

## השוואת אסטרטגיות רגולטוריות בפיקוח בנקאי: ניתוח טקסט במסמכי מדיניות



הוגש במסגרת קורס כריית טקסט, בית הספר למנהל עסקים –  
אוניברסיטת בר אילן

מגיש: אריאל מתן

מרצה: ד"ר רוני רמון-גובן

## **הגדרת הבעיה ומטרות**

רגולטורים בנקאיים ברחבי העולם מפרסמים מסמיכי מדיניות אסטרטגית שמגדירים את תחומי הפעולה המרכזיים שלהם ואת סדר העדיפויות הפיקוחי על הבנקים. מסמכים אלה הם מקור ידע חשוב אך בשל אורכם, ניסוחם הטכני והבדלי תחומי אחריות קשה לנתחם באופן שיטתי וערוך השוואה מסודרת בניהם.

בעידן של רגולציה שצריכה להיות דינמית ומערכת פיננסית משתנה בקצב מהיר, עולה הצורך של רגולטורים ומסודות בנקאיים ללמוד מהנעשה במדינות אחרות. להבין מהם הסיכונים שזוכים לסיקור והתייחסות רחבה, מהם הנושאים להם ניתנים משאבים גדולים יותר, וכיצד משתנה מיקוד הפיקוח בין רגולטורים שונים. ניתוח כמו בעבודה זו יכול לסייע בגיבוש מדיניות מקומית רלוונטית תוך זיהוי מגמות בינלאומיות והפקת לקחים מניסיונם של אחרים.

מטרת עבודה זו היא יישום טכניקות של כריית טקסט, על מסמכי מדיניות של רגולטורים במדינות מתקדמות בעולם. השיטה תאפשר לזהות מיקוד ונושאים מרכזיים בתוך מסמכי האסטרטגיה, להשוות בין רגולטורים, ולגלות הבדלים בין הגישות האסטרטגיות אותן הם מאמצים.

### **שאלות המחקר:**

אילו נושאים חוזרים ועולים במסמכי האסטרטגיה של רגולטורים שונים בעולם, וכיצד משתנה דגש האסטרטגיה בין מדינות?

האם ניתן לזהות מגמות משותפות או ייחודיות בהתמודדות עם סיכונים ויעדים רגולטוריים?

## סקירת ספרות

*Exploring the Space of Topic Coherence Measures (Röder, Both & Hinneburg, 2015)*

המאמר בודק מספר דרכים לבדוק האם נושא שנוצר על ידי Topic Modeling טוב להבנה ולקריאה אנושית. הוא מציע מדדים שעובדים טוב יותר, ומראים איך ניתן לבחור את מספר הנושאים (K) בצורה טובה יותר. המאמר עוזר לבחור ולהסביר את ההחלטה על K מסוים ולהראות שהנושאים יצאו קוהרנטיים והגיוניים.

*LDavis: A method for visualizing and interpreting topics (Sievert & Shirley, 2014)*

המאמר מדבר על כלי שנלמד בכיתה ונעשה בו שימוש בעבודה זו, הכלי מציג את הנושאים כבועות על מפה ומוסיף לכל נושא את המילים את המילים הכי חשובות בו. ישנה גם אופציה לאזן בין מילים שכיחות למילים ייחודיות. בצורה זו קל להבין על מה כל נושא מדבר ולתת לו את השם המתאים.

*Shocking Language: Understanding the Macroeconomic Effects of Central Bank Communication (Hansen & McMahon, 2016)*

המאמר בוחן את ההשפעה העצמאית של תוכן החלטת הריבית של בנק המרכזי בארה"ב (Federal Reserve) בנפרד מהחלטת הריבית עצמה. החוקרים מנסים להפריד בין ההחלטה עצמה – העלאת, השארת, הורדת הריבית ביחס לצפיות השוק לבין "זעזוע טקסטואלי" שהוא המסר וניסוח ההודעה עצמה, שיכולה לרמז על החלטות עתידיות. החוקרים מראים באמצעות מודלים כמותיים שגם כאשר הריבית משתנה או אינה משתנה כצפוי, המילים עצמן יכולות לגרום לתגובות משמעותיות בשווקים ולהשפיע על הכלכלה. המסקנה היא שלטקסטים רגולטוריים בפני עצמם ישנה השפעה על השוק, לכן יש ערך בניתוח שלהם כגורם נפרד.

*Transparency and Deliberation within the FOMC (Hansen, McMahon & Prat, 2018)*

בדיקה של תמלילי הוועדה הפדרלית לשוק פתוח בארה"ב מראה שככל שהשקיפות גדלה (למשל פרסום וידאו או תמלול של הדיונים), החברים נוטים לדבר בצורה דומה יותר ופחות להביע דעות שונות. כשמשווים מסמכי אסטרטגיה בין רגולטורים/מדינות, צריך לזכור שכללי שקיפות יכולים לשנות את הסגנון והתוכן—וזה יכול להסביר דמיון או שוני בנושאים. עם זאת, מאחר שאין לנו מידע על תמלולים או אינטרסים נוספים של המשתתפים, אין לממצא זה יישום משמעותי בעבודה הנוכחית.

## **נתונים**

### **הנתונים הרצויים**

לצורך המענה על שאלת המחקר, האידאל היה לאסוף מאגר של מסמכי אסטרטגיה רגולטורית עדכניים של רגולטורים מובילים בתחום הפיקוח על הבנקים ובעלי מבנה קרוב יחסית למבנה הרגולציה בישראל. המאגר הרצוי צריך לכלול מסמכים באנגלית בלבד אשר בהם תיאור מפורט של יעדי הרגולטור, תחומי הסיכון שבהם הוא מתמקד, והצעדים המתוכננים לטווח הקרוב והבינוני.

בנוסף, רצוי שיהיה מטא-דאטה על כל מסמך הכולל את שם הרגולטור, המדינה אליה הוא שייך, שנת פרסום לכל מסמך.

### **הנתונים בפועל**

הנתונים שנאספו בפועל מורכבים ממסמכי אסטרטגיה רגולטורית שהתקבלו מאתרי אינטרנט רשמיים של רגולטורים בנקאיים שונים, ביניהם:

- בריטניה – Prudential Regulation Authority (PRA)
- האיחוד האירופי – European Central Bank (ECB)
- ארה"ב – Office of the Comptroller of the Currency (OCC)
- ארה"ב – Federal Deposit Insurance Corporation (FDIC)
- קנדה – Office of the Superintendent of Financial Institutions (OSFI)

### **פערים בין הנתונים הרצויים לנאספים**

- כלל הרגולטורים מתעסקים בפיקוח על הבנקים. עם זאת, עקב מבנה רגולטורי שונה בין מדינה למדינה תחומי האחריות של הרגולטורים אינם זהים לחלוטין. ישנם רגולטורים ש-"משלימים זה את זה" וישנם רגולטורים בעלי אחריות מלאה לפיקוח על הבנקים בנוסף לפיקוח על גופים פיננסיים אחרים.
- טווח השנים אינו אחיד – עבור חלק מהרגולטורים קיימים מסמכים עדכניים מאוד, בעוד שעבור אחרים זמינים רק מסמכים בני מספר שנים. החדש ביותר- 2025, הישן ביותר- 2022.
- חלק מהמסמכים כוללים טבלאות ותרשימים שהומרו לטקסט פשוט.
- ישנם הבדלים באורך ובפירוט בין המסמכים, מה שעלול להשפיע על משקלם בחישובים שונים לאורך הניתוח.

### **המסד עצמו:**

- שורה = חלק ממסמך אסטרטגיה- נושא, מטרה וכו'
- עמודות:
- Global ID – מזהה ייחודי לכל מסמך.

- Country – שם המדינה.
- Regulator – שם הרגולטור.
- Year – שנת פרסום המסמך.
- Title – כותרת לטקסט.
- Text – גוף הטקסט הגולמי.
- Doc\_ID – מזהה מסמך לפי רגולטור.

## רלוונטיות לבעיה

המסמכים שנאספו מהווים תיעוד רשמי של סדרי העדיפויות, ההערכות והאתגרים המרכזיים כפי שמגדירים אותם רגולטורים בנקאיים. ניתוח טקסטואלי של מסמכים אלו באמצעות Topic Modeling מאפשר לזהות מגמות משותפות והבדלים בין רגולטורים, ולספק תובנות מבוססות נתונים לגבי התמקדות האסטרטגית של הפיקוח על הבנקים בעולם.

## ארגון וניקוי טקסט

הארגון והניקוי נעשה בסדר הבא:

1. הסרת קישורים וכתובות דוא"ל – אינם מידע רלוונטי לתוכן אותו חיפשתי בעבודה.
2. הסרת סימנים שאינם אותיות – אינם מסייעים לניתוח ואף מוסיפים רעש, הסרת רווחים תקל על בניית DTM טוב.
3. הפיכת עמודת טקסט לקורפוס – על להמשיך בתהליך הניקוי על ידי tm\_map בצורה טובה.
4. הקטנת אותיות – למניעת כפילויות של מילים וביצוע ספירה מדויקת יותר
5. Lemmatization – לאחר התלבטות בין אופציה זו ל- stemming, העדפתי לנסות לשמור על המשמעות ועל הדיוק. בנוסף, הבחירה בלמטיזציה תגרום להצגת תוצאות מובנות ומסודרות יותר.
6. הסרת מילות עצירה – הוסרו מילות ברירת המחדל בפונקציה "stopwords". הוסרו מילים נפוצות במסמכים רגילים וספציפית במסמכים רגולטוריים/אסטרטגיים כמו: "ensure", "consider" ועוד. בנוסף הוסרו שמותם של הרגולטורים שציינו הרבה את שמם במסמך האסטרטגיה.
7. הסרת סימני פיסוק – סימני הפיסוק אינם מוסיפים הקשר ואף יפריעו ליצירת TDM/DTM.
8. הסרת ספרות – על נמת לא ליצור הטיית תדירות או לשבש את תהליך ה- Topic Modeling. אומנם בחלק מהמסמכים הופיעו מספרים שמטרתם להמחיש מגמות אך רציתי להתמקד בתוכן לשוני.
9. רווחים נוספים – מיותרים וגורמים לרעש.

## משתנים (פיצ'רים)

### ייצוג טקסט – Bag of Words

הטקסט יוצג כ- BoW של Unigrams ו-Bigrams לאחר ניקוי ולמטיזציה. לכל מסמך נספרו הופעות מונחים ונבנה DTM. למודל LDA נעשה שימוש בספירה גולמית של Unigrams.

### סינון אוצר מילים

הוסרו Stopwords כלליות, מילים בהם יש שימוש גדול במסמכים רגולציה ואסטרטגיה, וכן מספרים ופיסוק. בוצעה למטיזציה לאיחוד נטיות לשוניות. הסינון נועד לצמצם רעש ולהדגיש מונחים נושאים המשקפים תכנים רגולטוריים.

### פיצ'רים מבוססי נושא

לאחר התאמת מודל LDA עם 7 נושאים, לכל מסמך הופקה התפלגות נושאים: שבעה מספרים בין 0 ל-1 שסכומם 1 המשקפים את חלקו היחסי של כל נושא במסמך. בוקטורי ההתפלגות הללו השתמשתי כפיצ'רים להמשך: חישבתי ממוצעים לפי מדינה כדי ליצור פרופיל נושאים לכל אחת, והצגתי אותם במפת חום להשוואה ויזואלית בין הקבוצות.

### פיצ'רים מבוססי FinBERT

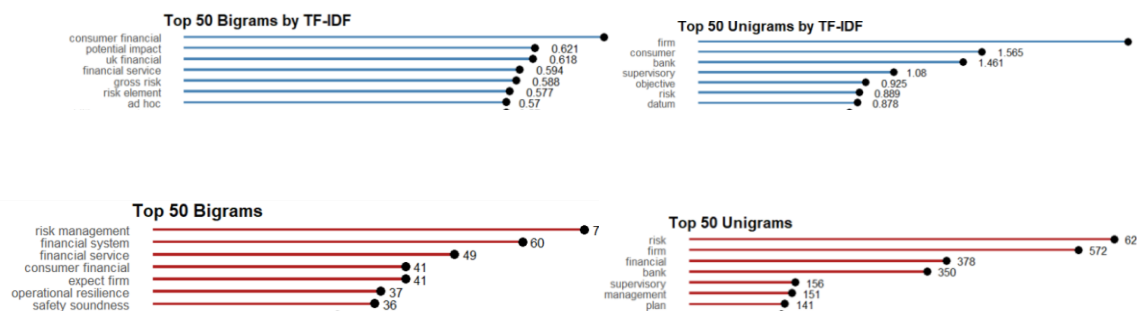
יצרתי Embeddings לכל מסמך באמצעות FinBERT. על בסיס מטריצת דמיון (קוסינוס) בין המדינות חושבו: ממוצע דמיון של כך מדינה לשאר המדינות, הזוג בעל הדמיון הגבוה ביותר, טווח גבהים בדנדוגרמה כמדד פיזור כללי, שיוך לאשכולות על בסיס הדנדוגרמה. בנוסף, הוצגו לאחר מכן דוגמאות של Unigrams + Bigrams אופייניים לכל בלוק לצורך הפרשנות שלי לתוצאות.

### מטה-משתנים לחיתוך

מדינה שימשה כמשתנה חתך להצגת התפלגויות אך לא הוזנו למודל כפיצ'רים.

## שיטות הניתוח ותוצאות

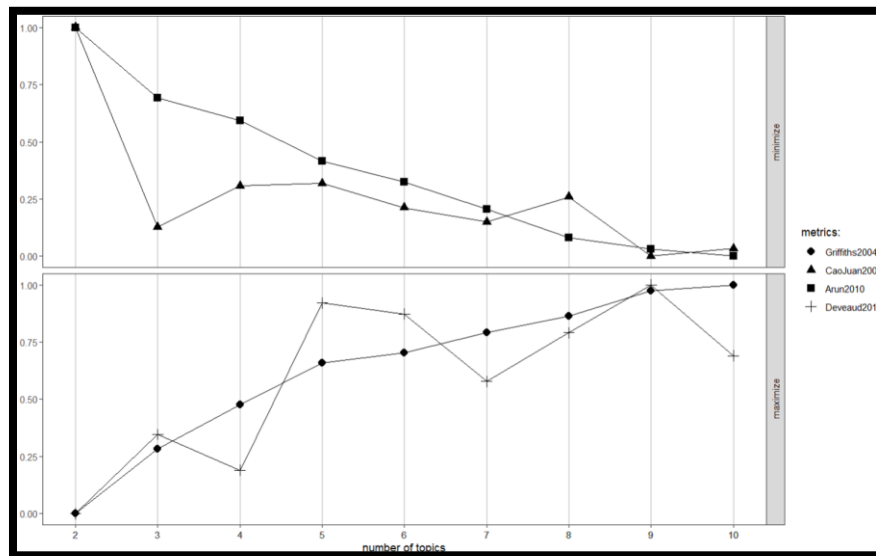
נבנו מטריצות DTM + TDM ביוניגרמים וביגרמים בשתי שיטות – BoW ו-TF-IDF והשוותה התפלגות המונחים בין השיטות. מצורפות טבלאות חלקיות-Top 7, הטבלאות המלאות מצורפות בנספחים.



לאחר שראיתי כי ישנם מילים בעלות או השורש שחוזרות על עצמן שוב ושוב הוחלט לבצע למטיזציה ולא סטמינג על מנת להבין ולהציג בצורה טובה יותר את המילים, מה שגרם לשינוי דרמטית במשקלי המילים, מילים חדשות שנכנסו לטבלת ה- Top 50 ולשינוי משמעותי בהשוואת מספר הנושאים בLDA. התוצאות המצגות למעלה הן לאחר למטיזציה.

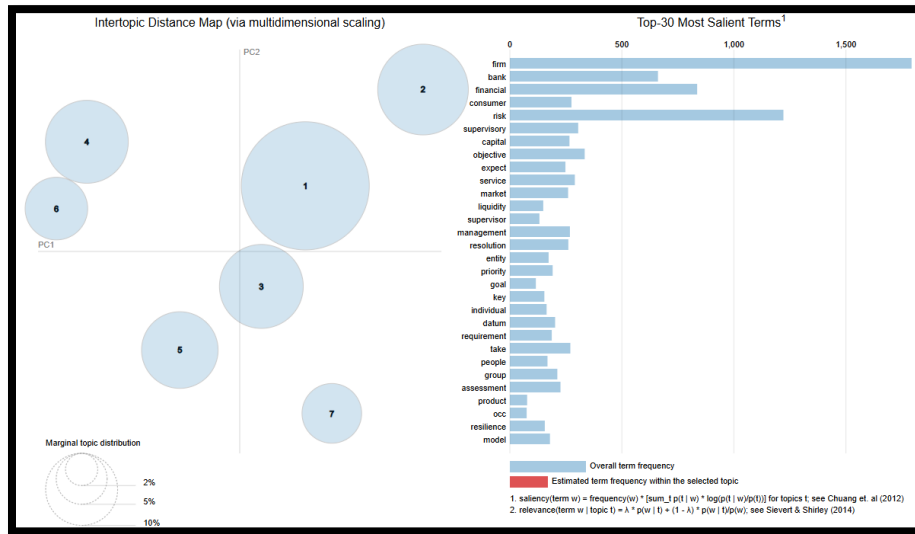
טבלאות Top 50 לפני ואחרי למטיזציה נמצאות בנספחים.

### השוואת מספר נושאים



נבחר  $K=7$  לאחר הרצה על טווח של שניים עד עשרה נושאים. בשבע נושאים ישנה התכנסות טובה בין המדדים למעט Deveaud2014. בחירה בתשעה נושאים יכולה לשפר עוד קצת את המדדים אך  $K=7$  נבחר על מנת להוציא מודל פשוט וקל יותר לפרשנות.

## מפת הנושאים (LDAvis) K=7



התרשים מציג מפה הממחישה את מרחק הנושאים על מפה דו ממדית על ידי הפחתת ממדים. הנושא בגדול ביותר הוא "ניהול סיכונים ותכנון כשלים". על פי תרשים נראה שיש בידול טוב בין כלל הנושאים למעט נושאים 4 ו-6, "פיקוח ואסטרטגיה" אל מול "רגולציה ומדיניות פרודנציאלית" שנראים קרובים אבל אינם חופפים- מה שמצדיק שמות נפרדים. התמונה מראה נושאים קוהרנטיים ללא כפליות, ומחזקת את הבחירה בשבעה נושאים. הוספת נושאים יכולה ליצור מורכבות והפחתת נושאים תיצור תמונה פחות ברורה על המתרחש בפועל.

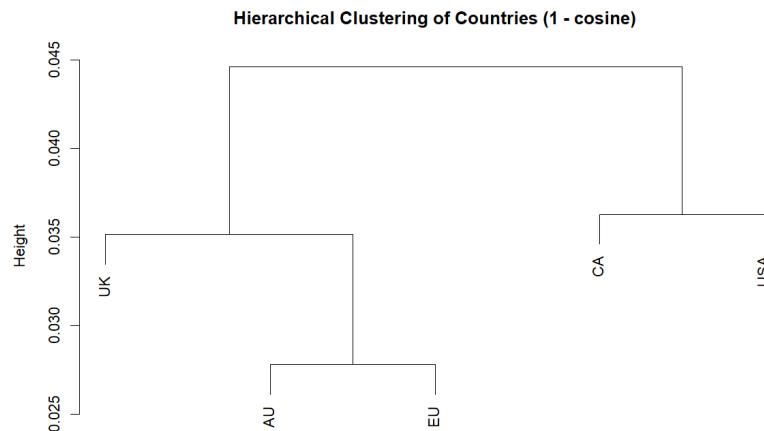
## מפת חום- השוואה בין מדינות לפי נושאי מיקוד





המפה מציגה את התפלגות הנושאים לפי מדינה, כפי שנגזרה ממודל LDA. השורות הן מדינות, עמודות הן נושאים. כל תא מציג את האחוז של אותו נושא מתוך כלל הטקסטים של אותה מדינה. גוון ככה יותר = חלק גדול יותר. בתאים נמצאים הערכים המדויקים באחוזים כאשר כל שורה מנורמלת ל-100%.

## קלאסטרिंग מדינות לפי דמיון קוסינוס באמצעות FinBERT



מודל השפה שנבחר הוא FinBERT, זו גרסה של מודל BERT שעברה אימונים על מסמכים פיננסיים כמו דוחות כספיים וכתבות פיננסיות, לכן הוא מתאים יותר לעולם הבנקאות והרגולציה. הוא מתאים לניתוח בעבודה זו מכיוון שהוא יודע לפרש משמעות בהקשר פיננסי. בנוסף, הוא מודל שלא נלמד בכיתה.

הניתוח נעשה על ידי יצירת Embeddings מהקורפוס בשפת פייתון, ולאחר מכן ניתוח שלהם בשפת R בסוף קובץ הקוד הראשי (קובץ הפייתון יצורף כנספח).

הדנדוגרמה מציגה קלאסטרिंग של המדינות.

ציר Y:  $1 - \text{cosine}$ , המרחק שבו אשכולות מתחברים. חיבור נמוך יותר = דמיון טקסטואלי גבוה יותר.

ציר X: תצוגתי, אין משמעות ממשית.

- ממוצע דמיון לשאר: CA 0.956, EU 0.955, AU 0.968, UK 0.961, USA 0.959
- הזוג הדומה ביותר: איחוד אירופי – אוסטרליה ( $\text{cos} = 0.972$ )
- טווח דמיון כללי (Height): 0.94 – 0.97, כולם דומים לכולם באופן יחסי.
- שני בלוקים עיקריים: USA-CA, UK-EU-AU

## הסקת מסקנות

### מפת החום:

- הבנק האירופאי המרכזי – מתמקד באופן מובהק בפיקוח ואסטרטגיה. ניתן להסביר את זה מתוקף תפקידו כבנק האחראי על כמות גדולה מאוד של בנקים והוא אף משמש כמפקח של מפקחים.
- בריטניה – מיקוד חזק בניהול סיכונים ובשירותים פיננסיים. לאחר הפרישה של בריטניה מהאיחוד האירופי המדינה מנסה למשוך השקעות על ידי מיצוב שוק אטרקטיבי, תחרותי, ובעיקר יציב.
- ארה"ב – מיקוד חזק בהגנת לקוחות, ציות ומעורבות קהילתית. בארה"ב ישנו מבנה רגולטורי מסובך יחסית אשר מחלק בו קיימים מספר גופים שונים. מאחר ונלקחו מסמכים של שני רגולטורים ואחד מהם אחראי על תחום הגנת הצרכן (CFPB), יש להניח שנוצרה הטיה לכיוון נושאים אלה שאינם עומדים במרכז סדר היום של רוב הרגולטורים האחרים בתחום.
- קנדה – מעורבות קהילתית, רגולציה ומדיניות פרודנציאלית (יציבות). היחידים שנתנו דגש על גיוון, אקלים ועוד.
- אוסטרליה – רגולציה ויציבות.

### קלאסטרנינג עפ"י מדד קוסינוס באמצעות FinBERT:

**פרשנות לגרף:** הבסיס המשותף לכלל הרגולטורים בסקטור הפיננסי הוא יציבות – ניהול סיכונים ורגולציה פרודנציאלית. ההבדלים בין המדינות בדנדוגרמה נובעים מהדגשים האחרים/משניים. האיחוד האירופי ואוסטרליה עסוקות כמעט אך ורק בנושא היציבות, ניהול סיכונים, נזילות, תפעול, מבחני לחץ. ארה"ב וקנדה מוסיפות לאסטרטגיה נושאים אחרים כמו הגנת הצרכן, הוגנות, גיוון והכללה, מעורבות קהילתית. בריטניה מדברת בעיקר על יציבות אך מדגישה גם תחרותיות וחדשנות – מה שמרחיק אותה מבחינת קרבה אל האיחוד האירופי ואוסטרליה.

דוגמאות תוכן – Unigrams, Bigrams לפי שני הבלוקים המרכזיים:

**EU/AU/UK – Prudential/Risk:** “prudential”, “capital”, “liquidity”, “risk management”, “stress testing”.

**USA/CA – Consumer/Community:** “consumer protection”, “disclosure”, “fair lending”, “complaints”, “DEI/diversity”, “community engagement”.

## מחקר המשך

- ניתוח ה-LDA בעבודה זו נעשה באמצעות Unigrams משיקולים של אמינות הספירות, ניסיון ליצור TDM/DTM פחות מדולל, ושיום נושאים ברור וקל יותר. המשך המחקר יהיה שימוש ב-Bigrams.
- התייעצות עם גורם מקצועי על מנת לבצע נרמול טוב יותר של המסמכים. בישראל למשל, חטיבת הפיקוח על הבנקים בבנק ישראל עוסקת בכלל הנושאים הקשורים לפיקוח על הבנקים, לעומת מדינות אחרות בהן עבודה זו היא רק חלק מתפקידו של רגולטור אחד או שאותה העבודה מבוצעת על ידי מספר רגולטורים. הנרמול מהווה אתגר משמעותי.
- הרחבת מסד הנתונים כך שיכלול רגולטורים ומדינות נוספות על מנת לייצר תמונה רחבה ואמינה יותר, כולל את ישראל.
- בעבודה זו הוחלט לצמצם את מספר הנושאים כך שיהיה מפורט אבל לא מבולגן. אך על מנת לראות תמונת מצב חדה יותר, לאחר הוספת מסמכים נוספים, נרחיב את מספר הנושאים.

אני חושב שהמחקר בשלב זה, הגיע ליעדים הנדרשים למעט ניתוח בין רגולטורי עקב כך שלא היו מספיק רגולטורים לכל מדינה. כאשר אוסיף עוד רגולטורים אוכל לעשות ניתוח בין רגולטורי טוב, בשלב זה לא ניתן מענה למטרת מחקר זו. האתגרים המרכזיים היו בניית מסד הנתונים בצורה טובה, נרמול הנתונים בתוך המסמכים עצמם וניסוי לכסות את כלל תפקידי הפיקוח על הבנקים ושיום הנושאים שלקח קצת זמן ולא היה ברור במבט ראשון.

## תרומת המחקר:

תרומת המחקר היא יצירת מפה השוואתית של המיקוד הרגולטורי בין רגולטורים מובילים בעולם, שמבליטה במה כל רגולטור מתמקד ומה חסר אצל אחרים—וכך מאפשרת לאתר תחומים שכדאי לאמץ או לחזק. בנוסף, המחקר מספק צפייה קדימה בזיהוי מגמות מתהוות והתפתחויות צפויות ברגולציה הפיננסית.

## מקורות

חומר שנלמד בכיתה: קוד, מצגות, הקלטות.

נעזרתי ב- ChatGPT, בעיקר בשיום הנושאים, תיקון קוד, וניתוח באמצעות FinBERT הן בשפת R והן בשפת פייתון ליצירת ה- Embeddings.

מסמכי האסטרטגיה של הרגולטורים עצמם:

RPA: [https://www.bankofengland.co.uk/prudential-regulation/publication/2024/april/pra-business-plan-2024-25?utm\\_source=chatgpt.com](https://www.bankofengland.co.uk/prudential-regulation/publication/2024/april/pra-business-plan-2024-25?utm_source=chatgpt.com)

ECB: [https://www.bankingsupervision.europa.eu/framework/priorities/html/ssm.supervisory\\_priorities202212~3a1e609cf8.en.html?utm\\_source=chatgpt.com](https://www.bankingsupervision.europa.eu/framework/priorities/html/ssm.supervisory_priorities202212~3a1e609cf8.en.html?utm_source=chatgpt.com)

OCC: [https://www.occ.gov/publications-and-resources/publications/banker-education/files/pub-occ-strategic-plan-2023-2027.pdf?utm\\_source=chatgpt.com](https://www OCC.gov/publications-and-resources/publications/banker-education/files/pub-occ-strategic-plan-2023-2027.pdf?utm_source=chatgpt.com)

FDIC: [https://www.fdic.gov/analysis/risk-review/2024-risk-review/2024-risk-review-full.pdf?utm\\_source=chatgpt.com](https://www.fdic.gov/analysis/risk-review/2024-risk-review/2024-risk-review-full.pdf?utm_source=chatgpt.com)

OSFI: [https://www.osfi-bsif.gc.ca/en/about-osfi/reports-publications/osfis-annual-risk-outlook-fiscal-year-2024-2025?utm\\_source=chatgpt.com](https://www.osfi-bsif.gc.ca/en/about-osfi/reports-publications/osfis-annual-risk-outlook-fiscal-year-2024-2025?utm_source=chatgpt.com)

חלקם של המאמרים של המאמרים בסקירת הספרות עזרו לחזק החלטות מסוימות:

Röder, M., Both, A., & Hinneburg, A. (2015). Exploring the space of topic coherence measures. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining (WSDM '15)* (pp. 399–408). ACM.

Sievert, C., & Shirley, K. (2014). LDAvis: A method for visualizing and interpreting topics. In *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces* (pp. 63–70). Association for Computational Linguistics.

Hansen, S., & McMahon, M. (2016). Shocking language: Understanding the macroeconomic effects of central bank communication. *Journal of International Economics*, 99(S1), S114–S133

Hansen, S., McMahon, M., & Prat, A. (2018). Transparency and deliberation within the FOMC. *The Quarterly Journal of Economics*, 133(2), 801–870.