**דוח אפיון:**

**מטרות הפרויקט:**

* לפתח מערכת המסוגלת לזהות תוכן חדשות מזויפות ברשתות החברתיות ולהזהיר מפני תוכן זה. אופן הביצוע הוא בצורה כזאת שהמשתמש מקבל אחוזי הסתברות מהימנות המידע ובכך יוכל לדעת אם המידע אמיתי או שקרי.
* יצירת מערכת אשר מסוגל להתלבש על מערכת קיימת בעזרת API חיצוני וקוד גמיש. בכך נוכל ליצור סביבה נקייה וסטרילית ממידע מזויף ומטעה. מוצר זה דרוש מכיוון שבמערכת מרובת משתמשים קשה מאוד לבצע את התהליך הזה באופן ידני ובאמצעות המערכת הדבר יתבצע באופן אוטומטי ובכך לחסוך את זמן העבודה היקר והמשאב האנושי שיבצע זאת.
* זיהוי משתמשים פוטנציאלים אשר עלולים לפרסם מידע מוטעה ומזויף. (עוד לא מידע מוטעה אבל עלול להוביל לכך - אחוז גבוה אבל לא מעל תנאי סף מסוים)
* יצירת אלגוריתם אשר יעבוד ויחשב את הקלטים בזמן אמת, אשר מנתר בכל עת את המידע הזורם באתר המיועד .
* יצירת אלגוריתם אשר יוכל להיות בשימוש מסחרי או מחקרי בשל הגדילה האקספוננציאלית של האינטרנט והרשתות החברתיות.

**רקע וסקירת ספרות:**

כיום, ברשתות חבריות כמו קבוצת META או פלטפורמת טוויטר ישנם אלגוריתמים דומים אשר מסוגלים לזהות Fake News (עם או בלי בינה מלאכותית).

בפייסבוק וכמו באתרים דומים לו, המערכת שואלת את המשתמשים שאלות מנחות במהלך יצירת פרופיל המשתמש שבאמצעות מידע זה נבנה פרופיל אלקטרוני שמספק מידע פונאציאלי לאלגוריתם לנטר או לאפיין את אותו משתמש.

ננתח לדוגמא את האלגוריתם בשימוש ברשת החברתית פייסבוק. בפייסבוק, קיים מנגון לזיהוי וסינון אוטומטי אשר מתבסס על אלגוריתמים מתקדמים שמבוססים על למידת מכונה כדי לזהות ידיעות כוזבות או מידע משובש. אותו מנגון מזהה דפים או פוסטים המכילים מידע מתעתע ומנסים לסנן אותו או להורידו מהאתר. אותו אלגוריתם מתייחס **לכמות** דיווחים של משתמשים אחרים וזה חלק בלתי נפרד מהשיקול שלו על מנת לדעת האם הדבר הוא אמת או שיש דיווחים כוזבים. דרך הפעולה נוספת היא השוואת המידע שפורסם בפוסט מסויים אל מול כמה מקורות שונים כדי לקבוע את רמת האמת של המידע.

נוכל לקחת דוגמא סותרת - טוויטר. בתקופה האחרונה הוסרה האכיפה והאלגוריתמיקה שמטרתם הייתה למנוע פרסום של פייק ניוז או לאפשר למשתמשים לדווח על פייק ניוז – דבר אשר מוביל ומשפיע על דעת קהל רחבה מאוד של משתמשים רבים שלא מסוגלים להבדיל בין אמת לשקר בגלל שטף המידע הרב שעובר ברשת.

בנוסף לכך, כיום קיימים **תוספי דפדפן** שונים אשר בהם ישנו שימוש באלגוריתם דומה. נוכל לקחת לדוגמא תוסף כרום שנקרא NewsGuard או תוספי כרום שונים שעושים שימוש בצ'אט של ChatGPT ובמוצרי מיקרוסופט רבים שמממשים מטרה דומה.

כמו כן, קיימים כלים אשר מסוגלים לנתח ולעבד וידאו בזמן אמת ולבצע השוואות מול מקורות מידע אחרים כדי להפחית את האפשרות של תמונות מעוותות או תמונות שקריות

מידע מוטעה שמתפשט מבלי כוונה להטעות נקרא מיסאינפורמציה - **Misinformation**. זה יכול לקרות בטעות או בגלל חוסר ידע, אך אין כוונה מתוך להטעות במכוון. לעומתו, דיסאינפורמציה - **Disinformation** הוא מידע מתוכנן ומכוון להטעות ומטרתו להשפיע על דעת קהל.

לצד הגדילה האקספוננציאלית של המשתמשים ברשתות החברתיות – התפתח צורך לכלים לזיהוי מידע שקרי בקנה מידה גדול שקשה מאוד לטפל בו באופן ידני – ולמשבצת זו בהחלט מתאים השימוש בAI ובאלגוריתמים מתקדמים, שכן גם הוא בתהליך גדילה אקספוננציאלית.

כיום, חלק בלתי מבוטל מהקהילה העולמית צורך מידע חדשותי כמעט לחלוטין מהרשתות החברתיות בלבד. הרשתות החברתיות מאפשרות לאדם הפרטי להשפיע באופן ישיר ואישי באמצעות פרסום ידיעות, מאמרים ודעות. עם זאת, לא כולם מורגלים או מודעים לכך שיכול להיות מצב שבו מפורסם מידע שקרי ולא מבצעים השוואה בין אתרים שונים. לדבר חשיבות רבה שאם לא יהיה מנגנוני הגנה כמו שהפרויקט שלנו בא להציע - דעת הקהל העולמית בנושאים שונים עלולה להיות שגויה ומבוססת על מידע שקרי.

**מפרט:**

* האלגוריתם ישתמש במודלי שפה קיימים כמו ChatGPT ומערכות AI אחרות.
* האלגוריתם יעבור סדרת אימון עם מידע מיועד שמכיל בתוכו מידע שקרי ומידע אמיתי מתוך אתרי אינטרנט ומקורות מידע המכילים מידע מסוג זה לצורכי אימון כמו למשל [Kaggle](https://www.kaggle.com/c/fake-news/data).
* האלגוריתם יתחשב **בחופש הביטוי –** לא יסווג מידע בתור fake news במידה והמידע מוצג בתור דעת (טור אינטרנטי למשל)
* האלגוריתם ישמור על פרטיותו של המשתמש – האלגוריתם יבצע בדיקת אמינות על המשתמש רק לצד המידע שאישר המשתמש וסיפק לאותה פלטפורמה.
* האלגוריתם יבצע למידה מתמשכת –האלגוריתם ילמד להסתגל לדפוסים חדשים של מידע או לדפוסי התנהגות של חדשים של משתמשים.
* המוצר מיועד לקהילות חברתיות, אתרי חדשות, אתרים אשר מאפשרים למשתמשים חיצוניים לפרסם מידע אתרי מאמרים, אתרי בלוגים ומקורות מידע כאלו או אחרים.
* נדרש ליצור ממשק API בין האלגוריתם לבין מודלי שפה של ChatGPT.
* נדרש ליצור חיבור API חיצוני על מנת לאפשר למשתמשים חיצוניים ליצור תקשורת עם האלגוריתם.
* נדרש לפתח אתר אינטרנט אשר יהיה ניתן להעלות אותו לכתובת ציבורית.
* נדרש לעצב את האתר בצורה נוחה ונגישה עם מדריך קצר לשימוש במוצר. האתר יכיל בתוכו קונסול – שם יוכל המשתמש להכניס מידע ולברר האם הוא fake news או אמת.
* האלגוריתם יספק תשובה אשר מבוססת על אחוזי הסתברות - על סמך המידע שיאסוף האלגוריתם.
* גרסה נוספת של המוצר תהיה תוסף לgoogle chrome – אשר ניתן יהיה לסמן טקסט שישלח בתור קלט לאלגוריתם ויספק פלט בצורה דומה כמו באתר עם הקונסול.
* זיהוי משתמשים פוטנציאלים להפצת fake news - באמצעות שירות צד שלישי שיפיק מסמך טקסט המכיל את כל הפוסטים של משתמש רצוי פלטפורמה מסויימת (פייסבוק ,טוויטר וכדומה). כל פוסט יוזרק כמידע לקונסול האלגוריתם, אשר יתן את אחוזי הזיהוי ל-fake news. לאחר מכן יחושב ממוצע על תוצאות ההסתברות למתן דירוג המשתמש וסיווגו האם הוא מפיץ מידע שקרי או לא.
* קבצי הפרוייקט ישמרו בסביבת GitHub משותפת לניהול גרסאות הפרויקט לשמירת הנתונים.

**גאנט פרויקט:**

| **מספר משימה** | **שם משימה** | **מועד התחלה משוערך** | **מועד סיום משוערך** | **תוצר/תיאור בסוף המשימה** |
| --- | --- | --- | --- | --- |
| 1.1 | בחירת נושא לפרוייקט ממאגר המוצע | 25.07.2023 | 01.08.2023 | קבלת נושא לפרוייקט ותפיסת הרעיון הכללי |
| 2.1 | חקר שוק | 01.12.23 | 30.12.23 | הבנת השוק וחשיבה על פתרונות שלא קיימים כיום. |
| 2.2 | איסוף מקורות מידע | 31.12.23 | 06.01.2024 | אוסף מאמרים, מקורות מידע, סדרות אימון ומידע שימושי לפרויקט. |
| 2.3 | כתיבת דו"ח אפיון | 07.01.2024 | 23.01.2024 | מסמך אפיון ראשוני המאושר ע"י מנחה הפרויקט |
| 3.1 | יצירת סביבת עבודה בGitHub | 24.01.2024 | 25.01.2024 | סביבת עבודה לניהול גרסאות הפרוייקט |
| 3.2 | יצירת חיבור לשירותי OpenAI | 26.01.2024 | 15.02.2024 | תקשורת תקינה עם OpenAI ובקשות בסיסיות תשלחנה ותתקבלנה + תכנון וקביעת פרוטוקול תקשורת ייחודי. |
| 3.3 | יצירת שלד האלגוריתם | 16.02.2024 | 28.02.2024 | בסיס לאלגוריתם המכיל בתוכו את כל חלקי הקוד המכיל בתוכו בסיס חיבורי לשירותי OpenAI |
| 0.1 | פגרה - תקופת מבחנים סמסטר א' | 01.03.2024 | 07.04.2024 | תקופת מבחנים סמסטר א' |
| 3.4 | יצירת פונקציות בסיסיות (הסתברות, אפיון, ממוצע, GT, ערבול) | 08.04.2024 | 18.04.2024 | האלגוריתם יכיל את הפונקציות הבסיסיות למימוש החיזוי. |
| 4.1 | סידור ופילטור המידע לסדרת אימון בהתאם לפרוטוקול התקשורת הייחודי | 19.04.2024 | 28.04.2024 | סדרת אימון מותאמת לקבלה בקלט לפי פרוטוקול התקשורת היחודי. |
| 4.2 | אימון המכונה עם סדרת האימון | 29.04.2024 | 12.05.2024 | העברת המידע מתבצעת בצורה תקינה תוך שימוש בפונקציות הבסיסיות. |
| 4.3 | הצלבת תוצאות מול GT | 13.05.2024 | 18.05.2024 | איסוף מידע אודות תוצאות האלגרותים אל מול התוצאות האמיתיות |
| 4.4 | תיקוני קוד ואלגוריתמיקה בהתאם לסדרת האימון | 19.05.2024 | 10.06.2024 | קבלת תוצאות תקינות אל מול הGT תוך הימנעות מ- over fitting. |
| 5.1 | יצירת WEB ראשוני לסביבת משתמש - הצבת קונסול | 11.06.2024 | 13.06.2024 | מתן אפשרות להכנסת קלט מהמשתמש וקבלת תשובה. |
| 5.2 | התחלת בדיקת האלגוריתם על סדרת מבחן - קלט מהמשתמש | 14.06.2024 | 24.06.2024 | אימות נכונות הפלט והסקת מסקנות לשיפור האלגוריתם. |
| 5.3 | תיקוני אלגוריתמיקה ומקצה שיפורים | 25.06.2024 | 30.06.2024 | שיפור יכולות האלגרותים ותוצאות הזיהוי תוך הסתגלות לדפוסים חדשים. |
| 0.2 | פגרה - תקופת מבחנים סמסטר ב' | 01.07.2024 | 01.08.2024 | תקופת מבחנים סמסטר ב' |
| 5.3 | תיקוני אלגוריתמיקה ומקצה שיפורים - המשך | 02.08.2024 | 15.08.2024 | שיפור יכולות האלגרותים ותוצאות הזיהוי תוך הסתגלות לדפוסים חדשים. |
| 6.1 | חיבור צד שלישי לייבוא פרסומי משתמש לקונסול האלגוריתם (פ'יצר לזיהוי משתמשים בעייתים פוטנציאלים) | 16.08.2024 | 23.08.2024 | קבלת מידע על פרסומי משתמש בתוצרת טקסט לצורך ניתוח בשלב הבא. |
| 6.2 | ניתוח המשתמש באמצעות האלגוריתם וחיזוי | 24.08.2024 | 30.08.2024 | הצגה ללקוח מידע סטטיסטי על הקלט שניתן - הצגת פוטנציאל המשתמש |
| 7.1 | שדרוג ועדכון WEB | 31.08.2024 | 05.09.2024 | שדרוג האתר - התאמה לקבלת משתמשים. הוספת שימוש נכון באלגוריתם, עיצוב הWEB ומימוש פתרונות נגישות. |
| 7.2 | בדיקת WEB סופית וסגירת גרסת production | 06.09.2024 | 09.09.2024 | גרסא סופית של המערכת בWEB. |
| 8.1 | יצירת Chrome extension | 10.09.2024 | 19.09.2024 | תוסף לדפדפן אשר מאפשר סימון טקסט והצגת חיזוי fake news |
| 9.1 | בדיקות סופיות של כלל המערכת | 20.09.2024 | 24.09.2024 | ניקוי הסביבה מכלל בעיות. |
| 10.1 | איסוף מידע לספר פרוייקט ומסמכי תיעוד | 25.09.2024 | 30.09.2024 | שמירת מידע מכל שלב בפרויקט לצורך כתיבתו בספר הפרויקט |
| 10.2 | כתיבת ספר פרויקט | 01.10.2024 | 09.10.2024 | כתיבת ספר פרויקט על פי דרישות המנחה. |
| 11.1 | סיום הפרויקט והיערכות להגנות | 10.10.2024 | 15.10.2024 | מוכנות להצגת הפרוייקט והצגתו. |

**רשימת מקורות לפרויקט:**

| **מספר מקור** | **שם המקור** | **שם המחבר** | **כתובת המקור** | **שימוש בפרויקט** |
| --- | --- | --- | --- | --- |
| 1.1 | Kaggle | Kaggle | [Kaggle: Your Machine Learning and Data Science Community](https://www.kaggle.com/) | לקיחת datasets מוכנים עם מידע חדשותי על מנת ללמד את האלגוריתם מה נכון ומה שקרי. |
| 1.2 | UCI Machine Learning Repository | UCI Machine Learning Repository | [Home - UCI Machine Learning Repository](https://archive.ics.uci.edu/) | לקיחת datasets מוכנים עם מידע חדשותי על מנת ללמד את האלגוריתם מה נכון ומה שקרי. |
| 1.4 | Open AI | Open AI | [ChatGPT (openai.com)](https://chat.openai.com) | ביסוס האלגוריתם ומפתחות. |