המרכז האקדמי למשפט ולעסקים

הפקולטה למערכות מידע ומדעי המחשב

שימוש ברשתות חברתיות לניבוי שוק ההון

ולדימיר אלקין

ת.ז 336463351

מאור ישראלי

ת.ז 319023057

מנחה: ד"ר יוסיפוף אברהם

שם הקורס: פרוייקט גמר

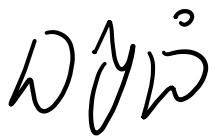
מס' הקורס: 6201

ו בטבת, תשפ"ג

30.12.2022

אישור המנחה

מאושר ד״ר אבי יוסיפוף



תודות

אנו מבקשים להקדיש דף זה להודעת תודה לכל אלה שתמכו ותרמו לפרויקט הגמר שלנו. בזכות התמיכה והעזרה שלכם, הצלחנו להשלים את הפרויקט בהצלחה ולהתמודד עם האתגרים השונים שלו. בבקשה קחו את הזמן וקראו כמה מילות תודה מאיתנו.

תודות רבות למנחה הפרויקט דוקטור אבי יוסיפוף על ההנחיה המובהקת והידע העשיר שתרמו לתהליך כתיבת הפרויקט. ההכוונה ההדרכות שקיבלנו ממך וניתנו בכל נדיבותך סייעו לנו לשפר את הפרויקט ולהשיג תוצאות טובות יותר.

תודה גם לדוקטור ג'וני אלון שעזר לנו להבין נושאים בפרוייקט, פקח את עיננו והכווין אותנו אל המטרה.

תודה למשפחה והחברים על התמיכה והסבלנות שהראיתם במשך כתיבת הפרויקט. התמיכה הנפלאה שלכם השפיעה על חשיבותו של הפרויקט והעניקה לנו כוח ועוצמה להמשיך ולהגיע לסיום.

תודה לכולם על התמיכה, העזרה וההשקעה. אתם הפכתם את התהליך לחוויה חיובית וללא תחרות. אנו מאוד מודים לכם על הזמן והמאמץ שהקדשתם עבורנו ועבור הפרויקט.

בברכה,

אלקין ולדימיר

ומאור ישראלי

תוכן עניינים

רשימת איורים, רשימת טבלאות וגרפים

תקציר בעברית

ידוע כי שוק המניות בעולם הינו שוק לא צפוי שקשה לחזותו מבעוד מועד ועובדה זו מקשה על משקיעים ובעלי הון לבחור באילו מניות וכן באיזה טווח זמן הכי נכון להשקיע את כספם, מתוך החשש שישנה סבירות גדולה כי בסופו של דבר הם לא יצליחו לקצור את הפירות מהשקעתם. לאור עובדה זו, רצינו למצוא דרך בה נוכל להקל על אותם המשקיעים בצורה כזו שיוכלו להשקיע את כספם בצורה מושכלת, בטוחה ונבונה יותר.

בפרויקט שלנו בחרנו לבדוק האם קיים קשר בין ציוצים ברשת החברתית "טוויטר" של אישי עולם בכירים, לבין השינויים המתרחשים במניות בבורסה העולמית. הסיבה שבגינה בחרנו לעסוק בתחום זה היא שבמידה ואכן נמצא כי יש קשר ישיר בין ציוצם של בכירי העולם בפלטפורמה זו, לבין שוק המניות, אזי שנוכל לבנות וליצור אלגוריתם אשר יהווה מודל לחיזוי וניבוי שוק המניות העתידי שיתבסס על סמך ציוצם של אישים משפיעים אלה. בניית אלגוריתם מסוג זה, יש בידו לעודד את השימוש בו בקרב אנשים שמשקיעים במניות בבורסה, מאחר והוא מקנה ביטחון רב לאותם המשקיעים אשר יוכלו לדעת בסבירות גבוהה מאוד מתי ובאילו מניות הכי נכון להשקיע את כספם.

בימינו, אכן ישנם מספר פתרונות לניבוי המניות אך פתרונות אלה אינם אידיאליים מאחר והם מתבססים על נתונים לא מספיק מהימנים. לכן, הפתרון שאנו מציעים הינו מודל שיתבסס גם על נתוני העבר , אבל בנוסף גם על נתוני העבר של ציוצים מרשת ה"טוויטר". על בסיס נתונים אלו, המודל יידע לנבא מה יקרה למניות ספציפיות תוך התחשבות בציוצים ספציפיים.

מטרת המחקר שלנו היא להשתמש בנתוניהם של ציוצים קודמים מרשת טוויטר אשר פורסמו על ידי אנשי עולם בכירים בעלי כוח השפעה, על מנת לבדוק את הקשר בין ציוצים אלה לבין תנודות ושינויים במניות בארה"ב ובמדדים מרכזיים. במידה ונמצא כי אכן ישנו קשר, נעסוק בבניית מודל חיזוי שיתבסס על קשר זה ויצליח לנבא באופן מיטבי את השינויים העתידיים של המדדים המרכזיים והמניות.

על מנת לבסס את מחקרנו, נעזרנו במספר מחקרים אקדמיים כדי להבין לעומק בצורה מיטבית יותר את הקשר שנבדק בין הציוצים לבין סטטוס המניות בשוק ההון. במחקרים אלו נלקחו ציוצים מסוימים של בעלי השפעה, לצורך דגימה ובדיקה לאמיתות הקשר. במחקרים אלה, נמצא כי על ידי שיטות ניתוח שונות על הנתונים, הצליחו להגיע למסקנה כי עם ניתוח נכון ניתן לראות כי אפשר לחזות את התנהגות שוק ההון ברמת דיוק גבוהה על סמך אותם הציוצים.

במהלך מחקרינו, הצלחנו לבנות מודל חיזוי שמותווה על ידי שישה שלבים עיקריים אשר מהווים יחד אלגוריתם שיוכל לנבא ולחזות את עתידן של המניות בשוק העולמי על סמך אותם הציוצים השונים. ששת השלבים הללו הם:

1. איסוף הנתונים הכוללים ציוצים ונתוני השוק
2. עיבוד מקדים של הנתונים לשם פירמוטם לצורך ניתוחם
3. ניתוח נתונים חקרניים כדי להפיק מידע חיוני ורלוונטי אודותיהם
4. בניית מודלים לצורך חיזוי השינויים בשוק המניות

5. הערכת המודל על ידי מדדים שונים
6. הסקת מסקנות אודות ניתוח התוצאות

במהלך מחקרנו השתמשנו בכלים ובשיטות שונות אשר היוו עבורנו בסיס חיוני אל עבר כתיבת האלגוריתם, וכיוצא בזאת את הכלים הדרושים לשם כך. חלק מהכלים בהם השתמשנו לשם יצירת האלגוריתם היא בין היתר שפת התכנות פייתון, אשר בעזרתה כתבנו את המודל לשם עיבוד הנתונים ופיתוחו בעזרתה. בנוסף חלק מהכלים שהיוו עבורנו את בסיס הנתונים הם Twitter API שעזר לנו בבחירת דגימת הציוצים לצורך בדיקתם, וכן גם ב API פיננסי על מנת לקבל גישה לנתוני המניות.

תקציר באנגלית

It is known that the stock market in the world is an unpredictable market that is difficult to predict ahead of time, and this fact makes it difficult for investors and capitalists to choose which stocks and in which time frame to invest their money, out of the fear that there is a high probability that in the end they will not be able to reap the fruits of their investment. Considering this fact, we wanted to find a way in which we could make it easier for those investors in such a way that they could invest their money in a more informed, safer, and wiser way.

Today, there are indeed several solutions for predicting stocks, but these solutions are not ideal since they are based on insufficiently reliable data. Therefore, the solution we offer is a model that will also be based on past data, including tweets from the "Twitter" network. Based on this data, the model will be able to predict what will happen to specific stocks while considering specific tweets.

The purpose of our research is to use the data from previous tweets from the Twitter network that were published by senior world figures with influence to test the relationship between these tweets and fluctuations and changes in US stocks and major indexes. If it is found that there is indeed a relationship, we will engage in building a prediction model that will be based on this relationship and will be able to optimally predict the future changes of the major indexes and stocks.

To base our research, we used several academic studies to better understand in depth the relationship between the tweets and the stock status in the capital market. In these studies, certain tweets of influential people were taken for the purpose of sampling and testing the truth of the relationship. In these studies, it was found that by using different analysis methods on the data, they were able to reach the conclusion that, with a correct analysis, it is possible to predict the behavior of the capital market with a high level of accuracy based on those tweets.

During our research, we were able to build a prediction model that is outlined by six main steps, which together constitute an algorithm that can predict the future of stocks in the world market based on those various tweets. These six steps are:

1. The collection of data, including tweets and market data.
2. Pre-processing of the data to format it for analysis.
3. Analyzing exploratory data to extract essential and relevant information about them.
4. Building models for predicting changes in the stock market.

5. Evaluation of the model by different indices.
6. Drawing conclusions about the analysis of the results.

During our research, we used various tools and methods that formed an essential basis for us in writing the algorithm and thus the tools necessary for this purpose. Part of the tools we used to create the algorithm is the Python programming language, with the help of which we wrote the model for processing the data and developed it. In addition, some of the tools that formed the database for us were the Twitter API, which helped us select the sample of tweets for testing, as well as a financial API to get access to the stock data.

מבוא

כולנו מכירים את הרשתות החברתיות ואנו נתקלים בהן כמעט בכל יום. אחת מהרשתות הבולטות והמשפיעות ביותר היא הרשת החברתית "טוויטר", אותה רכש עד לא מזמן היזם ואיש העסקים, אילון מאסק. יש לציין כי רשת זו אינה נחשבת לרשת החברתית הפופולארית ביותר או הרווחית ביותר מבין יתר מתחריה, ואף נחשבת לרשת הפסדית.

אולם, אחד מיתרונותיה הבולטים של טוויטר, הינו מידת ההשפעה שלה בתחומים רבים אשר נוגעים בחיי היומיום של כלל האזרחים, בין אם מדובר בפוליטיקאים אשר משתמשים בה כפלטפורמה להעברת מסרים, בין אם בתחום הכלכלי ועוד. בפרויקט שלנו בחרנו לעסוק בתחום ההשפעה של רשת חברתית זו על שוק ההון.

משום שרשת חברתית זו משפיעה באופן ישיר ובולט על העולם, נרצה לבחון בסוגייה הספציפית שלנו את רמת ההשפעה של אישים בכירים על הבורסות באמצעות ציוצים בטוויטר כאמצעי תקשורת בינם לבין העולם. כיוצא בזאת נבדוק האם ישנה בידינו היכולת לחזות את מידת השינוי של מניות בשוק ההון, שייגרם כתוצאה מציוצים מסוימים של אותם האישים, מבעוד מועד, וכך להסיק באיזה אופן יעיל ומיטיב יש לפעול עם אותן המניות אשר בהן משקיעים מעוניינים להשקיע את כספם.

הבעיה העיקרית שלנו הינה ששוק המניות הוא שוק בלתי צפוי וכן ישנן השפעות רבות הגורמות לשינויים בשוויה של המנייה, אשר מקשות לנבא את עתידה באופן יעיל. ההשפעות יכולות להיגרם מסיבות שונות, חלקן נגרמות בשל מצב פוליטי, חלקן נגרמות מתוך קשר ישיר אל מצב השוק, ואף עברה של המנייה יש בידו כדי להכריע בשוויה. אתגר נוסף בשוק ההון הוא לדעת האם מחיר המנייה מייצג את השווי האמיתי שלה או שמא לא. קיימות שלל תאוריות בקשר לשוויה של המנייה, ובכולן הוא נובע מכל תשואות העבר שלה, ולכן, אי אפשר להשתמש בנתוני העבר של המנייה שיש לנו כדי לבנות מודל שינבא על בסיס זה את העתיד הצפוי למנייה.

כבר כיום קיימים כמה וכמה פתרונות לחיזוי השינויים בשוק ההון, הפתרונות משתמשים במודלים של LSTM (long short-term memory networks) כ(Erkartal and Yılmaz, 2022) המשתמש בלמידה עמוקה כדי ללמוד את העבר ובכך לנסות לחזות את העתיד של המנייה על פי נתוני העבר. הפתרון הוא בעייתי משום שהבנו שבבורסה לפי השערת השוק היעיל בגרסתו החלשה (The efficient market hypothesis, EMH) אי אפשר להשתמש בנתוני העבר של מנייה כדי לחזות את עתידה כי השווי הנוכחי מכיל בתוכו כבר את עבר המנייה. פתרון נוסף הוא שימוש במודלים של ARIMA (autoregressive integrated moving average), שימוש במודל זה הוא בעצם שימוש בנתוני העבר ויצירת סטטיסטיקה שאיתה יהיה ניתן לחזות את עתיד המנייה, אך גם פתרון זה נתקל באותה הבעיה כמו הפתרון הקודם שכן אין אפשרות להשתמש בנתוני העבר כדי לנבא את עתיד המנייה.

הפתרון שאנו מציעים לחיזוי בשוק ההון הוא בניית מודל שיתבסס על נתוני העבר אבל החידוש בו יהיה שהמודל יתבסס בנוסף על נתוני העבר של ציוצים מרשת הטוויטר. במילים אחרות, מודל זה לא ינבא עתידה של המנייה רק על בסיס העבר שלה, אלא המודל ילמד אילו אנשים משפיעים הכי

הרבה ברשת החברתית, ואיזו סמנטיקה של אותם האנשים משפיעה הכי הרבה. על בסיס כל נתונים אלו, המודל יידע לנבא מה יקרה למניות ספציפיות לאחר התחשבות בציוצים ספציפיים

במילים אחרות: בפרויקט שלנו נחקור ונבין את הקשר בין הציוצים ברשת החברתית הפופולארית "טוויטר" על ידי בכירי העולם, לבין מידת השפעתם של אלה על המניות בשוק ההון. על מנת לבדוק קשר זה עלינו להיעזר בסדרה של פעולות אשר יהוו עבורנו אבני דרך עד למציאתו של אותו הקשר. לאחר שימוש מושכל באותן הפעולות שנציב בפנינו שיש לבצע, נוכל להגיע למסקנות נבונות אודות הקשר – האם קיים או לא. במידה ואכן נגלה כי ישנה השפעה של אותם הציוצים על המניות, נוכל לבנות מודל חיזוי מהימן אשר יוכל לנבא ולחזות את השינויים העתידיים של המניות בשוק ההון על ידי ציוצים עתידיים מטעמם של בכירי העולם. חיזוי זה יעניק לאנשים המשקיעים במניות יתרון רב, שכן באופן זה יהיה יותר קל וברור איפה ומתי הכי כדאי להשקיע את כספם במניות השונות. במילים אחרות, אופן חיזוי זה מעניק סיוע בניהול הסיכונים של אותם המשקיעים בשוק ההון ומגדיל את הוודאות של עתיד השקעתם, אם ירוויחו מעצם ההשקעה או שמא לא.

מטרות ויעדים

מטרת המחקר הינה להשתמש בנתוני פוסטים ובנתוני ציוצים קודמים מרשת טוויטר כדי לבדוק האם ישנה תלות בין ציוצים של אישי עולם בכירים ומשפיעים, לבין תנועות במניות בבורסה העולמית. במידה ונמצא כי אכן ישנו קשר, נתעסק בבניית מודל חיזוי וניבוי אשר יתבסס על קשר זה, כך שיצליח לנבא באופן מיטבי את השינויים העתידיים של המדדים המרכזיים והמניות.

במהלך המחקר, נדון בהשפעתם של אנשי ציבור ומנהיגים ידועים כדוגמת אילון מאסק, ביל גייטס, טים קוק, ג'ף בזוס, נשיא ארצות הברית ג'ו בידן ונשיא ארצות הברית לשעבר דונאלד טראמפ. בחרנו בדמויות ציבוריות אלה מאחר ולהם כוח רב בתחום הפוליטי, הטכנולוגי והעסקי, ויתכן כי ציוצם בטוויטר יכולים להוביל להשפעה בשוק המניות.

תקופת הזמן שבה המחקר מתמצאת הינו בין השנים 2017 ל-2020. מצאנו לנכון לסקור ולדון בתקופה זו ממטעמים של השפעת האירועים המדיניים, הכלכליים והטכנולוגיים הרבים שאירעו בה. מתוך כך, יתאפשר למחקר להתמקד בזמן מוגדר ובו נוכל למצוא קשרים בין הציוצים של אישים משפיעים אלה לבין התנועות בשוק המניות. במילים אחרות, הבחירה בזמן ספציפי עשוי להגביר את הדיוק ואת יכולות החיזוי של המודל.

המחקר יתמקד בשימוש בנתונים מהרשת החברתית "טוויטר" לצורך בניית מודל חיזוי שיתרום להבנת הקשרים בין ציוצים משפיעים ותנועות בשוק המניות ולמדדים מרכזיים כמו S&P 500. המחקר יבוצע בצורה אמפירית באמצעות התוכנות והכלים המתאימים לכך אשר עומדים לרשותנו.

במהלכו של המחקר, נתווה תהליך חקירה מקוון שיאפשר איסוף נתונים אודות טוויטר בצורה אוטומטית. תהליך זה יתאפשר בעזרת ממשקי תכנות (API) של טוויטר וכן כלים שונים לניתוח. נתונים אלו יחולקו למספר קבוצות על פי המשתנים הנבחרים. נבצע סינון של תגי חיפוש מקושרים לאנשי הציבור המבוקשים, למנהיגים ולמדדים המסקרים בתקופת הזמן שהגדרנו לצורך המחקר. תהליך זה יאפשר לנו להתמקד רק בנתונים הרלוונטים.

לאחר מכן, נבצע ניתוח סטטיסטי על הנתונים שנאספו, על מנת לזהות קשרים בין הציוצים לבין תנועות בשוק. ניתוח זה יוכל לעזור להבחין בקורלציות ושינויים המתחוללים מתוך הנתונים, ולהבין לאילו ציוצים של אישי עולם בכירים ישנה השפעה על כיווני השוק.

לאחר ניתוח הנתונים, במידה ונמצא כי אכן ישנו קשר בין ציוצים של אישיויות משפיעות בטוויטר לבין שינוי תנועות ושינויים בבורסה העולמית, נתעסק בבניית מודל חיזוי וניבוי אשר יתבסס על קשר זה. מודל חיזוי זה, יצליח לנבא באופן מיטבי את השינויים העתידיים של המדדים המרכזיים והמניות בשוק. עבור המשקיעים במניות, מודל זה עשוי להוות כלי אשר בידו לשפר את אופן ההתנהלות שלהם בבורסה מתוקף היותו מודל שמצליח לחזות תהליכים ושינויים עתידיים בשוק. במילים אחרות, מודל זה הינו כלי שימושי ויעיל להתנהלות מושכלת וחכמה יותר בתחום ההשקעות.

המשפיעים:

- אילון מאסק
- ביל גייטס
- טים קוק
- ג'ף בזוס
- נשיא ארצות הברית – ג'ו בידן
- נשיא ארצות הברית לשעבר – דונאלד טראמפ

מדדים:

- S&P 500

תקופות זמן:

- 2017-2020

סקירת ספרות

המחקר שלנו יתמקד במציאת דרך לחיזוי מוצלח של שוק ההון תוך שימוש ברשתות החברתיות ובאופן ספציפי תוך שימוש בציוצים של אישים חשובים בטוויטר. כיום יש כמה וכמה מחקרים בנושא אשר משתמשים בסמנטיקה של ציוצים, וברשתות למידה עמוקה כדי לנסות להבין את הקשר שבין הסמנטיקה של הציוץ לבין שוק ההון, אם קיים בכלל קשר כזה.

יש מחקר אשר משתמש ברשתות למידה עמוקה לצורך ביצוע ניתוח סמנטי של ציוצים מחשבון הטוויטר של אילון מאסק, המחקר משתמש ברשת למידה עמוקה הנקראת LTSM שהיא רשת נוירונים רפלקסיבית, המשמשת לתחומים רבים, הרשת דוגמת את הציוצים של אילון מאסק ומנתחת את הרגש של הציוץ, מסקנות המחקר הן שהרשת מצליחה בדיוק גבוה להגדיר מהי הסמנטיקה של הציוצים (Erkartal and Yılmaz, 2022.)

ניתוח סמנטי נוסף של ציוצים בטוויטר התבצע במחקר אחר שבו גם בדקו האם סמנטיקה של ציוצים יכולה לחזות את מחירי המניות. במחקר השתמשו בעצי החלטה ובמחלץ מאפיינים, המחקר הציג כי למרות שטוויטר מכיל המון מידע, אם עושים ניתוח נכון של ציוצים ויודעים להתעלם מרעשי הרקע, ניתן יהיה למצוא קשר בין השינויים בשוק ההון האמריקאי לבין הציוצים (Kordonis et al., 2016), גם במחקר של Mendoza-Urdiales et al. 2022 חוקרים את הקשר בין תגובות הציבור בטוויטר לבין ביצועי המניות בשוק ההון, במחקר זה החוקרים עשו שימוש בשיטות של Transfer Entropy ו- Exponential Generalized Autoregressive Conditional Heteroskedasticity (EGARCH), כדי לזהות את הקשר, EGARCH היא שיטה המשלבת בין הסתברויות וסטטיסטיקות ומיועדת לנתח תנודות של מחירים.

במחקר אחר שנערך על הבורסה בברזיל החוקרים בדקו האם קיימת השפעה בין ציוצים של חברות ברזיליות בטוויטר לבין מדדי הבורסה בברזיל, גם במחקר זה השתמשו במודל LSTM כדי לנתח את הרגשות ואת הסמנטיקה של הציוצים, במסגרת המחקר הוכחה השפעה בין ציוצי החברות לבין מדדי הבורסה הברזילאית (de O. Carosia et al., 2019).

מחקר נוסף שנערך על ציוצים מחשבון הטוויטר של נשיא ארצות הברית לשעבר דונאלד טראמפ (המחקר נערך בזמן כהונתו כנשיא), בדק את השפעתו של טראמפ על שוק המניות, המחקר השתמש במודל הנקרא (DIEN Deep Information Echoing Network) המשלב שלוש רשתות נוירונים שונות: רשת נוירונים אחת שמטרתה ללמוד את הפעילות הקודמת של המשתמש, שבמקרה הזה הוא טראמפ. רשת נוירונים שנייה שמטרתה ללמוד על מה שקורה בשוק ופעולות מסחר של משתמשי הטוויטר, ורשת נוירונים שלישית שמטרתה ללמוד על ההשפעות של טראמפ בכללי ובשוק המניות בפרט. המחקר משתמש גם באלגוריתם (GBDT Gradient Boosting Decision Tree) כדי לזהות את המאפיינים החשובים ביותר של הציוצים ולבנות מודלים לחזיית השוק. מסקנות המחקר הן כי המודל (DIEN) המבוסס על נתוני הציוצים בטוויטר יכול לחזות את התנהגות שוק ההון ברמת דיוק גבוהה, המודל מוכיח שהציוצים של טראמפ קשורים באופן ישיר להתנהגות המניות בבורסה באמצעות מאפיינים שונים.
במחקר נמצא כי ציוצים הזוכים להרבה ציוצים מחדש וסימוני אהבתי משפיעים יותר מאשר ציוצים שאינם זוכים לחשיפה כזאת, בנוסף נמצא כי ציוצים הנשלחים בשעות השוק בדרך כלל

משפיעים יותר מאשר ציוצים הנשלחים בשעות שאין בהן פעילות בבורסה, עוד נמצא שציוצים בעלי תוכן שלילי או שנוי במחלוקת משפיעים לרעה כל מחירי המניות. מאפיין נוסף הוא הסיקור התקשורתי של ציוצי טראמפ, ציוצים בעלי סיקור תקשורתי שלילי או סיקורים בעלי ביקורת על טראמפ יכולים להעצים את השפעת הציוצים על תשואות השוק. המחברים מצאו גם כי ההשפעה של ציוצים שונה בהתאם לתעשייה שעליה נכתב הציוץ, ציוצים על תעשיית הבריאות השפיעו הרבה יותר מאשר ציוצים על תעשיית הטכנולוגיה.(Yuan et al., 2020)

מהלך העבודה

כעת נפרק את השלבים של מהלך העבודה שלנו (תרשים 1). הסדר הוא כזה:

1. איסוף נתונים
2. עיבוד מקדים
3. ניתוח סנטימנט
4. יצירת דגמים
5. הערכה
6. השוואה וניתוח תוצאות
7. פרשנות

בשלב הראשון של איסוף הנתונים - באמצעות ה-API של Twitter ומשאבים פתוחים, אנו אוספים נתונים על מדד S&P 500 והציוצים של המשפיענים שלנו.

בשלב השני, אנחנו עושים את העיבוד מקדים לנתונים שאיתם אנו עובדים. כלומר, ציוצים יצטרכו להתנקות מרעשי מידע ומדברים שיכולים להשפיע לרעה על הערכת הסנטימנט. אנו משנים את האימוג'י כדי לתאר את האימוג'י בטקסט כדי לשמור על הנתונים אך להפוך אותם לקריאה יותר. בעזרת ביטויים רגולריים אנו מסירים מהטקסט קישורים, תמונות מקודדות וציוצים המכילים רק אימוג'י וללא טקסט. היה גם שלב שעוקב אחר ניתוח הסנטימנט בפועל, אבל יכול להיחשב גם כעיבוד מקדים. הוא משלים על הימים החסרים. מה זה אומר: לאחר ניתוח סנטימנטים, אנו לוקחים את הממוצע של כל הציוצים שנוצרו על ידי משפיע ביום. ומספר זה ישמש במודל לאחר מכן. אבל לא כל משפיע שיש לנו הוא כמו דונלד טראמפ ויש לו נתונים לכל יום. אז מילאנו את הימים ההם ברגרסיה ליניארית.

השלב השלישי הוא לנתח את הסנטימנט של כל ציוץ בנפרד באמצעות ספריית ויידר. כל ציוץ יקבל ניקוד על שיפוע מ-1 עד 1 כאשר -1 הוא ציוץ שלילי לחלוטין ו-1 הוא ציוץ חיובי לחלוטין. לאחר מכן יילקחו הציוצים של היום ומהם יערך הערך הממוצע של הסנטימנט לאותו יום.

בשלב הרביעי ייווצרו מודלים LSTM. ראשית, נחפש היפרפרמטרים מתאימים למודל. נחפש אותם באמצעות מודלים של LSTM המשתמשים בנתונים פיננסיים בלבד.

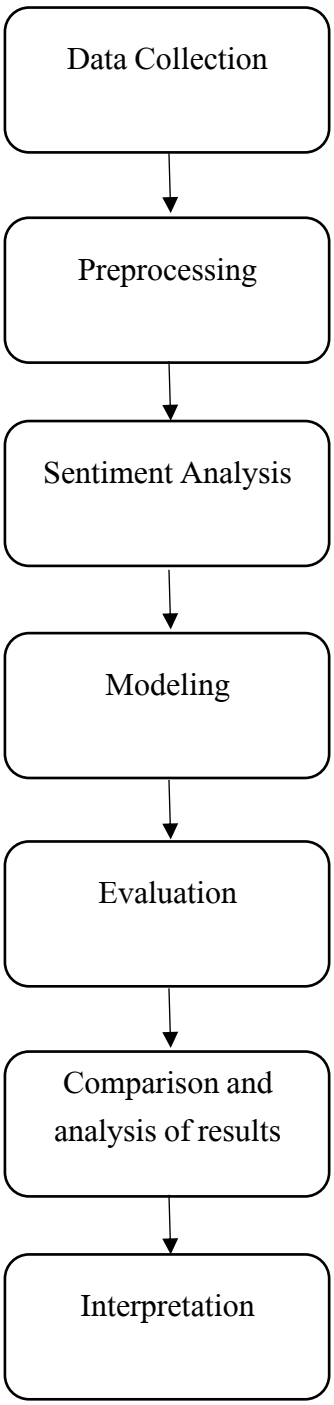
על ידי הקמת קו בסיס עם מודלים של "ללא סנטימנט", נוכל להגדיר היפרפרמטרים שעובדים טוב רק עם נתונים פיננסיים. זה מספק השוואה הוגנת ומבודד את ההשפעה של ניתוח סנטימנטים על ביצועי המודל, שכן כל שיפור שנראה במודלים מודעים לסנטימנט עשוי לנבוע מהוספת נתוני סנטימנט במקום פשוט לכוונון היפרפרמטרים. נקבל גם מודל ללא השפעת סנטימנט להשוואת התוצאות ולהערכת השפעתם של משפיענים.

לאחר מכן, אנו יוצרים מודלים המבוססים על הסנטימנט של כל משפיע. כלומר - בסך הכל יהיו לנו 6 מודלים LSTM.

בשלב הבא נבחן את הביצועים של כל אחד מ-6 הדגמים לפי המדדים שנבחרו.

בשלב השישי אנו מנתחים את התוצאות, מעריכים את המדדים של כל המודלים ומשווים אותם אחד עם השני ועם המודל ללא השפעת סנטימנט.

השלב האחרון בעבודתנו הוא הפרשנות של התוצאות. בהתבסס על הנתונים שהתקבלו, עלינו להסיק מסקנות ולהציע פיתוח עתידי של הנושא ושיפורים אפשריים לגישה זו.



תרשים 1 מהלך הפרוייקט

שיטות

בפרוייקט יש מספר אבני דרך עיקריות שנסקור כעת (תרשים 1):

1. <u>איסוף הנתונים:</u> אבן הדרך הראשונה של הפרויקט היא איסוף הנתונים, הנתונים שאנו משתמשים בהם בפרוייקט זה הם נתוני הציוצים של הידוענים שאותם נבדוק ונתוני המנייה שאחריה אנו עוקבים (s&p500) כדי לאסוף את נתוני המנייה נשתמש בAPI של חברת יאהו כדי להוריד את הנתונים ולהעביר אותם לקובץ csv (נספח 1), את נתונים הציוצים נוריד מאתר github אחרי שהם הוצאו מהטוויטר על ידי שימוש בAPI של טוויטר (נספח 2) הנתונים שלנו יתחילו מפברואר 2017 ועד סוף שנת 2020

2. <u>עיבוד מקדים:</u> השלב השני הוא שלח העיבוד המקדים, בגלל שהנתונים שיש לנו הם גולמיים אנחנו צריכים "לנקות" אותם ולהוציא את כל מה שאין לנו צורך או את כל מה שיכול להפריע ולכן בשלב העיבוד המקדים בתחילה אנחנו הופכים את כל הנתונים לאחידים כלומר כל הקבצים נשמרים באותה התבנית, משנים את שמות העמודות כדי שיהיו זהות בין כל המשפיענים שיש לנו (נספח 3). בנוסף יש צורך בניקוי כל הקישורים שיש לנו והמרה של כל האימוג'י הקיימים בציוצים לטקסט שאפשר להוציא ממנו סנטימנט (נספח 4), המרת כל הזמנים לתאריכים בתבנית זהה, ושמירה של כל קבצי הציוצים (נספח 5).

3. <u>ניתוח נתונים חקרניים:</u> אבן הדרך השלישית היא ביצוע ניתוח נתונים חקרני, בשלב זה ננתח את הנתונים ואת הסנטימנט של הנתונים כדי לקבל ציון מספרי על כל יום של ציוצים, בתחילה נוציא סנטימנט של כל ציוץ בעזרת TextBlob (נספח 6) וגם בעזרת Vader (נספח 7) בהמשך נבחר רק בשיטה אחת אחרי יצירת סנטימנט על כל ציוץ בנפרד ניצור ממוצע יומי של סנטימנט בכדי שעל כל יום יהיה לנו רק ציון יומי אחד (נספח 8), הבעיה שכעת אנחנו נתקלים בה היא חוסר של ציון יומי, בגלל שקיימים ימים שבהן אין אף ציוץ לאותו המשפיען אז אין לנו ציון סנטימנט יומי של ציוצים לכן אחרי בדיקה הדרך שהחלטנו למילו החוסרים היא באמצעות רגרסיה לינארית באותם הימים, בעיה נוספת היא הימים שבסוף ובתחילת הנתונים, בימים האלו אם חסר ציוצים אז אין דרך לעשות רגרסייה לינארית לכן הוחלט למלא את הימים הללו בנתונים הקרובים ביותר אליהם (נספח 9). לאחר האנליזה של הנתונים וקבלת הסנטימנט היומי נבדוק איזו שיטת ציון סנטימנטים היא העדיפה, הוצאנו גרפים של ניקוד הסנטימנטים של כל ידוען באמצעות כל שיטה (תרשימים 12-23) ולבסוף בחרנו בשיטת Vader משום שקיים בה פיזור יותר גדול של הציונים על רוב הידוענים שבחרנו (טבלה 3).

4. <u>בניית מודלים:</u> אבן הדרך הרביעית היא בניית מודל חיזוי שיכול לחזות את השינויים בשוק המניות בהתבסס על הציוצים. נשתמש ברשתות עצביות LSTM, שהן סוג של אלגוריתם למידה עמוקה שיכולה למדל נתונים רציפים כמו סדרות זמן. בתחילה אימנו כמה מודלים עם פרמטרים שונים רק על נתוני המנייה כדי לראות מה הפרמטרים הטובים ביותר למודל

ומה ציון הבסיס שאותו אנחנו שואפים לשפר, אחרי שמצאנו את המודל עם הציון הטוב ביותר (תרשים 24 וטבלה 4) הוספנו לעשות ובדקנו גם את הפרמטר שאומר למודל כמה ימים אחורה להסתכל כדי לתת את החיזוי והגענו למודל אפילו טוב יותר (תרשים 25 וטבלה 5), כעת כשיש לנו פרמטרים לרשת הלמידה העמוקה ויש לנו ציון בסיס שאותו אנחנו מעוניינים לעבור אפשר להתחיל לעבוד על המודל העיקרי שהוא מודל LSTM שמשתמש גם בסנטימנט של הידוען.
כדי לבדוק כמה ימים אחורה אנחנו צריכים להסתכל במודל עם הסנטימנט השתמשנו במבחן סטטיסטי הנקרא granger causality המבחן הנ"ל בודק את הקשר בין סדרות זמן וגם את ה"דילייי" שבינהן והגענו לתוצאות שהן לרוב 1 (טבלה 6) ולכן במודל עם הסנטימנט נסתכל יום אחד אחורה.
אחרי שיש לנו את כל הנתונים ואת כל הפרמטרים אפשר להכניס אותם למודל ולהוציא מודל חיזוי על כל המשפיענים שיש לנו סך הכל 6 מודלי חיזוי (טבלה 7 ותרשימים 26 - 31).

5. <u>הערכה</u>: אבן הדרך החמישית היא להעריך את המודל החזוי באמצעות מדדים שונים כגון טעות מוחלטת ממוצעת, שגיאה ממוצעת בריבוע ו R-squared score. נשווה את הביצועים של המודל עם מודלים בסיסיים ועם המודלים ללא הסנטימנט ונבדוק האם התוצאות שקיבלנו כעת טובות יותר מהתוצאות האחרות והאם ניתן לשפר עוד את התוצאות.

6. <u>פרשנות</u>: אבן הדרך האחרונה היא לפרש את תוצאות הניתוח ולהסיק מסקנות. נזהה את הגורמים המשפיעים על השינויים בבורסה ואת הקשר שלהם עם הציוצים. כמו כן, נדון במגבלות המחקר ונציע כיוונים עתידיים למחקר.

כלים ושיטות:

1. שפת תכנות: Python נשתמש ב-Python לאיסוף נתונים, עיבוד מוקדם של נתונים, ניתוח נתונים ופיתוח מודלים. לפייתון יש מערכת עשירה של ספריות כמו פנדה, numpy, matplotlib, seaborn, sikit-learn ו-tensorflow, המספקות כלים רבי עוצמה למניפולציה של נתונים, הדמיית נתונים ולמידת מכונה.

2. טוויטר API: נשתמש ב-Twitter API כדי לאסוף ציוצים על סמך מילות מפתח, האשטאגים וחשבונות משתמש ספציפיים. ה-API של Twitter מאפשר לנו לגשת לארכיון הציוצים המלא ולסנן אותם לפי קריטריונים שונים.

3. API פיננסי: נשתמש בממשק API פיננסי כדי לאסוף נתוני שוק המניות כגון מחירי מניות יומיים, מחזורי מסחר ויחסים פיננסיים. ה-API הפיננסי מספק נתונים בזמן אמת והיסטוריים עבור מכשירים פיננסיים שונים כגון מניות, אופציות, חוזים עתידיים ומטבעות.

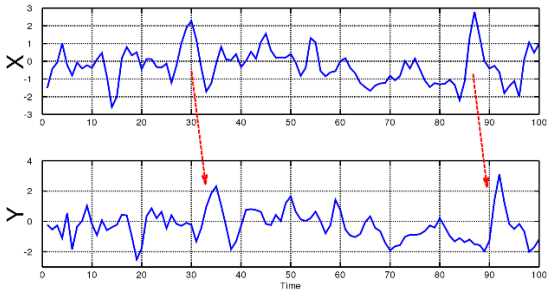
4. רשתות עצביות: LSTM נשתמש ברשתות עצביות LSTM כדי למדל את הקשר בין הציוצים לשינויים בשוק המניות. רשתות עצביות LSTM הן סוג של רשת עצבית חוזרת (RNN) שיכולה ללכוד את התלות ארוכת הטווח בנתונים עוקבים (תרשים 5), המשמשת בדרך כלל במשימות עיבוד שפה טבעית (NLP) כגון תרגום מכונה, סיכום טקסטים ומענה לשאלות.
הרשת עובדת באמצעות מנגנוני שער כדי לשלוט בזרמית המידע דרך הרשת, היא מורכבת משלושה מנגנוני שער: שער כניסה, שער שכח ושער יציאה.
שער הקלט שולט על כמות המידע החדש הנכנס אל הרשת ומתווסף למצב תא, שער השכח שולט על כמות המידע שנשכח ממצא התא ונמחק מהרשת ושער היציאה שולט על כמות המידע היוצא ממצב התא.
בגלל המנגנון שבו בנויה רשת LTSM הרשת מסוגלת ללמוד תלות ארוכת טווח מכיוון שהיא יכולה לזכור מידע במשך תקופה ארוכה בניגוד לרשתות עצביות אחרות כמו RNN שיכולות לזכור מידע לפרקי זמן קצרים.

5. כלים להדמיית נתונים: נשתמש בכלים להדמיית נתונים כגון matplotlib ו- seaborn כדי ליצור תרשימים וגרפים שיעזרו לנו לחקור את הנתונים ולהעביר את התוצאות. הדמיית נתונים היא חלק חשוב בניתוח נתונים מכיוון שהוא עוזר לנו להבין את הדפוסים.

6. TextBlob: ספריית פייתון לעיבוד נתונים טקסטואליים, משמשת בעיקר למשימות עיבוד שפה טבעית, כגון ניתוח סנטימנטים, תיוג חלקי דיבור וזיהוי ישויות. בפרויקט שלנו נשתמש בספרייה הזו כדי לנתח את הסנטימנטים של הטקסט ובכך להבין אם הציוץ חיובי או שלילי ולסווג את הציוצים לפי הסנטימנט שלהם.

7. Vader: שיטה לביצוע sentiment analysis השם של השיטה הוא ראשי תיבות של Valence Aware Dictionary and sEntiment Reasoner השיטה מבוססת על מילון השמור מראש בשיטה שבו לכל מילה יש ניקוד במדד ועל ידי כך השיטה מדרגת את הסנטימנט, לשיטה זאת קיימת גם ספריית פייתון שהשתמשנו בה בפרוייקט זה ודרכה ניתן לקבל ציון סנטימנט על טסקסטים.

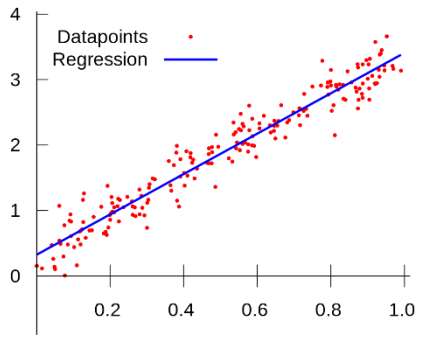
8. granger causality: עיקרון סטטיסטי המשמש לניתוח היחס בין שני סדרי זמן ולקביע האם קיים בכלל קשר ביניהם. השיטה מנסה לבדוק האם סדר זמן אחד יכול לשמש לחיזוי עתידי של סדר הזמן השני, והאם ההיפך נכון. זהו כלי נפוץ בתחומים רבים, ובפרט באקונומטריקה ובמחקרי התנהגות.
כאשר מכנים את היחס "Granger causality", הכוונה היא שהתהליכים בשני סדרי הזמן מתנהגים כאילו יש להם קשר גורם, אפילו אם לא ניתן לקבוע במדויק מהו הקשר האמיתי ביניהם.
עם זאת, התוצאות יכולות להתרחש מקשרים סטטיסטיים שאין באמת משמעות ביניהם. חשוב להיות זהיר ולהבין את ההשלכות של הניתוח ולא לסמוך עליו בלבד. במחקרים מדעיים, מומלץ לשלב אותו עם ניתוחים נוספים על מנת להגיע למסקנות מבוססות יותר.



תרשים 2 granger causality

9. Keras: ספריית Keras הינה ספרייה ללמידת מכונה פופולרית בפייתון, הספרייה מאפשרת לבנות בקלות רשתות נוירוניות ולאמן אותן.

10. רגרסייה לינארית: שיטה מתמטית למציאת הפרמטרים של הקשר בין שני משתנים, בהנחה שהקשר ביניהם הוא קשר לינארי.
השיטה משמשת לניתוח מדגמים סטטיסטים והנוסחה שלו היא נוסחת הקו הישר העובר דרך נקודות המדגם, לרוב כל הנקודות לא נמצאות על הקו ולכן הקו מחושב בצורה כזאת שסכום ריבועי המרחקים של הנקודות מהקו הוא הקטן ביותר.



תרשים 3 רגרסייה לינארית

11. R2 score: מדד המשמש להערכת ביצועי מודל רגרסיה וקובע את טיב המודל, ברגרסיה מקדם הקביעה R2 הוא מדד סטטיסטי של עד כמה תחזיות הרגרסיה מתקרבות אל הנתונים האמיתיים, ככל שהמקדם יותר קרוב ל1 התוצאות טובות יותר.

$$R^2 = 1 - \frac{SS_{RES}}{SS_{TOT}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

משוואה 1 R2 score

12. Mean Absolute Error: מדד המשמש להערכת ביצועי מודל רגרסיה. מודד את הפער בין ערכי היעד הצפויים לבין הערכים שניתנו על ידי המודל. יתרון ה-MAE שהוא מתייחס לשונות בצורה פוזיטיבית בלבד, כלומר, מחשב את הפער הממוצע בין הערכים בלבד, בלי להשוות ערך משלילי לערך חיובי.

$$\mathbf{MAE} = \frac{1}{n}\sum_{i=1}^{n} |\boldsymbol{x}_i - \boldsymbol{x}|$$

משוואה 2 Mean Absolute Error

13. Root Mean Squared Error: שגיאה המודדת את ההבדל הממוצע בין הערכים החזויים של המודל הסטטיסטי לבין הערכים שקיבלנו בפועל. מבחינה מתמטית זוהי סטיית התקן של המרחקים בין הנקודות לבין קו הרגרסיה, ככל שהמספר נמוך יותר כך אנחנו יודעים שהמודל שלנו מדייק יותר.

$$RSME = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{N - P}}$$

משוואה 3 Root Mean Squared Error

בסיס הנתונים

בפרק זה, נתמקד בהצגת הנתונים המרכזיים של הפרוייקט. נבין איך הנתונים נאספים, אילו כלים משמשים בניתוח, וכיצד הם משפיעים על המסקנות וההמלצות שלנו. נעמוד על החשיבות בעבודה עם נתונים ונדגים שיטות לניקוי ועיבוד הנתונים לצורך קבלת תוצאות מדויקות. בפרק זה, תקבל הקדמה ראשונית לנתונים המרכזיים, שמהם יוצאים התובנות והממצאים המרכזיים של הפרוייקט.

בתחילת נאסוף את נתוני המניות באמצעות הAPI של אתר יאהו פיננסים, את המידע אנחנו אוספים מ01-02-2017 ועד תאריך 31-12-2020 סך הכל 986 ימי מסחר (נספח 1). את נתוני הציוצים של הידוענים אספנו מאתר Github לכל ידוען יש קובץ עם טווח שנים אחר (נספח 2), לקחנו את אותו הטווח כמו טווח המניות אלא שבציוצים יש טווח של 1429 ימים משום שבימים שאין מסחר עדיין יש ציוצים, התאמנו את נתוני הציוצים לימי המסחר והחסרנו את הימים שבהם היו ציוצים אבל לא היה מסחר בבורסה (נספח 5).

טבלה 1 כמות ציוצים לפי משפיען בטווח

משפיען	כמות ציוצים בטווח שלנו
Trump	22855
Biden	5044
Jeff	154
Tim	540
Bill	1095
Musk	9658

אחרי שלב איסוף הנתונים ואחרי שצמצמנו את כל הנתונים שיש לנו לטווח המסויים שעליו אנחנו עובדים בפרוייקט הגיע השלב של ניקוי הנתונים, זהו שלב אחד לפני האנליזה של הנתונים, השלב הזה כולל שינוי שמות העמודות ככה שעמודות הזמן והציוצים יהיו באותו השם אצל כל הידוענים (נספח 3) וזריקה של שאר העמודות שיש בקבצים.

כעת, אנו מוחקים את כל הציוצים המכילים רק קישורים, בנוסף אנחנו ממירים את האימוג׳ים הקיימים בציוצים לטקסט באמצעות ספרייה הקיימת בפייתון (נספח 4).

אחרי ניקוי הנתונים הגענו לשלב האנליזה עצמו, בשלב זה נשתמש בשתי שיטות לאנליזת סנטימנטים השיטה הראשונה היא TextBlob (נספח 6) והשיטה השנייה היא Vader (נספח 7), ובהמשך נחליט עם איזו שיטה להמשיך את הפרוייקט.

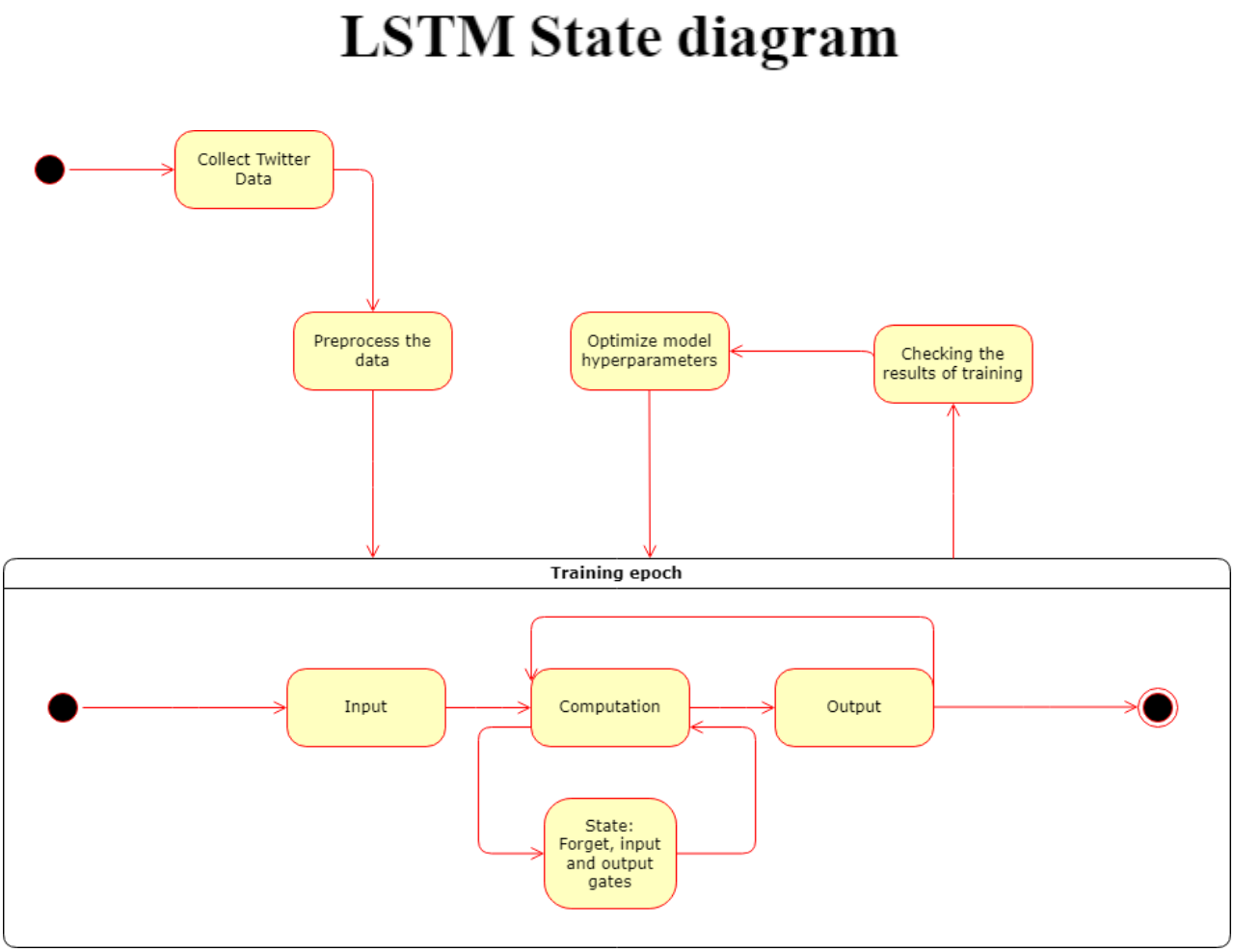
בשתי השיטות אנחנו עוברים על כל הציוצים של כל ידוען ונותנים ניקוד של סנטימנט לכל ציוץ ולבסוף אחרי שנתנו ניקוד לכל הציוצים של אותו הידוען מקבצים את הציוצים ועושים ממוצע יומי כך שיש ציון יומי לכל ידוען, הניקוד ששתי השיטות נותנות נע בין 1- ל1 כך ש1- סנטימנט שלילי מאוד ו1 הוא סנטימנט חיובי מאוד.

טבלה 2 ממוצע סנטימנטים

משפיען	ממוצע כללי TextBlob	ממוצע כללי Vader
Trump	0.114	0.139
Biden	0.144	0.233
Jeff	0.257	0.586
Tim	0.367	0.735
Bill	0.243	0.435
Musk	0.122	0.182

עכשיו כשבידנו כל הנתונים הסטטיסטים של מאגר הנתונים אפשר לראות שכאשר עושים אנליזות לפני 2 השיטות בשיטה אחת מקבלית סטיית תקן גבוהה יותר אצל רוב הידוענים, כלומר הפיזור של הנתונים לפי שיטת הסנטימנט Vader היא הרבה יותר מגוונת במרחב שבין 1- ל1 ולכן זאת השיטה שנמשיך איתה מהסיבה הפשוטה שככל שהנתונים מפוזרים יותר בטווח ולא אחידים סביב הממוצע נוכל ללמד את המודלים שלנו יותר טוב והסנטימנט ילקח בחשבון על ידי המודל בצורה טובה יותר.

תרשים מעבר מצבים / STD למודל-LSTM:

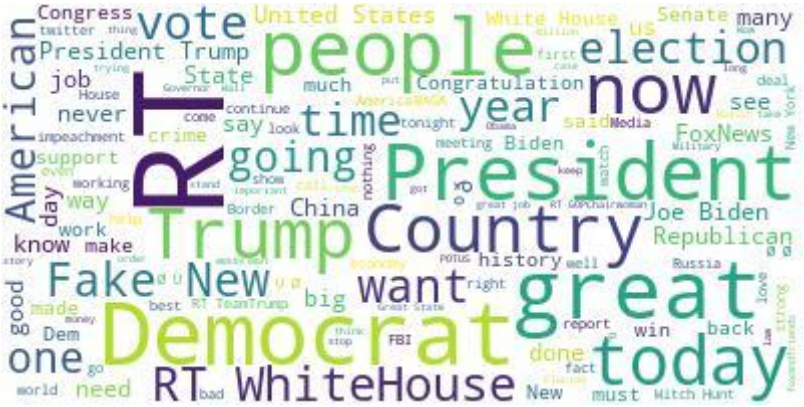


תרשים 4 מעבר מצבים

Word clouds

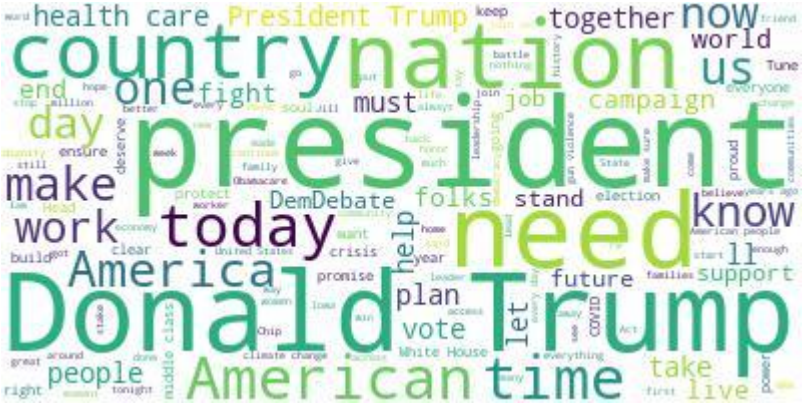
ענן מילים הוא ייצוג חזותי של נתוני טקסט בהם מילים מסודרות בגדלים שונים בהתאם לתדירותן בטקסט הקלט. מילים תכופות יותר מופיעות גדולות יותר ובולטות, בעוד שמילים תכופות פחות נראות קטנות יותר. הוא מספק דרך מהירה ואינטואיטיבית לזהות את מילות המפתח, הנושאים או הנושאים החשובים ביותר בטקסט, מה שהופך אותו לכלי פופולרי לסיכום והצגה של מידע טקסטואלי.

ניתוח ענן המילים (תרשימים 6-11) עבור כל אחד מהמשפיעים שלנו הוא מעניין מכיוון שהוא מספק תיאור מהיר וויזואלי של המילים הנפוצות ביותר שלהם, ומאפשר לנו לזהות נושאים וחוזרים ונושאים שמופיעים באופן בולט בציוצים שלהם. זה מאפשר לך לקבוע בצורה ויזואלית ופשוטה את הכיוון האידיאולוגי של הציוצים של כל משפיע.

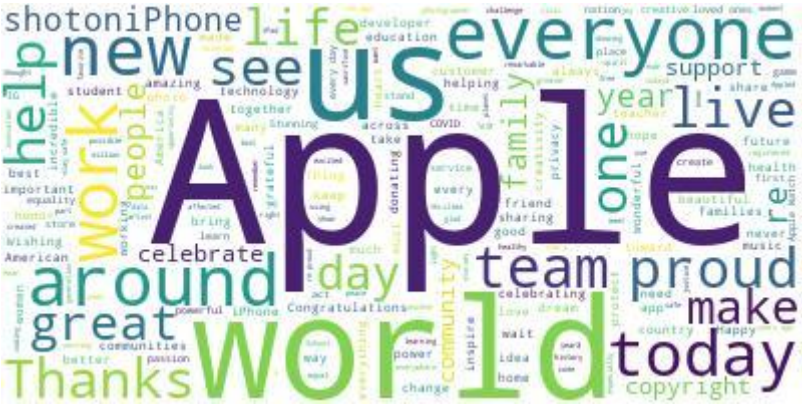


תרשים 5 ענן מילים דונלד טראמפ

תרשים 6 ענן מילים אילון מאסק



תרשים 7 ענן מילים ג'ו בידן



תרשים 8 ענן מילים טים קוק

תרשים 9 ענן מילים ג'ף בזוס

תרשים 10 ענן מילים ביל גיטס

עיבוד מקדים

בשלב העיבוד המקדים, המטרה היא לבצע עיבוד ראשוני ומהיר על הנתונים הגולמיים, כדי להכין אותם לשלבי העיבוד הבאים בתהליך. בשלב זה, אנחנו מבצעים סדרת פעולות כדי לטפל בנושאים כמו ניקוי הנתונים, הכנתם לעיבוד נוסף, ויצירת תבניות יחידות על פני כל הנתונים. השלב כולל את השלבים הבאים:

1. איחוד ותיקון הנתונים: נעביר את כל הנתונים לפורמט אחיד ונתקן פורמטים שונים של מידע. כך שכל הקבצים יהיו באותה התבנית. זה כולל תיקון שמות עמודות, טיפוח וסדר מאגרי הנתונים (נספח 3). נבצע סדר ובניית טבלאות אחידות כדי להבטיח שהנתונים יהיו קלים לניתוח ולעיבוד נוסף.
2. ניקוי וסינון: במהלך זה, אנחנו בודקים את הנתונים הגולמיים ומזהים רעשים, תווים מיותרים או מידע שאינו רלוונטי למטרת העיבוד. אנחנו מבצעים ניקוי כללי על הטקסטים ומסירים קישורים לא רלוונטיים ותווים מיותרים (נספח 4).
3. שמירת הנתונים: בסיום שלב העיבוד המקדים, נשמור את כל הנתונים על פורמטים מתאימים, כמובן תחת שמירה בטוחה ומאובטחת (נספח 5).

שלב העיבוד המקדים משמעותו להכין את הנתונים לשלבים הבאים בתהליך, בצורה יחידה ומטופחת כך שיהיה קל ויעיל להמשיך בטיפול וניתוח המידע בשלבים הבאים.

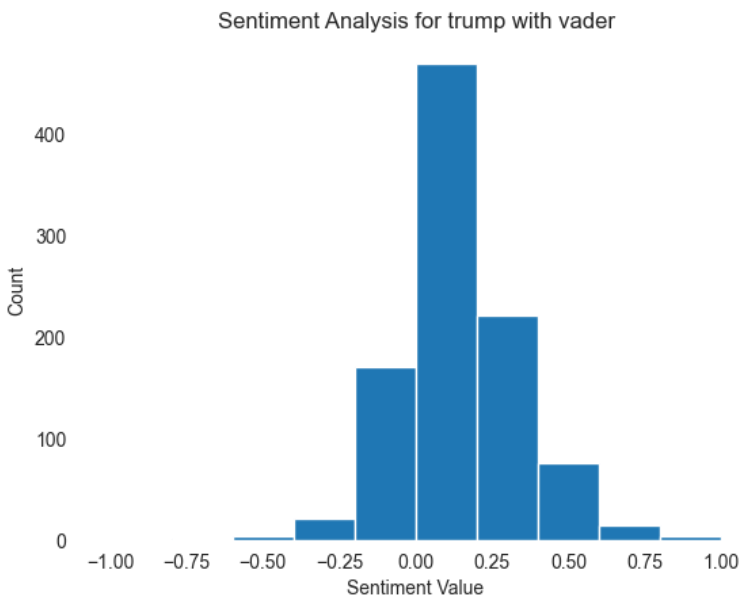
sentiment bar graphs

לאחר עיבוד מוקדם של הנתונים, נעבור לשלב ניתוח הסנטימנטים של הציוצים. כפי שהוזכר קודם לכן, באמצעות ספריית Vader, אנו מוצאים את הסנטימנט של כל ציוץ בנפרד ואז מחשבים את הממוצע היומי של ערכי הציוצים של אותו משפיען.
לשם השוואה, עשינו את אותו התהליך גם עם ספריית TextBlob, שנחשבת נאיבית יותר ולא נוצרה במיוחד עבור מדיה חברתית.

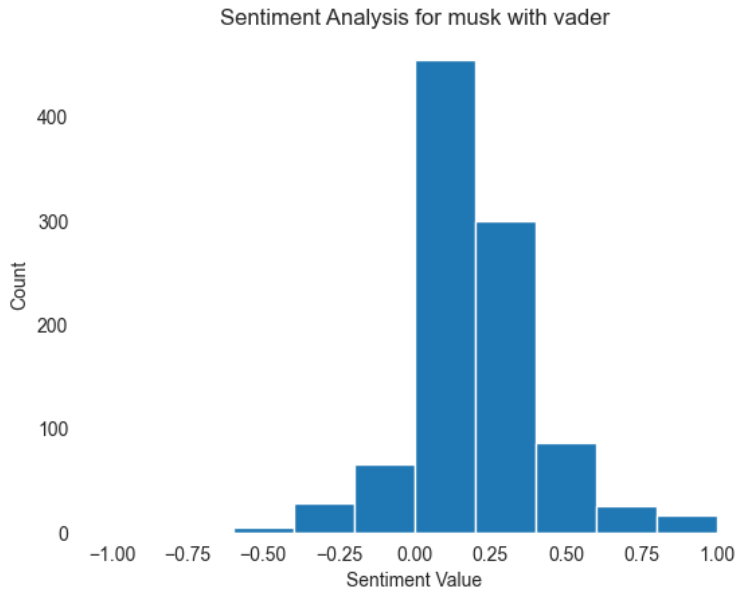
לפי הנתונים שהתקבלו משתי הספריות, ניתן לראות שההבדל בממוצעים ובסטיות התקן של ספריית Vader גדול יותר (טבלה 3). כמו כן, ל-TextBlob יש לעתים קרובות 1 ו-1- מוחלטים. טבלה זו מראה ש-Vader נותן מגוון רחב יותר של נתוני ניתוח סנטימנטים.

להלן גם כל תרשימי העמודות של כל אחד מהמשפיענים. אפילו מבחינה ויזואלית, אפשר להבחין כמה יותר מגוונת ופחות ריכוזית הערכת הסנטימנט היא בספריית Vader.

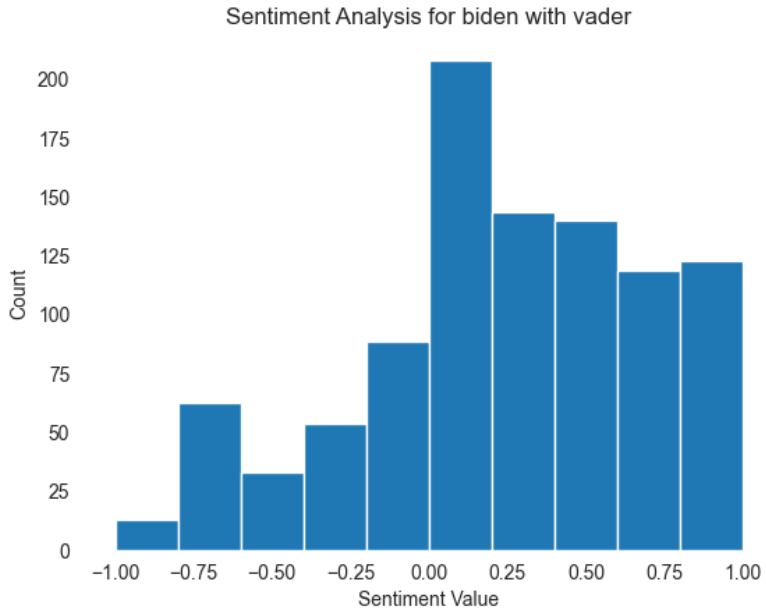
<u>Vader:</u>



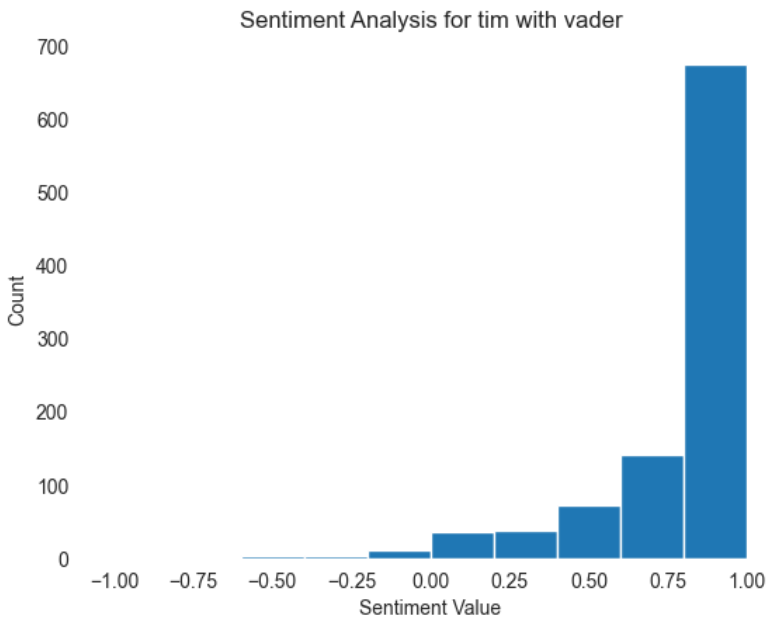
תרשים 11 סנטימנט לפי Vader לדונלד טראמפ



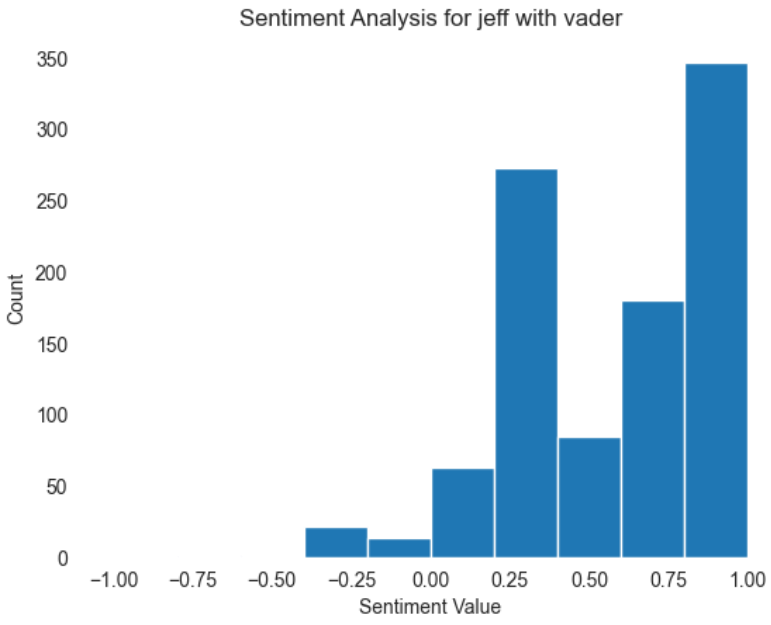
תרשים 12 סנטימנט לפי Vader לאילון מאסק



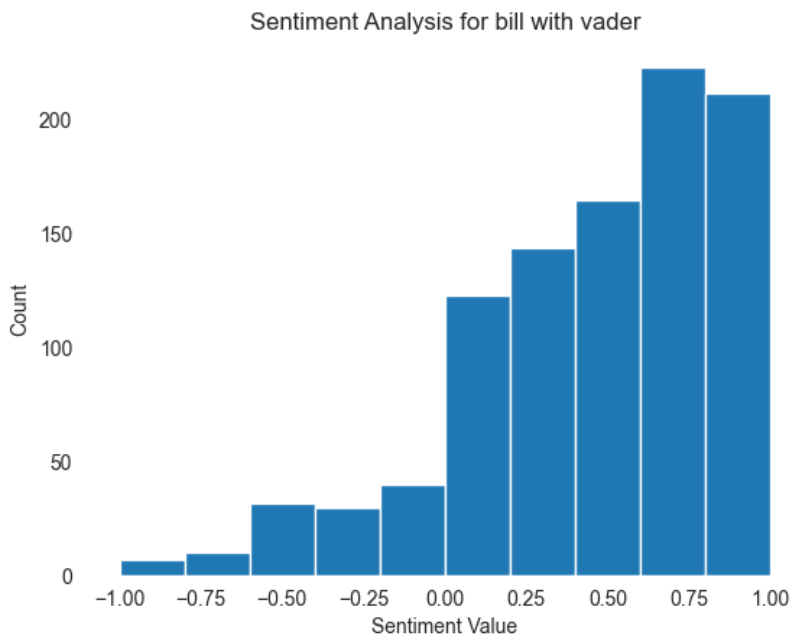
תרשים 13 סנטימנט לפי Vader לגיו בידן



תרשים 14 סנטימנט לפי Vader לטים קוק

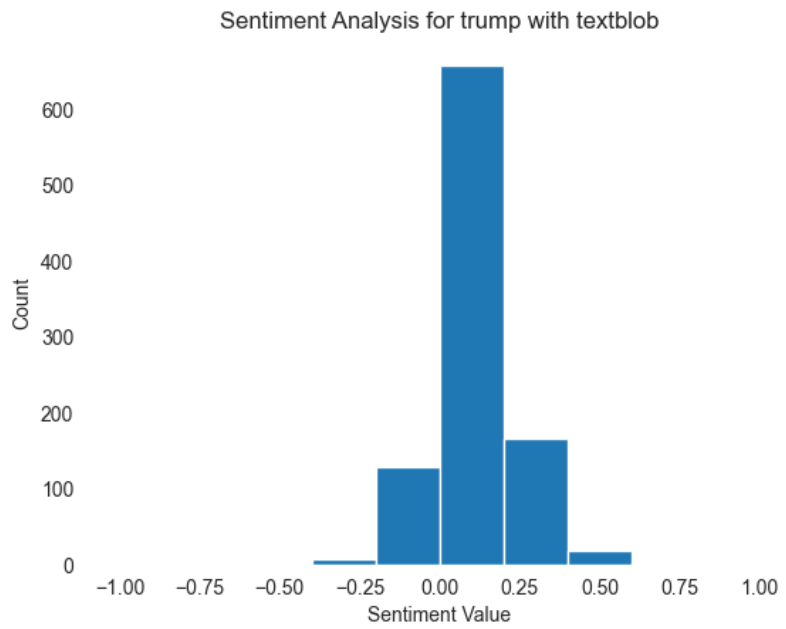


תרשים 15 סנטימנט לפי Vader לג'ף בזוס

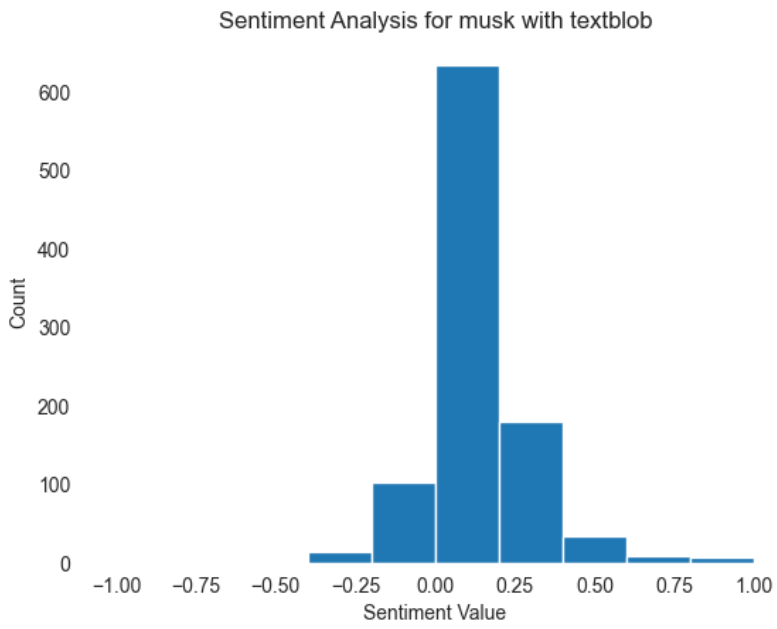


תרשים 16 סנטימנט לפי Vader לביל גייטס

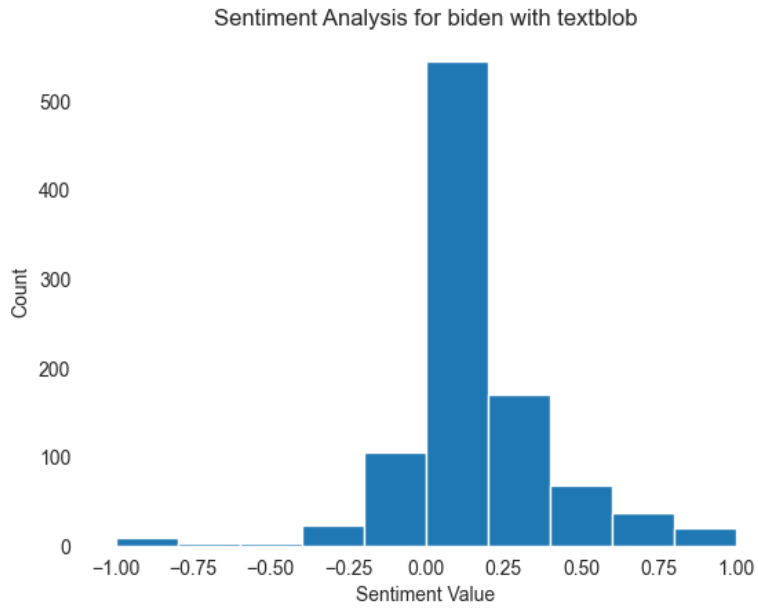
TextBlob:



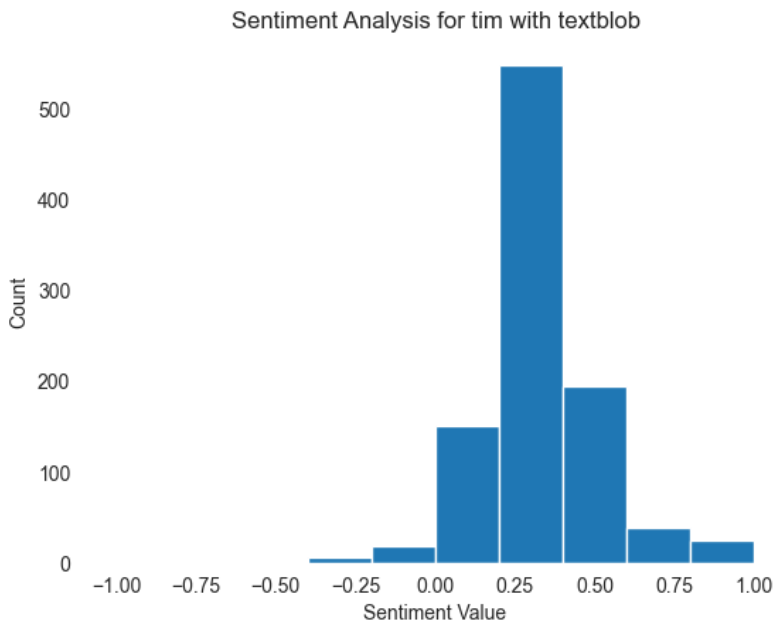
תרשים 17 סנטימנט לפי TextBlob לדונלד טראמפ



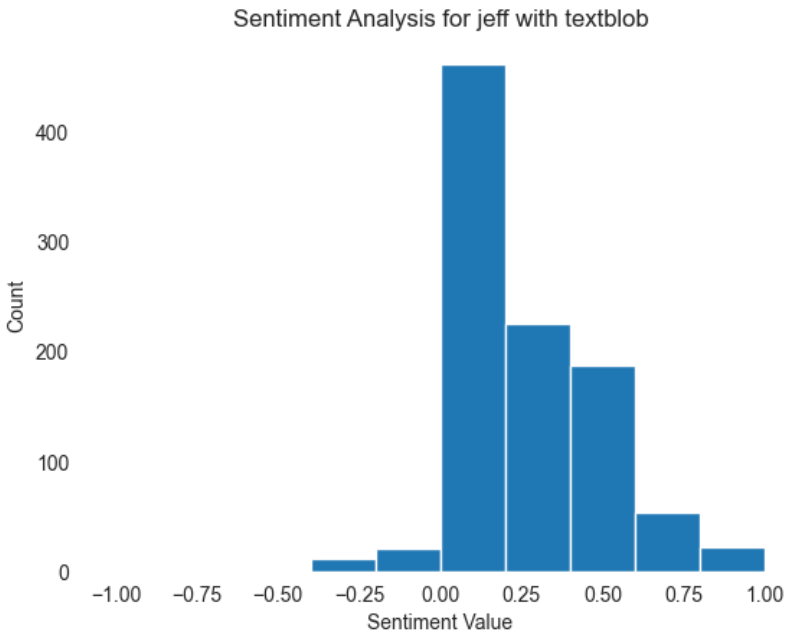
תרשים 18 סנטימנט לפי TextBlob לאילון מאסק



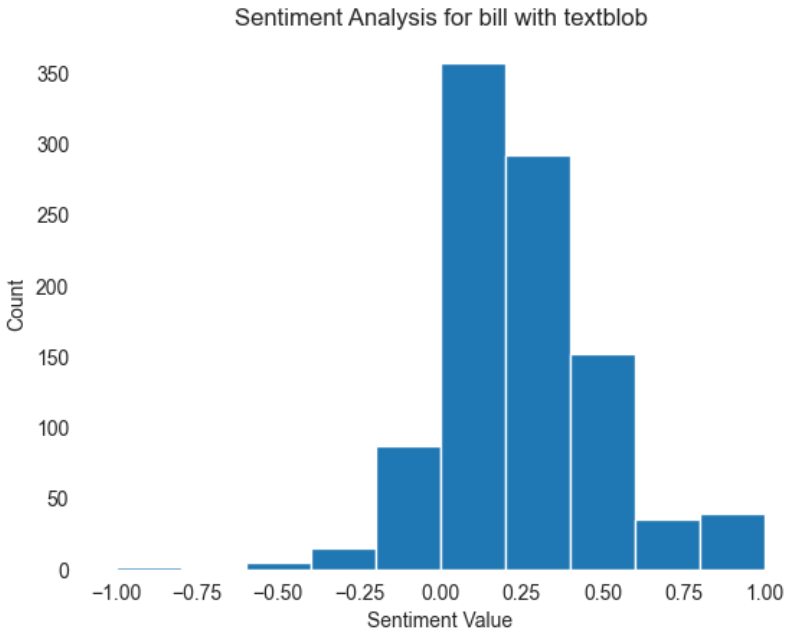
תרשים 19 סנטימנט לפי TextBlob לגיו בידן



תרשים 20 סנטימנט לפי TextBlob לטים קוק



תרשים 21 סנטימנט לפי TextBlob לג'ף בזוס



תרשים 22 סנטימנט לפי TextBlob לביל גיטס

טבלה 3 סטטיסטיקה של סנטימנטים

	Text Blob				Vader			
	Mean	Min	Max	סטיית תקן	Mean	Min	Max	סטיית תקן
Joe Biden	0.144	-1	0.9	0.248	0.233	-0.973	0.978	0.456
Elon Musk	0.122	-0.516	1	0.169	0.182	-0.765	0.921	0.215
Donald Trump	0.114	-0.427	1	0.129	0.139	-0.772	0.932	0.194
Tim Cook	0.367	-0.312	1	0.190	0.735	-0.726	0.968	0.252
Bill Gates	0.243	-1	1	0.258	0.435	-0.962	0.962	0.403
Jeff Bezos	0.257	-0.25	1	0.229	0.586	-0.690	0.981	0.326

מודל ללא סנטימנט

בפרק זה, נציג וננתח את התוצאות של מודל חיזוי באמצעות LSTM לקביעת מגמת המניות. המודל מבוסס רק על נתוני העבר של המניות ומשתמש בשיטת ה-LSTM (Long Short-Term Memory) כדי לנבא את מחירי המניות בזמן הבא. תהליך החיזוי מתבצע באמצעות למידת מניית הנתונים במהלך תקופות זמן קודמות, והמודל לוקח בחשבון במודל 6 ימים אחורה. המספר של 6 ימים נבחר בשלב זה באופן שרירותי כדי למצוא את הפרמטרים הטובים ביותר למודל.

החלוקה של הנתונים מתבצעת כך ש80 אחוזים מהנתונים ילכו אל האימון ו20 האחוזים הנותרים אל סט הבדיקה, את סט האימון נחלק שוב ל80 אחוזים של אימון ועוד 20 אחוזים של ולידציה. ניסינו 7 מודלים שונים כדי לבחון שכבות LSTM שונות (נספח 11) ואת ההשפעה שלהן על התוצאות ולבסוף לקחנו את המודל עם התוצאה הטובה ביותר.

טבלה 4 תוצאות מודלים ללא סנטימנט

Model parameters		R^2	MAE	RMSE
layer 1 (נוירונים)	layer 2 (נוירונים)			
120	100	0.931	0.040	0.051
128	128	0.926	0.041	0.053
256	256	0.868	0.054	0.071
200	100	0.917	0.040	0.050
256	128	0.914	0.044	0.057
120	105	0.923	0.038	0.049

אפשר לראות לפי התוצאות שהמודל הטוב ביותר עם ההסתכלות השרירותית שלנו ל 6 ימים אחורה הוא מודל 1 עם שכבה ראשונה של 120 נוירונים ושכבה שנייה של 100 נוירונים, המודל קיבל את הציון הטוב ביותר ולכן נמשיך לעבוד איתו ולשפר אותו.

משוואת המודל:

$$St = f(SP500t - i)$$

משוואה 4 מודל 1

כאשר $\boldsymbol{St}$ הוא הערך החזוי של הערך היומי של S&P 500 בזמן 't, $\boldsymbol{SP500t}$−i הן הערכים היומיים מהעבר בפיגור של i.
במודל שנבחר ערך ה i-הוא 8 (ימים)

בתרשים הבא (תרשים 24) ניתן לראות את תוצאות החיזוי של המודל כאשר מחיר המנייה האמיתי הוא באדום והמחיר שהמודל חוזה הוא בכחול.

פרק הזמן שעבר מתחילת המודל בימים

תרשים 23 גרף החיזוי של מודל 1

Layer (type)	Output Shape	Param #
lstm_8 (LSTM)	(None, 2, 120)	60000
lstm_9 (LSTM)	(None, 100)	88400
dense_4 (Dense)	(None, 1)	101

Total params: 148501 (580.08 KB)

Trainable params: 148501 (580.08 KB)

Non-trainable params: 0 (0.00 Byte)

תרשים 24 פרמטרי מודל 1

כעת נתמקד בבדיקת המודל המוצלח שבחרנו, ונבחן את יכולתו לנבא את מחירי המניות בהסתכלות אחורה בתקופות שונות. עד כה השתמשנו בהסתכלות אחורה של 6 ימים בלבד, אך כעת נרחיב את התצוגה ונבצע בדיקה על תקופה ארוכה יותר – מיום אחד של הסתכלות אחורה ועד ל14 ימים.

המטרה של הבדיקה הזו היא לראות כיצד תוצאות המודל משתנות ככל שזמן ההסתכלות אחורה משתנה ומציאת הפרמטר הטוב ביותר להסתכלות אחורנית במודל ללא סנטימנט וכך נוכל להשוות באמת את המודל עם התוצאה הטובה ביותר ללא סנטימנט עם שאר המודלים שנגיע אליהם ובהם אנחנו כן משתמשים בסנטימנטים.

המודל שבחרנו לבחון פה הוא המודל בעל התוצאה הטובה ביותר שמצאנו עד כה, מודל 1 בעל שכבה ראשונה של 120 נוירונים ושכבה שנייה של 100 נוירונים ואלה התוצאות שקיבלנו עם הסתכלות אחורה שונה:

טבלה 5 תוצאות מודל 1 לפי ימי הסתכלות

ימי הסתכלות אחורנית	R^2	MAE	RMSE
1	0.917	0.048	0.061
2	0.899	0.051	0.065
3	0.912	0.047	0.060
4	0.904	0.049	0.062
5	0.928	0.042	0.053
6	0.931	0.040	0.051
7	0.937	0.038	0.048
8	0.939	0.037	0.047
9	0.939	0.036	0.046
10	0.935	0.037	0.047
11	0.927	0.039	0.049
12	0.926	0.039	0.049
13	0.929	0.037	0.047
14	0.925	0.038	0.048

לפי הטבלה אפשר לראות שמספר ימי ההסתכלות בעל התוצאה הטובה ביות הוא 8 ולכן בהמשך הפרוייקט כאשר נתייחס לתוצאות של מודל ללא סנטימנט ניקח את התוצאות של מודל 1 עם 8 ימי הסתכלות אחורנית.

בתרשים הבא (תרשים 25) ניתן לראות את תוצאות החיזוי של המודל עם 8 ימי הסתכלות אחורנית כאשר מחיר המנייה האמיתי הוא באדום והמחיר שהמודל חוזה הוא בכחול.

תרשים 25 מודל 1 ללא סנטימנט עם 8 ימי הסתכלות אחורנית

granger causality

Granger Causality הוא מושג סטטיסטי הקובע אם סדרת זמן אחת יכולה לחזות שינויים בסדרת זמן אחרת. בהקשר של מודלי LSTM, הוא עוזר לזהות האם ערכי העבר של משתנה אחד (למשל, מחירים היסטוריים או סנטימנט של משפיענים) משפיעים באופן משמעותי על התנועות העתידיות של משתנה אחר (למשל, מחיר מדד S&P 500).

פיגור זמן בתוך מודל LSTM מתייחסים למספר שלבי הזמן או הימים שעברו שהמודל מחשיב לקלט. על ידי ביצוע בדיקת Granger Causality, נוכל לבחור את משכי הזמן המתאימים המציגים קשרים סיבתיים משמעותיים בין נתוני העבר לבין משתנה היעד (מחיר מדד S&P 500). שילוב משכי הזמן הרלוונטיים הללו כמאפייני קלט במודל ה-LSTM מאפשר לו ללכוד את ההשפעה של מחירי העבר והסנטימנט של משפיענים, מה שמוביל לשיפור התחזיות וביצועי תחזיות טובים יותר.

הטבלה הבאה מציגה את תוצאות הניתוח הסטטיסטי הזה ביחס לכל אחד מהמשפיעים (תוצאות מלאות בנספח 10).

טבלה 6 תוצאות Granger Causality

משפיען	מספר אופטימלי של פיגור (ימים)
Joe Biden	1
Elon Musk	2
Donald Trump	1
Tim Cook	1
Bill Gates	5
Jeff Bezos	1

לאחר המבחן הסטטיסטי של Granger Causality, לפי הרוב החלטנו להשתמש ב-1 כמספר הימים לאימון מודלים של LSTM עם ניתוח סנטימנט.

תוצאות

באופן מפתיע, כפי שנצפה מתוצאות כל המודלים, זה שמבוסס על הציוצים של אילון מאסק, למרות שיש לו גישה לכמות ניכרת של נתונים מדמות בולטת הקשורה קשר הדוק לחברות גדולות ולתעשיית ה-IT, המודל של מאסק הפגין תוצאה גרועה במעט מהמודל המבוסס על הציוצים של דונלד טראמפ. לשני המשפיעים היו ציוצים בשפע הזמינים לניתוח, מה שמצביע על כך שכמות הציוצים לבדה לא יכולה להיות הסיבה העיקרית להצלחת החיזוי במודלים של LSTM שלנו עם ניתוח סנטימנטים של הציוצים.

עם זאת, הגילוי המסקרן ביותר היה שהמודל שמשתמש בציוצים של ג'ף בזוס, למרות שיש לו מעט נתונים לעבוד איתן ביחס למודלים האחרים, הניב את התוצאות הטובות ביותר.

לכן, בעוד שההקשר והמשמעות של ציוץ אינדיבידואליים עשויים שלא להשפיע ישירות על המודלים הספציפיים של LSTM אלה, המוניטין הכולל של המשפיע, מומחיות התעשייה ודפוסי התקשורת עדיין יכולים להשפיע בעקיפין על הסנטימנט היומיומי. יתרה מכך, הרלוונטיות והאיכות של הציוצים עשויות למלא תפקיד בעיצוב הסנטימנט היומי, גם אם לא נלקח בחשבון באופן מפורש במודלים

למודלים של LSTM אין מדד דיוק פשוט כמו מודלים של סיווג, מכיוון שהם מנבאים משתנים מתמשכים ולא מחלקות קטגוריות. במקום זאת, אנו מסתמכים על R-squared, MAE ו RMSE- כמדדי הערכה עבור משימות טבלה, המציעות תובנות חשובות לגבי הדיוק והביצועים החזויים של המודל. בטבלה למטה נוכל לראות השוואה ישירה של כל המודלים שיצרנו לפי המדדים שנקבעו.

טבלה 7 תוצאות המודלים

model	R^2	MAE	RMSE
Jeff Bezos	0.971	0.026	0.035
Tim Cook	0.970	0.027	0.036
Donald Trump	0.970	0.027	0.036
Joe Biden	0.970	0.028	0.036
Bill Gates	0.961	0.033	0.041
Elon Musk	0.957	0.034	0.043
LSTM without sentiment	0.939	0.037	0.047

מהשוואה זו, מתברר כי המודל המבוסס על הציוצים של ג'ף בזוס השיג את הציון הגבוה ביותר בריבוע (R^2) R של 0.9715, מה שמצביע על כך שכ-97.16% מהשונות במדד המניות של S&P 500 ניתנת להסבר על ידי תחזיות של מודל זה. בנוסף, הוא השיג את ערכי השגיאה הממוצעת המוחלטת (MAE)ו Root Mean Squared Error (RMSE)-הנמוכים ביותר, מה שמציג את יכולתו לחזות את מחירי המניות עם סטייה מינימלית מהערכים בפועל.

משוואת המודלים:

$$St = f(SP500t - i, SIt - i)$$

כאשר $\boldsymbol{St}$ הוא הערך החזוי של הערך היומי של S&P 500 בזמן 't', $\boldsymbol{SP500t}$−i הן הערכים היומיים מהעבר בפיגור של הערך i, $\boldsymbol{SIt}$−i הם ערכי אינדקס הסנטימנט מהעבר, בפיגור של i. ערך ה i-שנבחר למודלים שלנו הוא פיגור של 1 ימים.

את הגרפים של מודלי החיזוי ניתן לראות בתרשימים הבאים (תרשימים 26-31) כאשר הגרף האדום הוא מחיר המנייה האמיתי והגרף הכחול הוא החיזוי של המודל, תוצאות סטטיסטיות של הגרפים ניתן לראות בטבלה 7.

תרשים 26 גרף חיזוי מודל עם סנטימנט ג'ף בזוס

תרשים 27 גרף חיזוי מודל עם סנטימנט טים קוק

תרשים 28 גרף חיזוי מודל עם סנטימנט דונלד טראמפ

תרשים 29 גרף חיזוי מודל עם סנטימנט ג'ו בידן

תרשים 30 גרף חיזוי מודל עם סנטימנט ביל גיטס

תרשים 31 גרף חיזוי מודל עם סנטימנט אילון מאסק

דיון

במחקר רצינו לבדוק האם קיים קשר בין הציוצים בפלטפורמה החברתית "טוויטר" של אישי עולם בכירים, לבין שינויים המתרחשים במניות בבורסה העולמית. כאמור, השערת המחקר הייתה שקיים קשר כזה. לאורך התהליך המחקרי, אספנו וניתחנו את המידע הרלוונטי עבור היתכנות הקשר, ומהממצאים שנאספו צלחנו בביסוס ואישוש השערתנו, שכן נמצא כי ישנו קשר בין הציוצים בטוויטר לבין שינויים שקורים בבורסה העולמית .

בנוסף, ניסחנו השערת מחקר נוספת והיא שככל שאישיות פופולארית מפרסמת כמות גדולה יותר של ציוצים בטוויטר, כך תגדל ההשפעה שלה על שוק המניות בבורסה. אמנם התבדינו לקשר זה מן הממצאים שנאספו, ראינו כי יש מקרים בהם רשת הנוירונים של בכירים בעולם שלא צייצו הרבה הינה מדויקת יותר וכי ה R-square score הינו גבוה יותר, מאשר אישים אחרים שכמות הציוצים שלהם הייתה גבוהה יותר. ניתן להסיק, כי השערה זו לא אוששה מאחר ויש יותר חשיבות לתוכן הציוץ מאשר כמות הציוצים.

תוצאות המחקר הראו כי ישנו קשר חיובי בין ציוצים בטוויטר לבין התנהלות המניות בבורסה. מאמרינו הציע מודל חיזוי מהימן אשר מקנה לאנשים המשקיעים במניות בבורסה עולמית, דרך יעילה ובטוחה לדעת כיצד ומתי הכי נכון להשקיע במניות בצורה כזאת שתחזיר את השקעתם בצורה המיטבית ביותר. מודל החיזוי משמש ככלי יעיל לסייע למשקיעים במניות לזהות תחומים פוטנציאליים להשקעה וכן להתמקד בתוכן של ציוצים ספציפיים שעשויים להיות קריטיים לשוק.

מחקרינו תרם להבנה ולחיזוק הקשר אשר מתקיים בין הציוצים לבין תנועות שונות בשוק המניות, ומבטא את החשיבות הגדולה שיש להבנה מעמיקה של הקשר שבין הרשתות החברתיות לבין השוק הפיננסי. בנוסף, המחקר יכול לעזור להבין בצורה טובה יותר את התהליכים שמתרחשים בשוק המניות ולהבין אילו גורמים עשויים להשפיע על תנועותיהן של המניות בשוק.

המקר פתח את הדלת לפיתוח מודלים חדשים ומתקדמים בתחום זה, כולל רשתות נוירונים ומתודולוגיות סטטיסטיות חדשות. המחקר יכול לעזור להעמיק במידה ולבצע ניתוחים מקיפים ומדויקים יותר לצורך הבנת ההשפעות וההתמודדויות עם מגמות בשוק המניות.

אנו מציעים לאנשים אשר מעוניינים לחקור תחום זה במחקרי המשך, לבדוק האם ישנם תחומים פיננסיים מסוימים בהם ההשפעה חזקה יותר מצד פעילותם של אישי עולם בכירים ברשתות החברתיות. כמו כן, ניתן לבדוק כיצד מאופיין קשר זה בחלוף הזמן וכן האם ישנם גורמים אחרים בסביבה הפיננסית אשר יכולים להשפיע על קשר זה.

<u>מגבלות המחקר :</u>

1. המחקר מתבצע בעבר ומתבסס על מידע המציין את הקשר בין ציוצים בטוויטר לבין התנהלות המניות בבורסה בעבר, אך לא ניתן להבטיח שהקשר יישמר גם בעתיד ובתנאים משתנים.
2. המחקר מתמקד בפלטפורמה החברתית "טוויטר" בלבד, ולא ניתן להכיר את כל הגורמים האחרים שעשויים להשפיע על השוק הפיננסי.
3. כמות המידע והמשתנים במחקר עשויים להיות מוגבלים, מה שיוצר אתגרים בהבנת כל הממצאים ומתן תמונה מקיפה ומדויקת.

המלצות למחקרי המשך:

1. מומלץ להתמקד בחקר השינויים בשוק המניות על רקע אירועים ספציפיים או משתנים פיננסיים נוספים כדי להבין את גורמי ההשפעה ולהעמיק בהבנת הקשר.
2. עדיף להרחיב את המחקר ולכלול פלטפורמות חברתיות נוספות ואישיות בכירות נוספות כדי לבדוק את תפקידם והשפעתם על השוק הפיננסי.
3. כדאי לבצע מחקרים מתקדמים יותר באמצעות מודלים נוירוניים ומתודולוגיות סטטיסטיות יותר מתקדמות לצורך ניתוח מדויק ומעמיק של הקשר בין הציוצים לבין התנהלות המניות.

השלכות המחקר:

1. המחקר יכול להעניק יתרון תחרותי למשקיעים במניות באמצעות פיתוח מודלים חיזוי מתקדמים.
2. הבנה עמוקה יותר של הקשר בין רשתות החברתיות לבין השוק הפיננסי יכולה לסייע בהפחתת סיכונים פיננסיים ולקידום כלל השקעות פוטנציאליות.
3. המחקר יכול להביא לפיתוח ויישום של כלים חדשים לניתוח ולהבנת השוק הפיננסי ולמדידת ההשפעות של אירועים ופעולות פוליטיות עליו.

סיכום ומסקנות

תוצאות המחקר שלנו היו בלתי צפויות ומעניינות בתיעוד ההשפעה שיש לציוצים שפורסמו על ידי דמויות משפיעות על מדד ה500 S&P. למרות שהתוצאות מבטיחות, ישנן דרכים למחקר עתידי שיש לשקול.

ראשית, המורכבות של שוק המניות ושלל הגורמים המשפיעים על מחירי המניות מרמזים על כך שמודלים של LSTM עשויים שלא להקיף באופן מלא את כל מגוון הגורמים הפוטנציאליים. בחינת השילוב של טכניקות למידת מכונה מתקדמות אחרות, יכולה להציע תובנות מקיפות יותר על יחסי הגומלין בין סנטימנט של פרסומים ברשתות החברתיות השונות והתנהגות שוק המניות.

שנית, בעוד שראינו תוצאות מסקרנות לגבי ההשפעה המוגבלת של נפח הציוץ על ההצלחה החזויה, עדיין חיוני לחקור את הסיבות הבסיסיות. ייתכן שרלוונטיות הציוץ ואיכות התוכן של דמויות משפיעות עלו על המשמעות של כמות הציוץ העצומה בדגמי ה-LSTM שלנו. מחקר עתידי יכול להתעמק בניתוח התוכן של ציוצים משפיעים ולחקור עוד יותר את הקשרים בין מאפייני הציוץ וביצועי המודל.

לסיכום, בעוד שהמחקר שלנו מדגיש את הפוטנציאל של מודלים של LSTM עם ניתוח סנטימנטים בחיזוי פיננסי, הוא גם מעלה שאלות מסקרנות לגבי המגבלות והמורכבות של התפקיד של שוק המניות ושל המדיה החברתית. מחקר עתידי יכול לחקור טכניקות מתקדמות, לשלב מידע נוסף ולנתח את יחסי הגומלין של מאפייני ציוץ שונים כדי לשפר את דיוק הניבוי. על ידי חידוד והרחבת ההבנה שלנו לגבי הקשר בין מדיה חברתית לשווקים פיננסיים, אנו יכולים לפתוח הזדמנויות חדשות לקבלת החלטות מונעות נתונים בעולם הפיננסים המתפתח ללא הרף.

ביבליוגרפיה

Brans, H., & Scholtens, B. (2020). Under his thumb the effect of president Donald Trump's Twitter messages on the US stock market. PloS one, 15(3), e0229931.

de O. Carosia, A. E., Coelho, G. P., & Silva, A. E. D. (2019, October). The influence of tweets and news on the brazilian stock market through sentiment analysis. In Proceedings of the 25th Brazillian Symposium on Multimedia and the Web (pp. 385-392).

Erkartal, B., & Yılmaz, A. (2022). Sentiment Analysis of Elon Musk's Twitter Data Using LSTM and ANFIS-SVM. Lecture Notes in Networks and Systems, 626–635.

Kordonis, J., Symeonidis, S., & Arampatzis, A. (2016, November). Stock price forecasting via sentiment analysis on Twitter. In Proceedings of the 20th Pan-Hellenic Conference on Informatics (pp. 1-6).

Kumar, P., Adhikari, S., Agarwal, P., & Sahoo, A. (2022, August). Stock Prices Prediction Based on Social Influence & Historic Data. In Proceedings of the 2022 Fourteenth International Conference on Contemporary Computing (pp. 276-283).

Mendoza-Urdiales, R. A., Núñez-Mora, J. A., Santillán-Salgado, R. J., & Valencia-Herrera, H. (2022). Twitter Sentiment Analysis and Influence on Stock Performance Using Transfer Entropy and EGARCH Methods. Entropy, 24(7), 874.

Nisar, T. M., & Yeung, M. (2018). Twitter as a tool for forecasting stock market movements: A short-window event study. The journal of finance and data science, 4(2), 101-119.

Teti, E., Dallocchio, M., & Aniasi, A. (2019). The relationship between twitter and stock prices. Evidence from the US technology industry. Technological Forecasting and Social Change, 149, 119747.

Yosipof, A., Inbar, G., & Aperstein2, Y. (n.d.). Long Short-Term Memory Model for Predicting the Stock Market Through President Donald Trump Tweets Sentiment [Review of Long Short-Term Memory Model for Predicting the Stock Market Through President Donald Trump Tweets Sentiment].

Yuan, K., Liu, G., Wu, J., & Xiong, H. (2020). Dancing with Trump in the stock market: a deep information echoing model. ACM Transactions on Intelligent Systems and Technology (TIST), 11(5), 1-22.

נספחים

נספח 1 API של יאהו:

```
# Define the start and end dates for the data.
start_date = '2017-02-01'
end_date = '2020-12-31'

# Download the historical data for the S&P 500 index using the ^GSPC ticker symbol.
sp500 = yf.download('^GSPC', start=start_date, end=end_date)

# The variable `sp500` now contains a DataFrame with the historical data for the
S&P 500 index.

# Drop the columns from the DataFrame
sp500 = sp500.drop(columns=['Volume', 'Adj Close'])
sp500_tb = sp500.copy()

target_directory = '.\stocks'
if not os.path.exists(target_directory):
    os.makedirs(target_directory)

csv_filename = 'sp500.csv'
sp500.to_csv(os.path.join(target_directory, csv_filename), index=True,
encoding='utf-8')
```

נספח 2 שמירת ציוצים:

```
# Define the URLs for the CSV files.
url_musk = 'https://raw.githubusercontent.com/maorisraelii/twitter-sentiment-
analysis/main/Musk(2014-2019).csv'
url_biden = 'https://raw.githubusercontent.com/maorisraelii/twitter-sentiment-
analysis/main/Biden(2007-2020).csv'
url_trump = 'https://raw.githubusercontent.com/maorisraelii/twitter-sentiment-
analysis/main/Trump(2017-2021).csv'
url_bill = 'https://raw.githubusercontent.com/maorisraelii/twitter-sentiment-
analysis/main/Bill_Gates.csv'
url_jeff = 'https://raw.githubusercontent.com/maorisraelii/twitter-sentiment-
analysis/main/Jeff_Bezos.csv'
url_tim = 'https://raw.githubusercontent.com/maorisraelii/twitter-sentiment-
analysis/main/Tim_Cook.csv'


# Read the CSV files into DataFrames.
musk = pd.read_csv(url_musk, encoding= 'unicode_escape')
biden = pd.read_csv(url_biden, encoding= 'unicode_escape')
trump = pd.read_csv(url_trump, encoding= 'unicode_escape',on_bad_lines= 'skip')
bill = pd.read_csv(url_bill, encoding= 'unicode_escape')
jeff = pd.read_csv(url_jeff, encoding= 'unicode_escape')
tim = pd.read_csv(url_tim, encoding= 'unicode_escape')
```

נספח 3 שינוי שמות עמודות:

```
# Convert 'time' column in the dataset to datetime format
trump['time'] = pd.to_datetime(trump['time'])
biden['time'] = pd.to_datetime(biden['time'], dayfirst=True)
musk['time'] = pd.to_datetime(musk['date'])
bill['time'] = pd.to_datetime(bill['time_stamp_UTC'])
jeff['time'] = pd.to_datetime(jeff['created_at'])
tim['time'] = pd.to_datetime(tim['created_at'])


# Convert 'tweet' column in the dataset
trump.rename(columns={'tweet': 'tweet'}, inplace=True)
biden.rename(columns={'tweet': 'tweet'}, inplace=True)
musk.rename(columns={'tweet': 'tweet'}, inplace=True)
bill.rename(columns={'tweet_text': 'tweet'}, inplace=True)
jeff.rename(columns={'text': 'tweet'}, inplace=True)
tim.rename(columns={'text': 'tweet'}, inplace=True)
```

נספח 4 ניקוי הציוצים:

```
# Define a function to replace emojis with words
def replace_emojis_with_words(text):
    # Replace each emoji with its corresponding description
    text_with_words = emoji.demojize(text)
    return text_with_words


# Define regular expressions to match links and emojis
link_pattern = r'https?://\S+'
emoji_pattern = re.compile("["
        u"\U0001F600-\U0001F64F"  # emoticons
        u"\U0001F300-\U0001F5FF"  # symbols & pictographs
        u"\U0001F680-\U0001F6FF"  # transport & map symbols
        u"\U0001F1E0-\U0001F1FF"  # flags (iOS)
        "]+", flags=re.UNICODE)

def clean_dataset(dataset, tweet_column = 'tweet', date_column = 'time'):
    dataset[tweet_column] = dataset[tweet_column].fillna('')  # Replace NaN values
with an empty string
    dataset = dataset[~dataset[tweet_column].str.match(link_pattern)]  # Remove
rows with tweets that contain only links
    dataset = dataset[~dataset[tweet_column].str.contains(emoji_pattern)]  # Remove
rows with tweets that contain emojis
    dataset[tweet_column] = dataset[tweet_column].apply(replace_emojis_with_words)
    dataset = dataset.dropna(axis=0, how='all')  # Drop rows with all NaN values
    dataset = dataset.dropna(axis=1, how='all')  # Drop columns with all NaN values
    dataset = dataset.reset_index(drop=True)  # Reset the index of the dataset
    dataset.dropna(subset=[date_column], inplace=True)  # Drop rows with NaN values
in the date column
    return dataset


# Clean datasets
musk = clean_dataset(musk)
biden = clean_dataset(biden)
trump = clean_dataset(trump)
bill = clean_dataset(bill)
jeff = clean_dataset(jeff)
tim = clean_dataset(tim)
```

נספח 5 המרת תאריכים ושמירה:

```
# Filter the dataset for the specified time range
trump = trump[(trump['time'] >= start_date) & (trump['time'] <= end_date)]
biden = biden[(biden['time'] >= start_date) & (biden['time'] <= end_date)]
musk = musk[(musk['time'] >= start_date) & (musk['time'] <= end_date)]
bill = bill[(bill['time'] >= start_date) & (bill['time'] <= end_date)]
jeff = jeff[(jeff['time'] >= start_date) & (jeff['time'] <= end_date)]
tim = tim[(tim['time'] >= start_date) & (tim['time'] <= end_date)]


target_directory = './tweets'
if not os.path.exists(target_directory):
    os.makedirs(target_directory)

trump.to_csv(os.path.join(target_directory,'trump.csv'), index=False,
encoding='utf-8')
biden.to_csv(os.path.join(target_directory,'biden.csv'), index=False,
encoding='utf-8')
musk.to_csv(os.path.join(target_directory,'musk.csv'), index=False, encoding='utf-
8')
bill.to_csv(os.path.join(target_directory,'bill.csv'), index=False, encoding='utf-
8')
jeff.to_csv(os.path.join(target_directory,'jeff.csv'), index=False, encoding='utf-
8')
tim.to_csv(os.path.join(target_directory,'tim.csv'), index=False, encoding='utf-8')
```

נספח 6 TextBlob:

```
# Function to perform sentiment analysis using TextBlob
def perform_sentiment_analysis_tb(df):
    analyzed_df = df.copy()
    analyzed_df['sentiment'] = ''
    analyzed_df['polarity'] = ''

    for index, row in analyzed_df.iterrows():
        tweet = row['tweet']
        sentiment, polarity = get_sentiment_label_tb(TextBlob(tweet).sentiment)
        analyzed_df.at[index, 'sentiment'] = sentiment
        analyzed_df.at[index, 'polarity'] = polarity

    return analyzed_df

# Function to get sentiment label, subjectivity, and polarity based on sentiment
score
def get_sentiment_label_tb(sentiment):
    polarity = sentiment.polarity
    subjectivity = sentiment.subjectivity

    if polarity > 0:
        sentiment_label = 'positive'
    elif polarity < 0:
        sentiment_label = 'negative'
    else:
        sentiment_label = 'neutral'

    return sentiment_label, polarity


# Perform sentiment analysis using TextBlob
analyzed_musk_textblob = perform_sentiment_analysis_tb(musk)
analyzed_biden_textblob = perform_sentiment_analysis_tb(biden)
analyzed_trump_textblob = perform_sentiment_analysis_tb(trump)
analyzed_jeff_textblob = perform_sentiment_analysis_tb(jeff)
analyzed_tim_textblob = perform_sentiment_analysis_tb(tim)
analyzed_bill_textblob = perform_sentiment_analysis_tb(bill)
```

נספח 7 Vader:

```
# Function to perform sentiment analysis using Vader
def perform_sentiment_analysis_v(df):
    analyzed_df = df.copy()
    analyzed_df['sentiment'] = ''
    analyzed_df['polarity'] = ''

    # Create an instance of the Vader sentiment analyzer
    analyzer = SentimentIntensityAnalyzer()

    for index, row in analyzed_df.iterrows():
        tweet = row['tweet']
        sentiment, compound_score =
get_sentiment_label_v(analyzer.polarity_scores(tweet))
        analyzed_df.at[index, 'sentiment'] = sentiment
        analyzed_df.at[index, 'polarity'] = compound_score

    return analyzed_df

# Function to get sentiment label and compound score based on Vader sentiment
scores
def get_sentiment_label_v(sentiment_scores):
    compound_score = sentiment_scores['compound']

    if compound_score >= 0.05:
        sentiment_label = 'positive'
    elif compound_score <= -0.05:
        sentiment_label = 'negative'
    else:
        sentiment_label = 'neutral'

    return sentiment_label, compound_score


# Perform sentiment analysis using Vader
analyzed_musk_vader = perform_sentiment_analysis_v(musk)
analyzed_biden_vader = perform_sentiment_analysis_v(biden)
analyzed_trump_vader = perform_sentiment_analysis_v(trump)
analyzed_jeff_vader = perform_sentiment_analysis_v(jeff)
analyzed_tim_vader = perform_sentiment_analysis_v(tim)
analyzed_bill_vader = perform_sentiment_analysis_v(bill)
```

נספח 8 אנליזה יומית:

```
# Group by the date and calculate the average sentiment for each day
day_grouped_musk_vader =
analyzed_musk_vader.groupby(analyzed_musk_vader['time'].dt.date)['polarity'].mean()
day_grouped_trump_vader=
analyzed_trump_vader.groupby(analyzed_trump_vader['time'].dt.date)['polarity'].mean
()
day_grouped_biden_vader =
analyzed_biden_vader.groupby(analyzed_biden_vader['time'].dt.date)['polarity'].mean
()
day_grouped_jeff_vader =
analyzed_jeff_vader.groupby(analyzed_jeff_vader['time'].dt.date)['polarity'].mean()
day_grouped_tim_vader =
analyzed_tim_vader.groupby(analyzed_tim_vader['time'].dt.date)['polarity'].mean()
day_grouped_bill_vader =
analyzed_bill_vader.groupby(analyzed_bill_vader['time'].dt.date)['polarity'].mean()


# Group by the date and calculate the average sentiment for each day
day_grouped_musk_textblob =
analyzed_musk_textblob.groupby(analyzed_musk_textblob['time'].dt.date)['polarity'].
mean()
day_grouped_trump_textblob =
analyzed_trump_textblob.groupby(analyzed_trump_textblob['time'].dt.date)['polarity'
].mean()
day_grouped_biden_textblob =
analyzed_biden_textblob.groupby(analyzed_biden_textblob['time'].dt.date)['polarity'
].mean()
day_grouped_jeff_textblob =
analyzed_jeff_textblob.groupby(analyzed_jeff_textblob['time'].dt.date)['polarity'].
mean()
day_grouped_tim_textblob =
analyzed_tim_textblob.groupby(analyzed_tim_textblob['time'].dt.date)['polarity'].me
an()
day_grouped_bill_textblob =
analyzed_bill_textblob.groupby(analyzed_bill_textblob['time'].dt.date)['polarity'].
mean()
```

נספח 9 מילוי נתונים ריקים:

```
sp500_tb.interpolate(method='linear', inplace=True)
sp500_tb.fillna(method='bfill', inplace=True)
sp500_tb.fillna(method='ffill', inplace=True)
```

נספח 10 Granger Causality:

```
Best Granger Causality between 'biden' and Close price:
Number of lags: 1
ssr based F test: F=1.6478171394068823, p=0.19955910400315524
ssr based chi2 test: chi2=1.6528512039876575, p=0.19857131084209903
likelihood ratio test: chi2=1.6514659934296105, p=0.19875951895439056
parameter F test: F=1.6478171394068823, p=0.19955910400315524

Best Granger Causality between 'musk' and Close price:
Number of lags: 2
ssr based F test: F=0.6768560326696601, p=0.5084499463153778
ssr based chi2 test: chi2=1.3606258143958319, p=0.5064584930611717
likelihood ratio test: chi2=1.3596859781246167, p=0.5066965430196093
parameter F test: F=0.6768560326696601, p=0.5084499463153778

Best Granger Causality between 'trump' and Close price:
Number of lags: 1
ssr based F test: F=19.996882553655333, p=8.66173803579489e-06
ssr based chi2 test: chi2=20.05797282622241, p=7.51296589388249e-06
likelihood ratio test: chi2=19.856479130683965, p=8.347930992855778e-06
parameter F test: F=19.996882553655333, p=8.66173803579489e-06

Best Granger Causality between 'tim' and Close price:
Number of lags: 1
ssr based F test: F=22.67822532058611, p=2.203467851394886e-06
ssr based chi2 test: chi2=22.74750706800126, p=1.847436435722862e-06
likelihood ratio test: chi2=22.488817755843684, p=2.113704797021565e-06
parameter F test: F=22.67822532058611, p=2.203467851394886e-06

Best Granger Causality between 'bill' and Close price:
Number of lags: 5
ssr based F test: F=0.4411283685147131, p=0.8198904283044126
ssr based chi2 test: chi2=2.230654275839017, p=0.8163960012770357
likelihood ratio test: chi2=2.2281220187185227, p=0.8167636809554155
parameter F test: F=0.4411283685147131, p=0.8198904283044126

Best Granger Causality between 'jeff' and Close price:
Number of lags: 1
ssr based F test: F=0.19839024661013532, p=0.6561215127238598
ssr based chi2 test: chi2=0.1989963267938203, p=0.6555322046550751
likelihood ratio test: chi2=0.19897622821190453, p=0.6555484771669544
parameter F test: F=0.19839024661013532, p=0.6561215127238598
```

נספח 11 בניית מודלים ללא סנטימנט:

```
models_list = []

models_list.append(build_and_test_model(120, 100, len(models_list)))
models_list.append(build_and_test_model(128, 128, len(models_list)))
models_list.append(build_and_test_model(256, 256, len(models_list)))
models_list.append(build_and_test_model(200, 100, len(models_list)))
models_list.append(build_and_test_model(256, 128, len(models_list)))
models_list.append(build_and_test_model(256, 256, len(models_list)))
models_list.append(build_and_test_model(120, 105, len(models_list)))
```