רשות החדשנות
Israel Innovation Authority

רשות ניירות ערך
Israel Securities Authority
سلطة الأوراق المالية

עיבוד שפה טבעית (NLP) ושוק ההון - **לאן?**

המחלקה הכלכלית - רשות ניירות ערך, רשות החדשנות | דצמבר 2021

תוכן עניינים

תקציר מנהלים 3

מבוא 5

פרק א: מהות הטכנולוגיה ומגמות התפתחות 7

פרק ב: יישומי עיבוד שפה טבעית (NLP) בשוק ההון 15

פרק ג: יישומי עיבוד שפה טבעית (NLP) ברגולציית ניירות ערך 24

סיכום 29

מונחון 31

ביבליוגרפיה 33

תקציר מנהלים

קידום החדשנות הטכנולוגית בשוק ההון הינו אחד מהיעדים המובילים בחזון רשות ניירות ערך. כנגזרת מיעד זה, הרשות תבחן בסדרת מאמרים (שזהו הראשון מבינם) את ההשלכות הפוטנציאליות על שוקי ההון של פיתוח טכנולוגיות חדשות. מימוש היעד לקידום החדשנות הטכנולוגית בשוק ההון נסמך בין השאר על שיתוף פעולה פורה עם רשות החדשנות המהווה שותפה לתוכניות עבודה ולכתיבת מאמר זה.

במוקד מאמר זה עומדת טכנולוגיית עיבוד שפה טבעית (NLP-Natural Language Processing) שמטרתה **הפקת משמעויות מטקסט באופן ממוחשב**. היכולת להפיק משמעויות ממאגרי טקסט גדולים באופן אוטומטי וללא מעורבות גורם אנושי טומנת בחובה תועלות בשני רבדים. ברובד הראשון, בדומה לשיפורים טכנולוגיים אחרים, היא גוררת **שיפור ביעילות** בעבודה עם מאגרי טקסט **וחסכון בעלויות**.

ברובד השני, העמוק יותר, היא מהווה **קפיצת מדרגה ליכולות חדשות**, שלא היו כלל אפשריות בלעדיה. דוגמא אחת לקפיצת מדרגה מסוג זה הינה היכולת להגיב באופן **מיידי** לפרסומים (תאגידיים או חדשותיים) ללא תלות באורך הטקסט שבפרסום. דוגמא שנייה היא היכולת להיחשף למגוון רחב של דעות של משקיעים שונים באופן מידי ו"לכרות" מידע מתוך **פרסומים ברשתות חברתיות**, בעיקר אלו המתמקדות במסחר בשוק ההון.

יכולות חדשות מסוג זה מהוות כמובן גם אתגרים בפני משקיעים ורגולטורים. כריית מידע, לדוגמא, עשויה לטמון בחובה פגיעה בפרטיות אף אם היא מתבצעת במרחב הציבורי. אכן, בהתייחסותו[1] של יו"ר ה–SEC לאירוע השורט סקוויז במניית גיימסטופ בינואר 2021, הוא ציין כי השימוש של משקיעים מתוחכמים בכלי עיבוד שפה טבעית לניתוח הפרסומים ברשתות חברתיות הינו תחום שהרשות האמריקאית תחקור לעומק. בהיבט הרחב יותר של המפגש בין טכנולוגיות חדשות לבין שוק ההון הוא הציב את השאלה המרכזית הבאה:

כאשר טכנולוגיות חדשות משנות את פני שוק ההון, כיצד ביכולתנו להמשיך ולממש את יעדי המדיניות הציבורית המרכזיים שלנו ולהבטיח שהשווקים מתפקדים כראוי בעבור משקיעים פרטיים?

בהתאם לשאלה מרכזית זו, המאמר בוחן ומפרט כיצד משפיעה ותשפיע הטכנולוגיה על שוק ההון בשתי פריזמות.

1 Testimony Before the House Committee on Financial Services, Chair Gary Gensler, 6.5.21.

בפריזמת המשקיעים, הטכנולוגיה מאפשרת לבצע אוטומציה וייעול של ניתוח פונדמנטלי, להרחיב את מנעד החברות והסקטורים אותם ניתן לסקור בזמן נתון ולהנגיש את התובנות הטמונות בניתוח הפונדמנטלי גם לכלי מסחר מבוססי ניתוחים כמותניים. בנוסף, הטכנולוגיה מאפשרת למשקיעים לזהות רגש המובע בזמן אמת בהודעות מיידיות של תאגידים, מידע חדשותי ופרסומים ברשתות חברתיות. מחד גיסא, בכך מתאפשרת "כריית דעות" יעילה על פני השוק והטמעת מידע יעילה יותר בתמחור. מאידך גיסא, נדון במאמר גם הסיכון בהטמעה מהירה יותר של ידיעות כזב (Fake News).

בהיבט השירותים המיועדים למשקיעים, גופים המספקים שירותי השקעה ופיננסיים מסתייעים בטכנולוגיה לצרכי שיפור השירות הניתן ללקוחות באמצעות תוכנות ייעודיות לקיום שיחה טקסטואלית עם אדם (צ'ט-בוטים). עם התפתחות הטכנולוגיה שבבסיסם, צ'ט-בוטים אלו עוברים מהיותם כלי טכני למתן משוב אוטומטי אל כלי הוליסטי שמשכלל את מאפייני הלקוח הספציפי ואת אסטרטגיות נותן השירות בהמלצותיו.

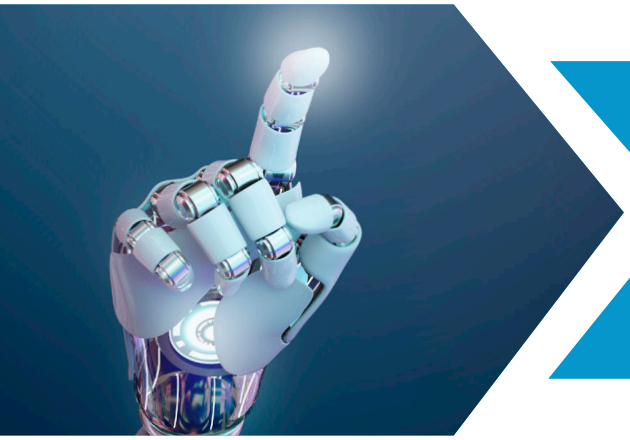
בפריזמת הרגולטור, נסקרים יישומי Suptech המבוססים על עיבוד שפה טבעית. המאמר כולל דוגמאות מן העולם ומהשוק המקומי, בינן פרויקטים המקודמים בשנים האחרונות במסגרת תכנית הפיילוטים (Data Sandbox) המשותפת של רשות ניירות ערך ורשות החדשנות, שמיועדת לקדם את פעילותן של חברות פינטק בישראל.

כלי NLP עשויים לייעל ולטייב את הזיהוי של דיווחי מפוקחים שאינם תקינים ושגילוי המידע בהם אינו מיטבי או שאינו מקיים את דרישות החוק. בכך הם משרתים את ליבת העיסוק הפיקוחי של רגולטורים פיננסיים. המעבר בשנים האחרונות לדיווח כספי מתויג (IXBRL) בשוק הגלובלי והמקומי, הינו גורם תומך לשימוש בכלי בינה מלאכותית בכלל ו – NLP בפרט, הודות לסטנדרטיזציה הטמונה בו, המסייעת לניתוח הדוחות על ידי מחשב.

כמו כן, הניתוח האוטומטי של פרסומים תאגידיים וחדשותיים טקסטואליים מאפשר לגופים האמונים על זיהוי סיכונים (רגולטורים פיננסיים, חברות דירוג אשראי וכו') לטייב את תהליכי זיהוי הסיכונים שלהם, זאת על בסיס מידע זמין שעקב היקפיו אינו בר מיצוי בכלים ידניים. עם זאת, המאמר שם גם דגש על חשיבות שימור הגורם האנושי בשרשרת ההחלטה, שכן על אף הייעול הטמון בטכנולוגיה, היא אינה חפה משגיאות אפשריות.

נציין כי השימוש בטכנולוגיה הינו בליבת הפעילות במגוון רחב של סקטורים (רפואה, טכנולוגיה וכו') ועל כן בא לידי ביטוי רב גם בפריזמת התאגידים. המיקוד במאמר זה הוא כמובן לגבי השלכות הטכנולוגיה הנובעות ממשק ציבור המשקיעים עם שוק ההון גרידא (בין אם באופן ישיר במסגרת פעילותם בשוק ההון ובין אם באופן עקיף דרך פעולותיו של הרגולטור) ועל כן פריזמת התאגידים לא נסקרת במאמר זה.

אנו מבקשים להודות לד"ר ארנה ברי, לפרופ' עלי בוקשפן ולפרופ' קרין נהון על תרומתם הרבה והתייחסותם המקצועית אשר סייעה לנו בגיבוש ובכתיבת מאמר זה.



מבוא

"רשות ניירות ערך תפעל לבסס ולהרחיב שוק הון ציבורי אטרקטיבי, הוגן, תחרותי וחדשני, במטרה לתרום לפיתוחה של הכלכלה הישראלית, הכול מתוך שמירת ענייניו של ציבור המשקיעים."

כנגזרת מחזון הרשות, קידום החדשנות הטכנולוגית בשוק ההון הינו אחד מיעדיה המובילים. הרשות אינה קופאת על השמרים בהיבט זה ומתוך מבט צופה פני העתיד תבחן בסדרת מאמרים (שזהו הראשון מבינם) את ההשלכות הפוטנציאליות של פיתוח טכנולוגיות חדשות על שוקי ההון. מימוש יעד קידום החדשנות הטכנולוגית בשוק ההון נסמך בין השאר על שיתוף פעולה פורה עם **רשות החדשנות** המהווה שותפה לתוכניות עבודה ולכתיבת מאמר זה.

מאמר זה עוסק **בטכנולוגיית עיבוד שפה טבעית** (NLP-Natural Language Processing) שמטרתה הפקת משמעויות מטקסט לא מובנה באופן ממוחשב. שוק ההון בפרט, והפעילות העסקית בכלל, מתבסס רבות על מקורות טקסטואליים ועל כן, יישומי טכנולוגיה זו צפויים להרים תרומה משמעותית לתחום.

בהתאם לכך, רגולטורים וגופים מובילים בשוק ההון בוחנים את השלכות הטכנולוגיה ואת הפוטנציאל הטמון בה, כפי שיוצג בהמשך הנייר:

- הרשות לניירות ערך האמריקאית (SEC) ציינה את פוטנציאל הטכנולוגיה לטייב את עבודת הרגולטור באיתור סיכונים ופירטה את היישומים הראשונים של הטכנולוגיה ברשות[2]. בנוסף, הרשות האמריקאית תקנה תקנות שיחייבו תאגידים באופן הדרגתי לדווח את דוחותיהם הטקסטואליים בפורמט קריא על ידי מחשב[3], במטרה להנגישן לטכנולוגיות חדשות, כגון NLP.
- FINRA בחנה את השלכות הטכנולוגיה על שוק ההון במסגרת סקירה רחבה הנוגעת לממשק שבין שוק ההון לטכנולוגיות חדשות[4].
- חברת S&P Global פרסמה מאמר סקירה[5] של הטכנולוגיה בתגובה ל"עניינם הגובר של משקיעים בטכנולוגיית עיבוד שפה טבעית".
- חברת Bloomberg מציינת כי לאורך העשור האחרון היא הגדילה את השקעותיה בטכנולוגיות NLP שבתורן מטייבות את יכולות מערכות המידע שלה[6].

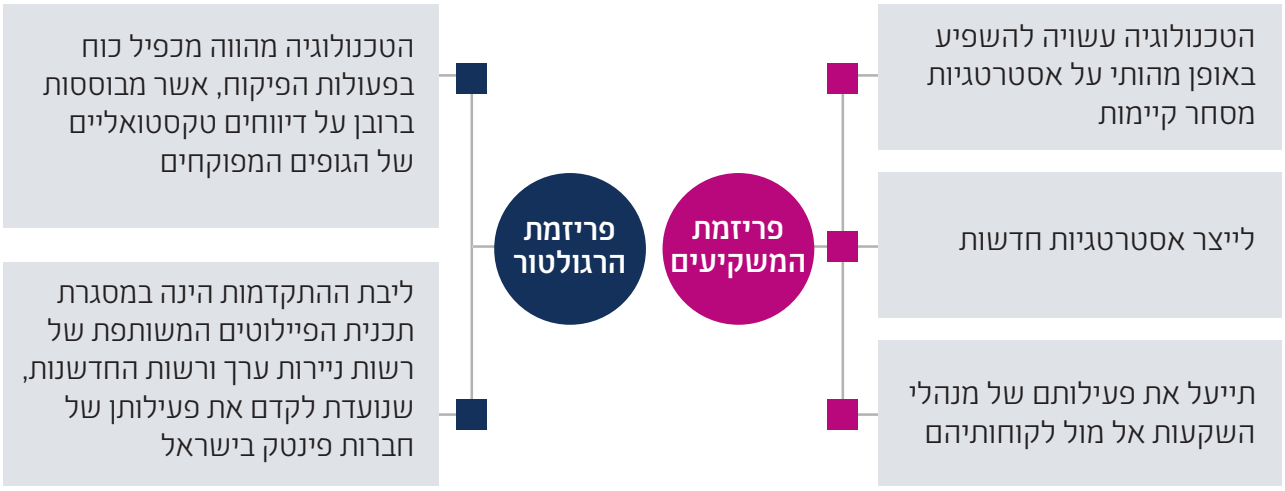
2 The Role of Big Data, Machine Learning, and AI in Assessing Risks: a Regulatory Perspective, SEC, 21.6.17

3 The Role of Machine Readability in an AI World, SEC, 3.5.18

4 Artificial Intelligence (AI) in the Securities Industry, FINRA, יוני 2020

5 Natural Language Processing, Part I: Primer, S&P Global

6 AI at Bloomberg, Bloomberg



בהמשך לעניין ההולך וגובר בטכנולוגיה, המאמר ייבחן וייפרט כיצד משפיעה ותשפיע הטכנולוגיה על שוק ההון בשתי פריזמות. בפריזמת המשקיעים יתואר כיצד הטכנולוגיה עשויה להשפיע באופן מהותי על אסטרטגיות מסחר קיימות ולייצר אסטרטגיות חדשות וכיצד היא תייעל את פעילותם של מנהלי השקעות אל מול לקוחותיהם.

בפריזמת הרגולטור, יתואר כיצד הטכנולוגיה מהווה מכפיל כוח בפעולות הפיקוח, אשר מבוססות ברובן על דיווחים טקסטואליים של הגופים המפוקחים. כפי שיתואר במאמר, רשות ניירות ערך יזמה בשנים האחרונות מספר פרויקטים הכוללים יישומי טכנולוגיית NLP. ליבת ההתקדמות הינה במסגרת תכנית הפיילוטים (Data Sandbox) המשותפת של רשות ניירות ערך ורשות החדשנות, שנועדת לקדם את פעילותן של חברות פינטק בישראל.

נציין כי השימוש בטכנולוגיה הינו בליבת הפעילות במגוון רחב של סקטורים (רפואה, טכנולוגיה וכו') ועל כן בא לידי ביטוי רב גם בפריזמת התאגידים. המיקוד במאמר הוא כמובן לגבי השלכות הטכנולוגיה הנובעות ממשק ציבור המשקיעים עם שוק ההון גרידא (בין אם באופן ישיר במסגרת פעילותם בשוק ההון ובין אם באופן עקיף דרך פעולותיו של הרגולטור) ועל כן פריזמת התאגידים לא תיסקר במאמר זה.

מבנה המאמר הינו כדלהלן: פרק א' יסקור ויסביר את **מהות הטכנולוגיה**, את מגמות ההתפתחות הצפויות בטווח הזמן הקצר ואת תמונת המצב בשפה העברית; פרק ב' יבחן את השפעות הטכנולוגיה על **יישומי מסחר** בהווה ובעתיד; פרק ג' יבחן את השפעות הטכנולוגיה בתחום **הרגולציה** וכפועל יוצא מכך על מבנה שוק ההון ועל המשקיעים.

פרק א
מהות הטכנולוגיה ומגמות התפתחות

ההתפתחויות הטכנולוגיות אותן אנו חווים, הביאו איתן, בין היתר, ייצור ועיבוד של כמויות עצומות של מסדי מידע ונתונים, מובנים[7] ושאינם מובנים, כגון: מידע אודיו-ויזואלי, מידע טקסטואלי וכו'. במקביל להתפתחות המהירה של הטכנולוגיה ושל נפחי הנתונים, מתעורר הצורך בהמרת המידע הרב לידע רלוונטי ואיכותי בצורה יעילה וזולה תוך שמירה נאותה על זכויות המשתמשים. המרה זו של מידע לידע, היא אחת מיעדי טכנולוגיות "**הבינה המלאכותית**" (AI- Artificial Intelligence) אשר מוגדרת כ"מערכת המסוגלת לפתור בצורה רציונאלית בעיות מורכבות או לנקוט בפעולות כדי להשיג את מטרותיה בנסיבות שונות בהן היא נתקלת בעולם האמיתי".[8] יישומים העושים שימוש בבינה מלאכותית נפוצים כיום במגוון רחב של תחומים. בפרט נכון הדבר לגבי יישומי **בינה מלאכותית צרה** אשר מתמקדים במגוון קטן של בעיות ובמשימות מוגדרות.

בעוד עיבוד נתונים מובנים הינו נוח לביצוע באמצעות מחשב, עיבוד מידע שאינו מובנה (מידע טקסטואלי וקבצי שמע[9] לדוגמא) והפקת תובנות ממנו הינם מורכבים יותר לביצוע באמצעות אלגוריתם. השלכות מורכבות זו גוברות כאשר במקרים רבים חלק הארי של המידע הרלוונטי ליישום נמצא במקורות הטקסטואליים.

אחת הנגזרות של עולם "הבינה המלאכותית" הנותנת מענה לאתגר זה, היא טכנולוגיית "**עיבוד שפה טבעית" (NLP)** שמטרתה לקרוא ולהפיק משמעות משפה אנושית. יישומי NLP עושים שימוש במגוון טכניקות חישוביות לניתוח וייצוג טקסטים על מנת לאפשר את עיבוד השפה האנושית עבור משימות שונות. טכנולוגיית ה-NLP מתבססת בין השאר על המחקר בתחום הבלשנות החישובית (Computational Linguistics). תחום זה עוסק בהבנת השפה הכתובה והמדוברת מנקודת מבט חישובית.

העולם העסקי בכללו ושוק ההון בפרט עושים שימוש נרחב במקורות טקסטואליים. ככאלה, קיימות בהם פעולות רבות הניתנות לטיוב באמצעות שימוש ביישומי NLP. פעולות מסוג זה (אשר יפורטו בהרחבה בהמשך הנייר) כוללות בין השאר ניתוח דו"חות כספיים והבנתם, עיבוד מידע חדשותי, טיוב תקשורת אוטומטית אל מול לקוחות ועוד.

טרום הדיון ביישומי הטכנולוגיה בפרקים הבאים, פרק זה יתאר את מהות הטכנולוגיה ומגמות עיקריות בהתפתחותה. בפרט, יודגמו האתגרים החישוביים בניתוח שפה טבעית (תת פרק א.1.), ייסקרו מגמות בהתפתחות התחום בהקשר הפיננסי (תת פרק א.2.) ויתוארו הרכיבים העיקריים בבסיס הטכנולוגיה (תת פרק א.3.). נסיים את הפרק בתיאור תמונת המצב בתחום בשפה העברית (תת פרק א.4.).

7 Structured. כלומר נתונים בעלי פורמט מוגדר מראש, לדוגמא מידע טבלאי.

8 רועי גולדשמידט, דו"ח בנושא "בינה מלאכותית", הכנסת - מרכז המידע והמחקר, 2018

9 בנייר זה יושם דגש עיקרי על מתודות עיבוד שפה טבעית הנוגעות לטקסט, כיוון שביישומים בשוק ההון מקור המידע העיקרי הינו טקסטואלי. נציין כי חלקים משמעותיים בעולם ה - NLP מתמקדים בעיסוק בנתוני שמע.

א.1. | המחשת האתגרים בעיבוד שפה טבעית

נפתח בדוגמא שמטרתה להמחיש את מורכבות האתגר ואת אופי הפתרון הניתן על ידי הטכנולוגיה. נניח כי אלגוריתם נדרש להפיק משמעות מן המשפט הבא:

"נפח ההשקעה הפרטית קטן יותר מאשר נפח ההשקעה המוסדית".

מטרת העל של האלגוריתם הינה להבין את תוכן המשפט ולהסיק ממנו מסקנות. באופן מופשט על האלגוריתם לבצע זאת על ידי חלוקת המשפט למס' ישויות (בדומה לפעולה האנושית):

יישות 1 (נושא)	יחס (נשוא)	יישות 2 (מושא)
נפח ההשקעה הפרטית	קטן יותר מ	נפח ההשקעה המוסדית

אם כן, **אתגר ראשון** שעומד בפני האלגוריתם הנו זיהוי החלקים השונים במשפט בצורה נכונה. לאחר חלוקה זו, על האלגוריתם לפענח את המשפט בצורה הבאה:

נפח ההשקעה הפרטית < נפח ההשקעה המוסדית

תובנה זו אמורה לסייע לאלגוריתם להסיק מסקנות נוספות. לדוגמא אם יוזן לו גם המשפט:

"נפח ההשקעה המוסדית קטן יותר מאשר נפח ההשקעה הזרה".

האלגוריתם יוכל להסיק את היחס הבא (מכללי לוגיקה פשוטים):

נפח ההשקעה הפרטית < נפח ההשקעה הזרה

נדגיש כי תובנה זו לא קיימת באף אחד מהמשפטים בנפרד אלא מהווה סינתזה של תובנות שונות על פני הטקסט. חיבור זה בין חלקים שונים בטקסט מהווה **אתגר שני**.

דא עקא, המורכבות אינה פוסקת כאן שכן ניתן לזהות מורכבות נוספת במשפט:

משמעות המשפט:

"נפח ההשקעה הפרטית קָטָן יותר מאשר נפח ההשקעה המוסדית".

שונה כמובן ממשמעות המשפט:

"נפח ההשקעה הפרטית קָטֹן יותר מאשר נפח ההשקעה המוסדית".

שאותו על האלגוריתם לפענח כ:

$\frac{d}{dt}$(נפח ההשקעה הפרטית) < $\frac{d}{dt}$(נפח ההשקעה המוסדית)

כלומר מהות המשפט מצויה בקצב השינוי של נפחי ההשקעה ולא בהיקפן הנוכחי.

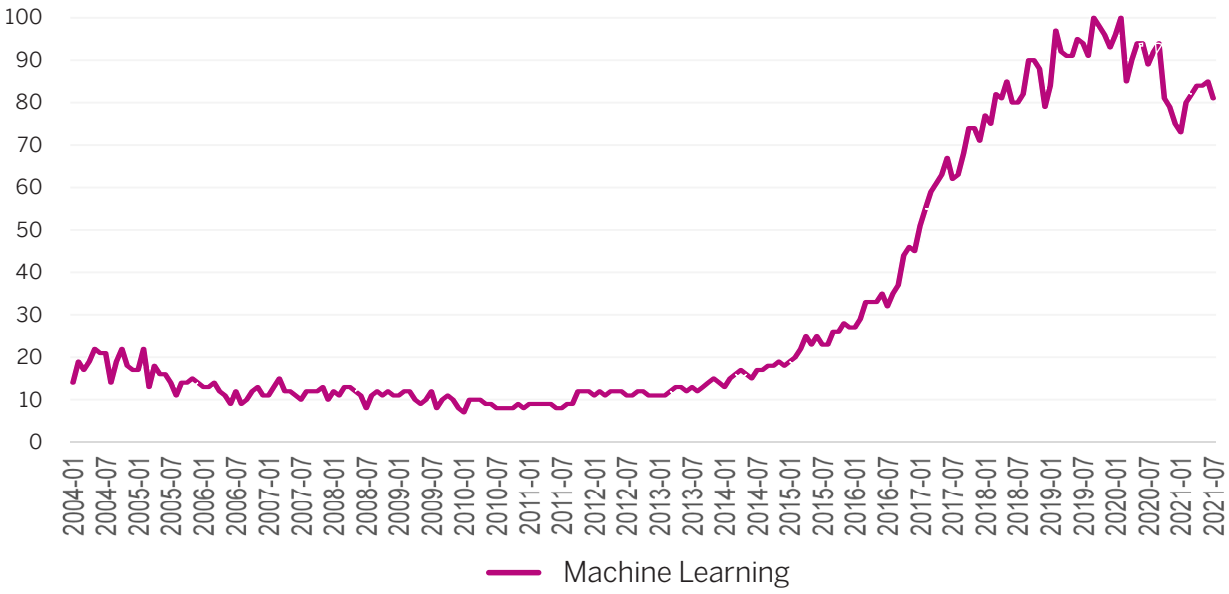
טקסט בשפה העברית המודרנית לרוב אינו מנוקד ולכן, את ההבחנה בין שתי המשמעויות האלגוריתם יוכל להסיק לדוגמה מדפוסים במשפטים נוספים בטקסט כגון התייחסות בטקסט לתקופה אחת בלבד (משמעות ראשונה) אל מול השוואה בין תקופות שונות (משמעות שנייה). על כן, ככל שכמות הפרטים הנלווים רבה יותר, כך פוטנציאל הסקת המסקנות הולך וגדל. **אתגר שלישי** שזיהינו הינו הכרעה בין מספר משמעויות אפשריות שונות בטקסט.

בתת הפרק א.3. נתאר את מרכיבי הטכנולוגיה וכיצד הם מספקים מענה לאתגרים שתוארו לעיל.

א.2. | מגמות התפתחות – NLP ופינטק

השימוש בכלי NLP לניתוח טקסטים פיננסים עומד בנקודת מפגש של שתי מגמות טכנולוגיות מובילות בעשור האחרון. המגמה הראשונה, שהינה רוחבית, היא הגידול בתפוצת השימוש בכלי למידת מכונה (Machine Learning - ML). למידת מכונה הינה תחום חישובי שמטרתו לאפשר לתוכנה ללמוד לבצע פעולות באופן עצמאי תוך התבססות על מסדי נתונים וכלים עם נתוני עתק (Big Data) סטטיסטיים. יכולת ההסקה הסטטיסטית של אלגוריתמי למידת מכונה תואמת את משימת ניתוח השפה האנושית עקב גמישות ומנעד ההבעה הרחב הקיים בשפה. זאת בניגוד לאלגוריתמים מבוססי כללים קשיחים (Rule Based) כגון "אם... אז" אשר לוקים בחסר בניתוח מבנים אמורפיים ומורכבים. התרשים הבא, שהופק מ-Google Trends, מתאר את הגידול בעניין בלמידת מכונה.

איור 1 | מגמת התפתחות העניין בלמידת מכונה



מקור: Google Trends[10]

הגרף מייצג את מגמות החיפושים במנוע החיפוש של גוגל של המילים 'MACHINE LEARNING'. נתוני הגרף הם משנת 2004 ועד יולי 2021 וכוללים את כלל החיפושים הגלובאליים. הנתונים מנורמלים לערך המקסימום על פני התקופה. ניתן ללמוד כי משנת 2015 החלה עלייה חדה במספר החיפושים של התחום.

המגמה הטכנולוגית השנייה, שהינה אורכית, היא הגידול בפיתוח יישומי טכנולוגיה פיננסית (Fintech, Finance Technology, פינטק). דוגמאות ליישומים השוכנים תחת קטגוריה זו ושהתפתחו במהלך העשור האחרון הנן: פלטפורמות מסחר דיגיטליות, שירותי תשלומים דיגיטליים, נכסים קריפטוגרפים וטכנולוגיית Blockchain, התאמה והנגשה אישית של שירותים פיננסיים ועוד. השימוש בכלי NLP בעולם הפינטק בא לידי ביטוי בעיקר בממשק שבין שירות פיננסי ללקוח או למקבל שירות ובניתוח מידע פיננסי טקסטואלי, כפי שיתואר בפירוט בהמשך נייר זה. שני תתי ענפים של עולם הפינטק בעלי חשיבות גבוהה למטרת מאמרנו זה: **Regtech** (Regulation Technology) שמתמקד בכלים טכנולוגיים המשרתים **גופים מפוקחים** ו- **Suptech** (Supervision Technology) שמתמקד בכלים טכנולוגיים המשרתים **רגולטורים**. על יישומי NLP בשני התחומים נרחיב בפרק ג'.

10 Google Trends

א.3. | תתי רכיבים בטכנולוגיה

כפי שהודגם בתת הפרק א.1., השפה האנושית היא מבנה מורכב, שהבנת משמעויות מתוכו דורשת היררכיה של תתי משימות נפרדות – הבנת המילה הבודדת, קישור יחסים בין מילים ומשפטים וכו'. בהתאם לכך, הכלים הטכנולוגיים העוסקים בניתוח השפה נדרשים לבצע סדרה ארוכה של שלבים שונים לצורך ניתוח טקסט. בתת פרק זה נתאר דוגמאות לשלבים אלו, מהרמה הגרעינית ביותר לרמה הכוללנית ביותר. בנוסף נדון במקורות המידע השכיחים ביותר ליישומי הטכנולוגיה.

נדגיש כי מטרת התיאורים בתת פרק זה (ובנייר זה בכללו) הינה ליצור הכרות ראשונית עם הטכנולוגיה ועל כן הם כוללים פישוט של דקויות טכניות.

שלב ראשון: הכנת הטקסט

השלבים הבאים מעבדים את הטקסט ברמה גרעינית, הנוגעת למשפטים בודדים ולמילים בודדות:[11]

1. **חלוקת הטקסט למשפטים:** מקורות טקסטואליים מורכבים לרוב ממאגרי קבצים גדולים, שבתורם מחולקים לפסקאות ולמשפטים. רכיב בסיסי אם כן בניתוח מקורות אלו הוא חלוקת הטקסט למשפטים בודדים. חלוקה זו מאפשרת את ניתוח משפטים אלו באופן בלתי תלוי ושיוך המשמעויות שמתקבלות אל רמת הפסקה והטקסט מהם הם נגזרו.
2. **חלוקה המשפטים לתמניות ("טוקנזציה"):** לאחר שהמשפט מחולק למשפטים בודדים, מבוצעת חלוקה של המשפטים לתַמְנִיות (Tokens) שהן אבני הבסיס עליהן מבוצעים עיבודים בכלי NLP. לרוב אבן בסיס (ועל כן התמנית) מוגדרת **כמילה בודדת**, אך קיימות גם הגדרות נוספות תלויות הקשר היישום.
3. **תיוג תחבירי:** לכל מילה, או תמנית, משויך תפקידה התחבירי במשפט. לדוגמא במשפט "איה קנתה מבועז מניה", התמנית "איה" תסווג כנושא המשפט והתמנית "קנתה" כנשוא המשפט. דוגמאות סטנדרטיות ומרובות של משפטים מתויגים תחבירית מאוגדות תחת "**קורפוס**" שמהווה מסד נתונים לייחוס עבור יישומי NLP שונים.
4. **המרת מילים לצורות בסיס (למיטיזציה):** בחלק גדול מן השפות מילים מוטות בהתאם למאפיינים תחביריים שונים (לדוגמא הטיית פעלים בהתאם לזמן הפעולה). בכדי להפיק משמעויות מן המילים בצורה אחידה, ישנה תועלת רבה בהמרת המילים לצורת בסיס אחת שלה מתווספות ההטיות התחביריות השונות ומשמעותן. לדוגמא, הצורות: "קָנָה, קוֹנֶה, נִקְנֶה" מומרות לצורת הבסיס שלהן במילון "קָנָה". צורת הבסיס המילונית מכונה גם "לֶמָה".
5. **זיהוי וסינון מילות קישור:** הקשר והיחסים בין למות שונות במשפט נגזרים ממילות קישור כגון "אֶת" ו"עם". על כן, מילים אלו מסומנות בנפרד מרכיבים תחביריים כגון נושא ונשוא המשפט.
6. **ניתוח התלות במשפט:** בהינתן כלל הרכיבים התחביריים במשפט כפי שתוארו לעיל, ניתן לבצע סינתזה מחודשת של הרכיבים. לכל משפט מוקצה מבנה תחבירי שנגזר מן הרכיבים ושלפיו מוגדרים הקשרים בין המילים השונות.
7. **זיהוי ישויות בשם (Entity Extraction):** שלב זה מתייג באופן חד חד ערכי ישויות כמו שמות של אנשים, שמות של חברות ומיקומים גיאוגרפיים המובאים בטקסט. יודגש כי אובייקט נדרש להיות מזוהה באופן אחיד אף אם שמו מובא בטקסט בצורות כתיבה שונות (לדוגמא "רשות ניירות ערך" ו"רנ"ע").

11 NLP in the Stock Market, Roshan Adusumilli, 2020

8. **יישוב הפניות משותפות (Coreference Resolution):** בטקסט בשפת אנוש מילים שונות במשפטים שונים עשויות להתייחס לאותה ישות ("המניה עלתה ב5%. היא הגיעה לערך שיא שנתי"). מטרת רכיב זה לזהות את ההפניות השונות כמשויכות לאותה ישות.
9. **שאיבת מידע מטקסט (Information Extraction):** מטלה זו מאגדת מגוון מתודות שמטרתן לחלץ מידע מובנה (לאמור - נתונים וההקשר שלהם) מתוך טקסט. כדוגמא, המטלה תשאף לייצר את שורת הנתונים הבאה:

שם מניה	תאריך	תשואה
X	3.8.21	1.2%

מתוך המשפט "המניה X עלתה אתמול ב – 1.2%" שמופיע בטקסט מתאריך 4.8.21. במסגרת ביצוע המטלה, האלגוריתם עושה שימוש, בין השאר, במגוון הכלים שתוארו לעיל.

לאחר יישום השלבים שפורטו לעיל על טקסט הקלט ניתן לנתחו בכלים שמטרתם למצות משמעויות מתוכו.

שלב שני: כלי ניתוח ויישום

הכלים והמטלות הבאים, המסתמכים על הכנת הטקסט, הם חלק מקבוצת יישומים ברמה גבוהה העוסקים בהפקת משמעות מטקסט בכללו:

1. **מידול נושאים מבוסס ניתוח סטטיסטי (Topic Modeling)**: כלי זה מיועד לזהות את הנושאים הנדונים במסמכים תוך התבססות על שכיחות הופעת מונחים מסוימים בטקסט. בנוסף מתבצע מידול נושאים על ידי השוואת מונחים המופיעים בטקסט לטקסטים אחרים על מנת לזהות את הנושאים הנדונים. לדוגמא, המערכת מבינה שטקסט עוסק ב"פוליטיקה" ו-"כלכלה" גם אם הוא אינו מכיל את המילים בפועל אלא מושגים קשורים כמו "בחירות", "דמוקרט", "דובר הבית" או "תקציב", "מס" או "אינפלציה".
2. **זיהוי רגש (Sentiment Analysis):** כלי זה מיועד לסווג טקסט לפי הרגש או הדעה המובעים בו ("חיובי", "שלילי", "נייטרלי" וכדומה) ולפי עוצמתם. זיהוי אוטומטי של רגשות בטקסט (על פני טקסטים רבים) מאפשר לדוגמא, להבין את דעת הקהל לגבי נושא מסוים בזמן אמת (בפרט מתוך רשתות החברתיות)[12] ואף עד כמה הנושא שנוי במחלוקת או מעורר עניין.
3. **זיהוי יחסים:** כלי זה נועד לחלץ יחסים בין שתי ישויות או יותר כפי שהם מובאים בטקסט.[13] יחס בין שתי ישויות עשוי לבטא, בין השאר, משמעויות הנוגעות ל -
   א. גודל יחסי - "חברה X בעלת שווי שוק גדול מחברה Y"[14].
   ב. קרבה - "מניה X ואג"ח Y שייכות לחברה Z".
   ג. סדר - "מניות X ו – Y הן, בהתאמה, הראשונה והשנייה במונחי שווי שוק במדד Z".
4. **סיווג ואשכול טקסטים:** בהינתן סט גדול של טקסטים שתוכנם לא ידוע מראש (לדוגמא מאגר פניות ציבור שאינו מקוטלג) ישנו צורך בזיהוי מתומצת של נושא הטקסט על בסיס קטגוריות נושאיות מוגדרות מראש. המענה לצורך זה נקרא סיווג טקסט. מטלה דומה (אשכול טקסט) כוללת מציאת קשרים בין טקסטים שתוכנם לא ידוע מראש, אך ללא קטגוריזציה מוקדמת של הנושאים האפשריים.

12 The basics of NLP and real time sentiment analysis with open source tools. Özgür Genç, 2019
13 Different ways of doing Relation Extraction from text, Andreas Herman, 2019
14 דוגמה נוספת מסוג זה פורטה בתת הפרק א.1 לעיל.

שיטות למימוש כלי הניתוח והכנת הטקסט

על מנת לפתח את רכיבי הטכנולוגיה שפורטו לעיל נעשה שימוש במגוון שיטות ללימוד מכונה. שיטות אלו ניתנות לסיווג תחת שתי מחלקות עיקריות: למידה מונחית (Supervised Learning) ולמידה בלתי מונחית (Unsupervised Learning). להלן נתאר בקצרה את גישות הפעולה שבבסיס שתי מחלקות אלו.

למידה מונחית מייצרת מודל להגדרת פלט מסוים עבור קלט מסוים, תוך התבססות על סט נתון של קלטים אפשריים ופלטים רצויים עבור קלטים אלו. מודל מסוג זה מבצע למעשה הסקת מסקנות מדוגמאות ("Learning by Example"). לצורך המחשת יישום של למידה מונחית בטכנולוגיות NLP, נתייחס למימוש של תיוג תחבירי (כפי שהוגדר לעיל). סט הקלטים והפלטים הנתון לצורך בניית מודל לתיוג תחבירי יכלול אוסף משפטים (שמהווים קלטים) ותיוגים תחביריים (שמהווים פלטים רצויים בעבור קלטים אלו). לרוב, תיוגי הדוגמא יינתנו על ידי תיוג אנושי מומחה שנכונותו מובטחת. המודל המיוצר בגישת הלמידה המונחית ילמד לבצע את התיוג באופן אוטומטי תוך התבססות על סט הדוגמאות הנתון.

למידה בלתי מונחית מתבססת על סט נתון של קלטים אפשריים בלבד ומסיקה קשרים ודפוסים מתוכם. לצורך המחשת יישום של למידה בלתי מונחית בטכנולוגיות NLP, נתייחס למימוש של אשכול טקסט (כפי שהוגדר לעיל). סט הקלטים האפשריים עשוי להיות מאגר של מסמכים שתוכנם אינו ידוע. תוך התבססות על פרמטרים סטטיסטיים שונים, המודל בגישת למידה בלתי מונחית מאתר דפוסי דמיון בין המסמכים השונים (לדוגמא, שכיחות דומה של מילים מסוימות) ועל ידי כך מייצר קשרים בינם.

גישות לפיתוח כלי NLP בעברית

יישומים קיימים בתחום ה - NLP לשפה העברית פותחו בשתי גישות עיקריות:

1. אלגוריתמי End to End: באלגוריתמים אלו מוזן קלט של מאגר מילים/שאלות/טקסט, בעוד הפלט הוא תשובה המתאימה לקלט ללא התחשבות במבנה תחבירי ומאפייני שפה ספציפיים. לדוגמא, הקלט עשוי להיות "אני מעוניין לקנות אוטו" והפלט המתאים: "מהו טווח המחירים הרצוי?". בהתאם, מדובר באלגוריתמים המאומנים למטרות ספציפיות. גישה זו דורשת מאגרי מידע גדולים לאימון ובעלת מגבלה עבור שיח הדורש מהפלטפורמה הבנה מעמיקה של השפה הנתונה.

2. פיתוח אלגוריתמים שמטרתם ללמוד את הפירוק של משפט למילים, המבנה התחבירי ומורפולוגיה (כפי שתואר לעיל). גישה זו דורשת סטים גדולים של משפטים מתויגים ומאגרי מידע רחבים לאימון המודלים. צעדים ראשונים בגישה זו בשפה העברית התבצעו במהלך העשור האחרון וממשיכים להתבצע כיום.

מקורות Data נפוצים ליישומי NLP

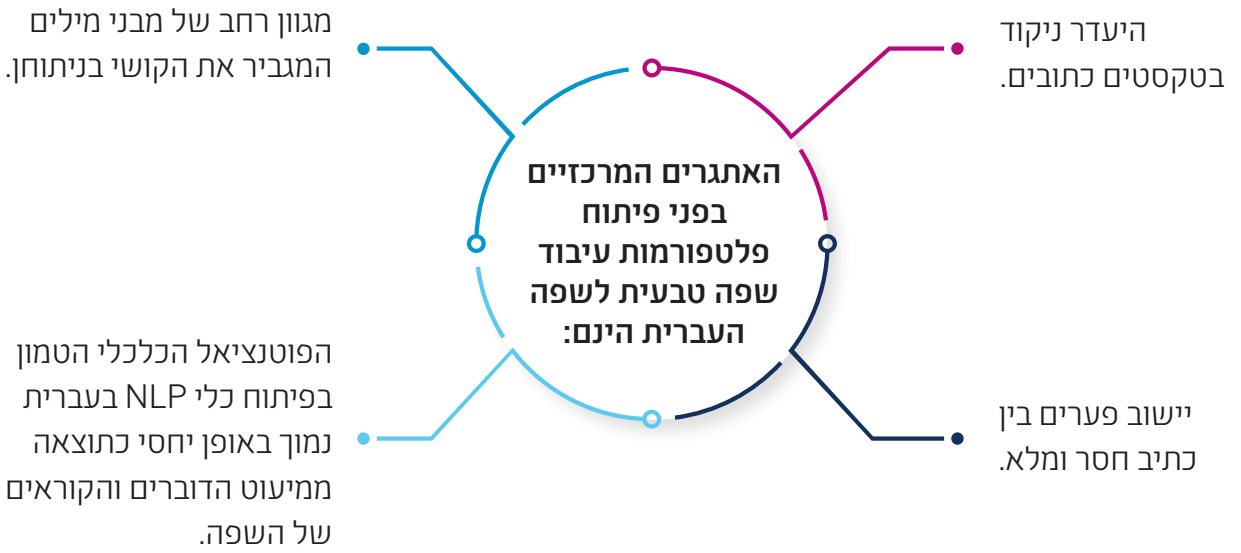
כאמור, יישומי NLP עושים שימוש ב-Data טקסטואלי שמטבעו אינו מובנה. מקורות ה-Data ליישומים אלו מגוונים וכוללים בין השאר: אתרי חדשות, רשתות חברתיות, ארכיונים ממוחשבים, אתרים מסחריים, שידורי המדיה, שיחות של מוקד לקוחות וכו'.

ככל שאמורים הדברים בעולם הפיננסיים יישומי NLP (כפי שיתוארו בהרחבה בפרקים הבאים) מתמקדים בעיקר בדיווחי תאגידים, אתרי חדשות פיננסיים (Bloomberg, Forbes, The Motley Fool וכו'), רשתות חברתיות המתמקדות בתחום הפיננסים (לדוגמא Stocktwits) ואתרי מסחר.

א.4. | תמונת מצב בשפה העברית

אתגרים

ליבת ההתפתחות הטכנולוגית בתחום ה - NLP כיום הינה בשפה האנגלית ורף התפתחות הטכנולוגיה עבור שפה אחת אינו מעיד בהכרח על מצב התפתחותה בשפה אחרת. זאת עקב הבדלים מהותיים בין שפות אנושיות שונות אשר מגבילים את היכולת להמיר כלים משפה אחת לאחרת. כדוגמא, למילים רבות בשפה העברית עשויות להיות משמעויות שונות אף אם הן כתובות באותו אופן, זאת בניגוד לשפות לטיניות ואחרות בהן מקרים אלו הם נדירים.



תהליכים מרכזיים בשנים האחרונות

בפברואר 2020 מינה יו"ר פורום תל"ם[15] ועדת בדיקה לבחינת הצורך בהתערבות ממשלתית לשם האצת התפתחות תחום הבינה המלאכותית ומדע הנתונים בישראל. בדצמבר 2020, פרסם פורום תל"ם את ממצאי הוועדה והמלצותיה. הוועדה המליצה[16] להשקיע בפיתוח קורפוסים (מסדי נתונים לייחוס עבור יישומי NLP) וכלי NLP לשפה העברית והערבית עבור התעשייה והמגזר הציבורי. להערכת הוועדה, לתחום זה חשיבות אסטרטגית והוא מהותי לצורך שימוש ביכולת בינה מלאכותית במשרדי הממשלה ובתעשיות נוספות.[17]

בשנת 2020 הקימה רשות החדשנות במשותף עם משרד הדיגיטל, את איגוד החברות לטכנולוגיות שפת אנוש, שיסייע בקידום הבנת השפה העברית והשפה הערבית במערכות ממוחשבות. האיגוד הוקם במטרה לתת לתעשייה להוביל את הגדרות הצרכים ולסייע בסגירת פערים טכנולוגים שיאפשרו לעשות שימוש במאגרי מידע לא מובנים בעברית ולהפיק על בסיסם תובנות שישמשו מנוף למוצרים ושירותים לחברות ישראליות.[18]

15 פורום תל"ם - הפורום לתשתיות לאומיות למחקר ולפיתוח ונחשב לחלק מהאקדמיה הלאומית הישראלית למדעים. מטרות פורום תל"ם הן לתאם בין גופים המרכיבים אותו בנושאי מחקר ופיתוח, לקיים התייעצויות, לאגם משאבים מתקציבי הגופים האלה ולקבוע אחריות ביצוע של אחד או יותר מגופי הפורום בנוגע לקידום תשתיות מחקר ופיתוח לאומיות. אתר פורום תל"ם

16 מיליארד שקל למחשב-על: כך תנסה ישראל לעלות על מפת הבינה המלאכותית, דצמבר 2020 ,TheMarker

17 ועדת בינה מלאכותית ומדע הנתונים, תל"ם, דצמבר 2020

18 משרד הדיגיטל הלאומי, הקמת איגוד חברות לטכנולוגיות שפת אנוש (NLP) בעברית ובערבית, 22.09.2020

תכנית העבודה של האיגוד מיועדת לייצר תשתית מו"פ שתציב בסיס אמפירי לא רק לזיהוי האלמנטים והדגמים המבניים המרכיבים את המערכת הלשונית, אלא גם למיפוי האופן שבו משתמשים במערכות האלו. על מנת לאפשר שיפורים מגוונים ורחבים ככל הניתן, הקורפוסים המתויגים בעברית ובערבית יהיו מתחומים מגוונים ורחבים ככל הניתן, בהם: חדשות, ארכיונים, סרטים, ספרים, מאמרים, שירות לקוחות, שידורי רדיו וטלוויזיה מתומללים ועוד. נוסף על כך, יבחן האיגוד את האפשרות להתאמת כלי צד ג' וכלים בקוד פתוח לבדיקות ושיפור איכות הבנת השפות עברית וערבית על ידי מערכות מחשוב שונות. באמצעות תשתית זו, ניתן יהיה לשפר ולהגביר את איכות הפתרונות השונים לזיהוי שפת אנוש בשפות אלו. האיגוד יקים את התשתית על גבי ענן שיאפשר שיתוף מאובטח של הקורפוסים והרצת מערכת ניהול ואלגוריתמים לכל השותפים באיגוד.

קבוצת המשתמשים שתבצע שימוש בתוצרי האיגוד תורכב הן מחברי האיגוד, המגיעים מתחומים שונים בתעשייה הישראלית ואלו יבצעו שימוש בתשתית לצורך פיתוח שירותים, יישומים ותוכנות לשיפור שירות לקוחות, ניהול ידע, קבלת החלטות ומימוש יישומים מתקדמים הדורשים הבנת שפה טבעית בעברית ובערבית.

בין החברות והמשתתפים בארגון נמצאות חברות המפתחות פתרונות תשתית (מחקר ופיתוח בתחומי הבנת שפה וחברות המפתחות אלגוריתמים המשמשים אבני בניין ליישומים שונים בתחום) וכמובן חברות העוסקות בפיתוח שירותים ומוצרים בתחומי הבנת שפה. הצרכנים הפוטנציאליים למוצרים ושירותים מבוססי טכנולוגיות זיהוי שפה טבעית מגיעים ממגוון נרחב של מגזרים ושירותים: הייטק, בנקאות, ביטוח, תקשורת, בריאות ועוד.

נכון לשנת 2020 בישראל פעלו 234 חברות סטארט אפ העוסקות בפיתוח יישומי טכנולוגיית NLP בתעשיות שונות.[19] במקביל לחברות טכנולוגיות מובילות הפועלות לפיתוח כלים ייעודיים לעברית, ישנן מספר מחלקות במוסדות אקדמיים העוסקות במחקר בתחומי AI ועיבוד שפה טבעית, בינן המרכז למחקר מדע הנתונים באוניברסיטת רייכמן (המרכז הבינתחומי הרצליה),[20] וקבוצות מחקר עיבוד שפה טבעית באוניברסיטת בר אילן[21] ובמכון וייצמן.

19 Start-Up Nation Central

20 המרכז למחקר מדע הנתונים, אוניברסיטת רייכמן

21 קבוצות מחקר, עיבוד שפה טבעית, המחלקה למדעי המחשב באוניברסיטת בר אילן

פרק ב

יישומי עיבוד שפה טבעית (NLP) בשוק ההון

הפעילות בשוק ההון מבוססת באופן נרחב על מידע המוכל במקורות טקסטואליים שכוללים, בין השאר, דו"חות כספיים, דיווחים מידיים של תאגידים, דיווחים בעיתונות הכלכלית, פרסומים ממשלתיים ובשנים האחרונות – מסרים במדיות חברתיות. בהתאם לכך שוק ההון מהווה כר פורה עבור יישומי NLP. בפרק זה, נסקור מספר יישומים של הטכנולוגיה בשוק ההון. יודגש כי הנדון בפרק זה מוכוון לאמוד את השפעות הטכנולוגיה (הזדמנויותיה ואתגריה) על **ציבור המשקיעים**.

ב.1. | אוטומציה של ניתוח פונדמנטלי

ניתוח **פונדמנטלי (Fundamental)** מבוסס על ניתוח דוחותיה הכספיים של חברה והשווקים בהם היא פועלת במטרה להעריך את שוויה של החברה. הערכת השווי מושתתת במידה רבה על **הגילוי הנאות** הניתן מטעם החברה, בין השאר בדוחותיה הכספיים. בנוסף לנתונים הכמותיים שנכללים בהם, הדוחות כוללים גם **מידע מהותי המובא בפורמט טקסטואלי**, בין השאר בפרק הביאורים. אם כן, מטלה בסיסית בהערכת שווי של חברה היא **גזירתו וכימותו של מידע כלכלי מהותי מן החלקים הטקסטואליים שבדוחות הכספיים**. כפי שנידון בפרקים לעיל, ביצוע אוטומטי של מטלות מסוג זה הינו בבסיס השימוש בכלי עיבוד שפה טבעית.

טכנולוגיית עיבוד שפה טבעית עשויה לייעל באופן דראסטי את ניתוח הדוחות הכספיים. בעוד הטכנולוגיה מאפשרת לנתח כמויות אדירות של טקסטים בזמן קצר, ניתוחם הידני דורש השקעת כוח אדם, זמן, כסף ומאמץ רבים. בנוסף, הטכנולוגיה מאפשרת להגדיל את היקפי סקירת הדוחות (בפרק זמן נתון) על פני תקופות ועל פני חברות. כך לדוגמא, משקיע יכול לסקור דוחות היסטוריים במהירות כדי לקבוע האם מתרחש שינוי מגמה בחברה אותה הוא מנתח.

יתרה מכך, הפירוט הטקסטואלי בדוחות הכספיים נמצא במתח תמידי שבין **עיקרון המהותיות** לעיקרון הגילוי המלא. בהתאם למתח זה, עשוי להיווצר מצב בו הפירוט בדוחות מוצג באופן סבוך ומסורבל כתוצאה מעודפי מידע עם רלוונטיות נמוכה. ניתוח פונדמנטלי אוטומטי בעל אפקטיביות גבוהה, יכול לאפשר למשקיע להסיק מהו המידע הרלוונטי ביותר עבורו, בפרק זמן קצר יותר, ובכך להגיע לתוצאות מהימנות יותר.

בעוד ניתוח פונדמנטלי בהגדרתו שואף לאמוד את שוויה של חברה ולסחור במניותיה אל מול מחירן בשוק, מנגנון האוטומציה בבסיס טכנולוגיית ה – NLP מאפשר לחבר את ההערכה הפונדמנטלית עם כלים **כמותניים (Quantitative)** שמתבססים על נתוני מסחר בזמן אמת ומערבים ומתמקדים בהיבטים כגון מומנטום, נפחי מסחר וכו'. בכך, ניתוח פונדמנטלי המבוסס על טכנולוגיות NLP שוכן

תחת אסטרטגיות **מסחר כמותניות – פונדמנטליות (Quantamental)**[24,23,22] שמערבות את שתי הגישות למסחר, לרוב באמצעות פלטפורמה ממוחשבת אחודה.

אחת הגישות הנפוצות לניתוח דוחות פיננסיים היא שימוש במילון הפיננסי של **לאוגורן ומקדונלד (Laughorn and McDonald, 2011)**,[25] שיצרו מילון בעל דגש ספציפי לתחום הפיננסי שנחשב למילון הבסיס בניתוח פיננסי באמצעות NLP. המילון (אשר מתעדכן באופן שוטף[26] ונגיש לשימוש חופשי) כולל רשימת מילים רלוונטית לדוחות השנתיים של חברות בארה"ב (המכונים 10-Ks על שם הטופס בו הם מוגשים ל - SEC)[27]. ברשימת המילון הסופית ישנן מעל ל-80,000 מילים, כאשר מתוכן מעל 350 מילים בעלות סנטימנט חיובי ומעל 2,300 מילים בעלות סנטימנט שלילי. המילון מבוסס על 50,115 דוחות 10-Ks מהשנים 1994-2008. ניתוח המילים וחלוקתן לקטגוריות (שלילי וחיובי) התרכז בניתוח פרק דיוני ההנהלה ואנליזה מתוך דוח 10-K שכן פרק זה הוא בעל הפוטנציאל הגבוה ביותר לחשיפת מידע על עמדת ההנהלה. הטבלה הבאה מדגימה את 30 המילים השליליות הנפוצות ביותר בדוחות ה-10-K. ניתן לראות מהן התדירויות של המילים מתוך סך המילים הנחשבות לשליליות.

22 Quantamental: What It Is & Why It Works, Leo Smigel, 2020

23 Natural Language Processing – Part III: Feature Engineering, Frank Zhao, S&P Global, 2020

24 Capital Markets Natural Language Processing - Part II: Stock Selection, Frank Zhao, 2018

25 Ks, Tim Loughran and Bill Mcdonald,-10 When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 2011 ,49 Page

26 University of Notre Dame, Software repository for accounting and finance

27 K, SEC website-10 Form

איור 2 | 30 המילים השליליות הנפוצות ביותר בדו"חות 10-K (משמאל) ובפרקי דיוני ההנהלה (מימין).

Panel B: Fin-Neg

Full 10-K Document				MD&A Subsection			
Word in H4N-Inf	Word	% of Total Fin-Neg Word Count	Cumulative %	Word in H4N-Inf	Word	% of Total Fin-Neg Word Count	Cumulative %
√	LOSS	9.73%	9.73%	√	LOSS	9.51%	9.51%
√	LOSSES	5.67%	15.40%	√	LOSSES	7.58%	17.10%
	CLAIMS	3.15%	18.55%	√	IMPAIRMENT	4.71%	21.81%
√	IMPAIRMENT	3.04%	21.59%		RESTRUCTURING	2.93%	24.74%
√	AGAINST	2.58%	24.17%	√	DECLINE	2.89%	27.62%
√	ADVERSE	2.44%	26.61%		CLAIMS	2.71%	30.33%
	RESTATED	2.09%	28.70%	√	ADVERSE	2.44%	32.77%
√	ADVERSELY	1.75%	30.45%	√	AGAINST	2.01%	34.78%
	RESTRUCTURING	1.72%	32.17%	√	ADVERSELY	1.94%	36.72%
	LITIGATION	1.67%	33.83%		LITIGATION	1.67%	38.40%
	DISCONTINUED	1.57%	35.40%		CRITICAL	1.63%	40.03%
	TERMINATION	1.35%	36.75%		DISCONTINUED	1.62%	41.64%
√	DECLINE	1.19%	37.93%	√	*DECLINED*	1.30%	42.94%
√	CLOSING	1.08%	39.01%		TERMINATION	1.06%	44.00%
√	FAILURE	0.97%	39.98%	√	NEGATIVE	0.96%	44.96%
	UNABLE	0.84%	40.82%	√	FAILURE	0.93%	45.89%
√	*DAMAGES*	0.82%	41.64%		UNABLE	0.91%	46.80%
√	DOUBTFUL	0.77%	42.41%	√	CLOSING	0.86%	47.65%
√	LIMITATIONS	0.75%	43.17%		*NONPERFORMING*	0.81%	48.47%
√	FORCE	0.74%	43.91%	√	IMPAIRED	0.81%	49.28%
√	VOLATILITY	0.73%	44.64%	√	VOLATILITY	0.79%	50.07%
	CRITICAL	0.73%	45.37%	√	FORCE	0.75%	50.82%
√	IMPAIRED	0.70%	46.07%	√	*NEGATIVELY*	0.73%	51.56%
	TERMINATED	0.70%	46.77%	√	DOUBTFUL	0.72%	52.27%
√	*COMPLAINT*	0.63%	47.39%	√	*CLOSED*	0.70%	52.97%
√	DEFAULT	0.57%	47.96%	√	DIFFICULT	0.69%	53.66%
√	NEGATIVE	0.51%	48.47%	√	*DECLINES*	0.63%	54.29%
√	*DEFENDANTS*	0.51%	48.99%	√	*EXPOSED*	0.60%	54.89%
√	*PLAINTIFFS*	0.51%	49.49%	√	DEFAULT	0.59%	55.48%
√	DIFFICULT	0.50%	50.00%	√	*DELAYS*	0.56%	56.04%

העמודה (% of Total Fin-Neg) מייצגת את יחס מספר הפעמים שמילה ספציפית מופיעה במסד מתוך כלל המילים השליליות. העמודה (Cumulative) מהווה סכימה של המילים שפורטו עד אותה שורה. ניתן לראות שעבור הדו"חות כ-30 מילים מרכיבות כ-50.00% מסך המילים המופיעות במסמכים. מעבר לכך, ישנה חשיבות לשימוש בערך משקלי לחשיבות של מילה והמשמעות שלה בטקסט, מאחר שמילה אשר מופיעה פעמים רבות לא בהכרח נחשבת ליותר אינפורמטיבית ממילה אחרת אשר מופיעה בתדירות גבוהה. דוגמא מהטבלה: המילה Loss, מרכיבה כ-9.73% מסך המילים השליליות המופיעות בדוחות.

בנוסף לרשימת המילים מסווגות הרגש, מילון לאוגורן ומקדונלד מכיל רשימות של מילות "נימה" שמטרתן לתפוס התדיינות, חוסר וודאות ומילות הדגש חלשות וחזקות, המתואמות להקשר הפיננסי. המילון מאפשר למכונות להעריך ממדים נוספים של הערות במסמך הדיווח.

בהתבסס על מילון זה, חברת S&P Global פיתחה מודל המאפשר לה לחזות את הביצועים של חברה על פי הדיווחים הרבעוניים של החברה ותמליל שיחות הרווחים הרבעוניים.[28] בעבור כל מניה ספציפית המודל משווה את היחס ההיסטורי בין השינוי בגישה הרגשית הנאמדת בדוחות החברה הנבדקת אל מול התוצאות של המנייה עד הדיווח הרבעוני הבא. רמת השליליות או החיוביות בדיווח נמדדת לפי מספר המילים (השליליות או חיוביות) בהן נעשה שימוש בדיווח ובשיחה הרבעונית בין המשקיעים להנהלה.

28 Natural Language Processing, Part I: Primer, S&P Global

בדומה למניות, ניתן לבצע ניתוח דיווחים טקסטואליים ברמת התעשיות השונות בכל רבעון, על פי סך המילים השליליות או החיוביות הנאמרות בדיווחים של כלל החברות שנכללות בכל תעשייה. דוגמא לניתוח ברמת התעשייה מוצגת בטבלה הבאה:

איור 3 | רגש מוגדר כאחוז המילים השליליות בתמליל שיחות הרווחים הרבעוניים בהתבסס על מילון לאוגורן ומקדונלד (2011).

Exhibit 4: S&P 500 Trends in Sentiment Change

GICS Industry Groups	Calendar Quarters Q3 2016	Q4 2016	Q1 2017	Q2 2017
Consumer Services	-4.1%	-4.0%	16.9%	20.1%
Software & Services	-0.2%	4.1%	-3.9%	19.2%
Transportation	-3.6%	-2.8%	14.6%	19.1%
Diversified Financials	5.4%	13.8%	16.5%	18.0%
Technology Hardware & Equipment	4.8%	7.1%	18.1%	17.5%
Banks	-3.7%	-0.4%	18.1%	16.8%
Capital Goods	-4.9%	9.1%	16.9%	15.5%
Commercial & Professional Services	7.6%	-4.4%	-6.2%	14.6%
Utilities	-4.6%	7.2%	7.0%	14.1%
Insurance	0.0%	14.4%	10.2%	11.9%
Energy	0.1%	13.6%	25.8%	10.3%
Real Estate	-11.5%	-2.4%	0.5%	5.5%
Media	-11.6%	-12.6%	5.0%	4.2%
Materials	5.7%	-1.8%	12.2%	1.4%
Semiconductors & Semiconductor Equipment	6.3%	-3.7%	14.5%	0.5%
Retailing	-18.4%	-0.3%	2.4%	-1.6%
Consumer Durables & Apparel	10.7%	0.2%	-4.7%	-2.6%
Pharmaceuticals, Biotechnology & Life Sciences	-5.7%	-3.6%	-1.3%	-2.9%
Household & Personal Products	-3.3%	-2.9%	-7.8%	-3.9%
Food & Staples Retailing	17.7%	-0.8%	-16.5%	-4.0%
Food, Beverage & Tobacco	2.2%	0.4%	-3.6%	-5.0%
Health Care Equipment & Services	-8.2%	-2.7%	-1.0%	-7.5%
Autos & Components	-14.6%	-4.3%	34.2%	-8.5%
Telecommunication Services	9.1%	-13.6%	-36.9%	-29.4%

Sentiment Improved

Sentiment Deteriorated

Note: Sentiment is defined as the proportion of negative words in an earnings call using Loughran and McDonald (2011). Sentiment changes are measured quarter-over-quarter from four quarters ago where the values are multiplied by -1 to make results easier to interpret. Industry group level values are rolled up equal-weighted from the stock-level. Source: S&P Global Market Intelligence Quantamental Research. Data as of 08/08/2017.

השינוי ברגש ברבעון נתון נאמד כגידול האחוזי ברגש באותו רבעון ביחס לרבעון שלפניו. הערכים מוכפלים ב – 1- בכדי שירידה במילים השליליות תיוצג באופן חיובי. ערכי התעשייה מחושבים על ידי ממוצע פשוט על פני מניות. מתוך: S&P Global Market Intelligence Quantamental Research.

מטבלה זו ניתן ללמוד על השינוי ברגש המובע בתעשיות השונות בין הרבעון השלישי בשנת 2016 לרבעון השני בשנת 2017, מתוך מניות S&P 500. ההשוואה מתמקדת בשינוי מרבעון לרבעון, כאשר הסנטימנט של כל ענף נמדד על פי מספר המילים השליליות שנאמרו מתוך סך המילים בשיחת הרווחים של מניה מאותו ענף. התחומים בירוק הם אלו שהשינוי ברגש בהם היה חיובי באותו הרבעון, כאשר השינוי ההדרגתי מירוק לאדום בגווני הצבע משקף את כיוון וגודל השינוי. בטבלה ניתן לזהות לדוגמא, שהערך הרגשי של תעשיית הבנקים השתפר משמעותית לאורך ארבעת הרבעונים. יודגש כי, בדומה לטכנולוגיות בינה מלאכותית רבות, המודלים לזיהוי רגש אינם חפים משגיאות אשר מהוות אתגר קיים בתהליך (כפי שיתואר גם בסוף פרק זה).

השימוש בכלי NLP לניתוח דיווחים כספיים מאפשר גם בחינה של סוגיות ספציפיות בגילוי התאגידי.

בחינה זו מתבצעת לרוב על ידי הגדרת מילון נושאי ייעודי לסוגיה הנדונה. כדוגמא, מחקר[29] משנת 2020 בחן את היקף ההתייחסויות בדו"חות כספיים לנושאים סביבתיים, חברתיים וממשל תאגידי (ESG - Environmental, social and corporate governance) בין השנים 2012 ל– 2015 על ידי בניית מילון ייעודי לתחומים אלו. המחקר מצא עלייה של כ–126% בגילוי הנוגע לנושאים סביבתיים ושל כ–7% בנוגע לנושאים חברתיים. הגידול בדיון בנושאי ממשל תאגידי לעומת זאת נשאר קבוע למדי לאורך התקופה.

מחקר נוסף בתחום האוטומציה של ניתוח דיווחים כספיים[30] מצביע על יכולת למיצוי אוטומטי של **קשרי סיבה ותוצאה** מטקסט כלכלי. הכלי שנבנה במסגרת המחקר נוסה על דיווחים כספיים בשפה היפנית ואפשר לייצר מתוך הדו"ח הכספי סט של קשרים בין אובייקטים שזוהו כסיבות לבין אובייקטים שזוהו כתוצאות. על ידי שרשור של קשרי סיבה ותוצאה מרובים (גם על פני דוחותיהן של חברות שונות) הכלי מוכוון להתחקות אחר תהליכי סיבתיות מורכבים. תובנות אלו לגבי סיבה ומסובב עשויות לספק ערך מוסף לתהליכי קבלת החלטות השקעה של משקיעים.

ב.2. | ניתוח זמן אמת – הודעות מיידיות, חדשות ורשתות חברתיות

בהמשך לסעיף הקודם, מחקרים ויישומים בתחום מובילים גם לניתוחים מבוססי NLP בזמן אמת[31] עבור **הודעות מיידיות** של תאגידים, מידע חדשותי ופרסומים ברשתות חברתיות. שיפור זמני התגובה והחיסכון בעלויות שנוצרים כתוצאה מאוטומציה של ניתוח טקסט הם בעלי חשיבות מכרעת עבור סוחרים הפועלים במסחר תוך יומי, בדגש על סוחרים אלגוריתמיים (Algo-Traders) אשר מוגדרים כסוחרים שפועלים במסחר תוך התבססות על קבלת החלטות וביצוע עסקאות באמצעות אלגוריתם.[32]

אחד הפרויקטים הראשונים בתחום זה הוא --The Thomson Reuters NewsScope Event Indi ces, המשמש מסגרת לשילוב חדשות בזמן אמת משירות המנויים של "תומסון רויטרס", עם פרוטוקולים שיטתיים של השקעות וניהול סיכונים. המסגרת מורכבת ממכלול מדדי אירועים בזמן אמת והיא נועדה לאתר אירועים חריגים ועל בסיסם לבנות תחזית.[33] באמצעות שימוש ביישום זה, התרגום והכימות של מידע טקסטואלי הופך נוח ומהיר יותר.

פלטפורמות רבות משקללות **מידע חדשותי** טקסטואלי בזמן אמת לצורך מודלי מסחר.[34] בנוסף, נעשה שימוש במחקרים וביישומים בניתוח דעות המובעות בפרסומים **ברשתות חברתיות** לצורך אנליזה של תנועות מחירים. אגרגציה מסוג זה של דעות המובעות על ידי מספר רב של סוכנים מכונה כריית דעות (Opinion mining). כדוגמא לכך, **במערכת המסחר של בלומברג** מיושמת פונקציה הנקראת

29 Environmental, social and governance reporting in annual reports: A textual analysis, Philipp Baier, Marc Berninger, Florian Kiesel, 2020

30 Economic Causal-Chain Search Using Text Mining Technology, Kiyoshi Izumi and Hiroki Sakaji, 2020.

31 NLP in the Stock Market, Roshan Adusumilli, 2020

32 Stock Price Forecasting by Combining News Mining and Time Series Analysis , Xiangyu Tang, Chunyu Yang, Jie Zhou, 2009

33 Managing real-time risks and returns: The Thomson Reuters NewsScope Event Indices, Alexander D.Healy & Andrew W.Lo

34 NLP in FinTech Applications: Past, Present and Future, Chung-Chi Chen, Hen-Hsen Huang, Hsin-Hsi Chen, 2019

"TREN" המבצעת סיווג רגשי לחדשות ולדיווחים ברשתות החברתיות.[36,35] המערכת מעדכנת בזמן אמת את ציון ה"רגש" של אותה מניה, הנע בטווח בין 1- (שלילי) ל-1 (חיובי) ואת ערך השינוי בנתון. בנוסף ניתן להשוות בין מספר חברות ותעשיות.[37]

הרשת החברתית Stock Twits, הדומה במאפייניה לטוויטר אך עוסקת בעיקר בשוק ההון, מאבחנת אף היא את דעת המשקיעים באמצעות כלי זיהוי רגש. המערכת של הרשת החברתית יכולה להגדיר האם "ציוץ" של משתמש הוא "דובי" (מביע דעה הצופה ירידת מחירים), "שורי" (מביע דעה הצופה עליית מחירים) או ניטרלי. משתמשים המעוניינים לאבחן את הדעה הרווחת על חברה מסוימת יכולים להעריכה באמצעות הרשת החברתית.[38]

ב.3. | התאמת שירות ללקוח בניהול השקעות ובשירותים פיננסיים

גופים מוסדיים, מנהלי תיקים ויועצי השקעות נבחנים בין השאר ביכולתם להתאים מאפייני השקעה למאפייני לקוח. כיום, פעילות זו מתבצעת ברובה בממשק שבין אנשי מקצוע (סוכני מימון וחשבונאות, יועצים פיננסים ואנליסטים למסחר) בגופי ניהול ההשקעות ובשירותים הפיננסיים לבין הלקוח. בהסתכלות עתידית, טכנולוגיות בינה מלאכותית הינן בעלות פוטנציאל להחליף פעולות רבות המבוצעות כיום באופן אנושי. בפרט, טכנולוגיות מבוססות NLP מאפשרות למכן את האינטראקציה של המוסד הפיננסי מול לקוחותיו בתחום קשרי הלקוחות.

כך, השירותים הניתנים על ידי יועצי השקעות ופיננסיים עשויים להיות מוחלפים על ידי תוכנה ייעודית (בינה מלאכותית צרה) לקיום שיחה טקסטואלית עם אדם (צ'ט-בוט). צ'ט-בוטים המצוידים ביכולת NLP מיועדים לפירוש דיבור או כתיבה אנושית ומתבססים כיום יותר ויותר על מתודות בינה מלאכותית. יכולות אלו מאפשרות להם לבצע קפיצת מדרגה מכלי טכני למתן משוב מבוסס חוקים (בדומה למענה קולי) אל כלי הוליסטי שכולל בתוכו גם את תהליך סיווג הלקוח הפוטנציאלי (לדוגמא - לאשר או לדחות בקשה) וניהול הקשר עם הלקוח הקיים (לדוגמא - לספק המלצות לפי אסטרטגיית המוסד הפיננסי).

אחד מהיישומים בתחום זה הינו **גרף הידע האישי** המוגדר כמקור לידע מובנה אודות ישויות והקשר ביניהן.[39] איתור הקשרים שבין לקוחות (קיימים או פוטנציאליים) בכלים מבוססי NLP[40] וניתוח רשתות חברתיות (Social Network Analysis) עשוי לאפשר התאמה טובה יותר של מוצרים פיננסיים ללקוח. נציין כי מתודולוגיה זו כבר מיושמת בהנגשת פרסומות בהתאם למאפייני לקוח.

דוגמא לגרף הידע מוצגת בתרשים מטה בו ניתן לראות קשרי גומלין שנבנו על סמך בסיסי נתונים שונים[41], ששילובם יכול לסייע לצ'ט-בוט פיננסי, להציע ללקוח מכשירים פיננסיים שונים. בגרף מתוארים הקשרים שבין המשתמשת (user, בעיגול ירוק) לבין אוסף של חברים המקושרים אליה דרך רשת חברתית (חברים א' – ג'). בנוסף מתוארים הקשרים בין המשתמשת לחבריה דרך ישויות משותפות נוספות (מוסד לימודים ומקום עבודה עכשווי).

35 AI at Bloomberg, Bloomberg

36 הרגש בשוק ההון ותגובות אתרי החדשות ורשתות חברתיות כלפי חברה כשלהי.

37 Numeral Understanding in Financial Tweets for Fine-grained Crowd-based Forecasting, Chung-Chi Chen, Hen-Hsen Huang, Yow-Ting Shiue, Hsin-Hsi Chen, 2018.

38 Stock Twits

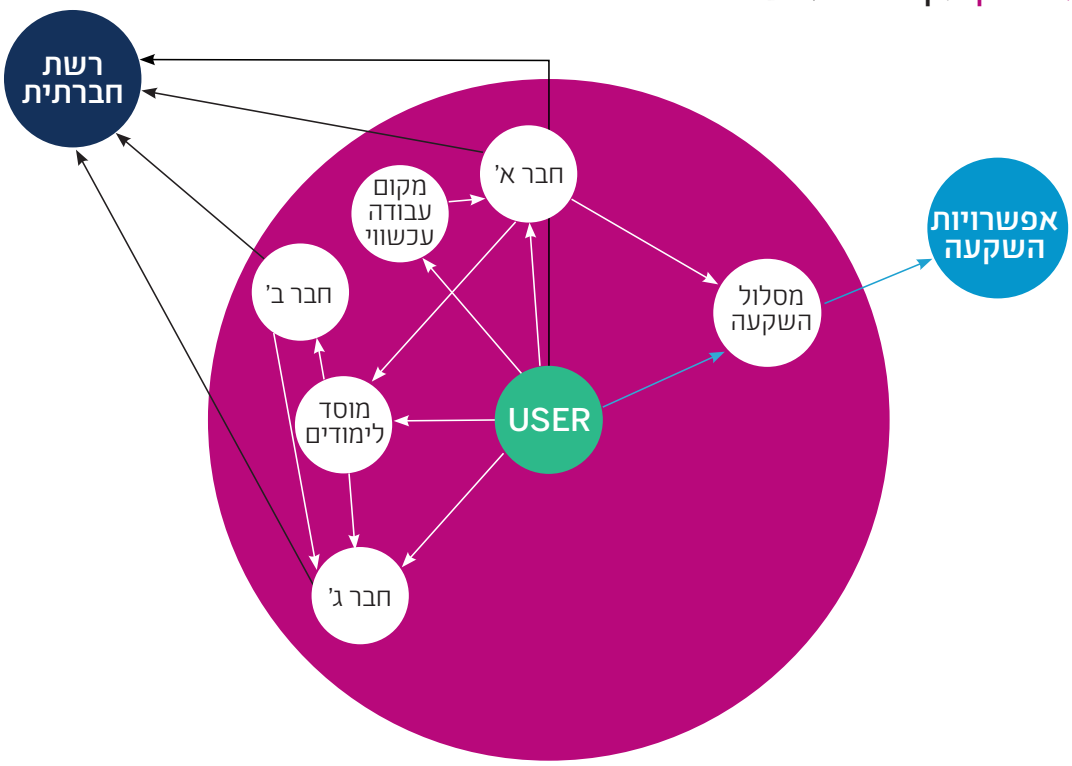
39 Personal Knowledge Graphs: A Research Agenda, Krisztian Balog & Tom Kenter, 2019

40 ראו "זיהוי יחסים" בפרק א'

41 נציין כי הדיון להלן מסתמך על ההנחה כי המשתמשים הנדונים מספקים את הרשאתם המלאה לצפייה בנתונים. אף אם ניתנת ההרשאה, איסוף הנתונים עשוי להעלות מספר שאלות אתיות שייבחנו בהמשך פרק זה.

על סמך הגרף, ניתן לראות למשל שחבר א' מקושר למשתמשת דרך הרשת החברתית, מוסד הלימודים ומקום העבודה העכשווי. בהתאם לכך ניתן ללמוד כי קיימת סבירות בלתי מבוטלת שאפיקי ההשקעה המועדפים על שניהם יהיו בעלי מאפיינים משותפים. כיוון שחבר א' מקושר אל מסלול השקעה מסוים, צ'ט-בוט המצויד בגרף הידע האישי עשוי להציע את מסלול ההשקעה אל המשתמשת (חצים כחולים) ובהמשך לפרוט בפניה אפשרויות השקעה נוספות.

איור 4 | גרף הידע האישי



בעוד צידה הראשון של התאמת השירות הוא הכרת הלקוח, צידה השני הוא אומדן אפקטיבי יותר של אפיק ההשקעה עצמו. אומדן זה ניתן לביצוע במתודות שתוארו בשני תתי הפרקים הראשונים בפרק זה.

יכולת נוספת בה ניתן להיעזר בטכנולוגיית NLP בקבלת החלטות השקעה, היא השימוש בעיבוד שפה טבעית על מנת להסביר את ההחלטה שהתקבלה בעזרת מערכת מבוססת בינה מלאכותית. מסחר באמצעות פלטפורמות בינה מלאכותית עלול להציג המלצות מבלי להסביר את סיבות קבלת ההחלטות. טכנולוגיית NLP מסוגלת לעבד את נתוני הקלט וההחלטה המשולבים כדי לייצר דוח המסביר כיצד התקבלה ההחלטה בתצורת טקסט אנושי. תהליך זה נקרא **ייצור שפה טבעית (NLG - Natural Language Generation)**. זהו תהליך תוכנה ההופך נתונים מובנים לשפה טבעית. ניתן להשתמש בו כדי לייצר תוכן מוסבר (מילולי) ממאגרי נתונים וטפסים ארוכים (מספריים). כדוגמא, במקרה בו מודלי חיזוי יצפו תת ביצועים בסקטור מסוים בעקבות נתוני מקרו חזויים המוזנים אליהם, יישום ה – NLG מסוגל לייצר כהסבר לתהליך את המשפט:

"ניתנה המלצה להקטנת החשיפה בסקטור X עקב נתונים חזויים Z,Y וביצועיהן ההיסטוריים של מניות הסקטור במצבים דומים".

פלט זה מאפשר למנהלי ההשקעות לבדוק, לאשר, או לפקח על המלצת המסחר. חברות יכולות להשתמש בפלט הטקסטואלי גם כדי לדווח ללקוחות או לרגולטורים על המניעים לביצוע פעולה.

במקרה זה NLP משמש כמערכת תמיכה לקבלת החלטות.[42]

ב.4. | אתגרים

בדומה לכל כלי אחר, השימוש בכלי NLP לצרכי ניהול השקעות אינו נטול חסרונות ואתגרים.

- **השפעה על כתיבת הטקסטים** - המחקר "איך לדבר כשמכונה מקשיבה"[43] לדוגמא, בחן את השאלה האם מנהלים מתאימים את שפת הדיווחים שלהם לתכונות כלי ה-NLP המוכרים. ממצאי המחקר מצביעים על כך שהגידול בקהל הסוחרים הנעזרים במכונות ובבינה מלאכותית ושמקורות המידע שלהם הם הדוחות הכספיים של החברות הנבדקות, מניע חברות להפיק דיווחים מוטים לניתוח ועיבוד מכונות. דוגמא להטיה היא הימנעות משימוש במילים הנתפסות כשליליות על ידי אלגוריתמים ומילונים מהסוג שתוארו בפרק זה.
- **ידיעות כזב** – זיהוי הרגש האוטומטי בדיווחים טקסטואליים שמקורם באתרי חדשות ובמדיה חברתית טומן בחובו סכנה בדמות קליטת מידע מוטעה שהופץ באופן מכוון (Fake News, ידיעות כזב). עסקאות בשוק ההון המתבצעות בעקבות שימוש בנתונים הקשורים לידיעות כזב, עלולות להוביל לשיבוש המסחר ולתהליך גילוי מחיר שגוי. סיכון זה צוין על ידי יו"ר ה – SEC בהתייחסותו **לאירוע השורט סקוויז במניית גיימסטופ** בינואר 2021. הוא ציין כי "... זיהוי רגש תופס מומנטום בשנים האחרונות ומכסה כיום גם קהילות מקוונות. למגמה זו מתלווה הסיכון ששחקנים בעלי כוונות זדון ינסו לשלוח אותות בכוונה להטעות את השוק. זהו תחום בו נעמיק את הבנתנו, הקצאת המשאבים שלנו ואת יכולותינו".

  לדוגמא, חוקרים מאוניברסיטת MIT בחנו את ההשפעה של ידיעות כזב על שוק ההון ומצאו כי העלייה בנפח המסחר בתגובה לידיעות כזב גבוהה יותר מאשר בתגובה לידיעות אמת. בהתאם לכך, ככל שהשימוש במערכות מבוססות NLP בעולם ההשקעות נפוץ, עולה החשיבות בכך שמערכות אלו יוכלו להבחין במהימנותו של המידע.[44]
- **סוגיות אתיות**[45] מהוות את אחד האתגרים המרכזיים בפיתוח טכנולוגיה ובשימושה. נתאר להלן מספר סוגיות אתיות הנוגעות לשימוש בטכנולוגיות NLP על ידי גופים מסחריים. כיום, השימוש ב"בוטים דיגיטליים" הולך וגובר בתחומים חברתיים רבים כגון שירות לקוחות, המלצות על מוצרים, תמיכה בחינוך, שירותים רפואיים, בידור, הסברה חברתית וארגון אישי. כפי שהוזכר, השימוש בבוטים אלו בא לידי ביטוי גם בעולם הפיננסי. צ'אט בוטים משתמשים בנתונים רבים על מנת להסיק מסקנות ולהגיש המלצות מתאימות למשתמש. סוגיה אתית אחת בהיבט זה נובעת מכך שלא ניתן להבטיח כי ההמלצה תתאם באופן מדויק לצרכי הלקוח.

  יתר על כן, סוגיות הנוגעות לפרטיות המידע עולות אף הן. ממחקר של אוניברסיטת MIT עולה כי כבר היום גופים פיננסיים ובפרט בנקים, החלו לפתח מודלים לזיהוי ומשיכת לקוחות פוטנציאליים,

42 2021 Making the investment decision process more naturally intelligent, Deloitte Insights, Feb

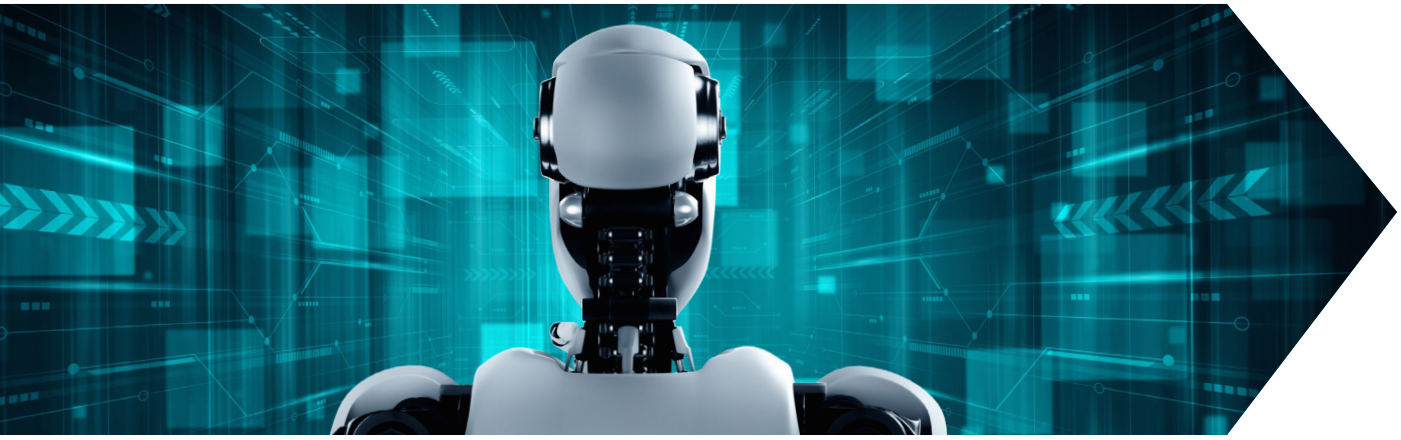
43 NBER Natioal ,14 How to Talk When a Machine is Listening: Corporate Disclosure in the Age of AI, Page 2020 ,Bureau of Economic Research

44 S. Kogan & T.J. Moskowitz & M.Niessner, MIT & Yale ,3 Fake News: Evidence from Financial Markets, Page 2018 ,schools of Management & AQR Capital Management

45 Ethical by Design: Ethics Best Practices, Jochen L.Leidner, Vassilis Plachouras, Thomson Reuters, Research 2017 ,& Development

ובפרט להציע מוצרים ללקוחותיהם בעזרת נתונים לא מובנים (כגון רשתות חברתיות, קשרים משפחתיים ומקום המגורים). כדוגמא לכך, נתונים על ילדיו של לקוח בבנק מאפשרים לבנק להציע מוצרים פיננסיים שונים (כגון הלוואות, תנאים חדשים ומוצרי חיסכון) על פי גיל ילדי הלקוח.[46]

- **אחריות לשגיאות** - התפקוד הפנימי של מערכות מבוססות NLP עשוי להעלות קשיים הנוגעים לשקיפות ואחריות על תוצאות המערכת. לא מן הנמנע כי מודל אשר נעזר בטכנולוגיית NLP יהיה מסוגל לנתח מאגרי מידע עליהם הוא אומן אך לא באופן חף משגיאות. במקרה של הזנת מידע שעלול להפיק תוצאות או המלצות לא תקינות למקרה הפרטני, עולה צורך בהתנהגות אחראית מצד המפתח בחשיפת מגבלות המערכת למשתמשים בה. העמקה באתגר זה בעולם הרגולציה, תובא בתת הפרק "אתגרים" בפרק ג'.



46 Tapping the power of unstructured data, MIT Sloan, 2021

פרק ג

יישומי עיבוד שפה טבעית (NLP) ברגולציית ניירות ערך

יישומי NLP מאפשרים לרגולטורים (באופן דומה לגופים עסקיים או ממשלתיים אחרים) להעלות תובנות ממסדי הטקסטים הזמינים להם כיום אך אינם ממוצים עקב נפחם הרב והתשומות הנדרשות לכך. בין השאר, כלים שנידונו בפרק א' כגון סיווג ואשכול טקסטים ושאיבת מידע מטקסט מאפשרים למקסם את יעילות השימוש במסמכים מורכבים.[47]

כפי שיתואר להלן, רשות ניירות ערך הניעה מספר פרויקטים בשנים האחרונות הכוללים יישומי טכנולוגיית NLP. ליבת ההתקדמות נעשית במסגרת **תכנית הפיילוטים (Data Sandbox)** המשותפת של רשות ניירות ערך ורשות החדשנות, שמיועדת לקדם את פעילותן של חברות פינטק בישראל.[48]

תכנית הפיילוטים מקדמת את היעד האסטרטגי שהציבה רשות ניירות ערך לקידום חדשנות פיננסית בשוק ההון הישראלי, הן כדי להטמיע טכנולוגיות פיננסיות שינגישו את שוק ההון לצרכן הישראלי והן כדי לייעל את אופן עבודת הרשות והגורמים המפוקחים. בתוך כך, התכנית מעניקה לחברות ולסטארט-אפים ישראליים אפשרות לפתח את המוצר שלהם למסחור ולחדירה לשוק, באמצעות תמיכת שתי הרשויות.

במסגרת התכנית, מוצגים בפני החברות והסטארט-אפים אתגרים, אשר עומדים על סדר יומו של שוק ההון ושל רשות ניירות ערך, במטרה לפתח מענה בדמות פיתוחים טכנולוגיים. אתגרים אלו, כוללים, בין היתר: עידוד הנפקה או רישום למסחר של חברות מקומיות וזרות, שיפור הציות של גופים מפוקחים, הנגשת מידע, הגברת הנזילות בבורסה, פיתוח כלים להנגשת הרגולציה הפיננסית לגופים המפוקחים, התגברות על השפה כמחסום להשקעה או הנפקה בבורסה, שילוב קהלים רחבים יותר בפעילות שוק ההון, הגברת התחרות בשוק ההון ועוד.

נתאר להלן מספר יישומי טכנולוגיית NLP ברגולציה הפיננסית, אשר מהווים רכיב אחד מתוך מגמת יישומי טכנולוגיה חדישה ברגולציה (Regtech).

ג.1. | מיכון ניתוח דיווחי מפוקחים

חלק משמעותי בעבודת הפיקוח של רגולטורים רבים (רשויות ניירות ערך בפרט) נוגע לניתוח דוחות טקסטואליים מטעם הגופים המפוקחים על ידם. בהתאם לכך, כלי ניתוח טקסט שתוארו בפרקים א' ו-ב' לעיל הם בעלי רלוונטיות גבוהה עבור רגולטורים. כך לדוגמא, היכולת למכן את ניתוח הדוחות התקופתיים והדיווחים המידיים היא בעלת חשיבות עבור רשויות ניירות ערך ועבור משקיעים כאשר האחרונים מסתייעים בה ברובם למטרות מקסום רווח בעוד הראשונות מסתייעות בה לצורך שמירת עניינם של המשקיעים.

47 Deloitte, Using AI to unleash the power of unstructured government data, 2019

48 רשות החדשנות ורשות ניירות ערך משיקות תכנית פיילוטים בתחום הפינטק בהשקעה של שישה מיליון שקלים

בנוסף, במסגרת שמירת עניינם של המשקיעים, רשויות ניירות ערך מבצעות פעולות פיקוח על דיווחיהן של חברות למשקיעים. בליבת פעולות הפיקוח, נמצאת העמידה על עיקרון **הגילוי הנאות** הניתן מטעם החברה שמהווה את "עקרון היסוד עליו מושתתים דיני ניירות ערך, ובכלל זה חוק ניירות הערך הישראלי..."[49]. הודות למנגנוני האוטומציה וההשוואה הסטטיסטית העומדים בבסיסם, כלי NLP מאפשרים לייעל ולטייב את הזיהוי של דיווחים שאינם תקינים ושגילוי המידע בהם אינו מיטבי או שאינו מקיים את דרישות החוק. כדוגמא, לצורך זיהוי חוסרים או הונאות בדיווחים פיננסיים של חברות, ה-SEC פיתח כלי הנקרא Corporate Issuer Risk Assessment (CIRA). CIRA הוא פלטפורמה בעלת לוח מחוונים של כ-200 מדדים המשמשים לאיתור דפוסים חריגים בדיווחי תאגידים. המערכת, שכוללת יישומי NLP ולמידת מכונה, מאפשרת ל -SEC לנתח ביעילות את הדיווחים הטקסטואליים ולזהות אילו מדווחים עלולים להיות מעורבים בדיווח לא תקין ובעבירה על החוק.[50]

מגמה מקבילה ותומכת ליישומי NLP בניתוח דיווחי מפוקחים הינה המעבר בשווקים רבים בעולם ל**דיווח כספי מתויג.** דיווח זה מבוצע על ידי סימון חד חד ערכי (תיוג) בפורמט סטנדרטי של נתונים חשבונאיים וטקסט המובאים בדיווח הכספי. כך לדוגמא, פסקאות בטקסט אשר עוסקות במדיניות חשבונאית מתויגות על ידי המדווח ככאלה. בהתאם לכך, הנגשת תיוגים אלו לכלי NLP מסייעת בשיפור האפקטיביות של הכלים. ב – 2018, ה - SEC תקננה תקנות שיחייבו תאגידים באופן הדרגתי לדווח את דוחותיהם הטקסטואליים בפורמט IXBRL[51] (שהינו פורמט התיוג המוביל בעולם) במטרה, בין השאר, להנגישם לכלים מבוססי NLP. בפרט , ציינו ברשות האמריקאית כי "... הצלחת הטכנולוגיה החדשה תלויה ביכולת המכונה לקרוא מידע שהינו רלוונטי לקבלת החלטות. ... לא נתונים מספריים בלבד, אלא כל סוגי המידע. בכלל זאת ... ניתוחים המובאים במילה הכתובה"[52].

רשות ניירות ערך מובילה אף היא מעבר לדיווח תאגידי בפורמט IXBRL[53] בשוק המקומי. דיווח זה ישרת בין השאר את יכולות מיכון הניתוח, תודות לתיוג נתונים כספיים וחלקי טקסט על פי פורמט מוסדר. כך לדוגמא, הביאורים, תיאור עסקי התאגיד, דוח הדירקטוריון וההצהרות (שהינם טקסטואליים) יסומנו בדיווח באופן מובנה המקל על כלים ממוחשבים לזהותם לצרכי ניתוח. בכך, יישום הדיווח בפורמט IX-BRL מהווה, בין השאר, הקלה משמעותית על השימוש בכלי NLP לניתוח דיווחי החברות.

בנוסף ליישומים בפיקוח על תאגידים מדווחים, ה-SEC מבצעת בקרה גם על דיווחי יועצי השקעות בשיטות מבוססות NLP.[54] בפרט, ברשות האמריקאית מצאו כי שימוש במידול נושאים (ראו פרק א'), זיהוי רגש המבוסס על מילון Loughran – McDonald (ראו פרק ב') ונתונים קודמים של חריגות מוכרות מאפשר לחזות בצורה טובה יותר סיכונים הנוגעים ליועצי השקעות ספציפיים. המנגנון שנבנה להערכת הסיכון בוחן את הנושאים הנדונים בדיווחיהם של יועצי ההשקעות ואת הרגשות המובעים כלפי נושאים אלו ומשווה את הממצאים אל מול דיווחים היסטוריים שבעקבות ניתוח אנושי הובילו לפעולות אכיפה כנגד יועץ ההשקעות. הממצאים מראים כי ניתוח אוטומטי זה בעל פוטנציאל רב לייעול תהליכי הסינון של דיווחים הנבחנים על ידי אנליסט אנושי.

---

49 פסק דין ע"א 90/5320 א.צ. ברנוביץ נכסים והשכרה בע"מ נ' רשות ניירות ערך.

50 Administrative ,23 Government by Algorithm: Artificial Intelligence in Federal Administrative Agencies, Page Conference of The United States, 2020

51 Inline XBRL, SEC

52 The Role of Machine Readability in an AI World, SEC, 3.5.18

53 יו"ר הרשות, ענת גואטה: "מעבר מדורג לדווח סטנדרט IXBRL יציב את שוק ההון הישראלי בשורה אחת עם שוקי ההון המתקדמים והמובילים באירופה ובארה«ב»

54 The Role of Big Data, Machine Learning, and AI in Assessing Risks: a Regulatory Perspective, SEC, 21.6.17

מחקר של חברת S&P משיק לנושא תת פרק זה ובוחן את השימוש ב"שפה גבוהה" בדיווחי חברות, בפרט במקרים בהם שימוש זה נועד להסתיר מידע המציב את החברה באור שלילי.[55] אכן, ניסוחים מפותלים ומונחי ז'רגון עשויים לחטוא למטרת הגילוי שבדיווחי החברות. בהתאם לכך הרגולציה האמריקאית מספקת כללים מנחים להיצמדות לשפה פשוטה ונהירה למשקיע הפרטי (Plain English Rule).[56]

המחקר בוחן האם ניתן לזהות חברות המנסות לעדן את הדיווח או לנסח אותו בצורה מסובכת. לצורך בחינת הסוגיה והערכת רמת הקריאות של הטקסט באנגלית, המחקר עשה שימוש במדד הבלשני Gunning fog שמטרתו לאמוד כמה שנות לימוד פורמאליות נדרשות לאדם על מנת להבין את הנאמר בטקסט. המדד מחושב על פי הנוסחה הבאה:

$$\text{Gunning fog index} = 0.4[\left(\frac{Words}{Sentences}\right) + 100\left(\frac{Complex\ Words}{Words}\right)]$$

כפי שניתן לראות מן הנוסחא, המדד כולל שני ביטויים (בסוגריים העגולות) אשר מבטאים היבטים שונים של סיבוכיות השפה בטקסט. הביטוי הראשון, $\frac{Words}{Sentences}$, מבטא את היחס שבין מספר המילים בטקסט לבין מספר המשפטים, כלומר את אורך המשפט הממוצע. ככל שהמשפט הממוצע ארוך יותר, המדד מייחס לטקסט רמת מורכבות גבוהה יותר. הביטוי השני, $\frac{Complex\ Words}{Words}$, מבטא את היחס שבין מספר המילים המסובכות בטקסט (שמוגדרות כבעלות לפחות שלוש הברות) לבין כלל המילים. לטקסט מיוחסת רמת מורכבות גבוהה יותר ככל שהוא מכיל מילים מסובכות בתדירות גבוהה יותר.

כאמור, לאחר שקלול הרכיבים, המדד מספק אומדן למספר שנות הלימוד הנדרשות להבנת הטקסט. ערך 16 במדד, לדוגמא, משמעו שעל אדם להיות בעל תואר ראשון על מנת להצליח להבין ולנתח את הטקסט. המחקר מציג ממצאים התומכים בטענה כי **דיווחים חיוביים נעשים בשפה הפשוטה להבנה בעוד דיווחים שליליים "מעורפלים" בשפה גבוהה.**

ג.2. | ניהול סיכונים ומגמות

כלי NLP מאפשרים לרגולטורים למכן תהליכי מעקב אחר סיכונים ומגמות בתחומים תחת פיקוחם. מחקר של רשות ניירות ערך האמריקאית (SEC) שבוצע בשנת 2009 ביקש להוכיח היתכנות של מתודה זו. המחקר בחן את התדירות בה הוזכרו חוזי החלף לסיכוני אשראי (Credit Default Swaps[57]) בתקופה שלפני המשבר הכלכלי בשנת 2008.[58] מהמחקר עלה, כי בשנים שלפני המשבר, הנושא עלה באופן תדיר יותר באתרי חדשות, בדיווחי חברות ובמאמרים, אך בשנת 2008 חלה עלייה חדה במספר המקרים בהם עלה הנושא בבתי השקעות ובבנקים. מכאן ניתן היה (לפחות בדיעבד) לזהות מגמה חשודה שעשויה להשפיע על ניהול הסיכונים של הרגולטור. הוכחת היתכנות זו הביאה להמשך מחקר בנושא יישומי NLP ברשות האמריקאית.

רשות ניירות ערך מזהה אף היא פוטנציאל בשימוש במתודות NLP לצרכי ניטור סיכונים ואבחון מגמות בזמן אמת. במסגרת זו, ובמסגרת מסלול הפיילוטים[59] (Data Sandbox) המשותף של רשות ניירות

55 Natural Language Processing, Part I: Primer, S&P Global

56 SEC, Plain writing initiative

57 Credit Default Swap, Wikipedia

58 21.6.17 ,The Role of Big Data, Machine Learning, and AI in Assessing Risks: a Regulatory Perspective, SEC

59 רשות החדשנות ורשות ניירות ערך בחרו שתי חברות פינטק למיזמי "ארגז חול" (Data Sandbox) במסגרת מסלול הפיילוטים.

ערך ורשות החדשנות, הרשות משמשת כאתר הרצה לפיילוט עם חברת BondIT המייצרת דירוג סינטטי לסדרות אג"ח ולתאגידים בהתבסס, בין השאר, על כלי NLP. דירוג זה יסייע לרשות לאבחן שינויים אפשריים בדירוגי חברות ואף שינויי מצב פיננסי בחברות מדורגות וחברות שאינן מדורגות.

נציין כי גם בקרב חברות דירוג האשראי מתבצע מחקר הנוגע ליישומי NLP לצרכי כימות הסיכון בתאגידים. Moody's Analytics, חברת בת של Moody's, פיתחה מודלי עיבוד שפה טבעית[60] ייעודיים למטרות הערכת סיכון פיננסי ומנגישה אותם כרכיבים אינטגרטיביים לתוכנות צד שלישי. בין השאר, המודלים מתבססים על זיהוי רגש וזיהוי ישויות בטקסט.

ג.3. | מיצוי תובנות מארכיונים ומרשת האינטרנט

ככלל, כלי NLP מאפשרים לגזור תובנות ממסדי טקסטים זמינים בנפחים גדולים. בהתאם לכך, הם מאפשרים למצות ארכיונים של מסמכי טקסט שכיום מאורכבים בלבד לצרכי בחינה ידנית עתידית. בנוסף, שילוב כלי NLP עם מתודות Web Scraping (סריקה אוטומטית של עמודי רשת) מאפשר חיפושים מורכבים שוטפים בתכנים הגלויים ברשת, אך המידע בהם אינו ממוצה.

צעד ראשון בכיוון זה נעשה בשנה האחרונה ברשות ניירות ערך במסגרת פיילוט נוסף בפרויקט ה- Data Sandbox עם חברת "זירה. קו בע"מ".[61] הפיילוט יתמקד באימות ואיתור של מידע אלטרנטיבי ואיכותי המיוחס לתאגידים מדווחים בישראל ברשת האינטרנט ובדיווחים ובחינה האם יש בו פוטנציאל לשיפור יכולות הפיקוח והאכיפה של הרשות. איתור המידע יתבצע על ידי סריקה אוטומטית של מקורות מידע פומביים וניתוח הטקסט בהם בכלים מבוססי טכנולוגיית NLP. טיוב תהליך איתור המידע צפוי להיות מתורגם לבסוף לגילוי אפקטיבי יותר בשוק ההון ולתמונת מצב מפורטת יותר בעבור משקיעים.

ג.4. | אתגרים

חשיבות מהימנות הכלים לעיבוד שפה טבעית, כפי שתוארו בסופו של פרק ב' לעיל עולה אף יותר עבור שימושים המבוצעים על ידי רגולטורים וגופים ממשלתיים.

- **חשיבות מכרעת לאמינות מלאה** - חובתם המנהלית של גופים אלו לאמינות ומקצועיות מלאה מהווה חסם מסוים להטמעה של טכנולוגיות חדשות שעל אף יעילותן הרבה עשויות להיות מלוות בתוצאות חיוביות ושליליות כוזבות. בהתאם לכך, השימוש בטכנולוגיה מבוצע לרוב ככלי תומך ומקדים לחוות דעתו של משתמש מומחה אנושי. דיון בנושא זה מובא גם בנאומו[62] של הכלכלן הראשי בפועל של ה – SEC משנת 2017 (תרגום חופשי): "... בעוד מכונות ימשיכו להחליף את המוח האנושי במשימות רבות, איני מאמין שאי פעם יהיה ניתן לוותר על שיקוליו בכל הקשור לרגולציה בשוק ההון".
- **שקיפות והסברה**- ככל שכלים מבוססים טכנולוגיות AI מתקדמים ונהפכים ליותר ויותר מתוחכמים, הם גם מהווים אתגר לשקיפות הנדרשת מרגולטורים בהחלטותיהם (חובת ההנמקה שבמשפט המנהלי). תוצאות אלגוריתמים שאינם נגזרים מכללים נוקשים, אינן ברות הסברה (Explainable) לעתים. חלק הארי של יישומי ה-NLP מבוסס על כלים סטטיסטיים ועל כן עשוי שלא להיות בר

60 Moody's ML Model Catalog

61 רשות החדשנות ורשות ניירות ערך בחרו חמש חברות פינטק למיזמי Data Sandbox במסגרת מסלול הפיילוטים

62 The Role of Big Data, Machine Learning, and AI in Assessing Risks: a Regulatory Perspective, SEC, 21.6.17

הסברה. הפרלמנט האירופאי, לדוגמא, קבע בשנת 2016 תקנה הקובעת כי תהליכים אוטומטיים המעורבים בהחלטות לגבי אדם נדרשים להיות ברי הסברה ("הזכות להסברה").[63] תקנה מסוג זה, עשויה בפועל לפסול את השימוש בכלי NLP מסוימים, לפחות בשלב התפתחותם הנוכחי. ההשפעה של אתגר זה על היישום בפועל של כלים משתנה כמובן לפי החקיקה והתקינה המקומית.

על מנת לספק מענה לאתגר ההסברה, מפותחים כלים ייעודיים שמטרתם איתור המשתנים הקריטיים בקבלת החלטה או בתוצאות מודל והצגתם בצורה ויזואלית מונגשת למשתמש. בנוסף, כפי שתואר בפרק הקודם, קיימת מגמה במסגרתה גופים שואפים לעשות שימוש ייעודי בטכנולוגיות NLG במטרה לספק הסברה למערכות מבוססות AI.

ככלל, ממשלות וארגונים בינלאומיים החלו לבחון את ההשפעה של כלים מבוססי בינה מלאכותית (בכללם כלי NLP) על ההיבטים החברתיים שבין הטכנולוגיה לאדם. למגמות הפיתוח המהירות בתחום הטכנולוגי עשויות להיות מצד אחד השפעות מסייעות בהיבט זה ומנגד השפעות מעכבות. כל זאת כנגזרת מכיוון ההתפתחות הטכנולוגית (כלים שקופים יותר או מעורפלים יותר) שאינו ניתן לחיזוי מלא בנקודה זו. בהתאם לכך, הצורך בפיקוח אתי מצד מפתחי הטכנולוגיה והמפקחים כאחד מהווה אתגר שריר שצפוי ללוות את שני הצדדים בעתיד הנראה לעין.

63 "European Union Regulations on Algorithmic Decision-Making and a "Right to Explanation

סיכום

נייר זה פתח חרך לפוטנציאל הטמון ביישומים בשוק ההון של טכנולוגיית עיבוד שפה טבעית (NLP) שמטרתה לקרוא ולהפיק משמעות משפה אנושית. יישומים אלו מאפשרים לגזור ולכמת מידע כלכלי מהותי ממאגרי טקסט גדולים (דיווחי תאגידים, אתרי חדשות, רשתות חברתיות וכו') בנפחים ובמהירות שאינה אפשרית בעבור הקורא האנושי.

השפה האנושית מעמידה מגוון אתגרים בפני ניתוחה על ידי מחשב: זיהוי תחבירי מדויק של חלקי משפט, סינתזה של תובנות מחלקי טקסט שונים, הכרעה בנוגע למשפטים הכוללים דו משמעויות וכו'. יישומי NLP עושים שימוש במגוון טכניקות חישוביות וסטטיסטיות בכדי לספק מענה לאתגרים אלו. פרק א' בנייר סיפק היכרות בסיסית עם המטלות הנדרשות מן הטכנולוגיה לצורך הפקת משמעות מטקסט, מהרמה הגרעינית ביותר (לדוגמא תיוג תחבירי של משפט) לרמה הכוללנית ביותר (לדוגמא זיהוי רגש המובע בטקסט).

כמו כן נסקרה בפרק הפעילות בתחום בישראל - בשנת 2020 הקימה רשות החדשנות במשותף עם משרד הדיגיטל, את איגוד החברות לטכנולוגיות שפת אנוש, שיסייע בקידום הבנת השפה העברית והשפה הערבית במערכות ממוחשבות. האיגוד הוקם במטרה לתת לתעשייה להוביל את הגדרות הצרכים ולסייע בסגירת פערים טכנולוגים שיאפשרו לעשות שימוש במאגרי מידע לא מובנים בעברית ולהפיק על בסיסם תובנות שישמשו מנוף למוצרים ושירותים לחברות ישראליות.

פרק ב' סקר יישומים והשפעות של הטכנולוגיה הנוגעים למשקיעים בשוק ההון. הטכנולוגיה מאפשרת לבצע אוטומציה של ניתוח פונדמנטלי ובכך לייעל אותו, להרחיב את מנעד החברות והסקטורים אותם ניתן לסקור בזמן נתון ולהנגיש את התובנות הטמונות בו גם לכלי מסחר מבוססי ניתוחים כמותניים. בנוסף, הטכנולוגיה מאפשרת לזהות רגש המובע בזמן אמת בהודעות מיידיות של תאגידים, מידע חדשותי ופרסומים ברשתות חברתיות. מחד גיסא, בכך מתאפשרת "כריית דעות" יעילה על פני השוק והטמעת מידע יעילה יותר בתמחור. מאידך גיסא, נדון בפרק גם הסיכון בהטמעה מהירה יותר של ידיעות כזב (Fake News).

בהיבט השירותים המיועדים למשקיעים, גופים המספקים שירותי השקעה ופיננסיים מסתייעים בטכנולוגיה לצרכי שיפור השירות הניתן ללקוחות באמצעות תוכנות ייעודיות לקיום שיחה טקסטואלית עם אדם (צ'ט-בוטים). עם התפתחות הטכנולוגיה שבבסיסם, צ'ט-בוטים אלו עוברים מהיותם כלי טכני למתן משוב אוטומטי אל כלי הוליסטי שמשכלל את מאפייני הלקוח הספציפי ואת אסטרטגיות השירות בהמלצותיו.

פרק ג' בחן יישומים טכנולוגיים המשרתים רגולטורים (Suptech) ומבוססים על עיבוד שפה טבעית. הפרק כולל דוגמאות מן העולם ומהשוק המקומי, בינן פרויקטים המקודמים בשנים האחרונות במסגרת תכנית הפיילוטים (Data Sandbox) המשותפת של רשות ניירות ערך ורשות החדשנות, שמיועדת לקדם את פעילותן של חברות פינטק בישראל.

כלי NLP עשויים לייעל ולטייב את הזיהוי של דיווחי מפוקחים שאינם תקינים ושגילוי המידע בהם אינו מיטבי או שאינו מקיים את דרישות החוק. בכך הם משרתים את ליבת העיסוק הפיקוחי של רגולטורים פיננסיים. המעבר בשנים האחרונות לדיווח כספי מתויג (IXBRL) בשוק הגלובלי והמקומי, הינו גורם תומך לשימוש בכלי NLP, הודות לסטנדרטיזציה הטמונה בו, המסייעת לניתוח הדוחות על ידי מחשב.

כמו כן, הניתוח האוטומטי של פרסומים תאגידיים וחדשותיים טקסטואליים מאפשר לגופים האמונים על זיהוי סיכונים (רגולטורים פיננסיים, חברות דירוג אשראי וכו') לטייב את תהליכי זיהוי הסיכונים שלהם, זאת על בסיס מידע זמין שעקב היקפיו אינו בר מיצוי בכלים ידניים. עם זאת, המאמר שם גם דגש על חשיבות שימור הגורם האנושי בשרשרת ההחלטה, שכן על אף הייעול הטמון בטכנולוגיה, היא אינה חפה משגיאות אפשריות.

השימוש בכלי NLP לניתוח טקסטים פיננסים עומד בנקודת מפגש של שתי מגמות טכנולוגיות מובילות בעשור האחרון - מגמה רוחבית, שהינה הגידול בתפוצת השימוש בכלים מבוססי אינטליגנציה מלאכותית ומגמה אורכית, שהינה הגידול בפיתוח יישומי פינטק. חיזוי מלא של עתיד והשלכות מגמות אלו כמובן אינו אפשרי, אך בהסתכלות ממעוף הציפור הן צפויות להיתרגם בסופו של דבר לייעול תהליכי הטמעת המידע ולשוק הון יעיל יותר.

המאמר הבא בסדרה

כפי שפורט במבוא למאמר, נייר זה הינו הראשון **בסדרת מאמרים** הבוחנת את ההשלכות הפוטנציאליות על שווקי ההון של פיתוח טכנולוגיות חדשות. בדומה למאמר זה, הבחינה במאמרים הבאים תתבצע אף היא דרך הפריזמה הדואלית שבין המשקיעים לרגולטורים.

הגידול בשימוש בכלים מבוססי NLP מהווה, במידה רבה, ביטוי של התפתחות היכולות בעולם **למידת המכונה ככל שהן מיושמות על מאגרי נתונים טקסטואליים. המאמר הבא בסדרה יתמקד בהיבטים הכמותניים של מתודות למידת מכונה**, בדגש על יישומן על נתוני עתק (Big Data) אשר מהווים במקביל אתגר והזדמנות.

בפריזמת המשקיעים, ייסקרו המתודות הנוגעות למסחר האלגוריתמי אשר מתבססות על זרם נתונים ברזולוציות זמן גבוהות (ננו-שנייה בשווקים המובילים), בממדים מרובים (סקטורים, שווקים ואפיקים שונים) ובהיקפי נתונים רחבים. בנוסף ייבחנו יכולות החיזוי (בטווחי זמן ארוכים) והעלאת התובנות של כלי למידת מכונה ביישומם על נתונים פיננסיים ונתוני מקרו.

בפריזמת הרגולטור ייסקרו, בין השאר, השימוש בכלים מבוססי למידת מכונה לזיהוי אנומליות וחריגים בנתוני המסחר, יישומי למידת מכונה המייעלים תהליכי זיהוי וניהול סיכונים וכלים תומכי החלטה מבוססי למידת מכונה.

מונחון

מונח (עברית)	מונח (אנגלית)	קיצור	הגדרה
אשכול טקסטים	Text clustering		מציאת קשרים בין טקסטים שונים שתוכנם לא ידוע מראש
בינה מלאכותית	Artificial Intelligence	AI	מערכת המסוגלת לפתור בצורה רציונאלית בעיות מורכבות או לנקוט בפעולות כדי להשיג את מטרותיה בנסיבות שונות בהן היא נתקלת בעולם האמיתי
גרף הידע האישי	Personal Knowledge Graph		מקור לידע מובנה אודות ישויות והקשר ביניהן
דיווח כספי מתויג	Tagged Disclosure		סימון חד חד ערכי (תיוג) בפורמט סטנדרטי של נתונים חשבונאיים וטקסט המובאים בדיווח הכספי, מבוצע בין השאר בפורמט IXBRL
זיהוי יחסים	Relationship Extraction		כלי הנועד לחילוץ יחסים בין שתי ישויות או יותר כפי שהם מובאים בטקסט
זיהוי ישויות בשם	Entity Extraction		תיוג באופן חד חד ערכי של ישויות כמו שמות של אנשים, שמות של חברות ומיקומים גיאוגרפיים
זיהוי רגש	Sentiment Analysis		כלי המיועד לסיווג טקסט לפי הרגש או הדעה המובעים בו ("חיובי", "שלילי", "נייטרלי" וכדומה) ולפי עוצמתם
ייצור שפה טבעיות	Natural Language Generation	NLG	תהליך אוטומטי ההופך נתונים מובנים לשפה טבעית
יישוב הפניות משותפות	Coreference Resolution		זיהוי הפניות שונות כמשויכות לאותה ישות ("המניה עלתה ב5%. היא הגיעה לערך שיא שנתי")
כריית דעות	Opinion mining		אגרגציה של דעות המובעות על ידי מספר רב של סוכנים, לרוב תוך התבססות על כלי זיהוי רגש בפרסומים במדיות חברתיות
לֶמָה	Lemma		צורת הבסיס המילונית (בהסרת הטיית זמן וכו')
למידה בלתי מונחית	Unsupervised Learning		מחלקת שיטות בלמידת מכונה המסיקה קשרים ודפוסים מתוך סט של קלטים אפשריים
למידה מונחית	Supervised Learning		מחלקת שיטות בלמידת מכונה המייצרות מודל להגדרת פלט מסוים עבור קלט מסוים, תוך התבססות על סט נתון של קלטים אפשריים ופלטים רצויים עבור קלטים אלו

מונחון (המשך)

מונח (עברית)	מונח (אנגלית)	קיצור	הגדרה
למידת מכונה	Machine Learning	ML	תחום חישובי שמטרתו לאפשר למחשב ללמוד לבצע פעולות באופן עצמאי תוך התבססות על מסדי נתונים וכלים סטטיסטיים
מידול נושאים מבוסס ניתוח סטטיסטי	Topic Modeling		כלי המיועד לזיהוי נושאים הנדונים במסמכים תוך התבססות על שכיחות הופעת מונחים מסוימים בטקסט
מידע מובנה	Structured Data		נתונים בעלי פורמט מוגדר מראש, לדוגמא מידע טבלאי
מידע שאינו מובנה	Unstructured Data		נתונים ללא פורמט מוגדר מראש, לדוגמא טקסט
מסחר כמותני	Quantitative Trade		גישת מסחר המבוססת על נתוני מסחר בזמן אמת ומערבת, בנוסף להערכת התמחור, גם היבטים של מומנטום, נפחי מסחר וכו'
מסחר כמותני – פונדמנטלי	Quantamental		גישת מסחר המערבת את היבטים פונדמנטליים וכמותניים, לרוב באמצעות פלטפורמה ממוחשבת אחודה
ניתוח פונדמנטלי	Fundamental Analysis		ניתוח השואף לאמוד את שוויה של חברה ולסחור במניותיה אל מול מחירן בשוק
סופטק	Supervision Technology	Suptech	תת ענף בתחום הפינטק המתמקד בכלים טכנולוגיים יעודיים לרגולטורים
סיווג טקסטים	Text Categorization		זיהוי מתומצת של נושא הטקסט על בסיס קטגוריות נושאיות מוגדרות מראש
עיבוד שפה טבעית	Natural Language Processing	NLP	טכנולוגיה שמטרתה הפקת משמעויות מטקסט באופן ממוחשב
פינטק	Finance Technology	Fintech	יישומי טכנולוגיים בתחום הפיננסיים
קורפוס	Corpus		מסד נתונים לייחוס עבור יישומי NLP שמאגד דוגמאות סטנדרטיות ומרובות של משפטים מתויגים תחבירית
רגטק	Regulation Technology	Regtech	תת ענף בתחום הפינטק המתמקד בכלים טכנולוגיים יעודיים לגופים מפוקחים למטרות ציות
שאיבת מידע מטקסט	Information Extraction		מטלה המאגדת מגוון מתודות שמטרתן לחלץ מידע מובנה (לאמור - נתונים וההקשר שלהם) מתוך טקסט
תַּמְנִית	Token		אבן הבסיס עליה מבוצעים עיבודים בכלי NLP (לרוב מילה בודדת)
תיוג תחבירי	Part of Speech Tagging		שיוך התפקיד התחבירי במשפט (נושא, נשוא וכו') לכל תמנית

ביבליוגרפיה

AI at Bloomberg, Bloomberg

Artificial Intelligence (AI) in the Securities Industry, FINRA, יוני 2020

Capital Markets Natural Language Processing - Part II: Stock Selection, Frank Zhao,2018

Credit Default Swap, Wikipedia

Deloitte, Using AI to unleash the power of unstructured government data, 2019

Different ways of doing Relation Extraction from text, Andreas Herman, 2019

Economic Causal-Chain Search Using Text Mining Technology, Kiyoshi Izumi and Hiroki Sakaji, 2020.

Environmental, social and governance reporting in annual reports: A textual analysis, Philipp Baier, Marc Berninger, Florian Kiesel, 2020

Ethical by Design: Ethics Best Practices, Jochen L.Leidner, Vassilis Plachouras, Thomson Reuters, Research & Development, 2017

European Union Regulations on Algorithmic Decision-Making and a "Right to Explanation"

Fake News: Evidence from Financial Markets, Page 3, S. Kogan & T.J. Moskowitz & M.Niessner, MIT &Yale schools of Management & AQR Capital Management, 2018

Form-10 K, SEC website

Google Trends

Government by Algorithm: Artificial Intelligence in Federal Administrative Agencies, Page 23, Administrative Conference of The United States, 2020

How to Talk When a Machine is Listening: Corporate Disclosure in the Age of AI, Page 14, NBER Natioal Bureau of Economic Research, 2020

Inline XBRL, SEC

Making the investment decision process more naturally intelligent, Deloitte Insights, Feb 2021

Managing real-time risks and returns: The Thomson Reuters NewsScope Event Indices, Alexander D.Healy & Andrew W.Lo

Moody's ML Model Catalog

Natural Language Processing – Part III: Feature Engineering, Frank Zhao, S&P Global, 2020

Natural Language Processing, Part I: Primer, S&P Global

NLP in FinTech Applications: Past, Present and Future, Chung-Chi Chen, Hen-Hsen Huang, Hsin-Hsi Chen, 2019

NLP in the Stock Market, Roshan Adusumilli, 2020

Numeral Understanding in Financial Tweets for Fine-grained Crowd-based Forecasting, Chung-Chi Chen, Hen-Hsen Huang, Yow-Ting Shiue, Hsin-Hsi Chen, 2018.

Personal Knowledge Graphs: A Research Agenda, Krisztian Balog & Tom Kenter, 2019

Quantamental: What It Is & Why It Works, Leo Smigel, 2020

SEC, Plain writing initiative

Start-Up Nation Central

Stock Price Forecasting by Combining News Mining and Time Series Analysis , Xiangyu Tang, Chunyu Yang, Jie Zhou, 2009

Stock Twits

Tapping the power of unstructured data, MIT Sloan, 2021

Testimony Before the House Committee on Financial Services, Chair Gary Gensler, 6.5.21.

The basics of NLP and real time sentiment analysis with open source tools. Özgür Genç, 2019

The Role of Big Data, Machine Learning, and AI in Assessing Risks: a Regulatory Perspective, SEC, 21.6.17

The Role of Machine Readability in an AI World, SEC, 3.5.18

University of Notre Dame, Software repository for accounting and finance

When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks, Tim Loughran and Bill Mcdonald, Page 49, 2011

אתר פורום תל"ם

המרכז למחקר מדע הנתונים, הבינתחומי הרצליה

ועדת בינה מלאכותית ומדע הנתונים, תל"ם, דצמבר 2020

יו"ר הרשות, ענת גואטה: "מעבר מדורג לדווח סטנדרט IXBRL יציב את שוק ההון הישראלי בשורה אחת עם שוקי ההון המתקדמים והמובילים באירופה ובארה"ב"

מיליארד שקל למחשב-על: כך תנסה ישראל לעלות על מפת הבינה המלאכותית, דצמבר 22.12.2020, TheMarker

משרד הדיגיטל הלאומי, הקמת איגוד חברות לטכנולוגיות שפת אנוש (NLP) בעברית ובערבית, 22.09.2020

פסק דין ע"א 5320/90 א.צ. ברנוביץ נכסים והשכרה בע"מ נ' רשות ניירות ערך.

קבוצות מחקר, עיבוד שפה טבעית, המחלקה למדעי המחשב באוניברסיטת בר אילן

רועי גולדשמידט, דו"ח בנושא "בינה מלאכותית", הכנסת - מרכז המידע והמחקר, 2018

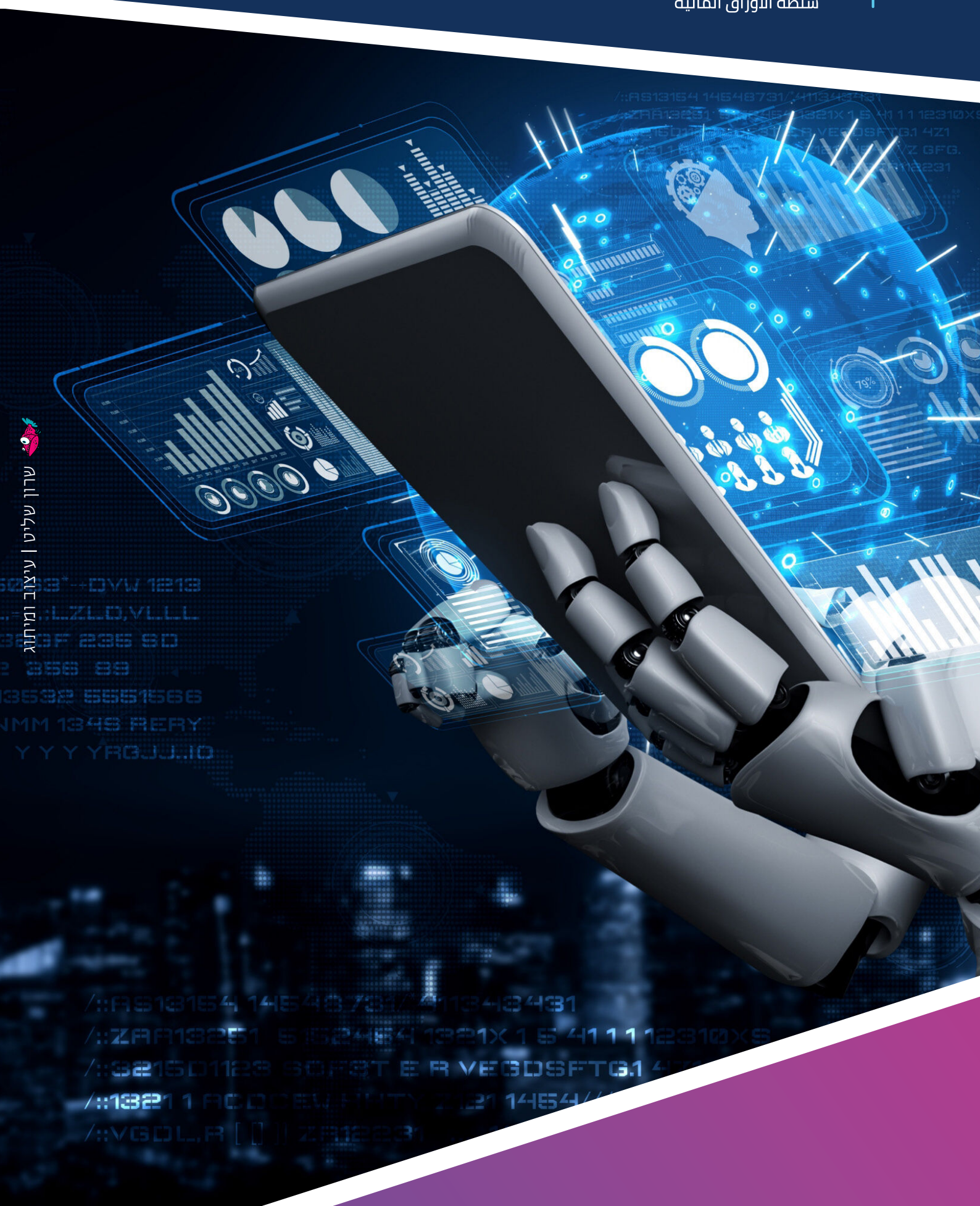
רשות החדשנות ורשות ניירות ערך בחרו חמש חברות פינטק למיזמי Data Sandbox במסגרת מסלול הפיילוטים

רשות החדשנות ורשות ניירות ערך בחרו שתי חברות פינטק למיזמי "ארגז חול" (Data Sandbox) במסגרת מסלול הפיילוטים.

רשות החדשנות ורשות ניירות ערך משיקות תכנית פיילוטים בתחום הפינטק בהשקעה של שישה מיליון שקלים

רשות החדשנות
Israel Innovation Authority

רשות ניירות ערך
Israel Securities Authority
سلطة الأوراق المالية



שרון שליט | עיצוב ומיתוג