**עבודה מסכמת – ניתוח נתונים פיננסים**

מגישים: משה גבאי- 314944661

חגי בן-יחיאל- 206638496

יאיר שפיר- 315494476

**חלק 1 ניקוי ועיבוד נתונים.**

ניקוי נתונים קובץ pickle:

אחרי העלאת הקובץ זיהינו Rick.pickle המרנו את הנתונים לדאטא פריים ולאחר בדיקה של סטטיסטיקה תיאורית ו - scatter plot למשתנים היה אפשר לזהות תצפיות חריגות בעמודת close. בנוסף עשינו בדיקה של ערכים חסרים וזיהינו שבעמודת volume חסרה תצפית אחת.

A graph of blue lines

AI-generated content may be incorrect.בשביל לזהות את התצפיות החריגות עשינו בדיקה של נתונים שסוטים פעמיים מטווח הבין רבעוני. באמצעות שיטה זו, הצלחנו לזהות ארבע תצפיות חריגות במחירי הסגירה, שכללו ערכים לא סבירים של 9999, 0 ו 20- , אשר ברור שאינם משקפים את שווי המניה האמיתי בימי המסחר הללו אלא כנראה טעויות בקליטת הנתונים או בעיות אחרות.

לאחר זיהוי הערכים החריגים והחסרים בנתונים, החלטנו לטפל בהם באמצעות חישוב ממוצע נע ל-20 ימי מסחר שקודמו לכל תצפית בעייתית. לשם מניעת השפעת הערכים החריגים על חישוב הממוצע הנע, ראשית שינינו את אותם ערכים לחסרים (NaN). בשלב הבא, חישבנו את הממוצע הנע בהתבסס על 20 הימים הקודמים, ולאחר מכן השלמנו את התצפיות החריגות בערכים המחושבים. לבסוף, בוצעה בדיקה באמצעות גרף ואימות נוסף של התצפיות, במטרה לוודא שהתיקון בוצע כהלכה ושאין עוד סטיות בלתי צפויות בנתונים.

גישה זו מנעה בצורה משמעותית את בעיית ה-Data Leakage, כיוון שלא נעשה שימוש בנתונים מהעתיד ביחס ליום שבו התגלתה החריגה – כלומר, לא "ניבאנו את העתיד". בנוסף, השימוש בממוצע נע המבוסס על 20 ימים בלבד מבטיח שהתיקון מתבצע בהסתמכות על נתונים עדכניים ורלוונטיים בלבד, מה שמסייע למזער את השפעתם של תוצאות עבר שאינן מייצגות את המצב הנוכחי.

שליפת נתוני טסלה:

ייבאנו את הת הספרייה YFINANCE בשביל לשלוף את הנתונים היומיים של מניית טסלה מאותו טווח תאריכים שביקשו ממנו כולל הדיוודנדים וספליטים. לאחר מכן בשביל לחשב את ה adjusted close חישבנו את מקדם השפעת הדיבידנד לכל יום על ידי חישוב של יחס של 1 פחות הדיבידנד חלקי מחיר הסגירה, אחרי חישוב יחס הדיבידנד עלינו לחשב את המכפלה המצטברת כך שכל יום יושפע מהימים שבאו אחריו. כשיש לנו את המכפלה המצברת אנחנו מכפילים את המחיר סגירה כל פעם במכפלה של אותה תצפית וככה נקבל את ה adjusted close. בשביל לחשב את כל שאר המחירים המותאמים אנחנו נמצא את היחס של התאמה עבור כל תצפית על ידי - מחיר סגירה \ מחיר סגירה מותאם, ואיתו נמצא את המחירים המותאמים של הפתיחה, סגירה והנמוך.

במקרה של מניית טסלה לא היה לה דיבידנדים באותה תקופה לכן המחירים לא השתנו באותה תקופה.

**חלק 2 - ניתוח טכני**

**איך UT BOT עובד:**

האינדיקטור UT-Bot מבוסס על שני פרמטרים עיקריים: ATR ו ־ KEY VALUE.

(ATR (Average True Range:

מדד זה מודד את התנודתיות בשוק, כלומר את עוצמת תנועת המחירים. הוא מחשב את ה־True Range עבור כל תקופה, כאשר ה־True Range הוא הערך המקסימלי מבין:

* (high – low) – הטווח הנוכחי של הנכס.
* (high – prev\_close) – הטווח בין המחיר הגבוה הנוכחי למחיר הסגירה של תקופה הקודם.
* |low – prev\_close| - הטווח בין המחיר הנמוך למחיר הסגירה של תקופה קודמת.

בברירת המחדל ATR מחושב לפי 10 תקופות - ממוצע הפער המקסימלי של 10 תקופות האחרונות כך שמתקבל משש שבודק את התנודתיות בשוק.

נוסחא: **ATR = (TR\_1 + TR\_2 + … + TR\_10)/10**

**KEY VALUE:**

פרמטר זה מאפשר לכוון את רגישות האינדיקטור. הוא משמש כגורם כפל שמגדיל או מצמצם את מרחק הביטחון (nLoss) שבו מתבצע החישוב.

חישוב nLoss:

נוסחא: nLoss = ATR \* KEY VALUE

חישוב סטופ לוס:

|  |  |
| --- | --- |
| trailing stop loss = current price - nLoss | פוזציות לונג |
| : trailing stop loss = current price + nLoss | פוזציות שורט |

חשוב לציין שסטופ לוס בפוזציות לונג מתעדכן רק כאשר המחיר עולה מעל המחיר הגבוה ובפוזיציות שורט הוא מתעדכן רק כאשר המחיר יורד מהמחיר הנמוך.

קביעת אותות קנייה ומכירה:

* כאשר המחיר עובר מעל לטריילינג סטופ, מדובר באות קנייה.
* כאשר המחיר עובר מתחת לטריילינג סטופ, מתקבל אות מכירה.

הרעיון כאן שנקבל אותות קנייה ומכירה כאשר הוא מזהה תנועות מחיר שחורגות וטווח התנודתיות הרגיל וכך אנחנו נקבל פריצות או שבירות מחיר אמיתיות.

איך STC עובד:

האינדיקטור STC מבוסס על שילוב של ממוצעים נעים אקספוננציאליים (EMA) ותהליך נורמליזציה והחלקה שמטרתו להפיק אוסצילטור המדגים את מחזוריות השוק ואת מהירות שינוי המגמה.

**הגדרת ממוצעים אקספוננציאליים (EMA):**

הקוד מחשב שני ממוצעים אקספוננציאליים על בסיס נתוני מחיר הסגירה (close)

**fastMA:** ממוצע מהיר עם תקופה קצרה של 27 (fast length).

**slowMA:** ממוצע איטי עם תקופה ארוכה של 50 (slow length).

פונקציית ה־EMA מיישמת את החישוב בצורה שמעניקה משקל גבוה יותר לערכים העדכניים.

נוסחה:

α = 2/(N + 1) - המשקל שניתן לנר הנוכחי בהתאם למספר התקופות המחושב

EMAₜ = Priceₜ × α + EMAₜ₋₁ × (1 – α)

**חישוב הפרש הממוצעים (diffEMA):**

לאחר קביעת הממוצעים, מחושב ההפרש ביניהם – fastMA פחות slowMA.

הפרש זה מסייע לזהות את עוצמת המגמה: ככל שהפרש גדול יותר, כך ישנה תנועה חדה יותר במחיר.

**נרמול לאחוזים בחלון זמן (length):**

בשלב זה לקחנו תקופות מהערך diffEMA מתוך 80 הנרות האחרונים ומצאנו את הערך המינימאלי והמקסימאלי. כדי לנרמל את diffEMA לאחוזים השתמשנו נוסחה הבאה:

ccccc = ((diffEMAᵢ – min(diffEMA)) / (max(diffEMA) – min(diffEMA))) × 100

פעולה זו מצביעה על מיקומו של הערך diffEMA בתוך טווח התונדתיות שבחלון length אותו הגדרנו ל- 80.

**ישום סינון ראשוני (ddd):**  
לצורך מיתון תנודתיות הערך הראשוני, ביצענו סינון באמצעות חישוב דינמי:

במקרה שאין ערך קודם, מאתחלים עם הערך הנוכחי. אחרת, מעדכנים לפי נוסחה:  
(ddd\_now = ddd\_prev + aaa \* (ccccc\_now – ddd\_prev

כאשר aaa הינו משקולת לקביעת רמת השפעת הערך הנוכחי מההפרש בין הערך האחוזי של diffEMA של התקופה הזאת לתקופה אחת קודם.

במידה ואין ערך קודם הגדרנו ddd = ccccc.  
ערך זה מהווה שכבת סינון ראשונה שעוזרת להחליק את השינויים החדים.

**נורמליזציה חוזרת על סדרת ddd:**

כעת עשינו פעם נוספת נרמול לאחוזים לקבלת ערך חדש אותו סימנו כ- dddddd לפי הנוסחה:

dddddd = ((dddᵢ – min(ddd)) / (max(ddd) – min(ddd))) × 100

**החלקה שנייה וחישוב הערך הסופי ל- STC:**

חזרנו על פעולת הסינון פעם נוספת באמצעות הנוסחה:

eeeeeᵢ = eeeee₍ᵢ₋₁₎ + α × (dddddd – eeeee₍ᵢ₋₁₎)

במידה ואין ערך קודם הגדרנו ש- eeeee = dddddd.

ערך זה הינו הערך הסופי המוצג כאינדיקטור

הרעיון המרכזי מאחורי STC הוא לשלב תגובה מהירה לשינויים (באמצעות חישוב diffEMA) עם סינון רעשים (באמצעות נורמליזציה והחלקות חוזרות), כך שמתקבל אינדיקטור לזיהוי מחזוריות ומגמות בשוק.

תיאור האסטרטגיה המסחר:

האסטרטגיה בוחנת את תנודתיות הנכס באמצעות UT-BOT בשתי דרכים: ראשית, היא מחשבת את ה-ATR לטווח הזמן המיידי, ושנית, היא מנתחת ATR לטווח ארוך (300 נרות בני חצי שעה אחורה). בהתאם לכך, היא מציבה Stop Loss דינמי המוגדר לפי הנוסחה: (מחיר נוכחי – 2\*ATR). האסטרטגיה מייצרת סיגנל קנייה כאשר המחיר חוצה כלפי מעלה את ה-Stop Loss הארוך, וסיגנל מכירה כאשר המחיר חוצה כלפי מטה את ה-Stop Loss המיידי.

לאחר מכן, מתבצע סינון נוסף באמצעות STC (Schaff Trend Cycle), שמוודא אם הסיגנל תואם למגמת השוק ומונע כניסה שגויה. בטווח הזמן המוגדר, האסטרטגיה מזהה את ערכי ה-MACD הגבוהים והנמוכים ביותר ומשתמשת בהם לחישוב הטווח. במידה והטווח חיובי, ה-STC מחושב כך שיהיה בטווח של 0-100 לפי הנוסחה:  
(ה-MACD הנוכחי פחות ה-MACD המינימלי) חלקי הטווח המקסימלי כפול 100.

כאשר ערך ה-STC נמוך מ-25 ("Green Line"), זוהי אינדיקציה למכירת יתר, או לחלופין כאשר ערך ה-STC הנוכחי גבוה מהקודם – סימן להיפוך מגמה כלפי מעלה. במקרים אלה, ואם קיים גם סיגנל קנייה מ-UT-BOT, מתבצעת קנייה. לעומת זאת, כאשר ה-STC גבוה מ-75 ("Red Line") או כאשר ערכו הנוכחי נמוך מהקודם (סימן להיפוך מגמה כלפי מטה), ובנוסף מתקבל סיגנל מכירה מ-UT-BOT, מתבצעת מכירה.

האסטרטגיה שואפת לצמצם סיכונים ולחזות מגמות מדויקות בנכס על ידי שילוב של מספר אינדיקטורים. רק כאשר מתקיימים כל התנאים הדרושים, מתבצעת עסקה. עם זאת, האסטרטגיה עובדת כאשר מאוד תנודתי. כאשר אין תנודתיות מספקת האסטרטגיה עלולה לייצר כמות גדולה של סיגנלים לא מדוייקים מה שעלול להוביל לכניסות ויציאות שגויות.

ביקורת על האסטרטגיה

על אף המורכבות והדיוק של האסטרטגיה, אנו סבורים שיש בה בעיות מהותיות:

1. אין בה אינדיקטור המעריך יחס סיכון-סיכוי, ולכן חסרה התייחסות לשאלה האם בכלל כדאי להיכנס לפוזיציה ובאיזה גודל. חוסר היכולת לתמחר את הפוזיציה מגביר את הסיכון.
2. האסטרטגיה מתעלמת מעלויות העסקה. כתוצאה מכך, היא עשויה לבצע מספר רב של עסקאות, אך העלויות המצטברות עשויות להפוך אותה ללא כדאית.

לסיכום, למרות הניתוח המעמיק של מגמות השוק, האסטרטגיה אינה מספקת פתרונות מלאים להתמודדות עם הסיכונים הכלכליים הכרוכים בביצועה.

תוצאות האסטרטגיה

מהבדיקה שערכנו על האסטרטגיה של המציג קיבלנו:

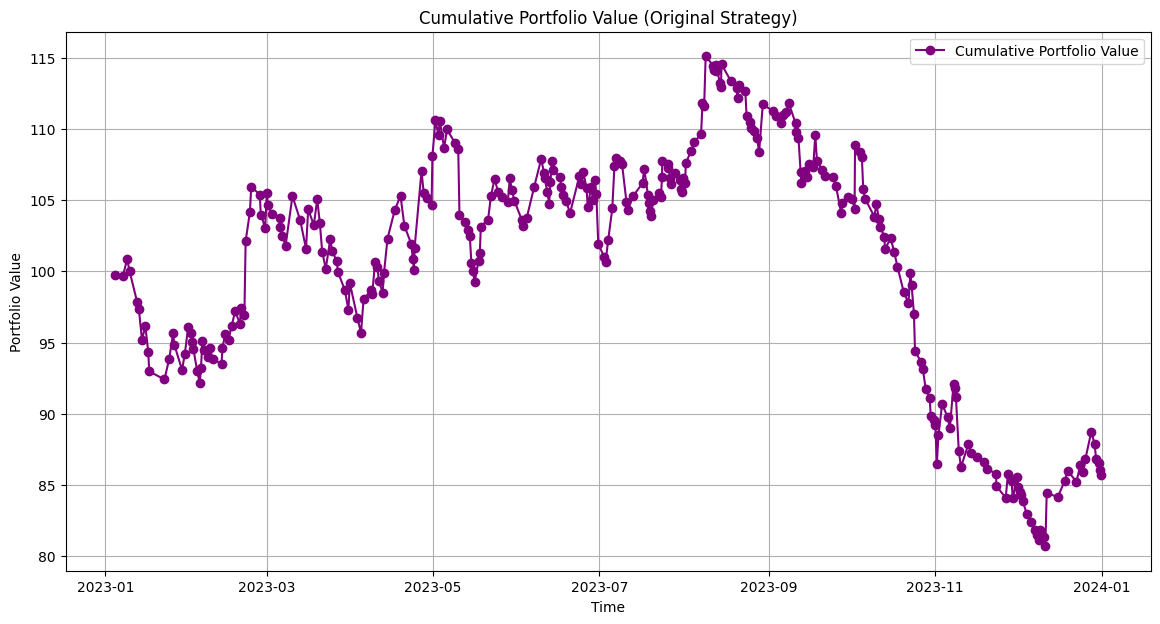
* בוצעו 301 עסקאות
* שיעור העסקאות הרווחיות – 33.22%
* ממוצע התשואה עבור עסקאות רווחיות – 1.60%
* ממוצע התשואה עבור עסקאות מפסידות – 0.86%-
* תשואה מצטברת – 14.28%-
* תשואה ממוצעת לעסקה - 0.04%-

\*חשוב לציין שהתשואות לא כוללת עלויות עסקה.

האסטרטגיה המוצגת עושה שימוש בפרמטר לקביעת ערך מפתח (2) המגדיר את רמת הרגישות ליצירת Stop Loss. ככל שערך המפתח גבוה יותר, כך מתקבלות פחות התראות לביצוע עסקה, מה שמפחית את תדירות הכניסות והיציאות מהשוק.

אנו ממליצים לבחור ערך מפתח גבוה יותר, שיביא להפחתת תדירות הסיגנלים לביצוע עסקה. גישה זו תאפשר לחכות לשינויי מגמה ברורים ומובהקים יותר בנכס. כתוצאה מכך, כמות העסקאות הכוללת צפויה לרדת, מה שיוביל לצמצום עלויות העסקה. יתרה מכך, גישה זו עשויה לשפר את שיעור העסקאות הרווחיות ובכך להגדיל את התשואה הנצברת לאורך זמן.

**פעולות האסטרטגיה לאורך התקופה:**

**שווי תיק ההשקעות לאורך זמן:**

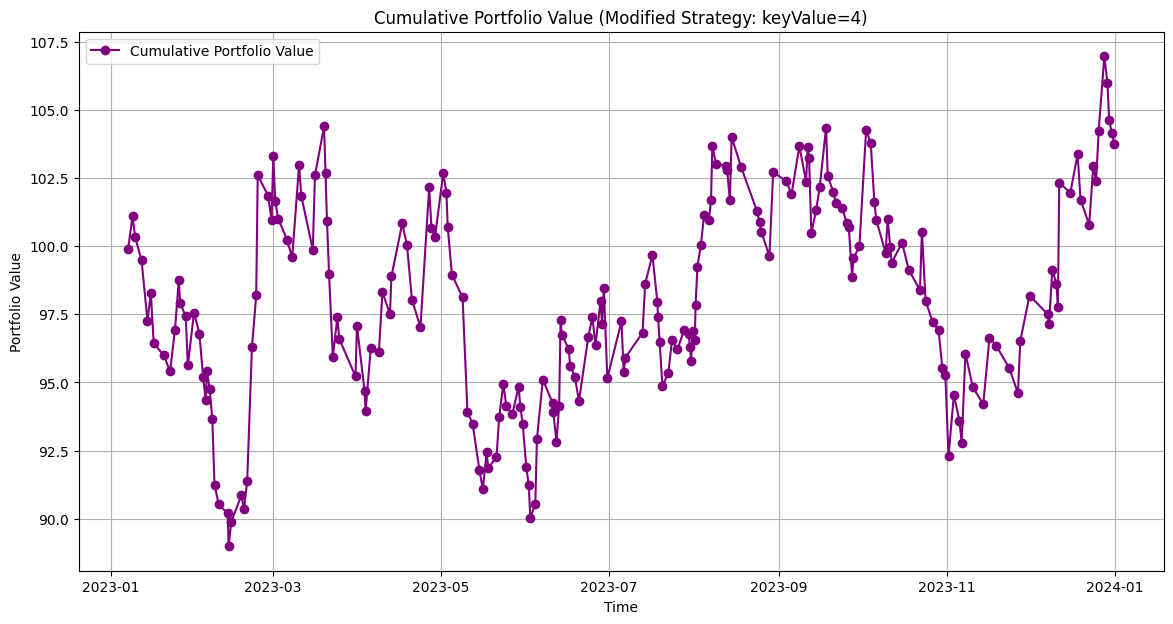
הצעת אסטרטגיה חלופית

כפי שהערכנו, לאחר שהעלינו את ערך המפתח ל-4, ביצועי האסטרטגיה השתפרו. אנו סבורים כי הסיבה לכך היא שהרגישות לסטופ לוס העליון והתחתון פחתה, מה שהוביל לצמצום מספר הסיגנלים שקריים של קנייה ומכירה.

לאחר ששינינו את המפתח ל- 4 במקום 2:

* בוצעו 212 עסקאות
* שיעור העסקאות הרווחיות – 36.32%
* ממוצע התשואה עבור עסקאות רווחיות – 1.84%
* ממוצע התשואה עבור עסקאות מפסידות – 1.00%-
* תשואה מצטברת – 3.75%
* תשואה ממוצעת לעסקה - 0.03%

**גרף המציג את האסטרטגיה לאורך התקופה**

**שווי תיק ההשקעות לאורך זמן:**

למרות השינוי לטובה לאחר שינוי המפתח, וקבלת רווח חיובי, אנו לא ממליצים על שימוש באסטרטגיה מכיוון שאחוז הרווח מקניה בתחילת תקופה ומכירה בסוף תקופה הוא גבוה יותר בצורה משמעותית (156%).

**חלק 3- Sentiment Analysis**

**חלק 1- Reddit:**

הקדמה:

בחלק זה של העבודה התמקדנו בבדיקת הסנטימנט של פוסטים ברדיט על ידי ניתוח של הכותרת והתוכן שלהם ביחד עם התייחסות לויראליות של הפוסט.  
ניתוח זה נעשה בעזרת מודל NLP שבחרנו distilroberta-financial-sentiment.

משיכת הפוסטים וביצוע סינון:

יצרנו משתמש באתר Reddit וקיבלנו API שבעזרתו ניתן למשוך פוסטים מהאתר.

התחברנו בעזרת מס' הלקוח והAPI היחודיים שלנו, ומשכנו פוסטים המסוננים על ידי מילת המפתח AAPL.

יצרנו דאטה פריים ריק שיהיה בתוכו מידע על הפוסט- כותרת, טקסט, תאריך פרסום, שם הכותב, כמות לייקים, כמות תגובות וכתובת קישור.  
הגבלנו את כמות המשיכה ל1000 פוסטים, מכיוון שלפעמים נמשכים פוסטים ריקים או כפולים, ורצינו להישאר עם 100 פוסטים יחודיים ותקינים לאחר סינונים של הפוסטים הלא רלוונטים.  
לאחר הסינון, אם נשאר יותר מ100 פוסטים רלוונטים, השתמשנו ב100 הפוסטים הראשונים.

ניתוח הסנטימנט לכל פוסט:

תחילה קראנו למודל distilroberta-financial-sentiment, הגבלנו את אורך הטקסט ל512 תווים והפעלנו אותו על הכותרת ועל הטקסט בנפרד.

לאחר ניתוח המודל רצינו ליצור ציון משוקלל לכותרת ולטקסט- לכותרת ניתן משקל של 0.1 מסך הציון הכללי, ולטקסט ניתן משקל של 0.9 מסך הציון הכללי.

שקללנו את תוויות הסנטימנט (חיובי - (1) / שלילי - (1-) / ניטרלי - (0)), וגם את ציון הביטחון של המודל בהשערתו (0-1).

**חישוב:**  
משקל כותרת \* תווית הסנטימנט כותרת \* ציון הביטחון כותרת +

משקל טקסט \* תווית הסנטימנט טקסט \* ציון הביטחון טקסט = ציון סנטימנט כולל

טווחי הציון הכולל בין 1 ל1-:

Positive - (0.5) ומעלה

negative - (-0.5) ומטה

neutral - (-0.5) - (0.5)

פיתוח מדד ויראליות:

כדי שנוכל להעריך את ההשפעה הפוטנציאלית של כל פוסט, פיתחנו מדד ויראליות על ידי יצירת עמודה בשם virality, שמחושבת כך: כמות לייקים\*1 + כמות תגובות\*0.3 = ציון ויראליות

המשקלים ניתנו כך מכיוון שלהערכתנו בפורום מהסוג הזה, לכמות הלייקים יש יותר חשיבות בניתוח של ויראליות/השפעת הפוסט מאשר תגובות.   
מכיוון שבתגובות יכולה להיווצר שיחה שתייצר כמות תגובות גדולה אך מנהולת על ידי מעט משתתפים.

לאחר חישוב הויראליות הפעלנו לוג בבסיס 100 על הציון, בשביל לקבל התפלגות מאוזנת יותר ולהפחית השפעה של פוסטים חריגים.

כל הציונים שמעל 1 נורמלו לציון 1 (ציון הויראליות המקסימלי), כדי שכל הציונים יהיו בטווח 0-1.

כלומר, כל פוסט שציון הויראליות המקורי שלו (לייקים\*1 + תגובות\*0.3) מעל 100, נורמל ל-1.

ציון סנטימנט-ויראליות:

ציון סנטימנט כולל \* 0.65 + ציון ויראליות מנורמל \* 0.35 = ציון סנטימנט-ויראליות.  
מתן המשקלים ניתן כך מכיוון שאנו רצינו לתת את רוב ההתייחסות לציון הסנטימנט.  
מכיוון שיכול להיווצר מצב שהפוסט אמנם לא כל כך ויראלי, אך הוא חיובי מאוד באופיו, וציון הויראליות יטה אותו באופן קיצוני.

חלק 2- ניתוחי כתבות מהרשת:

הקדמה:

חלק זה עוסק בניתוח סנטימנט של כתבות הקשורות לחברת אפל ולמניה שלה(AAPL).  
המטרה היא להבין את רמת/אופי הסיקור של המניה על פני תקופה מסוימת בעזרת שימוש במודלי NLP.

בחלק זה 3 שלבים עיקריים- איסוף מאמרים רלוונטים, חילוץ ועיבוד המידע, וניתוח הסנטימנט(כולל דיוק המדדים שבעזרתם אנו נחליט על ציון הסנטימנט).

איסוף מאמרים רלוונטים:

אסטרטגיית החיפוש שלנו הייתה לחפש מידע רלוונטי בעזרת הגדרת כמה מילות מפתח הקשורות לחברה (apple, iphone ועוד), בנוסף מיקדנו את האתרים בהם יתבצע החיפוש למספר אתרים פיננסים אמינים.

חילקנו את תקופת החיפוש למספר תתי טווחים כדי למקסם את כמות התוצאות, בגלל ההגבלה שיש לנו על כמות הפוסטים שאנו יכולים למשוך בו זמנית בעקבות השימוש בAPI חינמי.

חילוץ ועיבוד התוכן:

חילצנו את התוכן המלא של הכתבות באמצעות חבילת Trafilatura, שמטרתה קריאת וחילוץ הטקסט במלואו.  
סיננו כתבות עם תוכן חסר או שגוי, והסרנו כפילויות בכדי למנוע הטיה של התוצאות.  
לסיום בחרנו את 100 הכתבות הראשונות לאחר הסינונים.

ניתוח סנטימנט:

ניתחנו בנפרד את 3 המרכיבים של כל כתבה, ונתנו לכל חלק משקל שונה בציון הסנטימנט הסופי:  
- כותרת הכתבה: 0.1

- תקציר הכתבה: 0.15

- תוכן הכתבה: 0.75

בחירת המשקלים נעשתה על פי שיקול דעתנו, עם רצון לתת את רוב המשקל כמובן לתוכן הכתבה.

לניתוח השתמשנו במודל ייעודי לכתבות פיננסיות - (distilroberta-financial-sentiment)

טיפלנו בכתבות ארוכות באמצעות חלוקה וקיצור טקסט להתאמה למגבלות המודל (512 טוקנים, כל מילה- טוקן, אנו שינינו שכל 4 מילים יהיו טוקן).

שקלול התוצאות ויצירת מדד מורכב:

יישמנו את המשקל שניתן לכל חלק בכתבה (כותרת, תקציר, תוכן).  
שקללנו את תוויות הסנטימנט (חיובי(1)/שלילי(1-)/ניטרלי(0)), וגם את ציון הביטחון של המודל בהשערתו (0-1).

חישוב ציוני המדד הסופי weighted sentiment score(בין 1 ל1-):  
משקל כותרת \* תווית הסנטימנט כותרת \* ציון הביטחון כותרת +

משקל תקציר \* תווית סנטימנט תקציר \* ציון ביטחון תקציר +

משקל תוכן \* תווית סנטימנט תוכן \* ציון ביטחון תוכן = ציון סופי

סיווגנו את התוצאות המשוקללות ל5 קטגוריות:

very bullish - 0.75 ומעלה

bullish - 0.5 - 0.75

bearish - (- 0.5) - (- 0.75)

very bearish – (- 0.75) ומטה

neutral - (- 0.5) - (0.5)

אנו מאמינים כי פעולות הדיוק שעשינו ביצירת המדד החדש יעזרו לשקף את הלך הרוח בשוק בצורה טובה ומדוייקת יותר.