

30/10/202

ALMAGOR CLICK PRO

מנחה: ד"ר רינה אזולאי
מגישה: יעל יפת



תוכן עניינים

1. תקציר	2
1.1 תיאור הבעיה	2
1.2 מטרות הפרויקט	2
2. הקדמה	3
2.1 מבוא	3
2.2 סקירת הספרות	3
3. הנתונים	4
3.1 תיאור הנתונים:	4
4. שיטות	4
4.1 עיבוד מקדים (Preprocessing)	4
4.1.1 עיבוד מקדים -קטגוריות מסמכים	4
4.1.2 עיבוד מקדים מסמכים	5
4.2 איזון הנתונים	6
4.3 ייצוג טקסט (Text Vectorization)	6
4.4 בחירת מודלים ואימון	7
4.4.1 המודלים שנבחנו	7
4.5 הערכת ביצועים	10
5. תוצאות	10
6. דיון	11
7. מסקנות	11
8. המלצות לעתיד	12
9. ביבליוגרפיה	13
10. נספחים	14

1. תקציר

1.1 תיאור הבעיה

חברת אלמגור פיתחה מערכת ייחודית בשם "Almagor Click" המיועדת לסוכני ביטוח הרשומים בה. באמצעות מערכת זו, הסוכנים יכולים לבקש מידע על לקוחותיהם ישירות מחברות הביטוח. התהליך מתחיל בלחיצה על כפתור "בקשת מידע", והמערכת פונה לחברת הביטוח בבקשה לקבלת מידעים. חברות הביטוח מחויבות לספק את כל המידע הביטוחי הקיים על הלקוח המבוטח. לאחר קבלת הבקשה, הן שולחות הודעות המכילות קבצי מידע על הלקוחות למערכות Web של החברה או לאימייל ייעודי. תהליכים אוטומטיים במערכת עוברים על ההודעות שהתקבלו, שומרים את פרטי ההודעה בבסיס הנתונים ומאחסנים את הקבצים בתיקיות מתאימות. אנשי צוות של אלמגור נכנסים למערכת הניהול בה מוצגות כל ההודעות שהתקבלו מחברות הביטוח. כל הודעה שמכילה קובץ, נבדקת, הקובץ נקרא ומשויך באופן ידני ללקוח הרלוונטי ולקטגוריה המתאימה. לאחר סיום השייך, הסוכן מקבל הודעה שהמידעים התקבלו ויכול להיכנס לתיק הלקוח במערכת ולצפות במידע ששלחה חברת הביטוח ובסוג המידע שהתקבל בקובץ. כך, המערכת מביחה ניהול יעיל ומסודר של המידע הביטוחי, ומסייעת לסוכני הביטוח לספק שירות מהיר, מקיף ועדכני ללקוחותיהם.

1.2 מטרות הפרויקט

פיתוח מערכת "Almagor Click" משודרגת, שבה תהליך זיהוי תעודת הזהות והסיווג הקטגוריאלי של הקבצים מתבצע באופן אוטומטי, תוך חיסכון בזמן ומשאבים ושיפור היעילות של ניהול המידע הביטוחי. במערכת זו 2 משימות סיווג:

1. זיהוי אוטומטי של תעודת הזהות של הלקוח באמצעות ביטוי רגולרי והתאמה לת.ז בגינה הוגשה הבקשה לקבלת מידעים.
 2. סיווג אוטומטי של סוג המסמך: פיתוח מודל למידת מכונה שיזהה את סוג המסמך ויסווג אותו לקטגוריה המתאימה. הטמעת המודל במערכת כך שכל קובץ שנקלט יעבור תהליך סיווג אוטומטי לקטגוריה המתאימה.
- בדוח זה אתאר את תהליך בניית הפרויקט, שהתמקד במשימה השנייה – סיווג אוטומטי של המסמכים.

2. הקדמה

2.1 מבוא

תיוג מסמכים באמצעות למידת מכונה הוא כלי חשוב בעיבוד שפה טבעית, המאפשר סיווג אוטומטי של מסמכים לקטגוריות מוגדרות מראש. ככל שכמות המידע הטקסטואלי גדלה בעידן הדיגיטלי, יכולת סיווג אוטומטי זו הופכת להיות קריטית לניהול, חיפוש וניתוח מסמכים בצורה יעילה בארגונים. התיוג האוטומטי חוסך משאבים וזמן בהשוואה לתיוג ידני ומסייע בקבלת החלטות מבוססות נתונים על ידי הנגשת מידע רלוונטי. הטכנולוגיה כיום מאפשרת סיווג מסמכים מורכבים במגוון תחומים, תוך שימוש בעיבוד שפה טבעית ובראייה ממוחשבת. אף על פי שישנם אתגרים כמו דיוק הסיווג והתמודדות עם מסמכים מורכבים, פתרונות מבוססי בינה מלאכותית תומכים בניהול חכם ויעיל של כמויות מידע עצומות, ומשפרים את התפועול והיעילות הארגונית.

2.2 סקירת הספרות

תיוג מסמכים וטקסטים הוא תחום מרכזי בעיבוד שפה טבעית (NLP) ולמידת מכונה, המתמקד בסיווג אוטומטי של טקסטים לקטגוריות מוגדרות מראש. תחום זה התפתח משמעותית לאורך השנים, מגישות פשוטות ועד לשיטות מתקדמות המבוססות על בינה מלאכותית.

בתחילה, תיוג טקסטים התבסס על גישות מסורתיות, שכללו מודלים מבוססי חוקים. שיטות מוקדמות השתמשו בחוקים ידניים ומילות מפתח לסיווג טקסטים.

לאחר מכן, שיטות סטטיסטיות כמו Naive Bayes הפכו לפופולריות בשל יעילותן בסיווג טקסטים. אלגוריתמים נוספים כמו Support Vector Machines (SVM) הראו שיפור משמעותי בביצועים.

עם הזמן, התקדמו הגישות ללמידת מכונה, והחלו להופיע מודלים מתקדמים יותר ששיפרו את הדיוק והיציבות של הסיווג, כמו עצי ההחלטה, Random Forest, שיטות Boosting כמו AdaBoost ו XGBoost.

מהפכה נוספת בתחום התרחשה עם כניסת הלמידה העמוקה. רשתות נוירונים הפכו לכלי מרכזי בעיבוד טקסטים, רשתות קונבולוציה (CNN) הוצגו לסיווג משפטים. מודלים כמו LSTM תרמו לעיבוד רצפים טקסטואליים. במקביל לכך, פותחו שיטות לייצוג וקטורי של מילים, כמו Word2Vec ו GloVe, שאפשרו ייצוג טוב יותר של מילים ושיפרו את הבנת ההקשרים בטקסט.

במהלך השנים האחרונות, מודלים מבוססי טרנספורמרים הפכו לדומיננטיים בתחום ה NLP. ארכיטקטורת הטרנספורמר הובילה למהפכה בתחום זה, עם מודלים כמו BERT ו GPT-3 שהשיגו תוצאות פורצות דרך במגוון משימות NLP. כיום, גישות עכשוויות מתמקדות ב Transfer Learning, שיטה להעברת ידע ממודלים מאומנים מראש למשימות ספציפיות, ב Few-shot Learning, המאפשרת למידה ממספר מועט של דוגמאות ועוד.

עם כל ההתפתחויות הללו, עדיין קיימים אתגרים משמעותיים בתחום תיוג המסמכים והטקסטים. בין האתגרים ניתן למצוא את הצורך ביעילות חישובית גבוהה יותר, פרשנות ושקיפות של המודלים המורכבים, והצורך להתמודד עם הטיות במודלים. לסיכום, תחום תיוג המסמכים והטקסטים עבר התפתחות מרשימה מגישות פשוטות ועד למודלים מתקדמים של למידה עמוקה.

המחקר העכשווי מתמקד בשיפור היעילות והדיוק וביכולת להתמודד עם אתגרים מורכבים כמו מיעוט נתונים מתויגים ורב-לשוניות. ההתקדמות המהירה בתחום מבטיחה המשך חדשנות ושיפור ביכולות תיוג ועיבוד טקסטים.

3. הנתונים

3.1 תיאור הנתונים:

מאגר הנתונים כולל **224,192** מסמכי Pdf שהתקבלו מתשע חברות ביטוח שונות בדואר אלקטרוני או במערכות אינטרנט ייעודיות. כל מסמך שייך לקטגוריה מסוימת, ובכל חברה קיימת חלוקה קטגוריאלית שונה בהתאם לאופי הפעילות והצרכים שלה.

דוגמאות לקטגוריות השונות: פוליסת ביטוח חיים, דוח שנתי ביטוח בריאות, פרטי ביטוח, גילוי נאות קבוצתי שיניים ועוד. בפרק נספחים טבלה 1 מצורפת תמונת ההתפלגות לכל חברה של מספר המסמכים לכל קטגוריה.

4. שיטות

4.1 עיבוד מקדים (Preprocessing)

4.1.1 עיבוד מקדים -קטגוריות מסמכים

ניתוח הנתונים העלה שתי סוגיות מרכזיות:

1. יעילות מספר הקטגוריות: קיימות מספר רב של קטגוריות לכל חברה ולכן, יש לבחון האם הכמות הנוכחית של קטגוריות לכל חברה היא אכן הכרחית.
 2. פערים בהתפלגות המסמכים: נצפתה שונות משמעותית בכמות המסמכים בין הקטגוריות השונות. בעוד שחלק מהקטגוריות מכילות כמות גדולה של מסמכים, אחרות כוללות מספר זניח בלבד.
- בשלב הראשון בעיבוד המקדים, צמצמתי את כמות הקטגוריות במטרה להשאיר את הקטגוריות הרלוונטיות ובכך לשפר את איזון הנתונים ולהגדיל את דיוק המודל. מספר תובנות מרכזיות עלו מתוך תהליך זה:
- **איחוד קטגוריות עם משמעות דומה:** נמצא כי חלק מהקטגוריות מכילות מסמכים בעלי תוכן או מבנה דומים, מה שאפשר לאחד אותן ולצמצם את המורכבות ללא אובדן משמעותי של מידע.
 - **קטגוריות זהות המחולקות לפי שנים:** חלוקה לפי שנים יצרה הבדלים בין מסמכים דומים מאוד בתוכן אך שונים בהתייחסות לשנה, לדוגמא, הקטגוריות: דוח שנתי 2021, דוח שנתי 2022, דוח שנתי 2023 אוחדו לקטגוריה אחת, דוח שנתי. חלוקת הקטגוריות לשנים פחות נחוצה, ולכן איחדתי כדי להקל על הסיווג.
 - **קטגוריות עם מעט מסמכים:** זוהו קטגוריות שבהן כמות המסמכים נמוכה מ-30, לעומת קטגוריות אחרות המכילות אלפי מסמכים. הפערים הללו מעוררים חשש שהקטגוריות הקטנות עלולות להוביל להטיית המודל ולהגביר את אחוז הטעויות, ולכן החלטתי לאחד אותן עם קטגוריות דומות.

לסיכום, תהליך צמצום ואיחוד הקטגוריות נועד להפחית את מורכבות הנתונים ולשפר את ביצועי המודל. תהליך זה השיג את מטרתו והצליח להפחית באופן משמעותי את מספר הקטגוריות לכל חברה. הטבלה הבאה מציגה את מספר הקטגוריות לכל חברה לפני ואחרי תהליך העיבוד:

חברת הביטוח	מספר המסמכים	מספר קטגוריות לפני תהליך העיבוד	מספר קטגוריות אחרי תהליך העיבוד
הפניקס	69113	73	31
הראל	59400	28	16
כלל	32582	90	16
מנורה	27667	70	30
מגדל	20355	22	7
ביטוח ישיר	6044	18	13
איילון	5663	30	10
AIG	2825	14	10
הכשרה	532	3	3

4.1.2 עיבוד מקדים מסמכים

במהלך עיבוד מסמכי PDF נבחנו 2 סוגי מסמכים:

1. מסמכי **PDF** קריאים: מסמכים שניתן לקרוא את התוכן שלהם, השתמשתי בספריית Py2PDF, שמאפשרת שליפת טקסט בעברית באופן תקין מקבצי ה-PDF.
 2. מסמכי **PDF** מבוססי תמונה: מסמכים אלה מבוססים על סריקות או תמונות, ולכן התוכן שלהם אינו נגיש כטקסט באופן ישיר. עבור סוג זה השתמשתי בספריית Tesseract-OCR, שניתנת להתאמה לקריאת השפה העברית. Tesseract מבצעת המרה של תמונות לטקסט (OCR - Optical Character Recognition), כך שניתן לשלוף את התוכן הטקסטואלי גם ממסמכים מבוססי תמונה.
- לאחר שלב שליפת הטקסט מכל סוגי המסמכים, עברנו לשלב של ניקוי נתונים (data cleaning).
- ניקוי זה כלל מספר פעולות:
- הסרת תווים מיוחדים, רווחים ושורות מיותרות.
 - הסרת מילות עצירה (stopwords) שאינן תורמות להבנת התוכן באמצעות ספריית nltk והתאמתה לעברית.
- ביצוע אנונימיזציה:
- הסרת מספרי תעודות זהות באמצעות ביטויים רגולריים (Regex).
 - זיהוי והסרה של שמות אנשים, ערים ותאריכים ע"י מודל **Dicta** - מודל עיבוד שפה טבעית (NLP) לזיהוי שמות ישויות (NER) המודל מזהה ומסווג שמות של אנשים, מקומות, ארגונים וכדומה בטקסטים בעברית.

לאחר תהליך הניקוי, החלטתי להתמקד ב-500 התווים הראשונים מכל מסמך לצורך הסיווג, בדקתי שדי בתווים אלו כדי לקבוע את סוג המסמך. יצוין שגם בתהליך התיוג הידני מתמקדים בעיקר בקריאת העמוד הראשון של המסמך לצורך סיווג. בסיום התהליך, הנתונים שנאספו נשמרו בקובץ CSV, הכולל את השדות הבאים:

1. הטקסט הנקי, 2. התיוג של סוג המסמך, 3. מספר מזהה של ההודעה המקורית שממנה נלקח המסמך. כך, במידת הצורך, ניתן לחזור למסמך המקורי לצורך בדיקות או מענה על שאלות שיתעוררו בהמשך.

4.2 איזון הנתונים

חוסר איזון בנתונים (imbalanced data) מהווה אתגר משמעותי בפיתוח מודל מדויק ואמין. כדי להתגבר על אתגר זה, יש ליישם פתרונות ייחודיים הן בשלב עיבוד הנתונים והן במהלך פיתוח המודל. מטרת הפתרונות היא למנוע נטייה לא פרופורציונלית של המודל כלפי הקטגוריות הגדולות יותר, ובכך להבטיח דיוק ואמינות גבוהים יותר בתוצאות. במטרה להתמודד עם חוסר איזון בנתונים, בחנתי שתי אפשרויות במודלים השונים:

1. שימוש ב RandomOverSampler, - שמבצע דגימה חוזרת של קטגוריות המיעוט כדי לאזן את התפלגות הנתונים.
2. פרמטרים ייעודיים במודלים המסוגלים להעניק משקל יחסי לכל קטגוריה בהתאם לשכיחותה, כדוגמת הפרמטר `class_weight='balanced'`.

לאחר ניסויים, החלטתי להימנע משימוש בפרמטרים אלו, שכן הם השפיעו לרעה על ביצועי המודל ולא הביאו לשיפור התוצאות.

4.3 ייצוג טקסט (Text Vectorization)

בפרויקט זה, נעשה שימוש בטכניקות שונות לייצוג טקסט (Text Vectorization). מטרת ייצוג הטקסט היא להפוך את המידע הטקסטואלי לפורמט מספרי המאפשר למודלים של למידת מכונה לעבד אותו ולסווג את המסמכים בצורה מדויקת. בעבודה זו נעשתה השוואה בין שלוש שיטות מרכזיות לייצוג טקסט:

1. TF-IDF (Term Frequency-Inverse Document Frequency): שיטה זו מחושבת על בסיס השכיחות

היחסית של מילה במסמך לעומת השכיחות שלה בכל המסמכים. היתרון המרכזי של TF-IDF הוא היכולת להדגיש מילים ייחודיות במסמך מסוים ולהפחית את המשקל של מילים נפוצות בכל המסמכים. עם זאת, השיטה מתבססת על חישוב ליניארי של מילים ולכן אינה מתחשבת בקשרים המורכבים בין המילים או במשמעות הסמנטית שלהן.

2. Word2Vec: שיטה ליצירת הטמעות (Embeddings) של מילים המבוססת על רשתות נוירונים. Word2Vec

ממפה מילים לוקטורים מספריים, כאשר מילים בעלות הקשר דומה יהיו קרובות זו לזו במרחב הווקטורי. היתרון של Word2Vec הוא בכך שהוא שומר על ההקשר הסמנטי של מילים אך, בצורה שאינה תלויה בסדר שלהן במשפטים. שיטה זו דורשת כמות משמעותית של נתונים על מנת לאמן את המודל בצורה טובה.

3. Contextual Embeddings: מספקות ייצוג דינמי לכל מילה בטקסט, תוך התחשבות בהקשר של כל שאר

המילים במשפט. כך, מילה אחת יכולה לקבל ייצוג שונה בהתאם למשמעותה בהקשר מסוים. מבחינה טכנית, ייצוגים אלו מכילים מידע על המשמעות הסמנטית והמבנה הכללי של הטקסט. בפרויקט זה השתמשתי בהטמעות

ההקשריות של AlephBERT. למרות היתרונות הברורים של ייצוג זה, החיסרון שלו הוא זמן חישוב ממושך יותר ודרישות גבוהות למשאבי עיבוד.

בפרויקט זה, חקרתי את השפעתן של שיטות Word2Vec, TF-IDF ו-Contextual Embeddings על ביצועי מודלי למידת מכונה. כל אחת מהשיטות מציעה גישה ייחודית לייצוג טקסט, והשפעתה על ביצועי המודלים משתנה. ההשוואה שערכתי סייעה להבנה מעמיקה של היתרונות והחסרונות של כל שיטה, והציגה תובנות לגבי התאמת הייצוגים למודלים השונים. תוצאות ההשוואה מדגישות את החשיבות שבבחירת שיטת הייצוג המתאימה ביותר לנתונים ולמודל שנבחר.

4.4 בחירת מודלים ואימון

בחרתי לפתח מודל נפרד עבור כל חברת ביטוח. גישה זו נבחרה בשל מספר יתרונות משמעותיים:

1. התאמה מדויקת למאפיינים הייחודיים של כל חברה, כולל טרמינולוגיה, מבנה מסמכים וקטגוריות ספציפיות.
2. תחזוקה ועדכון קלים יותר של המודלים בעתיד, ללא השפעה על מודלים של חברות אחרות.
3. אפשרות להשוואה טובה יותר בין ביצועי המודלים עבור חברות שונות.

עבור כל חברת ביטוח, ערכתי השוואה בין מספר מודלים לסיווג טקסט ובחנתי את השפעת שיטות ייצוג טקסטואלי TF-IDF ו-Word2Vec על ביצועי המודלים, כדי לזהות את השילוב האופטימלי של מודל ושיטת ייצוג עבור כל חברה. תהליך פיתוח המודלים כלל כיוון של הפרמטרים. השתמשתי ב GridSearchCV - עבור מודלים כמו Logistic Regression ובכיוון ידני עבור מודלים מורכבים יותר. בחנתי מגוון רחב של ערכים אפשריים לכל פרמטר, במטרה להבטיח ביצועים אופטימליים עבור המאפיינים הייחודיים של נתוני כל חברת ביטוח. גישה זו אפשרה לא רק לבחור את המודל המתאים ביותר לכל חברה, אלא גם להבין את מאפייני הנתונים ואת השפעת שיטות הייצוג השונות על יכולת הסיווג. התוצאה היא פתרון מותאם ויעיל במיוחד עבור כל אחת מחברות הביטוח, המתחשב במורכבות ובייחודיות של המסמכים שלהן.

4.4.1 המודלים שנבחנו

K-Nearest Neighbors (KNN)

אלגוריתם למידת מכונה המסווג נקודות חדשות על סמך קרבתן לנקודות קיימות במערך האימון. עקרון הפעולה: האלגוריתם מזהה את k השכנים הקרובים ביותר לנקודה החדשה ומסווג אותה לפי הרוב מביניהם. יתרונו המרכזי של KNN הוא פשטותו והעובדה שאינו מניח הנחות מוקדמות לגבי התפלגות הנתונים. עם זאת, הוא עשוי לדרוש משאבי חישוב רבים כאשר מספר הדוגמאות גדל, מכיוון שהמרחק לכל דוגמה נמדד בזמן אמת.

פרמטרים ותצורה:

- נבחר ערך $k=15$ כלומר הסיווג של כל מסמך חדש יתבסס על 15 המסמכים הקרובים אליו ביותר מקבוצת האימון.
- משקלים: הוגדרו משקלים המבוססים על המרחק, כך ששכנים קרובים יותר משפיעים יותר על הסיווג. זאת בניגוד לברירת המחדל של משקלים אחידים.

- מדידת מרחק: עבור נתונים טקסטואליים שהומרו לווקטורים באמצעות TF-IDF או Word2Vec בחרתי במרחק קוסינוס (Cosine similarity) מדד זה נחשב מתאים יותר לטקסט, שכן הוא מודד את הזווית בין הווקטורים ומתמקד במבנה ולא באורך של המילים.

Random Forest

אלגוריתם למידת מכונה מבוססת אננסמבל, המשמש לסיווג ורגרסיה.
עקרון הפעולה: Random Forest פועל על ידי יצירת מספר רב של עצי החלטה, כאשר כל עץ מספק תחזית עצמאית. בתהליך הסיווג, התוצאה הסופית נקבעת על פי רוב הקולות מבין כל העצים. האלגוריתם משתמש בטכניקת Bootstrap Aggregating (bagging), המבצעת דגימה חוזרת של נתונים כך שכל עץ נבנה על בסיס תת-דגימה שונה. שיטה זו מפחיתה את השונות, משפרת את דיוק המודל ומסייעת במניעת Overfitting.

פרמטרים ותצורה:

- מספר עצים: 150 עצים, מה שמאפשר דיוק גבוה תוך איזון עם זמן העיבוד והמשאבים הנדרשים.
- עומק העצים: 20, מאפשר למידת דפוסים מורכבים תוך שמירה על איזון ומניעת Overfitting.
- מספר דוגמאות מינימלי לפיצול: נקבע ל-10, משפיע על מורכבות העצים ומסייע במניעת Overfitting.

XGBClassifier

אלגוריתם למידת מכונה המבוססת על טכניקת Gradient Boosting, המצטיינת בביצועים גבוהים גם על מערכי נתונים גדולים.

עקרון הפעולה: XGBoost פועל על ידי בניית סדרה של עצי החלטה, כאשר כל עץ מתמקד בתיקון השגיאות של העץ הקודם. תהליך זה מאפשר למידה הדרגתית של דפוסים מורכבים, מה שמוביל לתחזיות מדויקות יותר. האלגוריתם מצטיין בניהול אוטומטי של ערכים חסרים, כולל מנגנוני רגולריזציה למניעת Overfitting ותומך במגוון פונקציות אובדן להתאמה לבעיות סיווג שונות.

פרמטרים ותצורה:

- מספר עצים: נקבע ל-150, מאפשר למידה מעמיקה יותר בהשוואה לברירת המחדל של 100 עצים.
- עומק העצים: הוגדר ל-20, מאפשר זיהוי דפוסים מורכבים בנתונים.
- קצב למידה: נקבע ל-0.1, מאזן בין שיפור הדרגתי בביצועים ליציבות המודל.
- Subsample: מגדיר את אחוז הדוגמאות לאימון כל עץ, מסייע במניעת Overfitting.
- Colsample: משפיע על אחוז התכונות הנבחרות לכל עץ, תורם למגוון ושיפור הדיוק.
- פונקציית מטרה: נבחרה 'multi:softmax' מתאימה לבעיות סיווג רב-קטגוריות.

Logistic Regression

רגרסיה לוגיסטית היא שיטה סטטיסטית בלמידת מכונה המשמשת לסיווג, המבוססת על הפונקציה הלוגיסטית (סיגמואיד).
עקרון הפעולה: המודל מחשב את הסיכוי שדוגמה מסוימת תשתייך לקטגוריה מסוימת על ידי התאמת משקלים לתכונות השונות. אף שהמודל תוכנן במקור לסיווג בינארי, הרחבנו אותו לטיפול בבעיות סיווג מרובות קטגוריות.

פרמטרים ותצורה:

- רגולריזציה: נבחרה רגולריזציה מסוג L2 למניעת Overfitting, עם ערך $C=1.0$ המאזן בין רגולריזציה חזקה לחלשה.
- אלגוריתם אופטימיזציה: נבחר lbfgs, היעיל בפתרון בעיות בעלות ממדים גבוהים.
- מספר איטרציות מקסימלי: נקבע ל-100 להבטחת התכנסות המודל.
- טיפול בריבוי קטגוריות: (OVR) 'One-vs-Rest' multi_class המאפשר בניית מודל נפרד לכל קטגוריה.
- איזון קטגוריות: נעשה שימוש ב class_weight='balanced' להתמודדות עם חוסר איזון בין הקטגוריות.

MLP (Multi-Layer Perceptron)

סוג של רשת עצבית מלאכותית המורכבת משכבת קלט, שכבות חביות, ושכבת פלט, המתאימה במיוחד לפתרון בעיות סיווג.

עקרון הפעולה: MLP מאמנת את המשקלות של הקשרים בין הנוירונים באמצעות אלגוריתם Backpropagation, ממזערת פונקציית הפסד ומאפשרת למידה של דפוסים מורכבים בנתונים.

פרמטרים ותצורה:

- ארכיטקטורת הרשת: שתי שכבות חביות (100 ו-50 נוירונים) לאפשר למידת דפוסים מורכבים.
- פונקציית אקטיבציה, ReLU (Rectified Linear Unit): מתאימה לשכבות חביות ויעילה בעבודה עם קלטים חיוביים.
- אלגוריתם אופטימיזציה Adam: מספק קצב למידה מותאם אישית לכל משקל.
- רגולריזציה L2: עם ערך 0.0001 למניעת Overfitting.
- קצב למידה: קבוע על 0.001, מאזן בין מהירות אימון לדיוק.
- מספר איטרציות מקסימלי: 300.
- Early Stopping: עצירה לאחר 10 איטרציות ללא שיפור.

AlephBERT

מודל שפה מתקדם המותאם במיוחד לעברית מודרנית, המבוסס על ארכיטקטורת מודל BERT. מודל זה פותח באמצעות למידה עמוקה לא מפוקחת על מאגרי נתונים גדולים בעברית, במטרה לשפר משימות עיבוד שפה טבעית מגוונות.

פרמטרים ותצורה:

- קצב למידה: $5e-2$
- מספר אפוקים: 4
- רגולריזציה: שימוש ב weight decay למניעת overfitting.
- הערכה: בוצעה בסוף כל אפוק
- אופטימיזציה: שימוש ב gradient accumulation וחישובי FP16 לייעול האימון.

4.5 הערכת ביצועים

בפרויקט זה, הערכת ביצועי המודל לסיווג מסמכי ביטוח התבססה על מספר מדדים מרכזיים. הדיוק (Accuracy) שימש כמדד העיקרי, מספק תמונה כללית על יכולת המודל לחזות נכונה את הדוגמאות. חישוב הדיוק, המבוסס על היחס בין מספר התחזיות הנכונות לסך כל התחזיות, מדד זה מתאפיין בפשטות ובקלות לפירוש. אולם, נוכח חוסר האיזון בין הקטגוריות השונות בפרויקט, נדרש שימוש במדדים נוספים לקבלת תמונה מקיפה יותר. לכן, ניתוח הביצועים כלל גם שימוש במטריצת הבלבול (Confusion Matrix). כלי זה סיפק תמונה מפורטת של ביצועי המודל עבור כל קטגוריה, ואפשר זיהוי מדויק של נקודות חוזקה וחולשה במודל. השימוש במטריצת הבלבול היה קריטי בזיהוי טעויות ספציפיות, הערכת ביצועים לכל קטגוריה בנפרד, איתור הטיות אפשריות, וגיבוש תוכנות חיוניות לשיפור המודל. כלי זה תרם משמעותית להבנת ביצועי המודל ולקבלת החלטות מושכלות לגבי שיפורים עתידיים והתאמת המודל לצרכים הייחודיים של כל חברת ביטוח. בפרק נספחים בטבלה 2 מוצגות מטריצות הבלבול עבור המודלים שאומנו לחברת הביטוח מגדל. בנוסף למדדים אלו, נשמרו גם מדדים נוספים כגון precision, recall, ו-f1-score, כי הם לא היוו את המוקד העיקרי בהערכת המודל. שילוב כל המדדים הללו אפשר הערכה מקיפה ומדויקת של ביצועי המודל.

5. תוצאות

בגלל מגבלות אורך הדוח בחרתי להציג תוצאות מדד הדיוק של המודלים השונים עבור 2 חברות ביטוח שהוטמעו בסביבת ייצור, חברת מגדל וחברת כלל וכן את ניתוח זמני האימון והפרדיקציה של חברת מגדל:

- השוואת ביצועי מודלים שונים בעזרת מדד הדיוק Accuracy:

Company Name		KNN		Logistic Regression		Random Forest		XGB Classifier		MLP		Aleph Bert
		W2VEC	TF-IDF	W2VEC	TF-IDF	W2VEC	TF-IDF	W2VEC	TF-IDF	W2VEC	TF-IDF	
מגדל	Train	0.971	0.971	0.965	0.957	0.989	0.970	0.968	0.971	0.988	0.965	0.946
	Test	0.968	0.968	0.963	0.960	0.982	0.934	0.966	0.968	0.979	0.960	0.943
כלל	Train	0.983	0.983	0.921	0.958	0.928	0.984	0.927	0.979	0.926	0.928	0.966
	Test	0.961	0.961	0.923	0.961	0.932	0.975	0.928	0.963	0.931	0.926	0.952

- ניתוח זמני אימון, חיזוי ומשקל המודלים עבור חברת הביטוח מגדל:

Model	Training Time (Word2Vec)	Predict Time (Word2Vec)	Model Size (Word2Vec)	Training Time (TF-IDF)	Predict Time (TF-IDF)	Model Size (TF-IDF)
KNN	197.83	0.017	15.6MB	234.64	0.264	41.5MB
Logistic Regression	298.33	0.006	6KB	17.23	0.001	785KB
Random Forest	255.04	0.016	25.2MB	216.42	0.013	35.7MB
XGBClassifier	193.99	0.007	882KB	73.20	0.023	782KB
MLP	199.71	0.009	188KB	226.53	0.002	32.9MB

Model	Training Time	Predict Time	Model Size
AlephBert	6260.67	0.353	57MB

6. דיון

התוצאות שהתקבלו מצביעות על הבדלים בביצועי המודלים השונים עבור חברות הביטוח מגדל וכלל.

- בחברת כלל, שיטת הייצוג TF-IDF הוכיחה את עצמה כיעילה לסיווג מסמכים, ככל הנראה בשל מאפייני המסמכים שכללו מונחים ייחודיים, מה שהופך את TF-IDF למתאימה לזיהוי מילים משמעותיות. לעומת זאת, בחברת מגדל, שיטת Word2Vec הניבה תוצאות טובות יותר, כנראה בשל הצורך בהבנה מעמיקה יותר של הקשרים הסמנטיים בין המילים, שהייתה חיונית למסמכים מורכבים או מגוונים יותר.
- המודל Random Forest הציג את הביצועים הטובים ביותר בשתי החברות. ניתן לייחס את הצלחתו לשימוש בטכניקה המאפשרת חיזוי המבוסס על הצבעת רוב של עצים, וכך מעניקה למודל יכולת זיהוי של מבנים מורכבים בנתונים והתמודדות טובה עם רעש. בנוסף, המודל מצטיין בטיפול בנתונים לא מאוזנים, מציע עמידות גבוהה יותר לתכונות לא רלוונטיות או לשגיאות, ומסוגל לזהות אינטראקציות מורכבות בין תכונות שונות בצורה יעילה.
- מודל AlephBERT מציג את התוצאה נמוכה ביותר עבור חברת מגדל לעומת תוצאה טובה בחברת כלל. ההסבר לכך יכול להיות בגלל: מספר הקטגוריות או מידת ההבדל ביניהן שעשויות להשפיע על ביצועי המודל, לחברת כלל גודל מערך נתונים גדול יותר לאימון AlephBERT, טוב עם כמויות גדולות של נתונים. אפשר שהיה שוני משמעותי במורכבות הלשונית, או בשימוש במונחים מקצועיים בין שתי החברות.
- שיטת KNN השיגה תוצאות טובות, אך לא את התוצאות הטובות ביותר, וזמן החיזוי שלה ארוך יותר מפי 10.
- התוצאות של XGBoost משתפרות מעט עם TF-IDF אך הן עדיין לא מגיעות לרמות הביצועים של Random Forest.

7. מסקנות

בפרויקט זה פיתחתי מודלים לסווג מסמכים ביטוחיים של חברות ביטוח לקטגוריות מתאימות.

תהליך הפיתוח כלל מספר שלבים:

1. **איסוף ועיבוד נתונים** – איסוף מאגר מסמכים רחב, הכולל דוגמאות מכל סוגי המסמכים הנדרשים, ולאחר מכן עיבוד מקדים של הקטגוריות והנתונים.
2. **בחירת שיטות ייצוג טקסט** – במטרה לשפר את דיוק המודל, נבחנו מספר שיטות ייצוג טקסט.
3. **פיתוח ואימון המודל** – בניית מודל למידת מכונה שמטרתו לחלק מאפיינים ייחודיים של כל סוג מסמך ולסווג אותם לפי קטגוריות. בוצע אימון מודל על קבוצות נתונים מאוזנות, תוך שימוש במגוון מודלים כדי לבחון איזה מהם מספק את התוצאות המדויקות ביותר.
4. **הטמעה בסביבת יצור ובדיקת ביצועים** – שילוב המודל במערכת Almagor Click וביצוע בדיקות מקיפות, במטרה לוודא שהמודל מזהה ומסווג את המסמכים בדיוק גבוה. הדיוק נבדק על קבוצות אימות ונתוני אמת במערכת כדי להבטיח יישום מוצלח בסביבת ייצור שאכן הציג תוצאות מעולות: **93%** לחברת כלל ו**96%** לחברת מגדל (בתחילה התוצאות בחברת מגדל היו **86.9%**. לאחר בדיקה נמצא כי מגדל עשתה שינוי במסמכים ואיחדה 2 מסמכים יחד מה שגרם לטעות. המתייגים הונחו כיצד לתייג מסמך כזה ואחוז הדיוק עלה).

בהתבסס על התוצאות שהתקבלו, ניתן לראות כי מודלים קלאסיים של למידת מכונה, כגון Random Forest, MLP ו-KNN-הציגו ביצועים משופרים במספר מקרים בהשוואה למודל AlephBERT המבוסס על ארכיטקטורת Transformer מתקדמת.

ממצאים אלו מדגישים את היתרון הפוטנציאלי של שיטות למידת מכונה מסורתיות בסיווג טקסטים, במיוחד במשימה שאינה דורשת ניתוח מעמיק של קשרים לשוניים מורכבים. אף שמודלים מתקדמים כדוגמת BERT נחשבים לרוב כפתרונות מובילים בתחום עיבוד השפה הטבעית, הם אינם בהכרח האפשרות המיטבית לכל סוגי המשימות. במצבים מסוימים, מודלים פשוטים יותר עשויים להציג יעילות ודיוק גבוהים יותר, בפרט כאשר מדובר במשימות סיווג בסיסיות.

בהקשר הנוכחי, נראה כי שילוב של Word2Vec או TF-IDF עם מודל Random Forest מהווה גישה אפקטיבית לסיווג טקסטים.

תוצאות אלו מחדדות את הצורך בבחירה מושכלת של המודל ושיטת הייצוג המתאימים ביותר לכל משימה ספציפית, ומדגישות את החשיבות של בחינה אמפירית של מגוון גישות לפתרון הבעיה הנתונה.

8. המלצות לעתיד

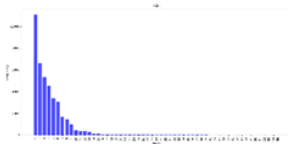
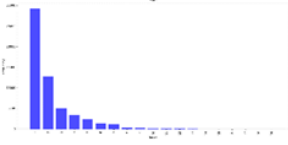
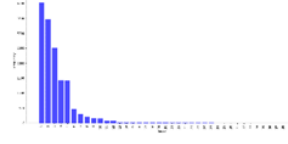
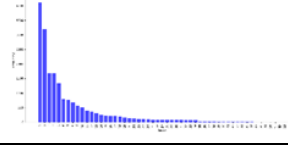
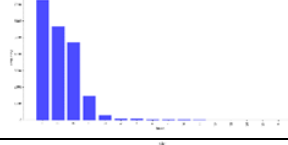
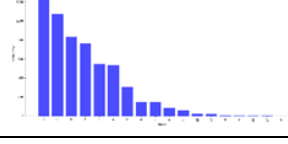
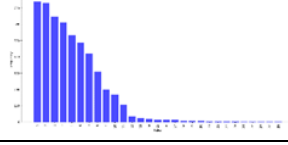
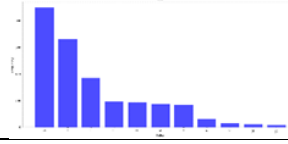
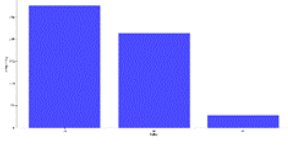
1. ניתוח שגיאות:
- ביצוע ניתוח מעמיק של השגיאות כדי לזהות דפוסים ולפתח אסטרטגיות לשיפור הדיוק בקטגוריות בעייתיות.
2. שמירת הסתברויות החיזוי:
- שמירת ההסתברויות שהמודל מייצר עבור כל קטגוריה, כדי לאפשר ניתוח מעמיק יותר ולזהות מקרים גבוליים.
3. שיפורים טכניים (לוגים והתראות):
- שפר את מערכת הלוגים וההתראות, כולל תיעוד פעולות המודל והגברת המעקב אחר ביצועיו.
4. סיווג ללא אימות אנושי:
- פיתוח מודל שיכול לסווג מסמכים ללא צורך באימות אנושי, תוך שימוש במדדי ביטחון לזיהוי מקרים בהם הוא בטוח בסיווג שלו.
- יישום המלצות אלו יכול לשפר משמעותית את ביצועי המודל, את יעילותו ואת האמינות שלו בסביבת עבודה ויאפשר שימוש נרחב יותר במודל בתהליכי העבודה, תוך הפחתת הצורך בהתערבות אנושית ושיפור הדיוק הכללי של תהליכי סיווג המסמכים.

9. ביבליוגרפיה

- Khan, A., Baharudin, B., Lee, L. H., & Khan, K. (2010). A Review of Machine Learning Algorithms for Text-Documents Classification. Journal of Advances in Information Technology, 1(1), 4-20.
- Zhang, X., Zhao, J., & LeCun, Y. (2015). Character-level Convolutional Networks for Text Classification. Advances in Neural Information Processing Systems, 28.
- Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv preprint arXiv:1810.04805.
- Scikit-learn Documentation. (2023). Text Classification with Scikit-learn. https://scikit-learn.org/stable/tutorial/text_analytics/working_with_text_data.html
- Towards Data Science. (2022). Advanced Techniques for Text Classification. <https://towardsdatascience.com/advanced-techniques-for-text-classification-f5f69d27b93a>
- Google Research. (2021). Comparative Study of Text Classification Models. <https://research.google/pubs/pub49969/>
- Dixon, J. (2023). Document-Classification. GitHub repository. <https://github.com/JDixonCS/Document-Classification>
- AltexSoft. (2023). Document classification with machine learning. <https://www.altexsoft.com/blog/document-classification/>

10. נספחים

- טבלה 1: תצוגת התפלגות המסמכים בקטגוריות השונות לכל חברת ביטוח:

חברת הביטוח	מספר המסמכים	מספר קטגוריות	התפלגות הנתונים (כמות מסמכים לכל קטגוריה)
הפניקס	69113	73	
הראל	59400	28	
כלל	32582	90	
מנורה	27667	70	
מגדל	20355	22	
ביטוח ישיר	6044	18	
איילון	5663	30	
AIG	2825	14	
הכשרה	532	3	

- טבלה 2: מטריצת הבילבול עבור המודלים השונים של חברת מגדל.

שם המודל	שיטת ייצוג WORD2VEC	שיטת ייצוג TF-IDF
Random Forest	<p>Confusion Matrix</p>	<p>Confusion Matrix</p>
MLP	<p>Confusion Matrix</p>	<p>Confusion Matrix</p>
KNN	<p>Confusion Matrix</p>	<p>Confusion Matrix</p>

