים בזמננו.

ניתן לסווג בינה מלאכותית גם לפי ארבעת הפונקציות השונות הבאות:

- **מכונות אינטראקטיביות:**

זהו הסוג הפשוט ביותר של בינה מלאכותית מכיוון שאין לו את היכולת ללמוד מניסיון העבר כדי לפתח עסקים עתידיים, כך שכאן הוא ייצור אינטראקציה עם הניסיון הנוכחי כדי לייצר את הדרך הטובה ביותר.

- **זיכרון מוגבל:**

בינה מלאכותית בקטגוריית זיכרון מוגבל יכולה לאחסן נתונים היסטוריים על המערכת הנוכחית לפרק זמן מוגבל, וגישת הנהיגה האוטונומית היא אחת הדוגמאות הטובות ביותר לסגנון זה, שכן היא חוסכת את המהירות האחרונה של מכוניות אחרות, הממוצע המרחק בין המכוניות הללו לבין ההגבלה. המהירות המותרת ומידע נוסף החיוני לנהיגה בנתיבי מעבר. [7]

- **תורת הנפש:**

סוג זה של בינה מלאכותית פירושו שהמכונה סופגת רגשות אנושיים, מקיימת אינטראקציה עם בני אדם ומתקשרת איתם, ויש לציין שלא מצאנו עד לרגע זה יישומים מעשיים על סוג זה של בינה מלאכותית.

- **מודעות עצמית:**

קטגוריית המודעות העצמית נחשבת לאחת מהתחזיות העתידיות שהבינה המלאכותית שואפת אליהן, והיא פועלת על פי עיקרון טכני וחושי מאוד מודרני לפיו המכונה יכולה לייצר ידע עצמי ורגשות משלה, מה שיהפוך אותה לאינטליגנטית יותר. מאשר האדם, ומה מושג זה עדיין אינו נוכח במציאות. [8]

תחומי משנה של בינה מלאכותית:

מדע הבינה המלאכותית מכיל תתי תחומים רבים, כגון: למידת מכונה, הכוללת מתן אפשרות למחשבים ללמוד באופן עצמאי מכל ניסיון קודם, כך שמחשבים יכולים לחזות לקבל את ההחלטה המתאימה במהירות, על ידי פיתוח אלגוריתם המאפשר מצב זה. יש לציין שמונח זה הוצע לראשונה על ידי ארתור סמואל בשנת 1959. להלן נתייחס לכמה מתתי התחומים המפורסמים ביותר של בינה מלאכותית כדלקמן:

- **חשיבה לוגית והסתברותית:**

חשיבה לוגית בבינה מלאכותית היא אחת מצורות ההנמקה השונות, מכיוון שעובדות מתקבלות על סמך הנתונים הזמינים. חשיבה לוגית תואמת למה שנקרא חשיבה הסתברותית, המשתמשת במושגים של הסתברות ואי ודאות בידע כדי להתמודד עם כל אי הוודאות העתידית של כל האירועים שעלולים לחשוד. [9]

- **למידת מכונה:**

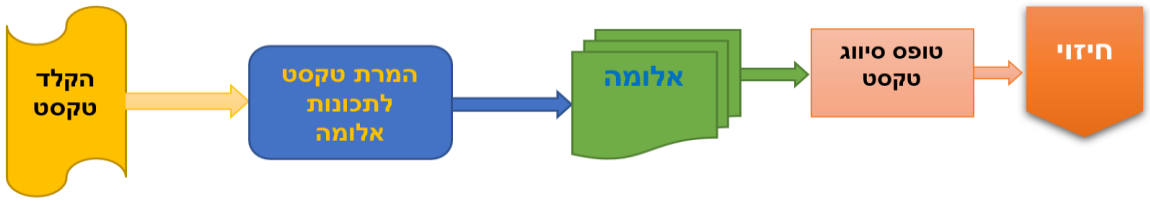
למידת מכונה היא ענף של בינה מלאכותית, הכולל תכנון ופיתוח של אלגוריתמים וטכניקות המאפשרות למחשבים להיות בעלי תכונות "למידה". באופן כללי, הלמידה מתחלקת לשתי רמות: אינדוקטיבית ודדוקטיבית, כאשר הגישה הדדוקטיבית מוציאה כללים ושיפוטים כלליים מביג דאטה.

שנית: הרעיון של למידת מכונה:

המשימה העיקרית של למידת מכונה היא לחלץ מידע בעל ערך מהנתונים, ולכן היא קרובה מאוד לכריית נתונים. למידת מכונה משמשת בתחום ניתוח הנתונים ומהווה שיטה לפיתוח מודלים מורכבים ואלגוריתמים מתאימים להפקת נתונים באמצעות תהליכי חיזוי ניתוח זה נקרא ניתוח חזוי. מודלים אנליטיים אלו מאפשרים לחוקרים ולמנתחי נתונים ללמוד החלטות ותוצאות אמינות ומסוגלים להבין נתונים מאוחסנים והקשרים ביניהם.

ניתן להגדיר מערכות למידת מכונה גם כמערכות המבצעות חיזויים על סמך מה שהנתונים הקודמים למדו. מערכות אלו זקוקות לאימון על דוגמאות רבות של טקסט וחיזויים (סימנים) הצפויים לכל אחת מהן. הנתונים המשמשים לאימון נקראים אימון מערך נתונים. נתונים אלו מסווגים מראש עם תכונות ובכל פעם ככל שסט האימונים מדויק יותר והתכונות שנבחרו מתאימות, כך תחזיות המסווג טובות יותר. כאשר מסווג מאומן בשיטת למידת מכונה, נתוני האימון חייבים להיות טובים יותר. המרה למשהו שהמכונה יכולה להבין. התכונות נשלפות ומומרות לאלומות (ייצוג טקסטים לפי מספרים) שיעזרו לה ללמוד מנתונים קיימים ולבצע תחזיות לגבי טקסטים עתידיים. [9]

המודל המאומן יכול לחלץ תכונות מהטקסט החדש ולחזות או לסווג את הטקסטים לפי מאפיינים ספציפיים באמצעות אלגוריתמים לסיווג נתונים כפי שמוצג באיור (1-1) להלן:



איור 3- מציג את המנגנון של סיווג טקסט באמצעות אלגוריתמים לסיווג נתונים.

שלישית: אלגוריתמים לסיווג נתונים:

ישנם מספר אלגוריתמים לסיווג נתונים המתאימים ליישומי כריית נתונים בטקסט בקלות לאחר עיבודם, והם גם קלים לאימון, אם עם כמויות גדולות או קטנות של נתונים מסופקים. להלן נסקור את האלגוריתמים המפורסמים ביותר של למידת מכונה. כדי לסווג נתוני טקסט ששימשו בפרויקט זה:

1- אלגוריתם מכונות וקטור-תמיכה SVM:

אלגוריתם זה מכונה בקיצור (SVM) אלגוריתם תחת למידת מכונה המסתמך על מערך נתונים עם תוצאות ידועות מראש (ערכת אימון) באימון האלגוריתם כך שיוכל לנתח ולסווג כל סט חדש של נתונים או לקבוע את הנטיות שלו. , אלגוריתם זה פותח על ידי שני המדענים ולדימיר פאבניק ואלכס שרבוניקס ב-1963, לאחר מכן פותח על ידי קורינה קורץ ופאבניק ב-1993 ופורסם ב-1995. [10]

2- אלגוריתם נאיבי של בייס (Bayes Naive)

אלגוריתם ה- Knife Base (NB) נחשב לאחד האלגוריתמים של למידת מכונה, והוא תלוי בכללי ההסתברות המותנית שגיבש המדען תומאס ביס [11] היכן הוא מחשב את ההסתברות באמצעות מספר איטרציות הערכים והאיטרציות והשילובים של ערכים בנתונים הידועים מראש בתוצאות (נתוני אימון). מפרצי סכין כקבוצה של מסווגים הסתברותיים פשוטים המבוססים על ההנחה הכללית שכל התכונות אינן תלויות זו בזו בהתאם למחלקה הספציפית, וכן עבור הקלות והמהירות של היישום של מסווג זה, הוא נחשב לקו הבסיס בסיווג טקסטים ונחשב יעיל בתחומים רבים, אם כי ישנם מספר מסווגים אחרים בעלי דיוק גבוה יותר כמו מודל SVM, שבו המודל של Bayes Naive מפיץ את טקסטים לכל מחלקה באמצעות מודל הסתברותי עם הנחות בלתי תלויות, שיטה זו פופולרית מאוד בתחום סיווג הטקסט, שכן המסווג הבינארי הוא אחת השיטות הידועות ביותר של המודל Bayes Naive שהשתמש בייצוג רדיאלי דו-ערכי של טקסטים.

3- אלגוריתם השכן הקרוב : K-Nearest Neighbor

אלגוריתם KNN יכול לשמש כמסווג פשוט ויעיל לסיווג טקסטים. למסווג KNN יש שני חסרונות: המורכבות החישובית אם הדגימות דומות, והביצועים שלו מושפעים בקלות אם דגימות האימון אינדיבידואליות. ניתן להפחית את המורכבות של ה-KNN על ידי שימוש בשלוש שיטות: או על ידי הגבלת ממדי הווקטור המיוצג על ידי הטקסט, על ידי הגבלת כמות דגימות האימון, או על ידי הגבלת מציאת השכן הקרוב, כלומר הערך של k.

KNN משתמש בסיווג טקסט על ידי חישוב המרחק בין הטקסט לכל הטקסטים במערך הנתונים של ההדרכה תוך שימוש במדד של הבדל או דמיון ביניהם, ולאחר מכן מציאת ה-K הקרובה ביותר מבין כל טקסטי ההדרכה ובחירה בשיעור הטקסט לזה עם המספר הגדול ביותר של טקסטים בשכנים הקרובים ביותר הם סקריפטים, וכמו אלגוריתמים אחרים, הם שופרו ביותר מדרך אחת. [7]

רביעית: החמץ אל-ואסל וקט '

בחלק זה של הצד התיאורטי, נעסוק ב-"همزة الوصل وهمزة القطع" בתחילת המילה למטה

העדות הבאות:

משמעות " الهمزة " (לשונית):

הוזכר במילוני לשון רבים, כולל ליסאן אל-ערב מאת אבן מנצור, " الهمزة " היא בשפה במובן של קריצה, מלמול ולחץ. [12]

המשמעות של " الهمزة " (ניב):

" الهمزة " במינוח הבלשנים יש כמה הגדרות, ביניהן ההגדרה של אל-אזהרי בספרו "תהדהיב אל-לוגה", שם הוא אומר: "דע ש" الهمزة " אינה כתיב עבורה, אלא כתובה מעל "אליף" פעם, מעל "וואו (و) "פעם , ומעל "יאא (ي)", והאליף הרך אין לו אות, אלא הוא דגש פתח לאחר

פתח, והאותיות הן עשרים ושמונה אותיות עם ה" וואו", "אליף" ו" יאא", עם " الهمزة " נהיו עשרים- ותשע אותיות, ו" الهمزة " היא כמו האות האמיתית, אלא שיש לה רישות מריכוך, מחיקה, החלפה והקלה... זה לא מאותיות שיוצאות מחלל הפה, אלא יוצאת מהגרון מקצה הפה"

מקטעים של " الهمزة " ה בתחילת המילה:

" الهمزة " בתחילת המילה מחולקת לשני חלקים:

- המזת אלוואסל (همزة الوصل):

המזת אלוואסל נקראת בשם זה, משום שהיא מביאה להגייה של האות העיצורית בתחילת הדיבור, שכן הכלל החשוב בכך הוא שהיא אינה מתחילה בעיצור ואינה מסתיימת בתנועה.

- המזת אלקאטע (همزة القطع):

המזת אלקאטע נקראת בשם זה כי היא חותכת ותופסת את האותיות שלפניה ומה שאחריה.

3.1 הגדרת הבעיה

הבעיה העיקרית של הפרויקט נעוצה בשימוש לא נכון ב-"همزتي الوصل أو القطع" במקום הלא נכון, במיוחד כשכותבים מילים המתחילות באות "أ", שכן רבים טועים בכתיבת "همزتي الوصل أو القطع, מה שמחליש את חוזק השפה. פרויקט זה תורם לפיתוח השפה הערבית, על ידי בניית מודל סיווג חכם עבור עורכים, מתרגלים וכל משתמשי השפה מהסיווג הנכון של "همزتي الوصل أو القطع". בפרויקט הזה זה ננסה לענות על השאלות הבאות לאור התועלת של אלגוריתמי למידת מכונה המשמשים בסיווג וחיזוי:

- כיצד נוכל להבחין בין "همزتي الوصل أو القطع" בטקסטים בערבית?
- מהם היתרונות שנסיג בעת סיווג "همزتي الوصل أو القطع"בטקסטים בערבית?
- מהם האלגוריתמים המתאימים של למידת מכונה לסיווג "همزتي الوصل أو القطع" בטקסטים בערבית?

3.2 האתגר הטכנולוגי

אתגר ראשון בפרויקט זה שהוא אחד מהפרויקטים החדשים בתחום המחקרים הדקדוקיים והטכניים העוסקים בשימוש באלגוריתמים של למידת מכונה לסיווג "همزتي الوصل والقطع" בטקסטים בערבית. למיטב ידיעתי, ובאמצעות מחקריי, לא מצאתי פרויקטים שעסקו בנושא זה בעבר בשל הקושי בשימוש בספריות תוכנה בהתמודדות עם טקסטים בערבית כראוי, אך יש דמיון רב וחפיפה למחקרים אחרים הקשורים לפרויקט לשימוש באלגוריתמים אלה ביישומים מקבלים ודומים, ואני הרווחתי מהם. האתגר השני היא בניית טבלת נתונים שאפשר להשתמש בה בפרויקט שלי ובדיקת קשר קורלציה כפי שרואים בטבלה(1) בין עמודות הנתונים .והאתגר השלישי קידוד טבלת הנתונים והפיכתה למספרים שאלגוריתמים ושפת תכנות יכולים לקרוא ולעבוד איתם.

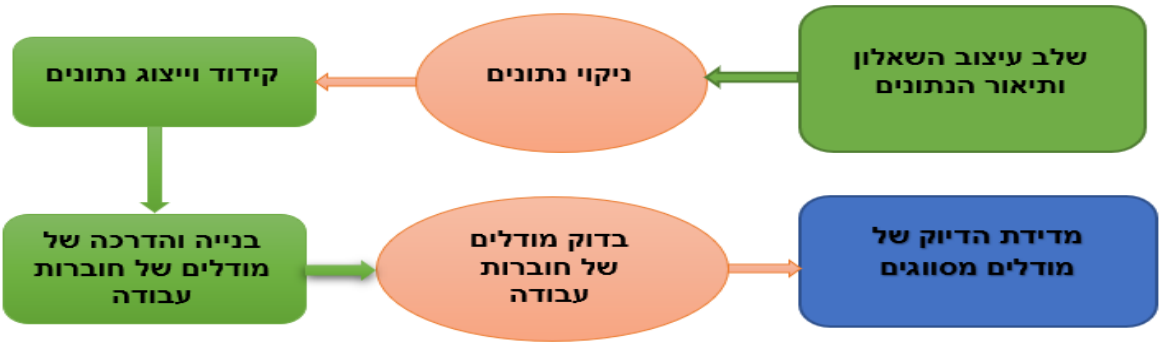
טבלה 1- טבלת קורלציה שמראה קשר בין עמודות הנתונים .

excel.ipynb > Cor (9, 9)

	index	Count	Noun	Verb	Letter	adject…	morpho…	The	diacri…	Outcome
0	Count	1	0.273505…	0.040448…	0.275463…	0.050252…	0.472810…	0.255007…	0.092353…	0.428777…
1	Noun	0.273505…	1	0.668436…	0.165126…	0.114964…	0.124209…	0.292223…	0.086618…	0.180214…
2	Verb	0.040448…	0.668436…	1	0.143876…	0.128155…	0.092737…	0.229672…	0.069620…	0.398138…
3	Letter	0.275463…	0.165126…	0.143876…	1	0.037011…	0.215230…	0.020144…	0.069959…	0.156864…
4	adjective	0.050252…	0.114964…	0.128155…	0.037011…	1	0.084649…	0.007658…	0.154404…	0.133641…
5	morpholo…	0.472810…	0.124209…	0.092737…	0.215230…	0.084649…	1	0.074084…	0.223871…	0.306152…
6	The	0.255007…	0.292223…	0.229672…	0.020144…	0.007658…	0.074084…	1	0.312592…	0.287442…
7	diacritic	0.092353…	0.086618…	0.069620…	0.069959…	0.154404…	0.223871…	0.312592…	1	0.072828…
8	Outcome	0.428777…	0.180214…	0.398138…	0.156864…	0.133641…	0.306152…	0.287442…	0.072828…	1

4 דרכי פתרון הבעיה

הנתונים בימינו גדלים מיום ליום ומקורותיהם מרובים, והדבר מביא לחשיפה של נתונים אלו לבעיות רבות המפחיתות את איכות הנתונים כמו ריבוי הנתונים החסרים ואי העקביות של הנתונים, אז בפרויקט הזה חילקתי את שלבי יישום מודלים של מסווג נתונים לשישה שלבים שהתחילו את השלב של עיצוב השאלון ואיסוף הנתונים והסתיים בשלב של מדידת הדיוק של מודלים מסווגים, כפי שמוצג באיור הבא (1-2):



איור 4- מציג את ששת השלבים של מודול אלגוריתמים סיווג הנתונים.

בחלק זה של הפרויקט אכסה את ארבעת השלבים הראשונים של יישום מודלים מסווגים ובפרק הבא אתייחס לשניים האחרונים.

- **שלב עיצוב השאלון ותיאור הנתונים:**

בפרויקט זה הסתמכתי על כמה ספרים בשפה וספרות הערבית [2] כדי למצוא ולבנות מאגר מילים בערבית המתחילות " بالهمزة " [3].המאגר הכיל 477 מילים המתחילות " بالهمزة " ונשאר לסווגן ל- "همزتي الوصل والقطع" על פי הכללים הדקדוקיים הנקטים בתהליך זה ,לכן שלחתי למומחים .

המאגר הועבר ונבדק על ידי שישה מומחים בתחביר בדרגים אקדמיים שונים. המספר הכולל של המילים המסווגות הגיע ל-477 מילים, ולאחר עיבוד ואי הכללה של המילים החוזרות, 101 מילים, השגתי 376 [4] . המילים שהופקו מהטקסטים לדוגמה סווגו למילים המתחילות בהמזה (الهمزة) המכונה בפרויקט זה (Wasl) (همزة الوصل) ומילים המתחילות בהמזה (الهمزة) המכונה בפרויקט זה(Gtaa) (همزة القطع) המוזכרות בפרויקט זה. המשתנה התלוי (Outcome) נקבע על פי השניים ערכים (Gtaa/Wasl).

- **שלב ניקוי הנתונים:**

אין ספק שכאשר איכות הנתונים נמוכה, הדבר ישפיע בהכרח על תוצאות הניתוח. בפרויקט זה נעשה שימוש במספר שיטות ניקוי נתונים על הטקסטים שנאספו.
שלב ניקוי הנתונים כלל את השלבים הבאים:

- התמודדות עם נתונים שאבדו
- מחיקת נתונים כפולים.

לאחר ביצוע פעולות ניקוי נתונים על הטקסטים שנאספו, ולאחר עיבוד והוצאת מילים כפולות, וגם השלמת הנתונים החסרים(מספרם 101),לכן בסוף קיבלנו 299 מילים תקפות ליישם את מודל חוברת העבודה.

- **שלב קידוד וייצוג נתונים:**

לאחר לימוד הכללים הדקדוקיים המציגים את מיקומם של "همزتي الوصل والقطع" בתחילת המילה, נקבעו היסודות והתכונות שניתן להסתמך עליהם בקביעת ערכו של המשתנה.

הערכים של הנתונים המילוליים מומרים לערכים מספריים כדי שהאלגוריתמים יוכלו להתמודד עם זה והנתונים הופכים למקודדים.

לאחר השלמת תהליך קידוד הנתונים, הנתונים מיוצגים באמצעות שפת Python והספריות (sklearn-scipy-numpy) מיובאות באמצעות עורך Vscode.

- **שלב בניית והדרכת דגמי מסווגים:**

הנתונים המיוצגים חולקו לנתוני אימון ולנתוני ניסוי כהקדמה לבניית מודל נתוני אימון באמצעות אלגוריתמי הסיווג שנבחרו: אלגוריתם תמיכת וקטור (SVM), אלגוריתם (NB) ואלגוריתם השכן הקרוב ביותר (KNN). המודל נבנה באמצעות פונקציות הספרייה (Learn-Sk).

5 תוצר מצופה מהפרויקט

מטרת הפרויקט שהושלם זה היה לבנות מודל חכם שמסווג את "همزتي الوصل والقطع" (בתחילת מילה) באמצעות שימוש באלגוריתמים של סיווג נתונים על מנת לקבוע קריטריונים מדויקים ונכונים בכתיבת טקסטים בערבית במדויק כדי לתרום ולעזור להתאים את הטכנולוגיה למדידת האיכות של אלגוריתמי הסיווג המפורסמים ביותר בהבחנה בין השירות של השפה הערבית, הפרויקט כוון גם בין "همزتي الوصل والقطع" בתחילת המילה. המודל החכם הזה שעוצב יכול לתרום לפיתוח השפה הערבית כדלקמן:

- ניתן להשתמש במודל החדש לסקור מחקר מדעי ולוודא"الهمزة" כתובה בצורה נכונה במילים, מה שתורם להשלמת מרכיבי המחקר המדעי.

- ניתן להשתמש בטופס החדש לעיון במאמרי חדשות המתפרסמים במדיות החברתיות השונות על ידי התקנתו בחלק ההרחבות של הדפדפן על מנת להבטיח "الهمزة" כתובה כהלכה בכל המילים של הכתבות שפורסמו, מה שתורם לפיתוח תקשורת ערבית חדשה.

- ניתן להשתמש במודל החדש ביישומי סיור ולהיחשב כחלק מקורי ממערכות ההפעלה הסלולריות כדי להבטיח שהפקודות המתורגמות לערבית והתחילה "بالهمزة" ייכתבו בצורה נכונה.

- ניתן לשלב את המודל החדש עם מערכות תרגום המשמשות בכנסים, מטוסים ורכבות, שכן הוא מסייע בהצגת "الهمزة" בצורה נכונה.

לסיכום, אנו יכולים להשתמש במודל החדש ככלי תוכנה שניתן לשלב עם כל התוכנה והמערכות הטכניות המציגות טקסטים בערבית, קריאה או כתובה, באמצעות מערכות תוכנה לסיווג " الهمزة" במילים שמתחילות " بهمزة وصل أو قطع".

כלים וציוד

תוכנה:

- תכנות Python בסביבת פיתוח Vscode של חברת Microsoft.

שימוש בספריות:

- TensorFlow של גוגל
- Pandas
- Sklearn
- Numpy
- שימוש בבסיסי הנתונים

חומרה :

חומרה

- אפשרויות לבחירת יחידת עיבוד:
  - CPU
  - GPU
  - שירותי עיבוד חיצוניים המשלבים סוגי מעבדים שונים (Amazon,Google)

תכונה	CPU	GPU
מספר ליבות	יחידים	מאות - אלפים
מהירויות שעון	2.4 GHz-4.0 GHz	1.0-2.0 GHz
זיכרון RAM	אין	יחידים - עשרות של GB
עלות	500 - 8,000 ש"ח	1000 - 10,000 ש"ח
קלות שימוש בפרויקט	פשוט יותר, חלק ממערכת המחשוב	לעיתים דורש התקנה ושימוש חיצוניים למערכת המחשוב
שימושים עיקריים	מגוון רב של פעולות חישוביות	חישובים ווקטורים, בעיקר בגרפיקה.

איור 5- מציג כלים של חומרה והשוואה ביניהם

6 תיאור רעיון דומה שיכול להוות השראה

הבעיה העיקרית של הפרויקט נעוצה בשימוש לא נכון ב- "همزتي الوصل والقطع" במקום הלא נכון, במיוחד כשכותבים מילים המתחילות באות "أ", שכן רבים טועים בכתיבת "همزتي الوصل والقطع" או הפלתה, מה שמחליש את חוזק השפה הערבית.

קיבלתי את ההשראה הזו (כמה אנשים טועים בכתיבת "همزتي الوصل والقطع") מסרטונים שצפיתי בהם כשניסיתי לעזור לבת שלי שלומדת בכיתה ט' איך לדעת איך לבחור "همزتي الوصل والقطع" למילה בתחילת השורה בטקסטים בערבית.

מצחיק ..... זיהיתי שאני עצמי עד עכשיו הייתי כותב אותם לא נכון.

מצורף קישור לסרטונים:

7 סיכונים, אי וודאות ואילוצי הפרויקט

אני חושש מחוסר ההתאמה של שימוש בספריות תוכנה לפרויקט זה, שכן מחקר זה הוא אחד המחקרים החדשים בתחום המחקרים הדקדוקיים והטכניים העוסק בשימוש באלגוריתמים של למידת מכונה לסיווג "همزتي الوصل والقطع" בטקסטים בערבית.
בגלל הקושי בשימוש בספריות תוכנה בהתמודדות טובה עם טקסטים בערבית, לכן, על מנת להתגבר על הבעיה והחשש הזה, נגעתי בשימוש באלגוריתם SVM, כי זו אחת השיטות המפורסמות ביותר לסיווג אוטומטי התלויה על מציאת עקומה או רמה מפרידה, המפרידה בין הדגימות שהוזנו זו מזו, ומטרתה היא למצוא את ההבחנה בין חברים משני סוגים של נתוני אימון ,כפי שציינתי קודם, אחד המאפיינים שלו הוא דיוק גבוה בסיווג, והוא מיושם בתחומים רחבים, כולל הגדרת קטגוריות טקסט לפי סיווג התמונה.
אני גם חושש מאוד מכך שלא לבנות את טבלת הנתונים בצורה נכונה וטובה ובכמות הנדרשת לעבודה עם אלגוריתמים, כי הנתונים בימינו גדלים מיום ליום ומקורותיו מרובים, וזה מוביל ל- חשיפת נתונים אלו לבעיות רבות המפחיתות את איכות הנתונים, כגון מספר רב של נתונים חסרים ואי-עקביות לכן, בפרויקט זה חילקתי את שלבי יישום מודלים של מסווג נתונים לשלושה שלבים, החל משלב התכנון, הספרים ומקורות ואיסוף הנתונים וכלה בשלב מדידת הדיוק של מסווגי הנתונים.
אני יכול לומר לעצמי שאכן הצלחתי אם אני נתתי תשובות לשאלות הבאות לאור התועלת של אלגוריתמי למידת מכונה המשמשים בסיווג וחיזוי:
- כיצד אוכל להבחין בין "همزتي الوصل والقطع" בטקסטים בערבית?
- מהם האלגוריתמים המתאימים ללימוד מכונה לסיווג "همزتي الوصل والقطع" בטקסטים בערבית?
בפרויקט זה השתמשתי גם בגישה הסטטיסטית התלויה במקורות שהשתמשתי וחקרתי אותם ו-(ערכת אימון מילים) באמצעות מודל סיווג,זה היה מאתגר מאוד וקשה מבחינה טכנולוגית , לכן בפרויקט זה אשתמש באלגוריתמים של למידת מכונה לאחר בחינת החזרה על תנאים ומאפיינים (דקדוק הקשור למילים המתחילות "همزتي الوصل والقطع"),כדי להסיק את אלמנטים מודל הסיווג הבסיסי.

8 עבודות בנושא

יש דמיון רב וחפיפה למחקרים אחרים הקשורים לפרויקט לשימוש באלגוריתמים אלה ביישומים מקבלים ודומים, ואני הרווחתי מהם היא עזרה לי לגשת לפרויקט שלי, שני המחקרים הדומים הם:

מחקר שכותרתו:

חקר דעות במשפטים השוואתיים בערבית [13]

מחקר זה עסק בבעיית זיהוי תחום ההשוואה בחקירת דעות המשמשות בטקסט הערבי. החוקרת הזכירה כי קיים מחקר מסוים בתחום זה לגבי המשפטים של השפה האנגלית ושפות נוספות, אך לגבי המשפטים בערבית, זהו המחקר הראשון, ובמחקר נעשה שימוש בטכניקה המבוססת על סיווג לשוני ועוד. טכניקה המסתמכת על למידת מכונה.

מחקר שכותרתו:

מחקר השוואתי של אלגוריתמים של כריית דעות וניתוח רגשות ויישומיהם [14]

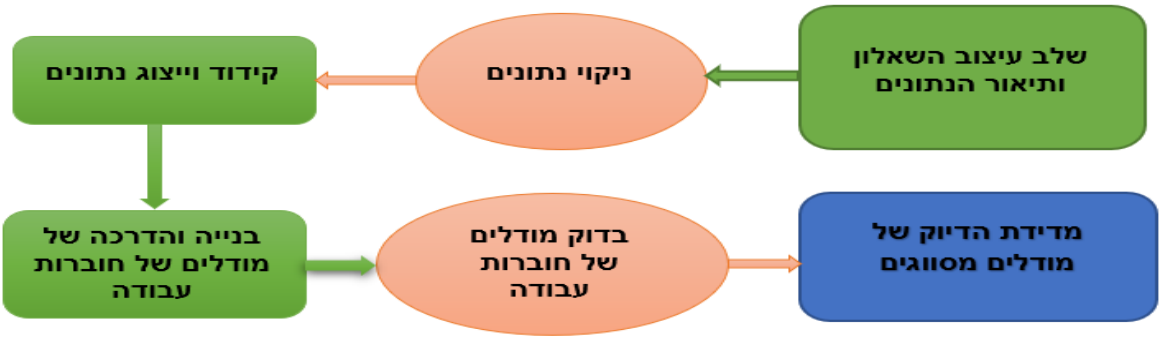
מחקר זה עוסק בבעיית ריבוי נקודות המבט של לקוחות המאוחסנות במאגרי הנתונים באינטרנט, אשר נתנה תשומת לב לכריית נתונים וניתוח סנטימנטים בשנים האחרונות. החוקר ציין כי אנשים הסתמכו על המכונה לסיווג ועיבוד נתונים, שכן הזמינות של כמויות צפיות עצומות על מוצר אחד מסייעת לחזות את תחושות הלקוח על ידי ניתוח הדעות שעוזרות לא רק להגדלת הרווחים אלא גם בשיפור מוצר. מחקר זה השווה בין הטכנולוגיות הזמינות כיום ובשימוש ביישומים שונים בתחום חקירת הדעות. לאחר סקירת המחקרים לעיל, הרעיון לקטלג את "الهمزة" עלה מהשימוש של החוקרים בשני המחקרים לעיל במושג חפירה בטקסטים בערבית בשיטת החפירה בטקסטים בערבית עם אלגוריתמי סיווג שונים.

9 מה בוצע עד כה הפרויקט

הפרויקט מורכב משלושה חלקים:

הנושא הראשון (מסגרת תיאורטית לפרויקט): עוסק במושג בינה מלאכותית, במושג למידת מכונה, אלגוריתמי סיווג "همزتي الوصل أو القطع".
הנושא השני (טיפול ויישום): עוסק בשלבי יישום מודלים של מסווגים הכוללים: לתאר, לנקות, לעבד, לקודד את הנתונים ולזהות את המשתנים הבלתי תלויים והתלויים בנתוני המדגם זה גם עוסק בבניית מודלים של סיווגים.
הנושא השלישי (תוצאות, תפוקות פרויקט): עוסק בבדיקת מודלים מסווגים, מדידת דיוק מודלים, תוצאות, תפוקות הפרויקט והתועלות הרצויות מכך.

הנושא הראשון (המסגרת התיאורטית לפרויקט) דיברתי והרחבתי בפרקים למעלה, אבל הנושא השני (יישום ויישום) אדבר וארחיב עליו עכשיו.
הנתונים בימינו גדלים מיום ליום ומקורותיהם מרובים, והדבר מביא לחשיפה של נתונים אלו לבעיות רבות המפחיתות את איכות הנתונים כמו ריבוי הנתונים החסרים ואי העקביות של הנתונים, אז בפרויקט הזה חילקתי את שלבי יישום מודלים של מסווג נתונים לשישה שלבים שהתחילו את השלב של עיצוב השאלון ואיסוף הנתונים והסתיים בשלב של מדידת הדיוק של מודלים מסווגים, כפי שמוצג באיור הבא (איור-2):



איור 6- מציג את ששת השלבים של מודול אלגוריתמים סיווג הנתונים.

בחלק השני של הפרויקט אכסה את ארבעת השלבים הראשונים של יישום מודלים מסווגים ויותר קדימה (בסעיפים הבאים) אתייחס לשניים האחרונים.

- **שלב בניית טבלת הנתונים ותיאור הנתונים:**

בפרויקט זה הסתמכתי על כמה ספרים בשפה וספרות הערבית [2] כדי למצוא ולבנות מאגר מילים בערבית המתחילות " بالهمزة " [3].המאגר הכיל 477 מילים המתחילות " بالهمزة " ונשאר לסווגן ל- "همزتي الوصل والقطع" על פי הכללים הדקדוקיים הנקטים בתהליך זה ,לכן שלחתי למומחים .

המאגר הועבר ונבדק על ידי שישה מומחים בתחביר בדרגים אקדמיים שונים. המספר הכולל של המילים המסווגות הגיע ל-477 מילים, ולאחר עיבוד ואי הכללה של המילים החוזרות, 101 מילים, השגתי 376 [4] . המילים שהופקו מהטקסטים לדוגמה סווגו למילים המתחילות בהמזה (الهمزة) המכונה בפרויקט זה (Wasl) (همزة الوصل) ומילים המתחילות בהמזה (الهمزة) המכונה בפרויקט זה(Gtaa) (همزة القطع) המוזכרות בפרויקט זה. המשתנה התלוי (Outcome) נקבע על פי השניים ערכים (Gtaa/Wasl).

באשר למשתנים הבלתי תלויים (המאפיינים), הם חולקו לשלושה מאפיינים לפי הכללים הדקדוקיים שאליהם מתייחסים בהמשך (טבלה 1).

הפורמט של קובץ הנתונים נבחר מסוג (csv) והוא הכיל את המאפיינים (משתנים בלתי תלויים), שהם:

- משתנה: (diacritic) משתנה מספרי באורך של מספר (1) המציין את תנועת האות אליף בתחילת המילה (פתחה, קאסרה, דמא) המיוצגת על ידי הערכים המספריים (1, 2, 3) בהתאמה.

- משתנה: (Count): משתנה מספרי המציין את מספר האותיות במילה, שכן המילה, עם ערכים המכילים שתי אותיות, שלוש אותיות, ארבע... וכו', מיוצגת מספרית על ידי הערכים (2, 3 או 4 .... 10) בהתאמה.

המשתנה (morphological): משתנה מספרי המציין את המשקל המורפולוגי של הפעלים

והאינפיניטיבים (פועל, עשה, עובר, ... פעיל) מיוצגים באופן מספרי על ידי הערכים (0, 1, 2, 3 ..... 11) בהתאמה.

- משתנה: (Noun) משתנה מספרי המציין שהמילה (לא שם עצם, שם עצם רגיל, אחד מעשרת שמות העצם, אחד מששת שמות העצם, שם עצם יחסי, שם עצם) ויוצגה על ידי מספרים (0,1,2,3..... 6) בהתאמה.

- משתנה: (verb) משתנה מספרי המציין שהמילה (לא פועל, זמן עבר, זמן הווה, פועל ציווי) מיוצגת באופן מספרי (0, 1, 2, 3) בהתאמה.

- משתנה: (adjective) משתנה מספרי המציין שהמילה (לא שם תואר, שם תואר) מיוצגת מספרית על ידי המספרים (0,1) בהתאמה.

- המשתנה (letter) הוא משתנה מספרי המציין שהמילה (לא אות, אות) מיוצגת במספרים מספרית (0,1) בהתאמה.

- המשתנה (the) הוא משתנה מספרי המציין שהמילה (לא מוגדרת על ידי al, מוגדרת על ידי al) מיוצגת על ידי מספרים באופן מספרי (0,1) בהתאמה.

טבלה 2- מתארת חלק מטבלת הנתונים הכללית שנתתי לה שם (Arabic_Word.csv).

290	افتتاح	6	اسم	ليس فعلاً	ليست صفة	ليس حرفاً	افتعال	غير مضافة	كسرة	Wasl
291	انكسر	5	ليس اسماً	ماضي	ليست صفة	ليس حرفاً	افتعل	غير مضافة	كسرة	Wasl
292	انكسِر	5	ليس اسماً	امر	ليست صفة	ليس حرفاً	افتعل	غير مضافة	كسرة	Wasl
293	انكسار	6	اسم	ليس فعلاً	ليست صفة	ليس حرفاً	افتعال	غير مضافة	كسرة	Wasl
294	انتهى	5	ليس اسماً	ماضي	ليست صفة	ليس حرفاً	افتعل	غير مضافة	كسرة	Wasl
295	انتهِ	4	ليس اسماً	امر	ليست صفة	ليس حرفاً	افتعل	غير مضافة	كسرة	Wasl
296	انتهاء	6	اسم	ليس فعلاً	ليست صفة	ليس حرفاً	افتعال	غير مضافة	كسرة	Wasl
297	ابتدأ	5	ليس اسماً	ماضي	ليست صفة	ليس حرفاً	افتعل	غير مضافة	كسرة	Wasl
298	ابتدِئ	5	ليس اسماً	امر	ليست صفة	ليس حرفاً	افتعل	غير مضافة	كسرة	Wasl
299	ابتداء	6	اسم	ليس فعلاً	ليست صفة	ليس حرفاً	افتعال	غير مضافة	كسرة	Wasl
300	احمرّ	5	ليس اسماً	ماضي	ليست صفة	ليس حرفاً	افتعل	غير مضافة	كسرة	Wasl
301	احمرِّ	5	ليس اسماً	امر	ليست صفة	ليس حرفاً	افعل	غير مضافة	كسرة	Wasl
302	احمرار	6	ليس اسماً	ليس فعلاً	صفة	ليس حرفاً	افتعال	غير مضافة	كسرة	Wasl
303	أشرف	4	أسم عادي	ليس فعلاً	ليست صفة	ليس حرفاً	افعل	غير مضافة	فتحة	Gtaa
304	أماكن	5	اسم	ليس فعلاً	ليست صفة	ليس حرفاً	افاعيل	غير مضافة	فتحة	Wasl
305	أشياء	5	اسم	ليس فعلاً	ليست صفة	ليس حرفاً	أفعال	غير مضافة	فتحة	Gtaa
306	أي	2	اسم	ليس فعلاً	ليست صفة	ليس حرفاً	فعل	غير مضافة	فتحة	Gtaa
307	إياك	4	اسم	ليس فعلاً	ليست صفة	ليس حرفاً	افعل	غير مضافة	كسرة	Gtaa
308	أيدي	4	اسم	ليس فعلاً	ليست صفة	ليس حرفاً	فعل	غير مضافة	فتحة	Gtaa
309	إياد	4	أسم عادي	ليس فعلاً	ليست صفة	ليس حرفاً	أفعال	غير مضافة	كسرة	Gtaa
310	أحسن	4	ليس اسماً	امر	ليست صفة	ليس حرفاً	افعل	غير مضافة	فتحة	Gtaa
311	أعظم	4	اسم	ليس فعلاً	ليست صفة	ليس حرفاً	افعل	غير مضافة	فتحة	Gtaa
312	إنّ	2	ليس اسماً	ليس فعلاً	ليست صفة	حرف	فعل	غير مضافة	كسرة	Gtaa
313	إنّ	2	ليس اسماً	ليس فعلاً	ليست صفة	حرف	فعل	غير مضافة	كسرة	Gtaa
314	أنّ	2	ليس اسماً	ليس فعلاً	ليست صفة	حرف	فعل	غير مضافة	فتحة	Gtaa
315	أن	2	ليس اسماً	ليس فعلاً	ليست صفة	حرف	فعل	غير مضافة	فتحة	Gtaa
316	إما	3	ليس اسماً	ليس فعلاً	ليست صفة	حرف	فعل	غير مضافة	كسرة	Gtaa
317	إلّا	3	ليس اسماً	ليس فعلاً	ليست صفة	حرف	فعل	غير مضافة	كسرة	Gtaa
318	ألّا	3	ليس اسماً	ليس فعلاً	ليست صفة	حرف	فعل	غير مضافة	فتحة	Gtaa
319	ألا	3	ليس اسماً	ليس فعلاً	ليست صفة	حرف	فعل	غير مضافة	فتحة	Gtaa
320	ألم	3	ليس اسماً	ليس فعلاً	ليست صفة	حرف	فعل	غير مضافة	فتحة	Gtaa

- **שלב ניקוי הנתונים:**

אין ספק שכאשר איכות הנתונים נמוכה, הדבר ישפיע בהכרח על תוצאות הניתוח. בפרויקט זה נעשה שימוש במספר שיטות ניקוי נתונים על הטקסטים שנאספו.
שלב ניקוי הנתונים כלל את השלבים הבאים:

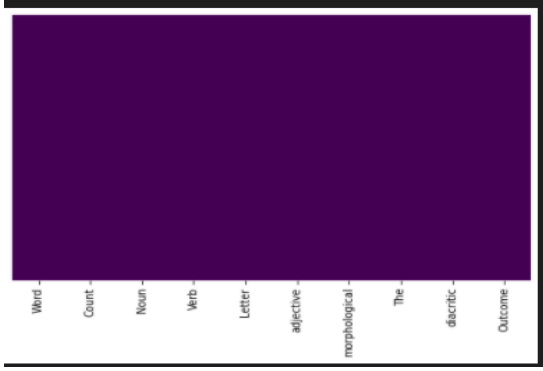
- התמודדות עם נתונים שאבדו
- מחיקת נתונים כפולים.

לאחר ביצוע פעולות ניקוי נתונים על הטקסטים שנאספו, ולאחר עיבוד והוצאת מילים כפולות, וגם השלמת הנתונים החסרים(מספרם 101),לכן בסוף קיבלנו 376 מילים תקפות ליישם את מודל חוברת העבודה. ביצעתי גם בדיקה בשפת python וקיבלתי שהנתונים תקנים ללא חיסרון אם היינו רואים חלק בצבע צהוב ,אז היה ברור שאין רצף בנתונים ויש נתונים חסרים או לא מוגדרים , אבל קיבלתי כל הגרף בצבע סגול שמראה שהנתונים תקנים ומוגדרים היטיב ורואים את זה באיור הבא:

```
from matplotlib import pyplot as plt
import numpy as np
import pandas as pd
import sys
import seaborn as sns
df = pd.read_excel (r'C:\Users\nizar.maarouf\Desktop\project2021\Arabic_Words.xlsm', sheet_name='data_encoding')
plt.subplots(figsize=(8,5))
sns.heatmap(df.isnull(),
            yticklabels=False,
            cbar=False,
            cmap='viridis')
df.head(376)
```

	Word	Count	Noun	Verb	Letter	adjective	morphological	The	diacritic	Outcome
0	8e117a32510d41082681de68937befd3f564c349067ec5...	4	1	0	0	0	2	0	2	0
1	b401e637ead22b50ca3012befa3c55901d2248781750ed...	5	1	0	0	0	2	0	3	0
2	c2132540fedae6dba455d57553753448f7685171ac60dd...	5	1	0	0	0	2	0	1	0
3	f0993ec4ddf41ffe8b87ecd35fa09be5a7d56ab0424875...	4	1	0	0	0	2	0	1	0
4	3fa07d2ebd61fb0384890778f359a7c873fc429d18a71b...	4	1	0	0	0	2	0	1	0
...	...	...	...	...	...	...	...	...	...	...
371	361c39212c58502882ea5e053143ecff7601344241a2db...	3	5	0	0	0	1	0	2	0
372	42bce38dbdd06fbcf6160676d26b2930bba39257f798c8...	4	5	0	0	0	1	0	2	0
373	6cdb0f2fed6538524c382826bf80dd166c67a4992f10f6...	5	5	0	0	0	1	0	1	0
374	ca275a27a116ff0e38aa5826c53e6c69bf57a3ee839b26...	2	5	0	0	0	1	0	1	0
375	8fe1ca77fcae9c127ec564803dfbb5948331f1ab697b99...	4	5	0	0	0	1	0	1	0

376 rows × 10 columns



איור 7- ביצוע בדיקה באמצעות שפת פיתון לבדיקה שיש רצף בנתונים ללא מחסור.

- **שלב קידוד וייצוג נתונים:**

לאחר לימוד הכללים הדקדוקיים המציגים את מיקומם של "همزتي الوصل والقطع" בתחילת המילה, נקבעו היסודות והתכונות שניתן להסתמך עליהם בקביעת ערכו של המשתנה(Outcome). הערכים של הנתונים המילוליים מומרים לערכים מספריים כדי שהאלגוריתמים יוכלו להתמודד עם זה והנתונים הופכים למקודדים. לאחר השלמת תהליך קידוד הנתונים, הנתונים מיוצגים באמצעות שפת Python והספריות (sklearn-scipy-numpy) מיובאות באמצעות עורך VScode.

טבלה 3- מסביר את הכללים המשמשים לקביעת מיקומי ה-"همزتي الوصل والقطع"

خيارات	ترميز	خيارات	ترميز	خيارات	ترميز	خيارات	ترميز	خيارات	ترميز	خيارات	ترميز	خيارات	ترميز	خيارات	ترميز		
ليس اسماً	0	ليس فعلا	0	ليس حرفاً	0	ليست صفة	0	فتحة	1	فعل	1	حرفين	2	غير مضافة	0	Gtaa	0
أسم عادي	1	ماضي	1	حرف	1	صفة	1	كسرة	2	افعل	2	ثلاثة احرف	3	مضافة	1	Wasl	1
الأسماء العشره	2	مضارع	2					ضمة	3	انفعل	3	أربعة احرف	4				
الأسماء الستة	3	امر	3					مده	4	افتعال	4	خمسة احرف	5				
أسماء الموصول	4									افاعيل	5	ستة احرف	6				
أسماء الأفعال	5									أفعال	6	سبعة احرف	7				
اسم	6									افتعل	7	ثمانية احرف	8				
أسماء الاشاره	7									استفعل	8	تسعة احرف	9				
اسم بلد	8									استفعال	9	عشرة احرف	10				
اسم استفهام	9									انفعال	10						
										فعل	11						

טבלה (2) מציגה את הכללים המשמשים לקביעת מיקומי "همزتي الوصل والقطع". הנתונים בטבלה (1) הוצפנו והערכים המילוליים מומרים לערכים מספריים כדי שהאלגוריתמים יוכלו להתמודד איתם בהתאם לטבלה (2), שמסבירה את הכללים המשמשים לקביעת מיקומי "همزتي الوصل والقطع". הנתונים הופכים לאחר מכן מקודדים כמו בטבלה (3) הבאה :

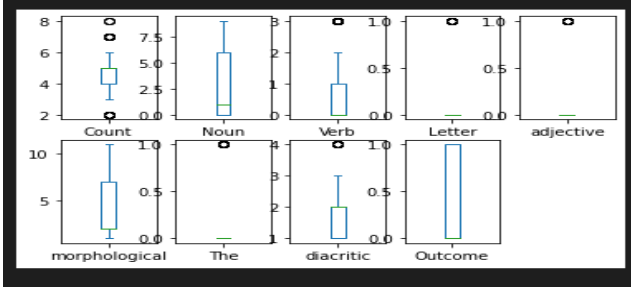
טבלה 4- קידוד נתונים בקובץ הנתונים (data_encoding.csv)

c2132540fedae6dba455d57553753448f7685171ac60dd64be0e691c16afff2c	5	1	0	0	0	0	2	0	1
f0993ec4ddf41ffe8b87ecd35fa09be5a7d56ab0424875a1e9c9002206793321	4	1	0	0	0	0	2	0	1
3fa07d2ebd61fb0384890778f359a7c873fc429d18a71bc6ad10f9fee2d080b3	4	1	0	0	0	0	2	0	1
50faa6a497e42d511ba571d20f5f1ceb122578eab3d7a756065298f19356fcda	5	1	0	0	0	0	2	0	1
6423ea67803e7caa19c41298bed41a265b9900e757ad3395b9f804a5305f94fb	2	3	0	0	0	0	2	0	1
244a9f5147173bd1a9be0053d4183daf092d9ab817a6c6dfc0bf39064bc25ad3	2	6	0	0	0	0	2	0	3
e54e0a389d25b651efec32f119e89f23e02cadb00b2ab34ed78aceebfef7fb15	2	3	0	0	0	0	2	0	3
8a6350153432a185ecd78404b8732fb34bc25efe61b65f3939481e46f46e9565	4	6	0	0	0	0	2	0	3
d2ee47ec7d0552ccb63459ef645de3e3c01449e2175bfd6c7113250f019e71d3	4	6	0	0	0	0	2	0	3
128ed0f7c9249594643e07505d32552b91f629ce0aa3f22a8680b77ff86cf49e	6	6	0	0	0	0	2	0	3
17a9f38e22b6672065c44ec7620a06edd76dbce9477ee74f2cf7734239c2993f	4	6	0	0	0	0	2	0	3
2f6fdef7d8a1e4190241efba8db6f45d8d93809bb750dc7f5aa619a4b713b82c	5	7	0	0	0	0	2	0	3
e416c08408d3e649104945edb82ff3125c8a22e1f5e241663f8a284d2ffe5479	5	6	0	0	0	0	2	0	3
98fcbf16cc64d8e6e9f40bd494b29a53e1c11a2b1842df3e3dc0aa28f46cfe54	5	6	0	0	0	0	2	0	2
ccf58d17ef4700a65a58a77d0552543e6a0a6792ead104fcaab407caf5904abe	5	6	0	0	0	0	2	0	2
aec17e2f7a418d8e358e48d819b07068906020cff205da9dec5a9144dc6a4863	5	6	0	0	0	0	2	0	2
32e65d8eaf588d4f742e6f0cf965e794f30f8e01f1d24ded08215ae30abd8b2f	5	6	0	0	0	0	2	0	2
f8ac19d897aa1101003c21210e4101b6cb44c678a35db882415b018386e36793	2	6	0	0	0	0	2	0	1
2fdd211020199da1ccccbf9527adc9e74247b30f81b1f99915899fba4bb1c91390	4	1	0	0	0	0	2	0	1
1209e2f9eecef8d9bf5aabb076eeab36df5e82c3a48b68eb7bb51b6517ab899e	5	6	0	0	0	0	2	0	2
61e702330149bf404f3c4a146bc56d8eb927fad8fa5983adc190f12b74a90d36	4	6	0	0	0	0	2	0	1
3817bd91e8d6e8df6a1330df1fdca75f97745eb2155037a8dcb7024a270c4e04	6	8	0	0	0	0	2	0	3

טבלה (3) קידוד נתונים בקובץ הנתונים (data_encoding.csv) לאחר השלמת תהליך קידוד הנתונים, הנתונים מיוצגים באמצעות שפת Python וייבא את הספריות (numpy-scipy-sklearn) באמצעות העורך (vscode) כמו באיור (4+5) להלן :

```
import warnings
warnings.filterwarnings("ignore")
import os
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.datasets import load_iris
data = pd.read_csv (r'C:\Users\nizar.maarouf\Desktop\project2021\Data_Encoding.csv')
data.plot(kind='box',subplots=True,layout=(2,5),sharex=False,sharey=False)
plt.plot( )
```

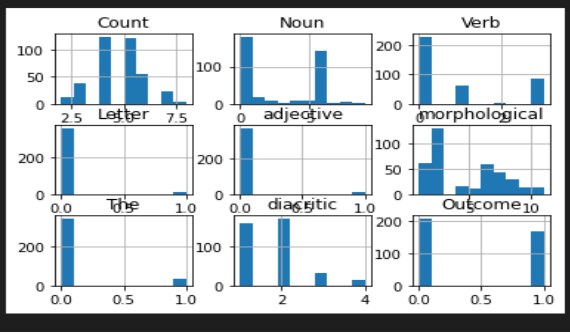
```
[]
```



איור 8- ייצוג נתונים באמצעות עורך Vscode

```
import warnings
warnings.filterwarnings("ignore")
import os
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.datasets import load_iris
from sklearn.neighbors import KNeighborsClassifier
data = pd.read_csv (r'C:\Users\nizar.maarouf\Desktop\project2021\Data_Encoding.csv')
data.hist()
plt.plot()
```

```
[]
```



איור 9- ייצוג נתונים באמצעות עורך Vscode

- **שלב בניית והדרכת דגמי מסווגים:**

הנתונים המיוצגים חולקו לנתוני אימון (67%) ולנתונים ניסיוניים (33%) מתוך כלל הנתונים (376) שנרשמו כהקדמה לבניית מודל אימון נתונים באמצעות אלגוריתמי הסיווג שבחרתי להשתמש בהם (אלגוריתם SVM - אלגוריתם NB Knifebase - אלגוריתם KNN), האלגוריתמים נבחרו כדי להתאים את גודל המדגם וערכי הנתונים עם האלגוריתמים לעיל. המודל נבנה באמצעות פונקציות הספרייה (Learn-Sk) כמו באיור (9) למטה:

```python
import pandas as pd
from sklearn.model_selection import train_test_split,cross_val_score
from sklearn.neighbors import KNeighborsClassifier
from sklearn.svm import SVC
from sklearn.naive_bayes import MultinomialNB
from sklearn.model_selection import StratifiedKFold

scoring= 'accuracy'
data = pd.read_csv (r'C:\Users\nizar.maarouf\Desktop\project2021\Arabic_Words.csv').head(376)
data.drop("Word", inplace = True, axis=1)
y = data['Outcome'].values
x = data.drop(["Outcome"], axis = 1)
x_train, x_test, y_train, y_test = train_test_split(x, y, test_size = None, random_state = None)
models= []
# models.append(('LR', LogisticRegression()))
# models.append(('RF', RandomForestClassifier()))
models.append(('KNN', KNeighborsClassifier(n_neighbors=8)))
models.append(('NB', MultinomialNB()))
models.append(('SVM', SVC(gamma='auto')))
results = []
names = []
for name, model in models:
    kfold = StratifiedKFold(n_splits=10, random_state=None)
    cv_results = cross_val_score(model, x_train, y_train, cv=kfold, scoring=scoring)
    results.append(cv_results)
    names.append(name)
    msg = '%s: %.2f (%.3f)' % (name, cv_results.mean(), cv_results.std())
    print('\n' + msg)
✓ 1.2s

KNN: 0.89 (0.042)

NB: 0.68 (0.076)

SVM: 0.94 (0.021)
```

איור 10- בניית מודלים לסיווג באמצעות העורך (Vscode)

סיכום

אני ביצעתי את הפרויקט בשלבים אחד אחרי השני להלן השלבים:

1- חקירה והבנת חשיבות ותרומת הפרויקט לשפה וספרות ערבית
2- אספת מקורות וספרים עם התייעצות עם המומחים
3- הכנת טבלאות הנתונים שכוללת את המילים ושיטת זיהוי באמצעות הדקדוק
4- בדיקת כל הנתונים כולל קורלציה בין עמודות הטבלאות
5- בדיקת שהנתונים אינם חסרים ומספיקות להמשך הפרויקט
6- קידוד כל הטבלאות כדי שיתאימו לשפת התכנות והאלגוריתמיים
7- חיפוש ובדיקת האלגוריתמים המתאימות לנתונים שלי בכלל ושפה ערבית בפרט
8- ביצוע כמה הרצות והשוואות בין האלגוריתמים

אני בשלבים הסופיים בפרויקט , מכין עוד בדיקות וכתיבת קודים והרצות , מכין תוצאות סופיות ומסכנות .....

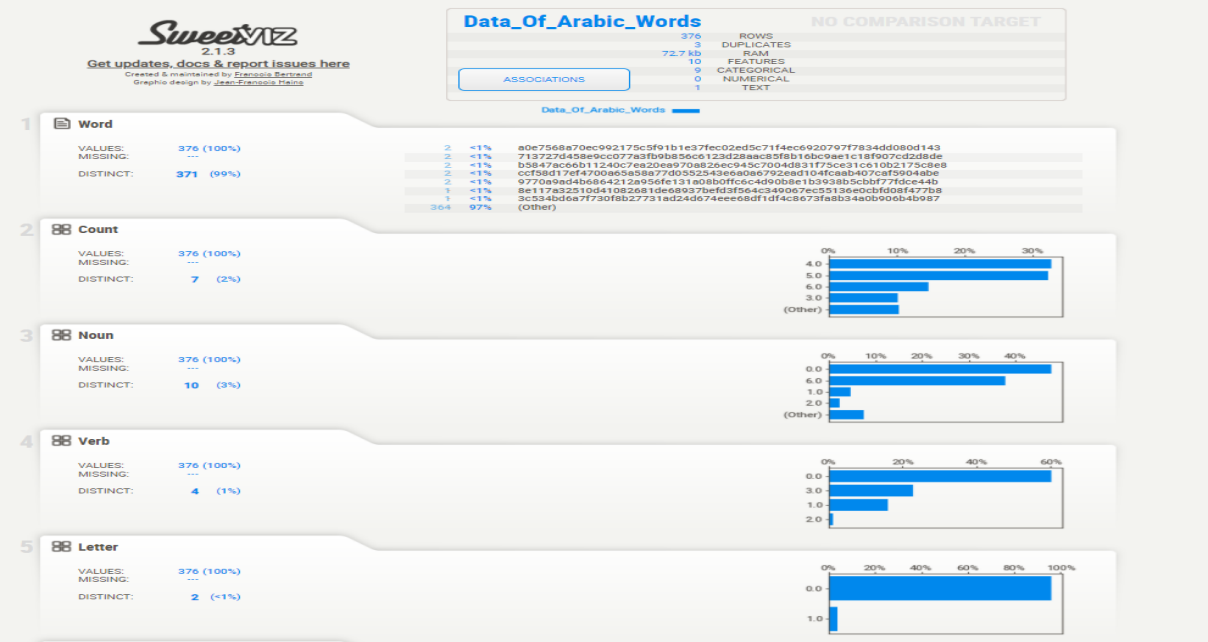
הערה: חלק מטבלאות שביצעתי הכל יהיה בשלב התוצאות (הנושא השלישי בפרויקט) .

data (474, 10)

index	Word	Count	Noun	Verb	Letter	adject…	morpho…	The	diacri…	Outcome
0	إسلام	4	1	0	0	0	2	0	2	0
1	أسامه	5	1	0	0	0	2	0	3	0
2	أسماء	5	1	0	0	0	2	0	1	0
3	أسيل	4	1	0	0	0	2	0	1	0
4	أنور	4	1	0	0	0	2	0	1	0
5	أبرار	5	1	0	0	0	2	0	1	0
6	آتْ	2	3	0	0	0	2	0	1	0
7	أم	2	6	0	0	0	2	0	3	0
8	آغ	2	3	0	0	0	2	0	3	0
9	أسره	4	6	0	0	0	2	0	3	0
10	أنثى	4	6	0	0	0	2	0	3	0
11	أكتوبر	6	6	0	0	0	2	0	3	0
12	أخرى	4	6	0	0	0	2	0	3	0
13	أولئك	5	7	0	0	0	2	0	3	0
14	أسبوع	5	6	0	0	0	2	0	3	0
15	إفاده	5	6	0	0	0	2	0	2	0
16	إنتاج	5	6	0	0	0	2	0	2	0
17	إهمال	5	6	0	0	0	2	0	2	0
18	إشراف	5	6	0	0	0	2	0	2	0
19	أغ	2	6	0	0	0	2	0	1	0
20	أريج	4	1	0	0	0	2	0	1	0
21	إداره	5	6	0	0	0	2	0	2	0
22	أداة	4	6	0	0	0	2	0	1	0

data (474, 10)

index	Word	Count	Noun	Verb	Letter	adject…	morpho…	The	diacri…	Outcome
0	8e117a32510d41082681de68937befd3f564c349067ec55136e0cbfd08f477b8	4	1	0	0	0	2	0	2	0
1	b401e637ead22b50ca3012befa3c55901d2248781750edb463802662103a0e12	5	1	0	0	0	2	0	3	0
2	c2132540fedae6dba455d57553753448f7685171ac60dd64be0e691c16afff2c	5	1	0	0	0	2	0	1	0
3	f0993ec4ddf41ffe8b87ecd35fa09be5a7d56ab0424875a1e9c9002206793321	4	1	0	0	0	2	0	1	0
4	3fa07d2ebd61fb0384890778f359a7c873fc429d18a71bc6ad10f9fee2d080b3	4	1	0	0	0	2	0	1	0
5	50faa6a497e42d511ba571d20f5f1ceb122578eab3d7a756065298f19356fcda	5	1	0	0	0	2	0	1	0
6	6423ea67803e7caa19c41298bed41a265b9900e757ad3395b9f804a5305f94fb	2	3	0	0	0	2	0	1	0
7	244a9f5147173bd1a9be0053d4183daf092d9ab817a6c6dfc0bf39064bc25ad3	2	6	0	0	0	2	0	3	0
8	e54e0a389d25b651efec32f119e89f23e02cadb00b2ab34ed78aceebfef7fb15	2	3	0	0	0	2	0	3	0
9	8a6350153432a185ecd78404b8732fb34bc25efe61b65f3939481e46f46e9565	4	6	0	0	0	2	0	3	0
10	d2ee47ec7d0552ccb63459ef645de3e3c01449e2175bfd6c7113250f019e71d3	4	6	0	0	0	2	0	3	0
11	128ed0f7c9249594643e07505d32552b91f629ce0aa3f22a8680b77ff86cf49e	6	6	0	0	0	2	0	3	0
12	17a9f38e22b6672065c44ec7620a06edd76dbce9477ee74f2cf7734239c2993f	4	6	0	0	0	2	0	3	0
13	2f6fdef7d8a1e4190241efba8db6f45d8d93809bb750dc7f5aa619a4b713b82c	5	7	0	0	0	2	0	3	0
14	e416c08408d3e649104945edb82ff3125c8a22e1f5e241663f8a284d2ffe5479	5	6	0	0	0	2	0	3	0
15	98fcbf16cc64d8e6e9f40bd494b29a53e1c11a2b1842df3e3dc0aa28f46cfe54	5	6	0	0	0	2	0	2	0
16	ccf58d17ef4700a65a58a77d0552543e6a0a6792ead104fcaab407caf5904abe	5	6	0	0	0	2	0	2	0
17	aec17e2f7a418d8e358e48d819b070680906020cff205da9dec5a9144dc6a4863	5	6	0	0	0	2	0	2	0
18	32e65d8eaf588d4f742e6f0cf965e794f30f8e01f1d24ded08215ae30abd8b2f	5	6	0	0	0	2	0	2	0
19	f8ac19d897aa1101003c21210e4101b6cb44c678a35db882415b018386e36793	2	6	0	0	0	2	0	1	0
20	2fdd211020199da1cccbf9527adc9e74247b30f81b1f9915899fba4bb1c91390	4	1	0	0	0	2	0	1	0
21	1209e2f9eecef8d9bf5aabb076eeab36df5e82c3a48b68eb7bb51b6517ab899e	5	6	0	0	0	2	0	2	0
22	61e702330149bf404f3c4a146bc56d8eb927fad8fa5983adc190f12b74a90d36	4	6	0	0	0	2	0	1	0



10 תכנון

10.1 טבלת אבני דרך

תכנון זמנים עתידי, תחת תנאי אי וודאות.

טבלת אבני דרך, המציגה רשימת אבני דרך ממוספרות להלן הפורמט המחייב:

מס' אבן הדרך	תיאור אבן הדרך	תאריך סיום	סה"כ שעות אדם	תוצר מדיד
1	דוח מכין	24/10/2021	42	דוח מכין
2	חקירה ראשונית +לימוד עצמי של קורסים באינטרנט הקשורים לפרויקט +הכנת DATABASE המתאים לפרויקט	1/1/2022	200	לימוד פיתון +אלגוריתמי לימוד מכונה קשורים לפרויקט
3	דוח התקדמות	16/1/2022	200	דוח התקדמות של 25 עמוד לפחות
4	לימוד אלגוריתם SVM לימוד אלגוריתם KNN לימוד אלגוריתם NB כתיבת התוכנה המתאימה + והרצת לימוד הנתונים בתוכנה	25/6/2022	100	בניית תוכנית והרצת נתונים וניתוח תוצאות
5	יום פרויקטים + הדגמה מעשית	12/7/2022	30	פוסטר + מצגת + POC
6	הכנת ספר הפרויקט הכנה והגשה של מסמך RED	1/9/2022	40	ספר עם לפחות 35 דפים כולל כל הדרישות
7	הגנות	לימודי ערב 15/9/2022	8	ספר פרויקט + פרויקט עובד

סה"כ :620 שעות

<u>תוצר מדיד:</u>

תוצר הוא מה שהסטודנט בוחר להציג - מה שנבחר כתוצר של אבן הדרך. התוצר הוא משהוא מדיד, שלפני הפרויקט לא היה קיים והסטודנט ייצר אותו תוך כדי עבודתו על הפרויקט. על הסטודנט להתמקד בתיאור היצירה החדשה שנוצרה. מומלץ לציין את החידוש, ביחס למה שהיה בפרויקט לפני אבן דרך זו, וכיצד זה מקדם את הפרויקט.

10.2 טבלת משימות

בפרויקט זה אני מבציע את הפרויקט כסטודנט יחיד ללא שותפים בהנחית דוקטור יורם סגל לכן המשימות שיש לבציע כדי לקיים ולסיים את הפרויקט הנ"ל אני מבציע בעצמי ,להלן המשימות (סה"כ שעות 620 שעות) :

1- דוח מכין (42 שעות)
1.1 חיפוש מקורות וחומר רלוונטי לפרויקט – 11 שעות
1.2 לימוד החומר הרלוונטי – 11 שעות
1.3 חיפוש ולימוד איזה שפת תכנות הכי מתאימה – 10 שעות
1.4 כתיבת דוח מכין – 10 שעות

2- חקירה ראשונית של הפרויקט (200 שעות)
2.1 לימוד עצמי של קורסים באינטרנט הקשורים לפרויקט – 40 שעות
2.2 לימוד נושא לימוד מכונה – 40 שעות
2.3 לימוד נושא אלגוריתמים של לימוד מכונה – 60 שעות
2.4 לימוד שפת פתון אחרי שמצאתי שהיא השפה הכי מתאימה לפרויקט שלי – 30 שעות
2.5 הכנת הנתונים המתאימים לפרויקט שלי – 30 שעות

3- דוח התקדמות (200 שעות)
3.1 בניית טבלאות הנתונים – 60 שעות
3.2 בדיקת טבלאות הנתונים , קורלציה, כמות, שלימות הנתונים..... – 40 שעות
3.3 קידוד טבלאות הנתונים – 40 שעות
3.4 מציאת האלגוריתמים המתאימים לטבלאות שבניתי – 40 שעות
3.5 כתיבת דוח התקדמות – 20 שעות

4- המשך לימודים וחקירה וכתיבת קודים בשפת פיתון (100 שעות)
4.1 לימוד אלגוריתם SVM (7 שעות)
4.2 לימוד אלגוריתם KNN (6.5 שעות)
4.3 לימוד אלגוריתם NB (6.5 שעות)
4.4 כתיבת הקודים המתאימים לניתוח הנתונים (10 שעות)
4.5 כתיבת הקודים המתאימים לבדיקת האלגוריתמים (10 שעות)
4.6 כתיבת הקודים לביצוע השוואה כללית של שלושת האלגוריתמים (15 שעות)
4.7 כתיבת קודים להרצאת אלגוריתמים אלה על הנתונים שלי ובדיקת האלגוריתם הכי מתאים (15 שעות)
4.8 תוצאות ותפוקות של הפרויקט (30 שעות)

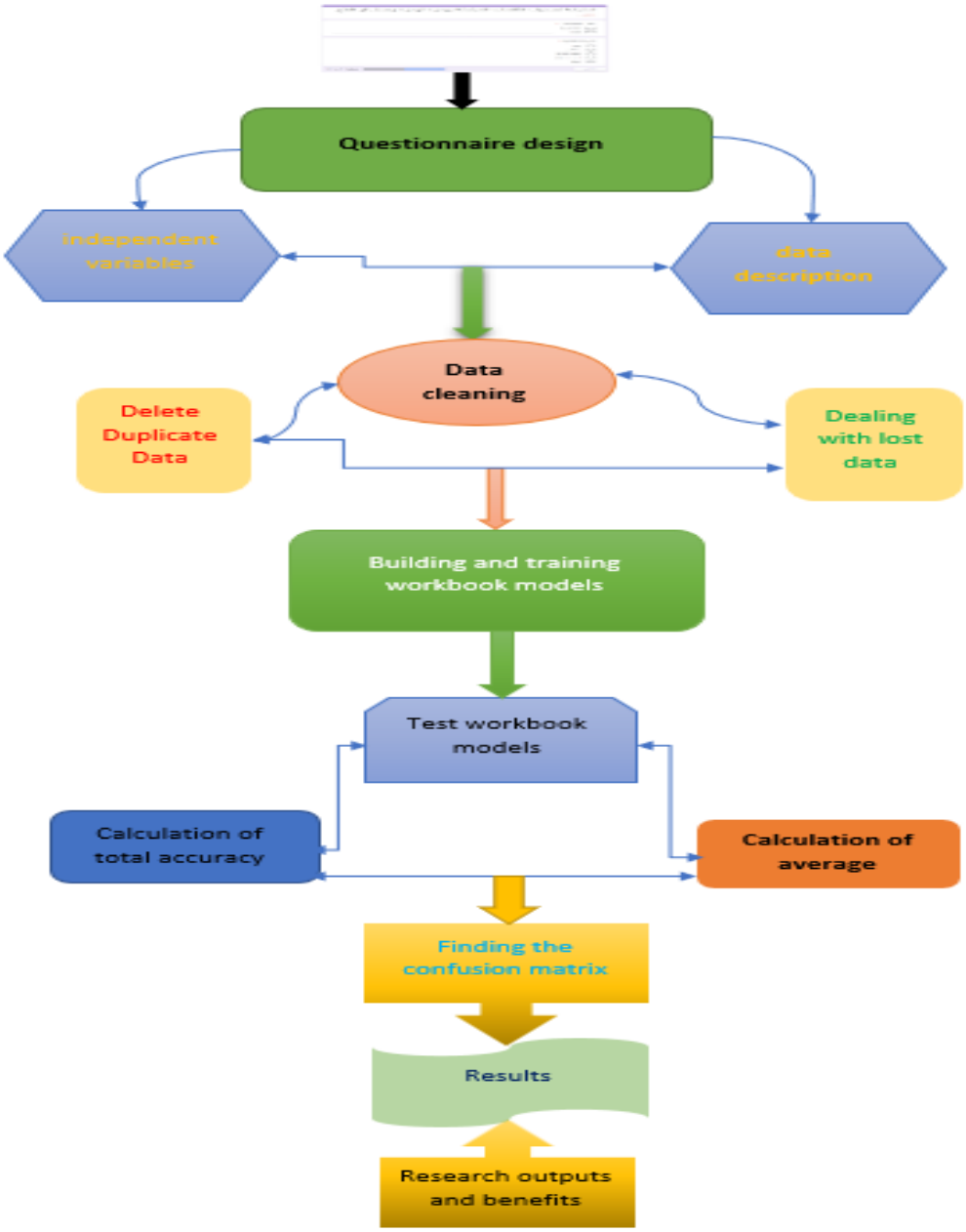
5- יום פרויקט והדגמה (30 שעות)
5.1 כתיבת הדוח הסופי כולל הנושא השלישי – 20 שעות
5.2 הכנת מצגת עבור הפרויקט – 8 שעות
5.3 הצגת והדגמת הפרויקט – 2 שעות

6- הכנת ספר הפרויקט (40 שעות)
6.1 כתיבת הספר ובדיקתו – 30 שעות
6.2 הדפסת הספר הסופי – 2 שעות
6.3 הכנת והגשת מסמך RED – 8 שעות

7- הגנות (8 שעות)
7.1 הבאת הספר והמצגת ומסמך RED – 2 שעות
7.2 הגנה על הפרויקט – 6 שעות

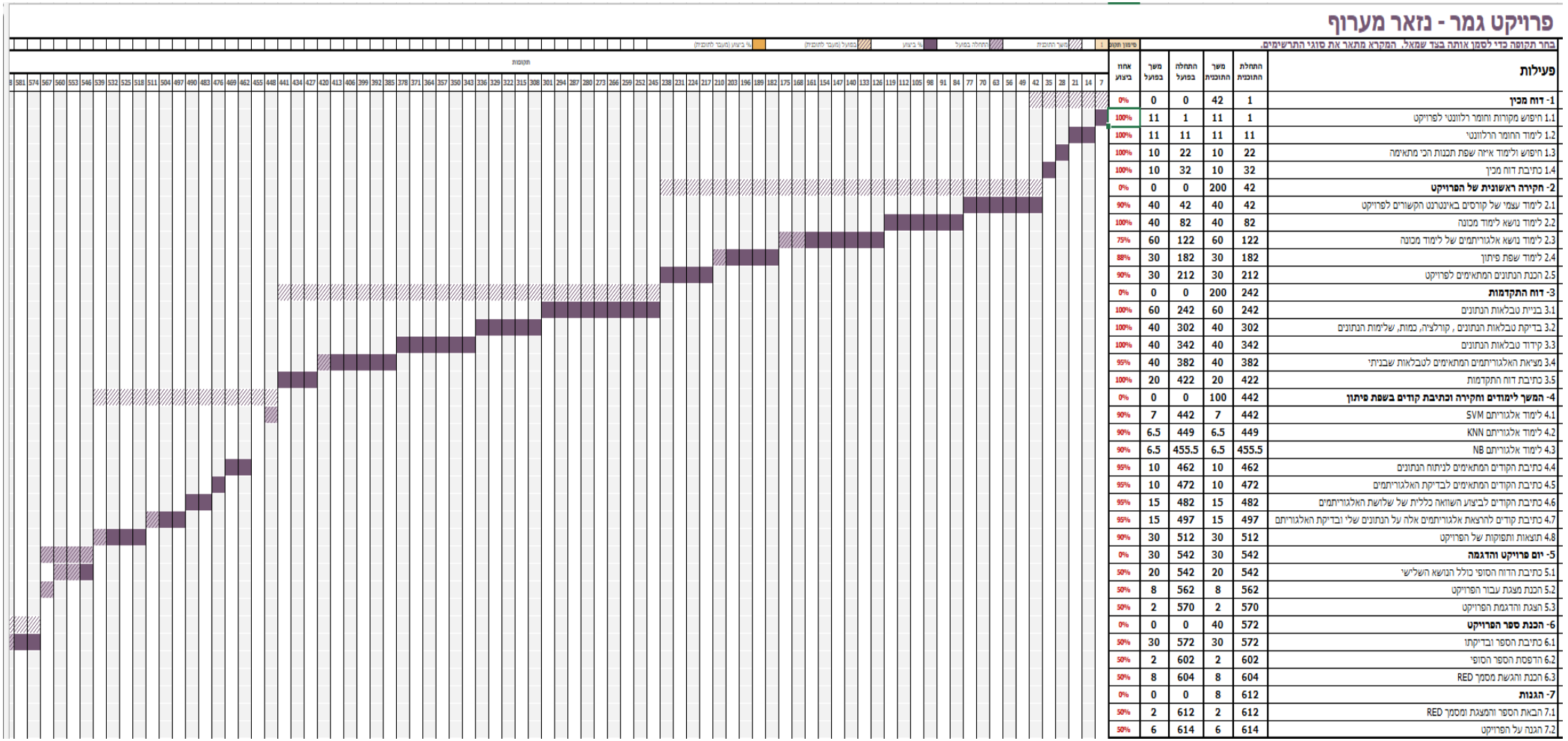
10.3 סכמת בלוקים

להלן דוגמה לסכמת בלוקים ,אך חילקתי את שלבי יישום מודלים של סיווג נתונים למספר שלבים, החל משלב עיצוב השאלון ואיסוף נתונים, עד שהסתיים בשלב של מדידת הדיוק של מודלים מסווגים, כפי שמוצג באיור הבא (1-4):



איור 11- סכמת בלוקים , חלוקת שלבי יישום מודלים של סיווג נתונים

10.4 גאנט



פעילות	התחלת התוכנית	משך התוכנית	התחלה בפועל	משך בפועל	אחוז ביצוע
1- דוח מכין	1	42	0	0	0%
1.1 חיפוש מקורות וחומר רלוונטי לפרויקט	1	11	1	11	100%
1.2 לימוד החומר הרלוונטי	11	11	11	11	100%
1.3 חיפוש ולימוד איזה שפת תכנות הכי מתאימה	22	10	22	10	100%
1.4 כתיבת דוח מכין	32	10	32	10	100%
2- חקירה ראשונית של הפרוייקט	42	200	0	0	0%
2.1 לימוד עצמי של קורסים באינטרנט הקשורים לפרויקט	42	40	42	40	90%
2.2 לימוד נושא לימוד מכונה	82	40	82	40	100%
2.3 לימוד נושא אלגוריתמים של לימוד מכונה	122	60	122	60	75%
2.4 לימוד שפת פיתון	182	30	182	30	88%
2.5 הכנת הנתונים המתאימים לפרויקט	212	30	212	30	90%
3- דוח התקדמות	242	200	0	0	0%
3.1 בניית טבלאות הנתונים	242	60	242	60	100%
3.2 בדיקת טבלאות הנתונים , קורלציה, כמות, שלימות הנתונים	302	40	302	40	100%
3.3 קידוד טבלאות הנתונים	342	40	342	40	100%
3.4 מציאת האלגוריתמים המתאימים לטבלאות שבניתי	382	40	382	40	95%
3.5 כתיבת דוח התקדמות	422	20	422	20	100%
4- המשך לימודים וחקירה וכתיבת קודים בשפת פיתון	442	100	0	0	0%
4.1 לימוד אלגוריתם SVM	442	7	442	7	90%
4.2 לימוד אלגוריתם KNN	449	6.5	449	6.5	90%
4.3 לימוד אלגוריתם NB	455.5	6.5	455.5	6.5	90%
4.4 כתיבת הקודים המתאימים לניתוח הנתונים	462	10	462	10	95%
4.5 כתיבת הקודים המתאימים לבדיקת האלגוריתמים	472	10	472	10	95%
4.6 כתיבת הקודים לביצוע השוואה כללית של שלושת האלגוריתמים	482	15	482	15	95%
4.7 כתיבת קודים להרצאת אלגוריתמים אלה על הנתונים שלי ובדיקת האלגוריתם	497	15	497	15	95%
4.8 תוצאות ותפוקות של הפרויקט	512	30	512	30	90%
5- יום פרוייקט והדגמה	542	30	542	30	0%
5.1 כתיבת הדוח הסופי כולל הנושא השלישי	542	20	542	20	50%
5.2 הכנת מצגת עבור הפרויקט	562	8	562	8	50%
5.3 הצגת והדגמת הפרויקט	570	2	570	2	50%
6- הכנת ספר הפרויקט	572	40	0	0	0%
6.1 כתיבת הספר ובדיקתו	572	30	572	30	50%
6.2 הדפסת הספר הסופי	602	2	602	2	50%
6.3 הכנת והגשת מסמך RED	604	8	604	8	50%
7- הגנות	612	8	0	0	0%
7.1 הבאת הספר והמצגת ומסמך RED	612	2	612	2	50%
7.2 הגנה על הפרויקט	614	6	614	6	50%

איור 12- תרשים גאנט למשימות של הפרויקט

11 מקורות קריאה

[1] ا. عبده, في التطبيق النحوي والصرفي, الاسكندريه: دار المعرفه الجامعيه, 1992.

[2] م. التونجي, المعين في الاعراب والعروض والاملاء وعثرات اللسان, بيروت: عالم الكتب, 1995.

[3] ا. ج. قنبس, معجم الاملاء العربي, بيروت: دار الوسام, 1994.

[4] ح. ا. ع. اسماعيل احمد عمايرة, مهارات اللغة العربية:اختبار تشخيصي في النحو والصرف والصوت والمعجم والعروض والاملاء والمعنى, عمان: دار وائل, 2000.

[5] ت. ج. ك. امل كاظم ميرة، تطبيقات الذكاء الاصطناعي في التعليم من وجهه نظر الجامعه، المجلد 1، العراق: الجامعه العراقية , بغداد، 2019، p. 250.

[6] 5. د. م. م. عبيد، "(.التحليل المتقدم وتنقيب البيانات،" *دار الفكر العربي*، المجلد الاول، رقم الاولى، p. 58، 2018.

[7] A. Burkov, The hundred-page machine learning, Canada:: Andriy Burkov., 2019.

[8] F. Gorunescu, Data mining: concepts, models and techniques, Berlin Heidelberg: Springer-Verlag., 2011, p. 56.

[9] G. a. J. A. Kalyani, "Lakshmi Performance assessment of different classification techniques for intrusion detection," *Journal of Computer Engineering,* vol. 5, no. 7, p. 156, Nov-Dec. 2012.

[10] J. B. a. G. S. L. Michael, Data mining techniques for marketing, sales, and customer relationship management,, Indianapolis, Indiana: USA.: Wiley Publishing, Inc., 2004.

[11] ا. ا. ت. ب. (. Bayes، "هو من قام بصياغة حالة خاصة من النظرية المشهورة والتي تحمل اسمه وهي نظرية بيز،" *رغم أنها لم تنشر في حياته وإنما نشرت بعد وفاته بواسطة ريتشارد برايس*، المجلد 1، رقم 1، p. 200، عاش خلال الفترة 1701 -1761م.

[12] م. ب. م. ب. ع. أ. ا. ج. ا. ا. م. ا. ا. الإفريقي, لسان العرب, يروت: دار صادر, 1993.

[13] د. ع. م. الهليس، "تنقيب الآراء في جمل المقارنة العربية،" *المجلة العربية الدولية*، المجلد 4، رقم 2، p. 78، 2013.

[14] ر. ز. ع. ا. ا. د. غيداء عبد العزيز الطالب، "دراسة مقارنة لخوارزميات التنقيب في آالراء وتحميل العواطف وتطبيقاتها،" *مجلة الرافدين لعلوم الحاسوب والرياضيات* ، المجلد 12، رقم 2، p. 11، 2018.

[15] Available: .[מקוון] .A. 2020, "Asquero," 16 8 2020 https://www.asquero.com/article/advantages-and-disadvantages-of-different-types-of-machine-learning-algorithms/. [התבצעה גישה ב- שבת 10 2020].

[16] London, United ,*Computational Logic* -ב "R. A. Kowalski, " Logic Programming Kingdom, The Boulevard, Langford lane, Kidlington, Oxford, 2014, p. 736.

מכתב חוות דעת עבור חשיבות ותרומה של פרויקט לשפה וספרות הערבית

השוואה וניתוח אלגוריתמים ללימוד מכונה כדי לסווג את "همزتي الوصل والقطع" בשפה הערבית
פתרון הבעיה של שימוש שגוי ב- "همزتي الوصل والقطع" בטקסטים בשפה הערבית

מנחה : דר' יורם סגל .
סטודנט : נזאר מערוף.

מה תרומת פרויקט זה לשפה הערבית ואיך אפשר לשלב אותו לדברים אחרים , עם קשר בשפה הערבית ?

מטרת הפרויקט שהושלם זה היה לבנות מודל חכם שמסווג את "همزتي الوصل والقطع " (בתחילת מילה) באמצעות שימוש באלגוריתמים של סיווג נתונים על מנת לקבוע קריטריונים מדויקים ונכונים בכתיבת טקסטים בערבית בצורה מדויקת כדי לתרום ולעזור להתאים את הטכנולוגיה למדידת האיכות של אלגוריתמי הסיווג המפורסמים ביותר בהבחנה של שירות השפה הערבית. דגם חכם זה שעוצב יכול לתרום לפיתוח השפה הערבית באופן הבא:

- המודל החדש יכול לשמש כדי לסקור את המחקרים המדעיים ולהבטיח כי כתיבת את "همزتي الوصل والقطع " במילים כראוי, תורם לשלומם של אלמנטים של מחקר מדעי.
- המודל החדש יכול לשמש כדי לסקור מאמרי חדשות שפורסמו ברשתות החברתיות השונות על ידי התקנתו כתוסף לדפדפן כדי לוודא כי "همزتي الوصل والقطع " נכתבת כראוי בכל המילים של המאמרים שפורסמו, אשר תורם להתפתחות התקשורת הערבית החדשה.
- ניתן להשתמש בדגם החדש עם אפליקציות בטלפון ולהתייחס אליו כחלק מקורי ממערכות ההפעלה לניידים כדי להבטיח שהפקודות המתורגמות לערבית ומתחילות " بهمزة " נכתבות בצורה נכונה.
- ניתן להתאים את הדגם החדש למערכות תרגום המשמשות בכנסים, מטוסים ורכבות, ומסייעות להציג את "همزتي الوصل والقطع" כראוי.

בשורה התחתונה, אנו יכולים להשתמש במודל החדש ככלי תוכנה שניתן לשלב עם כל המערכות הטכניות המציגות טקסטים בערבית באמצעות מערכות תוכנה לסיווג "الهمزة " במילים שמתחילות ב- "همزتي الوصل والقطع".

דר. שחאדה הארון
ראש מרכז השפות
המכללה האקדמית הערבית לחינוך בישראל- חיפה

איור 13- מכתב חוות דעת