הצעה למתווה התממה של טקסט רפואי בעברית

פרויקט "נְמַל מִבְטַחִים"

הדס ביטרן, מנהלת קבוצת טכנולוגיות הבריאות ומדעי החיים, מיקרוסופט ישראל מחקר ופיתוח

*רועי כהן, מנהל תחום ההתממה, פרויקט תמנע, משרד הבריאות*

*דצמבר 2021*

# מבוא

### רציונליזציה

התממה של מידע רפואי נדרשת כדי לאפשר שימוש נוסף במידע לטובת מחקר רפואי או אימון של מודלים מבוססי למידת מכונה, תוך הגנה על המידע ועל פרטיות האינדיבידואליים שאודותיהם המידע. על ההתממה לצמצם למינימום את הסיכון שזיהוי מחדש של אינדיבידואל יתרחש, בין אם מתוך מאגר הנתונים עצמו ובין אם מהצלבת נתונים ממאגר זה עם נתונים ממאגר זמין אחר. שיטה נפוצה לביצוע התממה היא הסרה של פרטים מזהים מהנתונים.

### האתגר בהתממת טקסט רפואי

כאשר מתמימים מידע מובנה (structured data), ניתן לקבוע באופן מושכל את אופן ההתממה הרצוי של השדות המזהים והניתנים לזיהוי מתוך המידע ולהסיר אותם. האתגר בהתממת טקסט רפואי חופשי (unstructured data) הוא גדול יותר, כיוון שלא קיימים שדות מוגדרים שניתן להתייחס אליהם. בנוסף יש להבין את משמעות והקשר המילים כדי להחליט אם צריך למחוק אותן. מן העבר השני, מחיקה מיותרת עלולה לפגוע במשמעות הקלינית של הטקסט , למנוע שימוש בו או ליצור הטייה במודלים המאומנים עליו.

פעולת ההתממה של טקסט רפואי חופשי דורשת זמן ניכר אם היא מתבצעת על ידי אדם, ועל כן לא ריאלי לבצעה על מאגרי מידע גדולים. עם זאת, אימון מודלים מבוססי למידת מכונה דורש כמות גדולה מאד של טקסטים רפואיים. על כן יש חשיבות רבה בהתממה על ידי אלגוריתם ויצירת מנגנונים מאזנים למקרים של טעויות: מנגנונים אשר מפחיתים את החשש לפגיעה בפרטיות באמצעים נוספים מלבד התממה (כמפורט להלן בהנחות העבודה), וכן מנגנונים להערכת איכות אלגוריתם ההתממה.

מטרת מסמך זה היא לשמש כהצעה לדיון עם כלל בעלי העניין בנושא התממת טקסט חופשי, על מנת להעמיק את השיח בין חוקרים העושים שימוש במידע בריאות ומומחי התממה ופרטיות, לקדם פיתוח של אלגוריתמים הנותנים מענה מיטבי לבעיות התממה לצורך קידום המחקר הרפואי בישראל תוך שמירה על פרטיות המטופלים, ולסייע בגיבוש מתווה מוסכם.

### הנחות עבודה

מתווה ההתממה המוצע להלן מבוסס על הנחות העבודה בהבאות:

1. המידע המותמם שייוצר כתוצאה מהרצת אלגוריתם יאפשר אימון מודלים של למידת מכונה והערכת הביצועים שלהם, אך לא יפורסם לציבור שום מידע טקסטואלי כחלק מפרסום מדעי.
2. חלק מהמידע המותמם שייוצר כתוצאה מהרצת אלגוריתם ההתממה יאפשר הערכת ביצועים של אלגוריתם ההתממה על ידי ארגון הבריאות (ראו מנגנוני הערכה לאיכות אלגוריתם ההתממה).
3. כל המידע המותמם שייוצר כתוצאה מהאלגוריתם יישאר בתוך גבולות ארגון הבריאות ו**לא** מעורבת מסירה של המידע המותמם אל מחוץ לגבולות ארגון הבריאות (trust boundary).
4. החוקרים המבצעים את אימון המודלים קיבלו אישור אתי מועדת הלסינקי.
5. התקבל אישור ועדת ענן במידה ונעשה שימוש בענן לצורך אימון המודלים, והשימוש מבוצע על מידע מותמם בסביבת הענן בשליטתו של ארגון הבריאות.

### המתווים ששימשו לכתיבת המסמך

המתווה שלהלן מבוסס על סטנדרט Safe Harbor האמריקאי המופיע כחלק מ - Health Insurance Portability and Accountability Act of 1996 (HIPAA), חקיקה המתייחסת לפרטי מידע שיש להסתיר על מנת להתמים מידע לצורכי מחקר רפואי. אולם, המתווה המוצע מרחיבה את ההגדרה האמריקאית עם התאמות לצרכים הספציפיים בישראל, ובכך מספק להערכתנו הגנה טובה יותר על פרטיות.

המתווה מסתמך גם על [טיוטת הנחיות להתממת מידע בריאות לצרכי שימוש מחקרי](https://www.health.gov.il/Services/Documents/H06102919.pdf) שפרסם משרד הבריאות ב 2019, דו"ח מבקר המדינה על [היבטים בהגנה על הפרטיות במאגרי מידע](https://www.mevaker.gov.il/sites/DigitalLibrary/Documents/69b/2019-69b-101-Hagana.pdf) מ 2019, על חוזר המנהל הכללי המשרד הבריאות בנושא [שימושים משניים במידע בריאות](https://www.health.gov.il/hozer/MK01_2018.pdf) מ 2018, [ופרדיגמת התממה מוצעת על פי I2B2 de-id challenge](https://www.sciencedirect.com/science/article/pii/S1532046415001173?via%3Dihub).

# המתווה המוצע לדיון

המתווה המוצע מחולק לשני חלקים, תחילה נציע כיצד אלגוריתם המבצע התממה לטקסט חופשי נדרש להתייחס לסוגי אלמנטים שונים העשויים להופיע בטקסט. לאחר מכן, נדון במנגנוני הערכה לאיכות אלגוריתם ההתממה.

# הגדרות פונקציונליות

בפרק זה יוגדרו סוגי אלמנטים של טקסט חופשי שדורשים התממה ושיטת ההתממה הנדרשת.

## שמות

שמות פרטיים, אמצעיים ושמות משפחה יותממו ויוחלפו ב <שם\_>.

## תאריכים

### *פורמט תאריכים*

הגדרות ההתממה הבאות רלוונטיות לתאריכים בכל פורמט, כולל תאריכים לועזיים ועבריים.   
דוגמאות לפורמטים השונים:

* 22 לנובמבר 2005
* כ"ו באלול תשמ"א
* כ"ט בנובמבר 1947
* ב' בכסלו 1980
* 22.11.2005 עם מפרידים שונים כגון נקודות, לוכסן קדמי/אחורי, מקף וכדומה.
* שנה בפורמט מקוצר למשל 22.11.05
* חודש ויום עם או בלי 0 מקדים, למשל 04.11.2000, 11.06.03

### *תאריכי לידה*

יש להתמים תאריכי לידה למעט שנה. למשל, התאריך 28.12.2015 יוחלף ב 2015.<חודש\_>.<יום\_>.

משיקולי פרטיות, עבור תאריכי לידה שידוע כי מייצגים גיל הגדול מ 89 שנים (למשל על פי תאריך המסמך), למשל 28/12/1917, התאריך יוחלף ב <שנה\_>/<חודש\_>/<יום\_>.

בשל המשמעות הקלינית, עבור תאריכי לידה שידוע כי מייצגים גיל הקטן משנה, יוחלף היום בלבד וישמר החודש. למשל, התאריך 28/12/2021 יוחלף ב 12/2021/<יום\_>.

בשל המשמעות הקלינית, כאשר מוזכר גיל המטופל (ללא תאריך לידה), אין להתמים את גיל המטופל בשנים (או חודשים אם מדובר בתינוק), אלא אם המטופל מבוגר מ 89 שנים ואז יש להחליף ב <גיל\_>.

### *תאריכי אירועים רפואיים*

סעיף זה מתייחס לאירועים כגון תאריך אישפוז, תאריכי בדיקות וכדומה.

בשל המשמעות הקלינית, עבור תאריכים שמייצגים אירועים רפואיים, יוחלף היום בלבד וישמר החודש. למשל, התאריך 28/12/2021 יוחלף ב 12/2021/<יום\_>.

## מידע גאוגרפי

### *כתובות*

יש להתמים כתובות המופיעות בטקסט כולל שם רחוב, מספר הבית, שם השכונה – אשר יוחלפו ב <מיקום\_>.   
יש להתמים מיקוד, בפורמט ישן קצר (5 ספרות) או חדש ארוך (7 ספרות) ולהחליפו ב <מיקום\_>.

בשל המשמעות הקלינית הפוטנציאלית אין להתמים את שם העיר או שם מקום היישוב. עם זאת, משיקולי פרטיות, מקום ישוב בו יש פחות מ 2000 תושבים יוחלף ב <מיקום\_>. רשימת הישובים הקטנים-מדי תילקח ממרשם האוכלוסין.

יש להתייחס לכל כתובת, בין אם מדובר במקום מגורים ובין אם זהו מקום בו התרחש אירוע (לדוגמא, המטופל נמצא מחוסר הכרה ליד רחוב ארלוזורוב 51 בת"א, או התאונה התרחשה בכיכר אוסישקין ברמה"ש).

מידע גאוגרפי המובע על ידי קואורדינטות (בדרך כלל בפורמט (Latitude, Longitude צריך להיות מותמם ויוחלף ב <מיקום\_>.

לארץ מוצא, ארץ עליה, או יעד נסיעה עשוי להיות ערך קליני או מחקרי. עם זאת, ערכים אלה עשויים לספק מידע מזהה על נשואי מידע פי מוצאם או מקומות בהם ביקרו. על כן, מוצעת ההתממה הבאה:

* ישראל, אתיופיה – להשאיר
* מדינות בריה"מ לשעבר – להחליף ב"ברית המועצות"
* אלג'יריה, מרוקו, תוניסיה, לוב, מצרים – להחליף ב"צפון אפריקה"
* מדינות מזרח רחוק כגון הודו, סין, תאילנד – להחליף ב"מזרח רחוק"
* יתר המדינות – להחליף ביבשות (כולל תתי יבשות), על פי טבלת אקסל המצורפת למסמך זה.

## אירגונים ומוסדות

יש להתמים פרטים על ארגונים ומוסדות המקושרים לאדם ולהחליף ב <ארגון\_>. סעיף זה מתייחס לדברים כגון:

* מקום עבודה, בין אם מוזכר בעברית או באנגלית (למשל מיקרוסופט או Microsoft), בין אם במגזר הפרטי ובין אם במגזר הציבורי, בין אם מועסק בתשלום ובין אם בהתנדבות.
* מקום לימודים כגון שם בית ספר, שם מוסד אקדמי, מכללה וכדומה (למשל: בית ספר גולן, בי"ס כרמל, מכללת עמק חפר, האוניברסיטה העברית, מתנ"ס קרית הצעירים).
* הארגון בו מוענק טיפול ו/או שירותי מגורים למטופל כגון בית חולים, בית חולים גריאטרי, בית אבות וכו'

## מספרי טלפון

יש להתמים כל איזכור של מספר טלפון, בכל פורמט, ולהחליפם ב <קשר\_>.

ההגדרה מתיחסת לטלפון נייח, טלפון נייד, מספר פקס, מספרי טלפון בבית ובעבודה, מוסדיים ופרטיים, בארץ ובחו"ל.

## פרטים מזהים

יש להתמים פרטים שעלולים לזהות אדם ולהחליפם ב <מזהה\_>, למשל מספרים מזהים.

רשימת סוגי הפרטים המזהים כוללת אך אינה מוגבלת ל:

* מספר ת.ז.
* מספר אישי
* מספר רשיון כגון רשיון נהיגה, רשיון רופא, רשיון טיס אן כלי שיט
* מספר תיק בביטוח לאומי או בכל מוסד ממשלתי
* מספר מנוי
* מספר עובד
* מספר רכב
* מספרי חשבון בבנק או מוסדות אחרים
* מספרי כרטיס אשראי
* מספרים סידוריים של מכשירים
* מספר רישיון עוסק
* מספר מזהה אצל המבטח או קופת החולים
* מספר תיק רפואי
* מספר אשפוז
* מספר מרשם
* שם משתמש או כינוי רשת אשר מזהה אדם באופן חד חד ערכי, כגון כינויו ב Twitter או ב GitHub

## כתובות אונליין

יש להתמים כל כתובת אימייל או כתובת אינטרנט ולהחליף ב<קשר\_>.

כתובות אינטרנט, בין אם כ IP address ובין אם כ URL גם הן חייבות להיות מותממות - להחליף ב<קשר\_>.

## מזהים חזותיים

מידע אשר כולל תמונות או מזהים נוספים צריך להיות מותמם: יש להסיר מהמידע כל תמונה שכוללת פנים או אלמנטים מזהים אחרים, וכן מזהים ביומטרים כגון טביעת אצבע. סעיף זה מובא לצורך שלמות ההגדרות והקבלתן ל Safe Harbor כיוון שאיננו צופים שטקסט רפואי יכיל מידע שכזה.

מזהים נוספים

כתלות במידע להתממה ובמטרות המחקר, נמליץ לשים לב לפרטים נוספים בטקסט החופשי שיכולים להיות מזהים חלקיים כמו מצב משפחתי או לאום ולהפעיל שיקול דעת לגבי התממתם.

# מנגנוני הערכה לאיכות אלגוריתם ההתממה

הערכת ביצועי האלגוריתם תתבצע על סמך best practices הנהוגים בתעשיה עבור אלגוריתמי למידת מכונה, הכוללים חישוב precision, recall, F1, F2 score .

עבור הערכת הביצועים ייבחר סט מידע (data set) מתוך כלל המידע של ארגון הבריאות שיכלול לפחות 100 מסמכים בני 10 משפטים לפחות בכל אחד, או 1000 משפטים לפחות, וישמש עבור הערכת ביצועי האלגוריתם. המידע שנבחר עבור הערכת ביצועי האלגוריתם צריך להיות מייצג את המידע המיועד להתממה מבחינת אופי המידע (למשל סוגי המסמכים), אך אינו יכול להיות מידע שבוצע בו שימוש לצורך אימון האלגוריתם.

המידע להערכת הביצועים ייחשב מייצג את כלל המידע אם הוא לקוח מאותם סוגי המסמכים המיועדים להתממה, באותו אורך טקסט ממוצע פלוס מינוס 100 תווים, מכיל את אותם סוגי מחלקות (למשל אם במקור יש מחלקה פנימית, גם במידע המייצג) ומאותו סוג של ארגון בריאות (למשל, בית חולים לעומת קופת חולים).

המידע הנבחר יתוייג על ידי מומחים שיקבעו אילו מילים ואיזה חלקים צריכים להיות מותממים ובאיזו צורה (expert determination) על פי מתווה הגדרות זה (התממה ידנית). הצוות יכלול לפחות שני מומחים, שהם רופאים, או סטודנטים לרפואה, צוות רפואי או בעלי תואר שני או שלישי במדעי הרפואה, ויכללו לפחות גבר אחד ולפחות אישה אחת. הגיוון בבחירת המומחים נדרש על מנת לנטרל הטיות. יש לוודא שהמומחים בקיאים בהגדרות ההתממה על ידי מתן הנחיות מפורטות על מנת שההתממה הידנית תהיה מדוייקת.

לאחר התיוג, יורץ האלגוריתם על אותו סט של מידע (ללא תיוגי המומחים כמובן) ותתבצע השוואה בין ההתממה שבוצעה על ידי המומחים ובין ההתממה שבוצעה על ידי האלגוריתם. ההשוואה תתבצע על פי דיסציפלינת [SemEval-Partial](https://aclanthology.org/S13-2056.pdf) ותיעשה עבור כל סוג מונח שהופיע בסט המידע וצריך להיות מותמם (שם, מיקום, טלפון וכדומה). בחירת דיסציפלינה זו מתעדפת את צורך ההתממה על פני הצורך בסוג המונח.

אלגוריתם התממה עבור ה intended use הזה שאינו מערב מסירה של המידע אל מחוץ לגבולות ארגון הבריאות יחשב באיכות מספקת אם ה F2 score המשוקלל יהיה מעל 0.9. הבחירה ב F2 score על פני F1מייחסת חשיבות גדולה יותר ל Recall על פני Precision, כלומר מתעדפת התממת-יתר על פני דיוק, וגורסת כי אם יש ספק לגבי מונח, עדיף כן להפעיל התממה /

ארגון הבריאות יוכל לקבוע רף שימושיות של המידע לאחר התממה, בין אם לצורך הערכת ביצועים ובין אם לצורך אימון מודלים. לדוגמא, מסמך או משפט אשר יותר מ 50% מהמילים בו הותממו יכול להיחשב כחסר ערך לאחר ההתממה כיוון שאיבד את משמעותו ולהיזרק מסט המידע או מסט הערכת הביצועים.

# דוגמאות

"סיכום אישפוז עבור ישראל ישראלי, ת.ז. 123456789, מנהל סניף חברת חשמל בחולון" ->"סיכום אישפוז עבור <שם\_>, ת.ז. <מזהה\_>, מנהל סניף <ארגון\_> בחולון"

"החולה הוא תלמיד בכיתה ג' בבית הספר גולן בכרמיאל" -> "החולה הוא תלמיד בכיתה ג' בבית הספר <ארגון\_> בכרמיאל"

"הנבדק משרת בצה"ל ומתגורר במצפה הילה" -> "הנבדק משרת ב<ארגון\_> ומתגורר ב<מיקום\_>"