המרכז האקדמי רופין
Ruppin Academic Center
המחלקה להנדסת חשמל ומחשבים
Department of Electrical and Computer Engineering

גירסה 25.8

24/10/2021

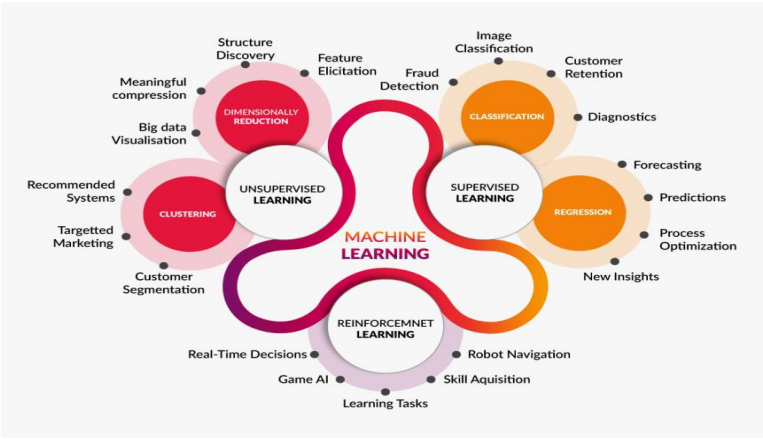
השוואה וניתוח אלגוריתמים ללימוד מכונה כדי לסווג את "همزتي الوصل والقطع"בשפה הערבית

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

קוד פרויקט: לדוגמא 21220570

דוח מכין - פרויקט גמר תשפ"ב

מנחה אקדמי : מר יורם סגל



מגיש

שם סטודנט: 025879123 נזאר מערוף

חתימת מנחים

תוכן עניינים

1 תקציר בעברית

השפה הערבית נחשבת לאורגניזם חי הגדל 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" 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.

פרויקט זה הסתמך על הידור של מילים בערבית המתחילות " بالهمزة " על ידי עיצוב שאלון דיגיטלי. המשימה של שאלון זה היא לאסוף את המספר הגדול ביותר של מילים המתחילות " بالهمزة " ולסווגן ל- "همزتي الوصل والقطع" על פי הכללים הדקדוקיים הננקטים בתהליך זה.

השאלון הופץ ברשת האינטרנט ומולא על ידי חמישים מומחים בתחביר בדרגים אקדמיים שונים. המספר הכולל של המילים המסווגות הגיע ל-400 מילים, ולאחר עיבוד ואי הכללה של המילים החוזרות, 101 מילים, השגנו 299 מילים תקפות ליישום במודל המסווגן, ובהתבסס על גודל וסוג הנתונים שנאספו 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 שבחרנו להשתמש בהם וקבלת התוצאות הסופיות, נחליט איזה אלגוריתם הוא הטוב ביותר לתהליך תיקון כתיבת "همزة الوصل وهمزة القطع" במאמרים בשפה הערבית .

מילות מפתח: אינטליגנטי מלאכותי, אלגוריתמי למידת מכונה, "الهمزة", "الوصل", "القطع", אלגוריתמי סיווג, שפה ערבית.

2 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 [1].

סוגי בינה מלאכותית:

על פי יכולותיה, ניתן לחלק את הבינה המלאכותית לשלושה סוגים [1]. 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 [2]

- **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. [3]

- **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. [4]

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. [5]

- **למידת מכונה:**

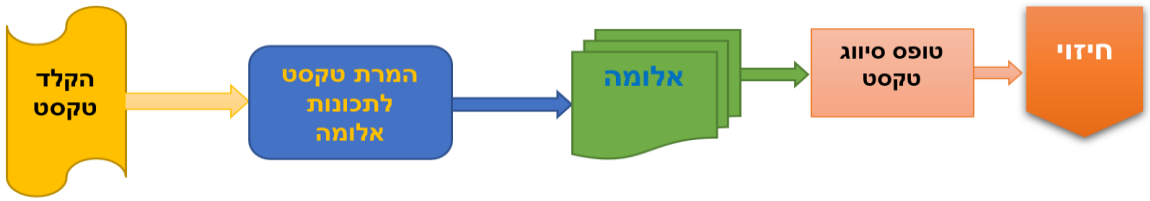
למידת מכונה היא ענף של בינה מלאכותית, הכולל תכנון ופיתוח של אלגוריתמים וטכניקות המאפשרות למחשבים להיות בעלי תכונות "למידה". באופן כללי, הלמידה מתחלקת לשתי רמות: אינדוקטיבית ודדוקטיבית, כאשר הגישה הדדוקטיבית מוציאה כללים ושיפוטים כלליים מביג דאטה.

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. [5]

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 (1-1): 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. [6]

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 [7] 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. [3]

רביעית: החמץ אל-ואסל וקט '

בחלק זה של הצד העיוני, נעסוק "همزة الوصل وهمزة القطع" בתחילת המילה.

2.1 כיצד קשור או משתלב הפרויקט לתחום כללי כלשהו

מטרת הטכנולוגיה של למידת מכונה היא לייעל את ביצועי המערכת בעת טיפול במקרים חדשים של נתונים באמצעות **לוגיקת תכנות** [8]מוגדרת על ידי משתמש עבור סביבה נתונה .כדי להשיג מטרה זו ביעילות, למידת מכונה נשענת רבות על סטטיסטיקה ומדעי המחשב .שיטות סטטיסטיות מספקות אלגוריתמים של למידת מכונה דרכים להסיק מסקנות מנתונים. יש הרבה **חסרונות** [9]שונים של סוגים שונים של אלגוריתמים של למידת מכונה .

החסרונות של אלגוריתמים של למידת מכונה בפיקוח:

- ☒ השיעורים עשויים שלא להתאים לשיעורים ספקטרליים.
- ☒ עקביות משתנה בשיעורים.
- ☒ עלות וזמן כרוכים בבחירת נתוני הכשרה.

חסרונות באלגוריתמים של למידת מכונה ללא פיקוח:

- ☒ המחלקות הספקטראליות אינן מייצגות בהכרח את התכונות בשטח.
- ☒ הוא אינו מתחשב ביחסים מרחביים בנתונים.
- ☒ זה יכול לקחת זמן לפרש את המעמדות הספקטרליים.

חסרונות האלגוריתמים של למידת מכונה בפיקוח למחצה:

- ☒ תוצאות האיטרציה אינן יציבות.
- ☒ זה לא ישים לנתונים ברמת הרשת.
- ☒ יש לו דיוק נמוך.

חסרונות באלגוריתמים של למידת מכונות חיזוק:

- ☒ למידה רבה מדי של חיזוק יכולה להוביל לעומס יתר של מצבים שיכולים להקטין את התוצאות.
- ☒ אלגוריתם זה אינו עדיף לפתרון בעיות פשוטות.
- ☒ אלגוריתם זה צריך הרבה נתונים והרבה חישובים.
- ☒ קללת הממדיות מגבילה את למידת החיזוק למערכות פיזיות אמיתיות.

השוואת אלגוריתמים של למידת מכונה (MLA) חשובה כדי לצאת עם האלגוריתם המתאים ביותר לבעיה מסוימת, לכן בפרויקט זה ננסה לזהות את אלגוריתמי הסיווג המתאימים ביותר בתוך אלגוריתמי למידת מכונה כדי לקבוע את המיקומים הנכונים לשימוש ב- "همزتي الوصل والقطع" בתחילת מילה. פתרון בעיית השימוש השגוי ב- "همزتي الوصل والقطع" בטקסטים בערבית. מדידת האיכות של אלגוריתמי הסיווג המפורסמים ביותר בהבחנה בין "همزتي الوصل والقطع" בתחילת מילה. אנו יכולים לעשות זאת באמצעות חבילת פיתון scikit-learn. נלמד כיצד להשוות מספר רב של אלגוריתמי לימוד מכונה בכל פעם באמצעות יותר מסטטיסטיקות התאמה המסופקות על ידי scikit-learn וגם ליצור עלילות להמחיש את ההבדלים.

2.2 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?

2.3 האתגר הטכנולוגי

פרויקט זה הוא 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. היא עזרה לי לגשת לפרויקט שלי, שני המחקרים הדומים הם:

מחקר שכותרתו:

חקר דעות במשפטים השוואתיים בערבית [10]

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.

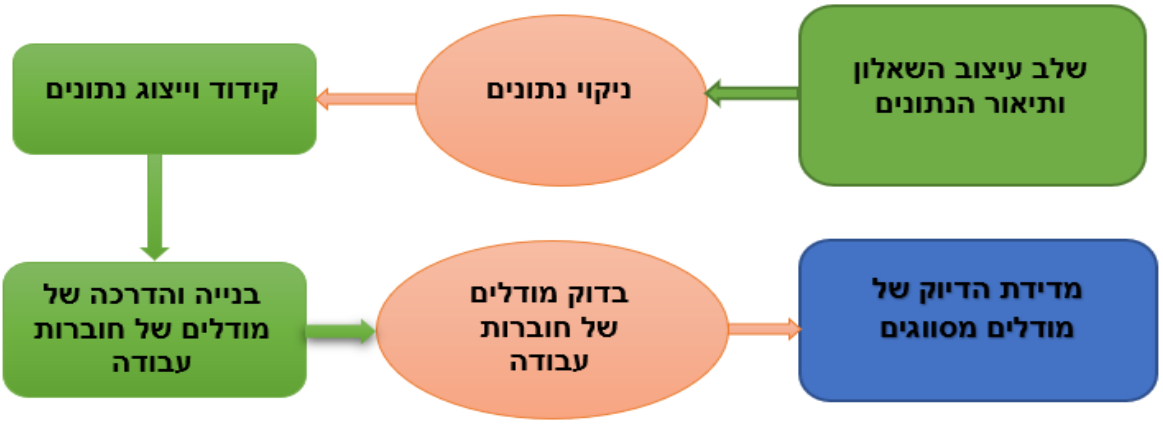
מחקר שכותרתו:

מחקר השוואתי של אלגוריתמים של כריית דעות וניתוח רגשות ויישומיהם [11]

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.

3 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 (2-1):



איור (2-1): 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:**

המילים הערביות המתחילות "بالهمزة" נאספו ע"י עיצוב שאלון דיגיטלי ראה (נספח ו') שעוצב באמצעות Google Forms וקיבל את שמו (השאלון לסיווג מילים המתחילות "بهمزة وصل أو قطع"). מומחים בשפה הערבית סיווגו אותו " لهمزتي وصل أو قطع" לפי כללים דקדוקיים ידועים. השאלון פורסם גם בכתובת האינטרנט הבאה:

https://docs.google.com/forms/d/e/1FAIpQLSdtwnflWQ7hoTdVbKfZUHDVix4fmXPE3grXMiQGIDem9Dsc0Q/viewform

50 מומחים בדקדוק בדרגות שונות נרתמו למילואי השאלון על מנת לקבל מדגם הומוגני, והמספר הכולל של מילים מסווגות הגיע ל-400 מילים. המילים שהופקו מהטקסטים לדוגמה סווגו למילים שמתחילות " بهمزة وصل" המכונה בפרויקט זה (Wasl) ולמילים המתחילות " بهمزة قطع" המכונה בפרויקט זה (Gtaa). המשתנה התלוי (Val) נקבע לפי שני הערכים (Gtaa/Wasl). באשר למשתנים (המאפיינים) הבלתי תלויים, הם חולקו לשלושה מאפיינים על פי הכללים הדקדוקיים שאליהם אתייחס איליו מאוחר יותר.

- **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 Jupyter.

- **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).

4 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".

ראשית: בדוק את המודלים של חוברות העבודה

לאחר השלמת שלב בניית מודלים מסווגים באמצעות אלגוריתמים (NB, SVM KNN), עברנו לשלב בדיקת המודלים הללו על מנת שנוכל למדוד את איכותם מאוחר יותר. המודלים נבדקו ע"י ביצוע תחזיות על המודלים לאחר הכשרתם באמצעות פונקציות ספרייה sklearn .

בפרק זה של הפרויקט אכסה את שני השלבים האחרונים כהמשך לארבעת השלבים שהזכרתי בפרק קודם.

שנית: מדידת הדיוק של מודלים מסווגים

שלב הערכת התוצאות של מודלים לכריית נתונים הוא אחד השלבים החשובים המאפשרים לנו להגדיר את המודל היעיל ביותר, ויעילות המודל נמדדת באמצעות דיוק התוכנית הרלוונטית.

אופי הנתונים המשמשים לבניית מודלים ממלא תפקיד מרכזי ביעילותם, וישנן שיטות סטטיסטיות רבות הבוחנות מודלים של סיווג, והחשובים שבהם הם כדלקמן:

- **דיוק ממוצע (Accuracy Average)**

זהו הממוצע האריתמטי של היחסים בין דיוק החיזוי הנכון עבור כל קטגוריה שסופק על ידי המודל למספר הדירוגים בפועל עבור אותה קטגוריה במערך נתוני הבדיקה.

- **חישוב דיוק כולל (Accuracy Total)**

זהו היחס בין סכום התחזיות הנכונות שנתן המודל לסכום הדירוגים בפועל במערך נתוני הבדיקה.

- **מציאת מטריצת הבלבול (Matrix Confusion)**

מטריצת הבלבול מציגה את מספר המקרים החזויים בצורה נכונה ואת מספר המקרים החזויים השגויים במערך המבחנים עבור כל פריט בהשוואה למספר המקרים בפועל עבור אותם פריטים. הדרגה של מטריצה זו היא n × n כאשר n הוא מספר הפריטים בעמודה של משתנה היעד (המשתנה התלוי).

שלישית: כלים וציוד

תוכנה:

- תכנות Python בסביבת פיתוח Pycharm של חברת JetBrain

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 (1-3): מציג כלים של חומרה והשוואה ביניהם

5 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:

6 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 מאך 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 לשישה שלבים, החל משלב התכנון. השאלון ואיסוף הנתונים וכלה בשלב מדידת הדיוק של מסווגי הנתונים.

7 מקורות קריאה

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

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

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

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

[5] 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.

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

[7] ا. 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.

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

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

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

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

[12] A. R. A. Rawan A., S. Muhammed N. and F. Polla, "A comprehensive study on sign languages recognition systems using (SVM, KNN, CNN and ANN)," in *Proceedings of the First International Conference on Data Science (DATA '18)*, Madrid, Spain, 2018.

[13] M. Xian, "A deep learning framework for assessing physical ו Y. Liao, A. Vakanski rehabilitation exercises," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, כרך 28, 2020, p. 468–477.

[14] K. Y. L. L. S. L. Jiefu Zhai, "A low complexity motion compensated frame interpolation method," ב- *IEEE International Conference: Circuits and Systems, 2005. ISCAS 2005.*, 2005.

[15] W. M. Y. H. Yanli Li, "A Spatial Prediction-Based Motion-Compensated Frame Rate Up-Conversion," January 2019.

[16] G. Hidalgo, "CMU-Perceptual-Computing-Lab/OpenPose," GitHub, 19 December 2018. [מקוון]. Available: https://github.com/CMU-Perceptual-Computing-Lab/openpose.

[17] T. Simon, H. Joo, I. Matthews ו Y. Sheikh, "Hand keypoint detection in single images using multiview bootstrapping," *InProceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 1145-1153, 2017.

[18] ع. م. اهلليس، "تنقيب اآلراء يف مجل املقارنة العربية،" *املجلة العربية الدولية للمعلوماتية، املجلد الثاين، العدد الرابع، 2013م*، المجلد 1، رقم 1، p. 7، Jul. 2013.

נספחים

רשימת נספחים

נספח א - סכמת בלוקים

נספח ב - טבלת אבני דרך ותוצרים

נספח ג - טבלת משימות (לא חובה במסגרת דוח מכין - לפי החלטת המנחה)

נספח ד- מסמכים רלוונטיים של המרכז האקדמי רופין

נספח ה - גאנט (אופציונלי בהתאם לשיקול דעתו של המנחה)

נספח א - 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 1-4: סכמת בלוקים , חלוקת שלבי יישום מודלים של סיווג נתונים

נספח ב - טבלת אבני דרך ותוצרים

מטרת נספח זה היא הוכחת כושר תכנון זמנים עתידי, תחת תנאי אי וודאות.

עליכם להכין במסגרת דוח המכין טבלת אבני דרך נפרדת, המציגה רשימת אבני דרך ממוספרות (7 אבני דרך בדיוק!!!) 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	60	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	200	בניית תוכנית והרצת נתונם 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 :580 שעות

תוצר מדיד:

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.

הערה – הצגת תוצר מדיד בטבלת אבני דרך במסגרת דוח המכין הוא מעיקרי הדוח.

נספח ג - טבלת משימות (לא חובה במסגרת דוח מכין - לפי החלטת המנחה)

במסגרת דוח ההתקדמות יש להכין טבלת משימות. (אם טבלת אבני הדרך היא רשימת פרקים, אזי טבלת משימות היא רשימת תתי פרקים). לכל אבן דרך, יש להציג את רשימת המשימות המובילות להשלמת אבן הדרך וקבלת התוצר המדיד (כולל הערכת משך הזמן הנדרש להשלמת כל משימה (טבלה המשימות היא טבלה נפרדת המכילה את כל המשימות יחדיו של כל אבני הדרך. מספור המשימות הוא בהתאם למספור אבני הדרך – 1.1, 1.2 .1.3 וכ"ו).

להלן רשימת העמודות שיש לכלול בטבלת המשימות:

- מספר משימה
- שם משימה
- מועד התחלה משוערך
- מועד סיום משוערך
- הקצעת שעות עבודה
- תוצר ביניים
- מועד סיום בפועל
- סה"כ שעות בפועל

נספח ד

במידת הצורך, יש לצרף חתימה על מסמכים רלוונטיים של המרכז האקדמי רופין. הנחיות פרטניות תימסרנה בעתיד.

נספח ה - גאנט (אופציונלי בהתאם לשיקול דעתו של המנחה)