

תקציר

בדוח זה נציג את שלבי מידול הנתונים והערכת ביצועים בפרויקט חיזוי מדד S&P 500 באמצעות שילוב של נתונים טכניים וניתוח סנטימנט של חדשות כלכליות. בתחילה נבחנו מודלים נפרדים GRU ו-LSTM לחיזוי סדרות זמן, ומודלים מסוג BERT לסיווג סנטימנט של כותרות חדשות. לאחר ניסויים מרובים נמצא שמודל GRU מציע דיוק גבוה, יציבות ויעילות חישובית, ולכן נבחר לשמש כמודל התחזיות המרכזי.



דוח מידול הנתונים

השקעות בסיכון נמוך זה אצלנו-NexTrade

מגישים – אריאל קריב ומשי בר



תוכן עניינים

2.....	בחירת טכניקות מידול	.1
4.....	תכנון בדיקות	.2
10	תיאור המודלים	.3
16	הערכת מודלים	.4

1. בחירת טכניקות מידול

בחירת טכניקות מתאימות

בשלב המידול בפרויקט, מטרתנו העיקרית היא לתכנן ולבנות מודל המסוגל לחזות את כיוון תנועת מדד S&P 500 לשבוע קדימה, בהתבסס על נתונים טכניים ושילוב של מידע חדשותי סובייקטיבי. המודל נועד לא רק לחזות את מחירי הסגירה הצפויים, אלא גם להפיק עבור המשתמש המלצת פעולה ישירה- האם כדאי לקנות (BUY), להחזיק (HOLD) או למכור (SELL)- תוך שימוש בשיעור השינוי הצפוי, גרף חזוי ותצוגה ויזואלית של מגמת התחזית.

תחילה שקלנו שימוש בשני מודלים נפרדים: מודל עיבוד שפה טבעית (NLP) מסוג BERT לניתוח סנטימנט של חדשות כלכליות, ומודל LSTM נפרד לחיזוי סדרות עתיות של מדד S&P 500. אכן פיתחנו את שני המודלים בנפרד, אך לאחר סדרת ניסויים, עלו מספר קשיים:

- מודל BERT דרש זמן עיבוד ארוך ומשאבים חישוביים גבוהים מאוד.
- מודל LSTM הציג תנודתיות רבה יותר בתוצאות ופחות יציבות יחסית בהשוואה למודל ששקלנו להשתמש בו בשם GRU.

מודל GRU (Gated Recurrent Unit) הציג יתרון ברור: ביצועים יציבים, מהירות חישוב גבוהה, פשטות יחסית במבנה, ותמיכה באינטגרציה חלקה של נתונים ממקורות שונים - טכניים וחדשותיים כאחד.

לפיכך, התקבלה החלטה אסטרטגית להתמקד במודל GRU יחיד, אשר יכלול את כל מאפייני הנתונים הטכניים שנבנו במיוחד לשלב זה (לדוגמה: MACD, RSI, אחוז שינוי יומי, ממוצעים נעים, סטיית תקן, יחס MACD/RSI, ועוד), יחד עם שילוב של פיצ'רים שמייצגים את מצב הסנטימנט הכלכלי בכל יום (כגון Sentiment, overall_sentiment_label ו־had_news).

המודל שנבחר אכן דורש חלוקה לסט אימון (Train), סט אימות (Validation) וסט בדיקה (Test) - תהליך אשר יושם בפועל בכל ניסוי במדויק. הנתונים שעמדו לרשותנו, בהיקפם ובאיכותם, אפשרו תהליך חלוקה תקין, והתאימו לדרישות המינימליות של מודלים לחיזוי סדרות עתיות.

בנוגע לאיכות הנתונים- בוצעו תהליכי ניקוי מקדימים, טיפול בערכים חסרים, נרמול, והנדסת פיצ'רים מתקדמת. סוג הנתונים (כמותיים רציפים + קטגוריאליים מקודדים) הותאם למבנה הרציף של המודל (GRU) באמצעות מניפולציות שמפורטות בסעיף הבא.

הנחות המודל והמניפולציות בנתונים

במהלך שלב זה ביצענו מספר הנחות מהותיות אשר עיצבו את מבנה המודל ואיכות תחזיותיו:

- הנחה מרכזית הייתה כי סנטימנט של חדשות כלכליות משפיע בצורה מובהקת על כיוון המדד. עם זאת, התברר כי לא קיימת חדשה כלכלית משמעותית בכל יום- מאחר שבחרנו להכניס רק חדשות מקרו-כלכליות רלוונטיות במיוחד למדד (למשל חדשות ריבית, תעסוקה, אינפלציה וכו'), אשר מתפרסמות אחת למספר שבועות.

לשם כך הוספנו עמודות סנטימנט חדשות, אך נתקלנו עם מצב שבו עבור חלק גדול מהימים לא הייתה חדשה כלל. בכדי להתמודד עם בעיית הערכים החסרים- מילאנו את ערכי הסנטימנט קדימה (Forward Fill) כך שכל ערך סנטימנט ממשיך להשפיע על המודל עד לפרסום החדשה הבאה.

- נוסף לכך, הנחנו כי לתזמון פרסום החדשות יש חשיבות קריטית להשפעתן על השוק. כלומר, לא רק תוכן החדשה משפיע, אלא עצם פרסומה באותו היום. בהתאם לכך יצרנו עמודה חדשה- עמודה בינארית אשר מקבלת ערך 1 כאשר פורסמה חדשה חדשה באותו היום, ו-0 כאשר מדובר בהמשך של הסנטימנט הקודם. פעולה זו נועדה לאפשר למודל הבחנה בין השפעה ראשונית של אירוע חדשותי, לבין ימים שמהווים "הדהוד" בלבד לאותו האירוע.

צורת מילוי זו סייעה למנוע הצפה של ערכי אפס מיותרים אשר עלולים לפגוע בדיוק החיזוי של המודל ולגרום לעיוותי מגמה.

כל הפיצ'רים שהוספו ושהוזנו למודל יוצגו ויפורטו בהמשך הדוח בפרק המוקדש לבחירת משתנים.

2. תכנון בדיקות

לפני בניית המודלים בפועל, הוגדר תהליך ברור ומבוסס לבחינת איכותם של המודלים ולבדיקת היכולת שלהם לספק תחזיות מדויקות, יציבות וברות-הכללה על נתונים שלא נראו קודם. שלב זה נועד לקבוע קריטריונים ברורים להצלחת המודל, לקבוע כיצד יחולקו הנתונים לבדיקות, ולבחור את המדדים שיעריכו את ביצועי המודל בפועל.

נתוני הקלט וההכנה לבדיקות למודלים LSTM ו-GRU

שני המודלים LSTM ו-GRU התבססו על אותה טבלת נתונים מקורית (SNP_DATA.csv) המכילה את משתני השוק הבאים:

- Open - מחיר פתיחה של המדד בכל יום.
- High - המחיר הגבוה היומי.
- Low - המחיר הנמוך היומי.
- Close - מחיר הסגירה של המדד (היעד לחיזוי).
- Volume - מחזור המסחר היומי.
- MACD - אינדיקטור טכני המודד ממוצעים נעים ומזהה מומנטום.
- RSI - אינדיקטור טכני המודד מצבי קנייה/מכירה יתרה.

לפני הבדיקות:

- ערכים חסרים בעמודת RSI הושלמו או הוסרו.
- LSTM עבד על הנתונים לאחר השלמה של RSI.
- GRU עבד על הנתונים לאחר הסרה מוחלטת של שורות עם NaN.
- כל המשתנים עברו נרמול באמצעות MinMaxScaler על מנת לשפר את מהירות ויציבות האימון.
- הנתונים אורגנו לחלונות זמן של 30 ימים, כדי לאפשר למודל ללמוד מגמות על פני טווח זמן ריאלי.

חלוקת הנתונים

במסגרת תכנון החלוקה לסטים שקלנו גם את האפשרות לבצע שיטות חלוקה נוספות, כמו חלוקה אקראית או שימוש ב-K-Fold Cross Validation עם זאת, מסיבות מתודולוגיות ומעשיות החלטנו שלא לבצע חלוקות נוספות, והבחירה בחלוקה הכרונולוגית נבעה מהשיקולים הבאים:

1. אופי הנתונים - סדרת זמן

מאחר והנתונים עוסקים בחיזוי מדד S&P 500 לאורך זמן, שימור הסדר הכרונולוגי קריטי. חלוקה אקראית הייתה מביאה לזליגת מידע מהעתיד לעבר, דבר שהיה פוגע בתוקף המודל וביכולת ההכללה שלו לנתונים עתידיים.

2. היקף נתונים מצומצם יחסית

עמדו לרשותנו כ-1,500 שורות בלבד - נתון שמציב מגבלה ממשית על פיצול נוסף של הדאטה מבלי לפגוע באיכות סט האימון או האימות. שימוש בשיטות כמו K-Fold עלול היה לייצר סטים קטנים מדי, ובכך לפגוע ביציבות ובאמינות התוצאות.

3. איזון בין אמינות לבדיקה מעשית

ביצענו מספר ריצות ניסוי תוך שינוי היפר-פרמטרים שונים, וכל אחת מהן נשענה על אותה חלוקה כרונולוגית. לא נרשמה מגמת Overfitting או ירידה באיכות התחזית, מה שמעיד על כך שהחלוקה שנבחרה מספקת תוקף פנימי גבוה.

לפיכך, לא בוצעה השוואה בין סוגי חלוקות שונים, לא מתוך חוסר מודעות אלא מתוך התאמה למבנה הנתונים ולאילוצי הפרויקט. חלוקה כרונולוגית נמצאה כמתאימה ביותר עבור חיזוי סדרות זמן בהיקף נתונים מצומצם.

בכדי להבטיח תהליך בדיקה מדויק, בוצעה חלוקה כרונולוגית של הדאטה לסטים שונים:

- סט אימון 70%- ראשונים של הנתונים: המודל לומד מהם את הדפוסים השגרתיים.

- סט ולידציה 15%- הבאים: משמש בזמן האימון לכוונון ולמדידת שיפור ביצועים בזמן אמת.

- סט בדיקה 15%- אחרונים: משמש לבדיקה סופית של המודל, תוך שמירה על נתונים "נקיים" שהמודל לא ראה מעולם.

חשוב לציין שהנתונים חולקו מבלי לערבב סדר כרונולוגי- מהותי במיוחד במודלים של סדרות זמן, כדי למנוע תופעת Data Leakage.

מדדי הצלחה למודלים מונחים

הערכת המודלים בוצעה לפי המדדים הבאים:

מדד	הסבר	מדוע נבחר
Loss / Validation Loss	שגיאת התחזית על סט האימון והוולידציה (Mean Squared Error)	מאפשר ניטור בזמן אמת של דיוק המודל והתכנסות הלמידה
RMSE (Root Mean Squared Error)	שורש ממוצע ריבועי השגיאות בין התחזיות לערכים בפועל	מעניש שגיאות גדולות, שימושי בזיהוי מתי המודל "מאבד שליטה"
MAE (Mean Absolute Error)	ממוצע ההפרשים המוחלטים בין התחזיות לערכים בפועל	מדד יציב פחות מושפע מערכים קיצוניים

כיצד מתבצעת בפועל הבדיקה?

1. לאחר אימון המודל, הוא מריץ תחזיות על סט הבדיקה (X_{test}).
2. התחזיות מושוות לערכי ה-Close האמיתיים.
3. לפי הפערים בין התחזיות לערכים בפועל- מחושבים המדדים RMSE ו-MAE.
4. נבנה גרף המשווה בין תחזית ל-אמת ומציג את דיוק החיזוי.

מספר איטרציות של שינוי פרמטרים

לכל אחד מהמודלים הוגדרו מראש עד 5 ניסויים לשינוי פרמטרים (כגון גודל חלון זמן, כמות שכבות, Dropout, קצב למידה וכו). מטרה זו נועדה למנוע אופטימיזציה יתר על דאטה ספציפי (Overfitting) ולהבטיח תהליך ניסויי סדור.

מסקנה כללית לתכנון הבדיקות

שיטת הבדיקה שנבחרה הותאמה במיוחד לאופי תחזיות סדרות זמן:

- שימוש ב-validation loss בזמן אמת מאפשר ניטור תהליך האימון ובחירת נקודת העצירה האידיאלית.
- שילוב בין מדדים מספריים (RMSE, MAE) וגרפים ויזואליים מחזק את הבטחון בביצועי המודלים.
- העובדה שהשיטה יושמה באופן עקבי בשני המודלים, מאפשרת השוואה הוגנת והסקת מסקנות לגבי המודל המתאים ביותר לפרויקט.

תכנון בדיקות עבור מודלים מבוססי Transformer לניתוח חדשות כלכליות

בשלב זה של הפרויקט, הוגדר תהליך שיטתי לבחינת ביצועי מספר מודלים מבוססי Transformer לצורך ניתוח סנטימנט של כותרות חדשות כלכליות. מודלים אלו נועדו לאפשר שילוב של ניתוח איכותי של טקסט חופשי כחלק מהמודל הסופי לניבוי מדד S&P 500.

המטרה של תכנון זה

מטרת השלב היא להשוות בין מודלים שונים לסיווג טקסטים - BERT רגיל, DistilBERT, RoBERTa ו- BERT sped-up על בסיס יכולתם לסווג בצורה מדויקת את הסנטימנט הכלכלי מתוך טקסט. המודל בעל הביצועים הטובים ביותר ייבחר בהמשך וישמש ככלי הפקת משתנים תומכים במודל התחזית הסופי.

סוגי המודלים הנבחרים

המודלים שנבדקו מבוססים על ארכיטקטורת Transformer - אך נבדלים בגודל, בעומק, ובמהירות החישוב:

- BERT - המודל המלא, דו-כיווני, עם דיוק גבוה אך איטיות יחסית.
- RoBERTa - גרסה משופרת ללא משימת Next Sentence Prediction מותאמת לזרימת טקסט רציפה.
- DistilBERT - גרסה "מרוכזת" ומהירה של BERT עם חצי מהשכבות אך דיוק כמעט זהה.
- BERT sped-up - גרסה המתאימה לאימון מהיר, תוך שמירה על דיוק בסיסי.

מדדים לבחינת איכות המודלים

לכל אחד מהמודלים נמדדו המדדים הבאים, לצורך הערכת הצלחתם כמודלי סיווג מרובה תוויות:

- Validation Accuracy - אחוז התחזיות הנכונות על סט ולידציה.
- Validation Loss - מדד שגיאה על סט שלא השתתף באימון.
- Training Time - זמן האימון בפועל, בשיקול של פריסה מעשית.

על אילו נתונים יתבצעו הבדיקות בפועל

הבדיקות יבוצעו על סמך קובץ הנתונים:

- economic_news_bert_ready_updated - קובץ זה כולל את כותרות החדשות (title) ותוויות סנטימנט מתויג אשר שימשו כקלט לאימון ובדיקה של המודלים.



החוג למערכות מידע

- הנתונים מהקובץ חולקו באופן אקראי לסט אימון (80%) וסט בדיקה (20%), תוך שמירה על איזון בין תוויות הסנטימנט (Stratified Split), לצורך מניעת הטיה בייצוג הקטגוריות.
- בנוסף, בוצעו מספר ריצות עם זרעים אקראיים שונים (Random States) כדי לוודא שהתוצאות אינן תלויות בחלוקה בודדת. הריצות הנוספות הדגימו עקביות במדדי הדיוק וה-loss של המודלים, דבר שחזק את המסקנות לגבי ביצועיהם.
- שיטה זו נבחרה כחלופה יעילה ל-K-Fold Cross Validation, שאינה תמיד ישימה במודלי Transformer עקב זמני הרצה ארוכים וצריכת משאבים גבוהה. הבחירה בריצות חוזרות עם חלוקות אקראיות מבוקרות סיפקה אינדיקציה מספקת ליציבות ואמינות תהליך ההערכה.

כיצד ייבחר המודל המתאים

- הבחירה הסופית תתבסס על מספר פרמטרים:
- ביצועים גבוהים במדדים שהוזכרו לעיל.
 - יציבות בתהליך האימון ללא Overfitting.
 - יעילות חישובית (משאבים נדרשים, זמן ריצה).
 - התאמה לצרכים של מודל התחזית המרכזי של המדד - כלומר, עד כמה המודל מצליח לזקק מידע איכותי מחדשות לצורך ניבוי תנועות בשוק ההון.

שימוש עתידי בתוצרי המידול

תוצרי המודל שייבחר לא ישמשו כתחזית סופית אלא, מתוכננים להוות קלט תומך למודל התחזית המרכזי. בהתאם לכך, ייגזרו ממנו משתנים חדשים שיתווספו לטבלת הנתונים הסופית (כגון אינדיקטור סנטימנט, ציון כמותי, אינדיקציה לקיום חדשות), אשר יוכלו לשפר את דיוק התחזית הכללית.

מודלים לא-מונחים

בשלב זה לא נעשה שימוש במודלים לא-מונחים מאחר שכל הבעיות המוגדרות בפרויקט מבוססות על נתונים מתויגים מראש.

3. תיאור המודלים

מודלים לחיזוי ערכי מדד S&P 500 ו-LSTM

בשלב המידול נבחנו שני מודלים נפוצים ממשפחת הרשתות העצביות החוזרות :

- מודל LSTM.
- מודל GRU.

שני המודלים שימשו במטרה לחזות את ערך הסגירה של מדד S&P 500 ליום הבא, על בסיס רצף כרונולוגי של 30 ימי מסחר קודמים.

נתוני הקלט

שני המודלים התבססו על אותו קובץ מקור :

- `SNP_DATA.csv` קובץ הכולל נתונים טכניים יומיים שהרחבנו עליהם לפניכן. חשוב לציין: הקובץ `snp500_lstm_ready.csv` לא שימש בפועל למידול, למרות שתוכן לשימוש מוקדם יותר.

עיבוד מקדים

- ניקוי נתונים: הוסרו 13 שורות ראשונות מאחר והכילו ערכים חסרים (NaN) באינדיקטורים MACD ו-RSI. השמטה זו הכרחית, שכן מודלים מסוג RNN אינם מסוגלים לקבל ערכים חסרים. מדובר בפחות מ-1% מהנתונים הכוללים, ולכן לא צפויה פגיעה במובהקות התוצאות.
- שימור סדר כרונולוגי: הנתונים חולקו לסטי אימון (70%), ולידציה (15%) ובדיקה (15%) לפי סדר תאריכים, תוך מניעת זליגת מידע.
- שימוש ב-Google Colab: המעבר לקולאב בוצע לאחר בעיות התקנה של TensorFlow ב-PyCharm.

מבנה והסבר על המודלים ורכיביהם

לשני המודלים מבנה זהה כמעט לחלוטין, למעט סוג השכבות:

הסבר	תיאור הפעולה	שכבה
בכל איטרציה, המודל מקבל חלון זמן של 30 ימים, ובכל יום יש 7 מאפיינים מספריים (טכניים וסנטימנטליים).	רצף של 30 ימים \times 7 מאפיינים	קלט
שכבה חוזרת הלומדת תבניות טמפורליות. האפשרות return_sequences=True מאפשרת לשכבה להחזיר רצף שלם של פלטים לכל צעד זמן.	GRU או LSTM עם 50 יחידות חישוב, return_sequences=True	שכבה ראשונה
משמש להפחתת Overfitting ע"י השבתה אקראית של נוירונים בזמן האימון.	Dropout בשיעור 0.2	Dropout
שכבה נוספת שסוכמת את כל המידע לרגע האחרון, (last hidden state) המשמש כבסיס לחיזוי.	GRU או LSTM נוספת, return_sequences=False	שכבה שנייה
שכבת הפלט מייצרת ערך מספרי בודד (תחזית מחיר), ללא פונקציית אקטיבציה לא ליניארית.	Dense עם נוירון יחיד ואקטיבציה ליניארית	שכבת פלט

פרמטרים שנבחרו לשני המודלים

פרמטר	ערך נבחר	הסבר
Optimizer	Adam	אופטימיזר יציב ויעיל ללמידה של סדרות זמן
Learning Rate	0.001	קצב עדכון משקלים סביר, מאפשר יציבות והתכנסות
Loss Function	MSE	מדד לרגרסיה שמתחשב במיוחד בשגיאות גדולות
Batch Size	32	איוון בין מהירות עיבוד לבין דיוק סטטיסטי
Epochs	50	מספיקים לזיהוי מגמות מבלי להוביל ל-Overfitting

תיאור תהליך האימון ובדיקות שבוצעו בפועל

- ריצה וחישוב זמן: המודלים רצו בהצלחה למשך 50 אפוקים כל אחד, ללא שגיאות טכניות.
- הערכת ביצועים: במהלך האימון נמדדו:
 - Loss - השגיאה על סט האימון.
 - Validation Loss - השגיאה על סט הוולידציה.
- לא נמצאו בעיות ערכים חסרים מלבד 13 השורות שהוסרו מראש.
- לא הופיעו תקלות חישוביות והמודלים התכנסו באופן תקין.

השוואת תוצאות והסקת מסקנות

- LSTM הציג ביצועים טובים אך עם תנודתיות גבוהה ב- Validation Loss (0.001–0.0019).
- GRU הציג יציבות גבוהה יותר והתכנסות עקבית לערכים נמוכים יותר של Validation Loss (0.00014–0.0003).
- יתרה מזאת GRU, כולל פחות פרמטרים לאימון, ולכן איטרטיבי ומהיר יותר.
- על בסיס ביצועים מספריים, יציבות ויעילות-לכן ככל הנראה נבחר מודל GRU כמודל התחזיות הסופי עבור חיזוי מדד S&P 500.

ניתוח חדשות כלכליות באמצעות מודלים מבוססי Transformers

נתוני הקלט- הקובץ ששימש לבניית כלל מודלי הסנטימנט הוא `economic_news_bert_ready_updated.csv`.

עיבוד מוקדם

- קידוד תוויות: הומרו לערכים מספריים בעזרת LabelEncoder והוסבו ל- One-Hot Encoding לצורך סיווג רב-קטגוריאלי.
- טוקניזציה: בוצעה לפי הדרישות הספציפיות של כל מודל.

מודלים שנבחנו

1. BERT Regular - מודל עוצמתי דו-כיווני עם 12 שכבות Transformer המאומן מראש על מיליוני דוגמאות בשפה האנגלית. מצטיין בדיוק גבוה, אך דורש זמן חישוב משמעותי.
2. RoBERTa Regular - גרסה משופרת של BERT ללא משימת ניבוי משפט הבא. מותאם לטקסטים רציפים, אך בפועל לא סיפק יתרון בביצועים על הדאטה שלנו.
3. DistilBERT Regular - גרסה "רזה" של BERT הכוללת 6 שכבות בלבד. מהירה פי 2 לערך מבניית BERT מלאה, עם שמירה על כ-97% מהדיוק. אידאלית לפריסה מהירה ומערכות בזמן אמת.
4. BERT Speed Up - גרסה קלת-חישוב של BERT מלאה, שהתמקדה בקיצור זמני עיבוד. הציגה דיוק נמוך יותר מהמודלים האחרים ולכן לא נבחרה.

מבנה המודלים ופרמטרים

מבנה כל המודלים כלל את השלבים הבאים:

- קלטים - `input_ids` ו-`attention_mask`
- שכבת Pretrained - אחת מהשכבות `TFBertModel` / `TFRobertaModel` / `TFDistilBertModel`
- שכבת Dropout - עם יחס של 0.3 למניעת Overfitting.
- שכבת פלט Dense - עם softmax לסיווג ל-5 תוויות.

פרמטר	ערך	הסבר
Optimizer	Adam	יצב ומהיר, אידיאלי ל fine-tuning על מודלים מוכנים
Learning Rate	2e-5	קצב למידה אופטימלי עבור fine-tuning של מודלי Transformer
Loss Function	Categorical Crossentropy	מתאימה לבעיות סיווג רב-קטגוריאליות
Batch Size	32	איזון טוב בין משאבים לדיוק
Epochs	3	מספיקים לשיפור ביצועים מבלי לגרום ל-Overfitting

פלט המודלים

כל אחד מהמודלים מחזיר וקטור softmax של 5 ערכים, כאשר כל ערך מייצג את ההסתברות של הכותרת להשתייך לאחת מהקטגוריות הסנטימנטליות. הקטגוריה בעלת ההסתברות הגבוהה ביותר משמשת כתחזית הסופית של המודל.

מדדי השוואה

מודל	Accuracy	Validation Loss	Training Time
BERT Regular	0.5228	1.0739	13.5 דקות
DistilBERT	0.5222	1.0648	4.5 דקות
RoBERTa Regular	0.5190	1.1591	9.5 דקות
BERT Speed Up	0.4985	1.0914	4.25 דקות

מסקנות ובחירה במודל הסופי

למרות שמודל BERT Regular הציג את הדיוק הגבוה ביותר, זמן האימון הארוך שלו פי שלוש מ-DistilBERT וה- Validation Loss הגבוה יחסית הפכו אותו לפחות אטרקטיבי.

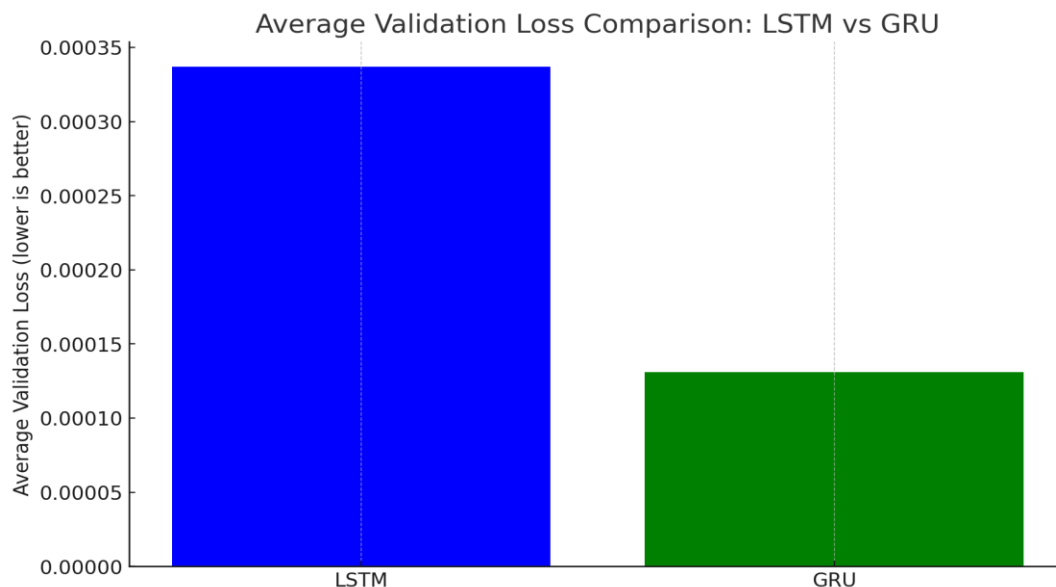
מודל DistilBERT הפגין איזון מיטבי בין דיוק, יציבות, פשטות חישובית ומהירות. הוא שומר על ביצועים כמעט זהים ל-BERT אך עם צריכת משאבים נמוכה יותר, זמן ריצה מהיר יותר והתאמה מעשית למערכות בזמן אמת.

לכן, נבחר מודל DistilBERT לשמש את מערכת החיזוי הכללית, והוא יספק את תוויות הסנטימנט הסופיות לקידוד בטבלת הנתונים המאוחדת.

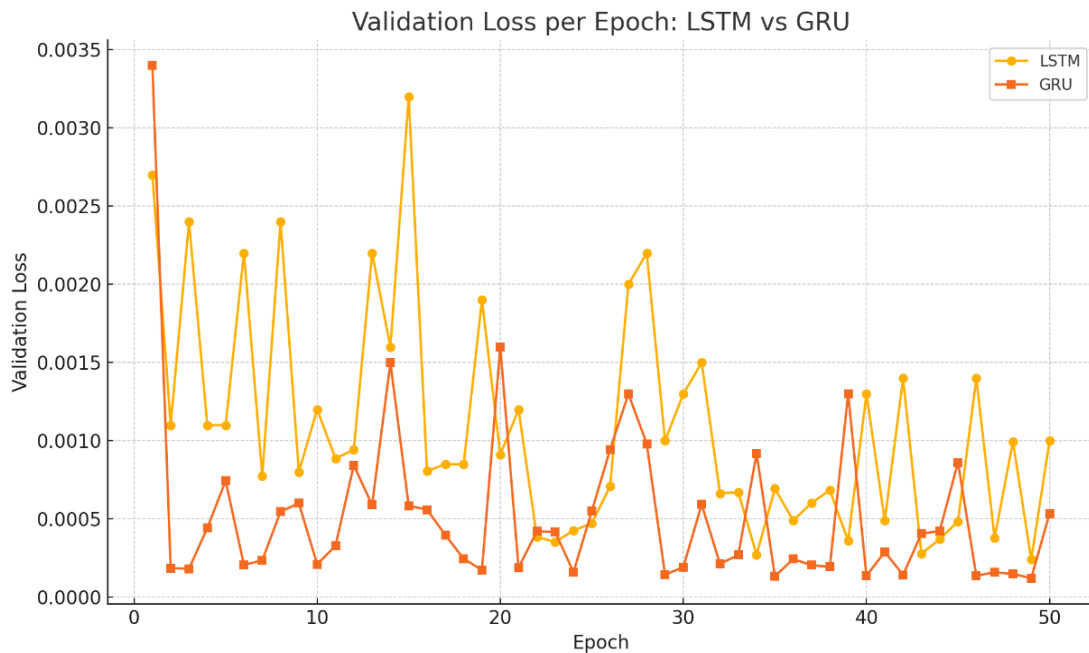
4. הערכת מודלים

הערכת מודלים לחיזוי ערכי מדד S&P 500

לצורך חיזוי מגמות במדד S&P 500 נעשה שימוש במודלים לזיהוי תבניות סדרתיות : LSTM ו-GRU. גם כאן נבחנו המודלים לפי שגיאת תחזית, קצב התכנסות, מספר פרמטרים, ויציבות לאורך זמן.



מתוך הגרפים והפלטים שהופקו, ניתן לראות כי המודל GRU הגיע לערכי שגיאה נמוכים יותר לאורך כל שלבי האימון. שגיאת הוולידציה הממוצעת של GRU עמדה על כ-0.000131 לעומת 0.000337 במודל LSTM.



בנוסף GRU, מתכנס מהר יותר לערכים נמוכים, תוך שמירה על יציבות לאורך כל 50 האפוקים - מה שמעיד על יכולת הכללה טובה יותר ועל סיכון מופחת ל-Overfitting.

במישור החישובי GRU, הצריך פחות פרמטרים 24,201 לעומת 31,851 ב-LSTM ולכן גם היה קל יותר לביצוע ולשחזור. הגרפים הראו קווים מקבילים של train loss ו-val loss מה שמעיד על כך שהמודל אינו לומד יתר על המידה את הנתונים, אלא משמר הכללה טובה.

על סמך ניתוח זה GRU, נבחר כמודל המתאים ביותר לחיזוי סדרות זמן של מדדי שוק.

שיקולים איכותיים

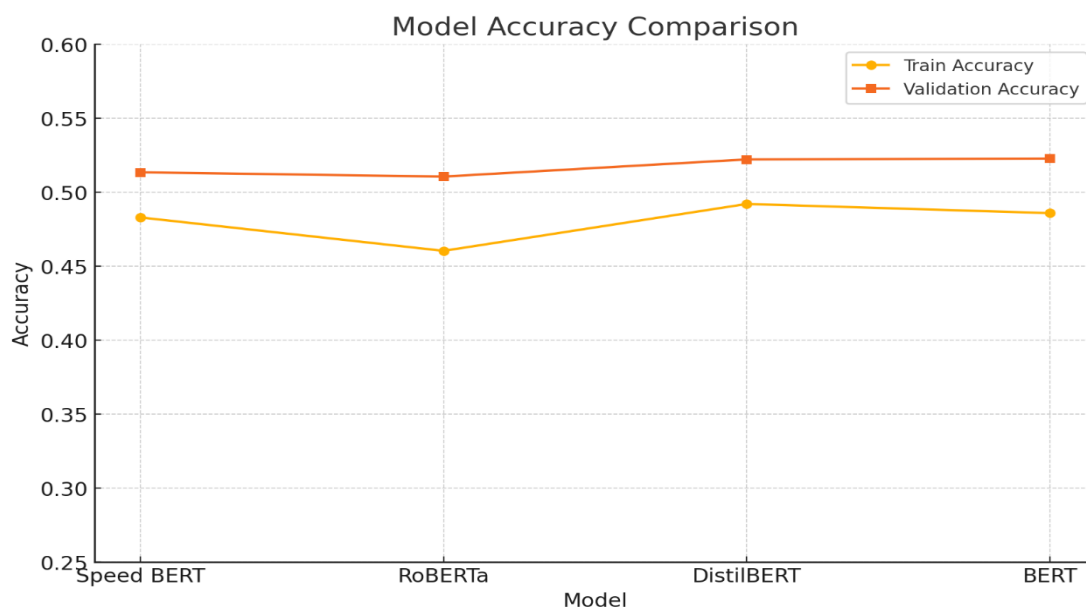
בנוסף למדדים הכמותיים, נבחנו גם שיקולים איכותיים במטרה לוודא שהמודלים שנבחרו לא רק מדויקים, אלא גם פרקטיים לשימוש בעולם האמיתי. נבדקו היבטים כגון:

- עד כמה התוצאה שהמודל מחזיר נראית סבירה והגיונית.
- האם למודל יש נטייה לתחזיות קיצוניות או לא עקביות.
- האם קל להסביר את דרך הפעולה של המודל.
- פשוט לשלב את המודל בתשתית כגון מערכת BI או דשבורד תחזיות.

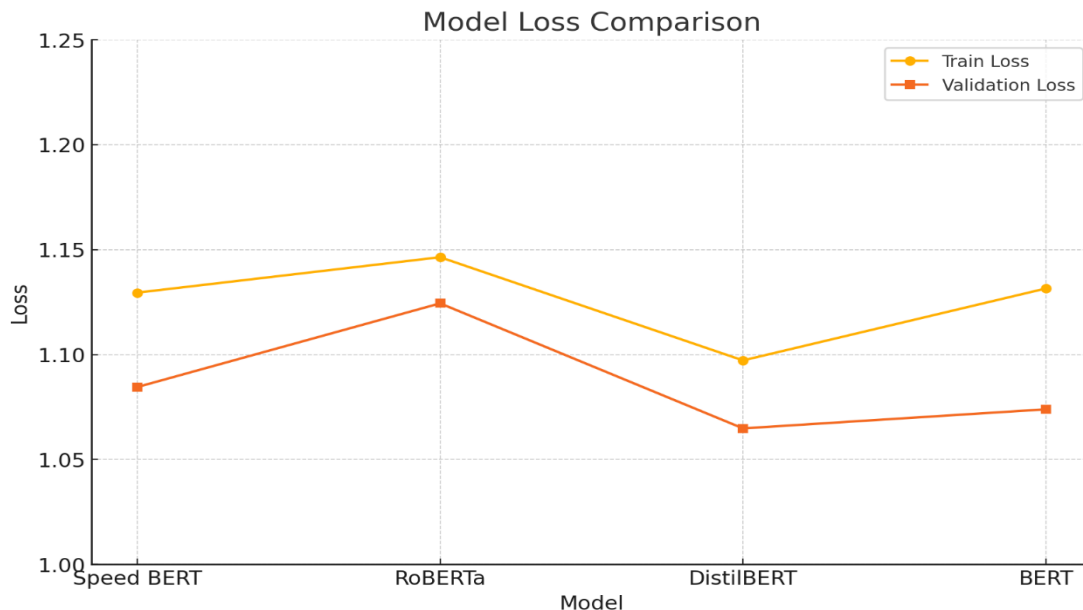
הערכת מודלים למשימת סיווג סנטימנט (על חדשות כלכליות)

למשימה זו נבחנו ארבעה מודלים BERT, DistilBERT, RoBERTa ו-SpeedBERT. המודלים הוערכו על סמך מדדי דיוק, פונקציית הפסד, משך זמן ריצה, צריכת משאבים ואיזון בין ביצועים למורכבות.

במונחי דיוק, נמצא כי המודל DistilBERT השיג את התוצאה הגבוהה ביותר באימון (דיוק של 49.2%), כאשר מודל BERT הרגיל עקף אותו בהפרש קטן מאוד בנתוני הוולידציה (52.3% מול 52.2%).



עם זאת DistilBERT, הציג שגיאה נמוכה יותר (Loss) הן באימון (1.097) והן בולידציה (1.064), דבר המעיד על איכות ניבוי גבוהה יותר ויציבות.



היבטים טכניים נוספים מחזקים את הבחירה ב-DistilBERT הוא מודל קל מבחינה חישובית, עם פחות שכבות ופרמטרים ולכן רץ מהר יותר, צורך פחות זיכרון, וניתן לפריסה נוחה יותר- הן בענן והן מקומית. לעומתו RoBERTa, היה האיטי והכבד ביותר, בעוד ש-SpeedBERT לא הציג יתרון משמעותי במהירות על פני המודלים האחרים.

בהתבסס על שילוב הקריטריונים- ביצועים כמותיים, פשטות מבנית ויעילות, מודל DistilBERT נבחר להיות המודל הסופי למשימת סיווג סנטימנט.

לאור תהליך ההשוואה המעמיק בין המודלים, הוחלט להתמקד לבסוף במודל GRU כמודל המרכזי לחיזוי ערכי מדד- S&P 500 זאת לאחר שהוכח כיעיל, מדויק, מהיר ויציב יותר ממודל LSTM. לאחר בחירה זו, בוצעה הרחבה משמעותית למודל על ידי הוספת חישובים טכניים נוספים שנועדו לדייק את תחזיותיו- כגון תנודתיות, אחוזי שינוי יומיים, ממוצעים נעים ועוד, אשר מרוכזים בטבלה המופיעה בתחתית העמוד. לצד זאת, רק בשלב מאוחר של תהליך ההרצה ולאחר בחינת מודלים מתקדמים לניתוח סנטימנט כגון DistilBERT הבנו שניתן לשלב את השפעת החדשות הכלכליות ישירות בתוך מודל GRU אחד, באמצעות המרת מאפייני הסנטימנט לעמודות מספריות חדשות (Sentiment_Score, sentiment_label, Is_News_Day ועוד). תובנה זו חסכה את הצורך באינטגרציית מודלים מורכבת, והפכה את GRU למודל סופי מאוחד, המכסה גם היבטים טכניים וגם רגשיים של השוק- ביעילות מרבית ונתוניו מרוכזים בטבלה אחת בשם SNP_NEWS_FINAL_GRU_3.

העמודות החדשות שנוספו לקובץ הנתונים הסופי למודל GRU

שם העמודה	תיאור הפיצ'ר	כיצד הוא משפיע על דיוק המודל	איזו בעיה/חיסרון העמודה פתרה או עזרה למזער
Pct_Change	שיעור השינוי היומי באחוזים בין סגירה של יום קודם ליום נוכחי.	מאפשר למודל להבין מגמות מיידיות-עליה או ירידה יומית חדה.	תיקון חוסר ברזולוציה יומית בנתונים המקוריים.
MA_5	ממוצע נע של מחירי הסגירה ב-5 הימים האחרונים.	מייצב את האינדיקציה הכללית לכיוון השוק על פני זמן קצר.	מזער השפעת רעשים בודדים ונקודתיים בנתוני סגירה.
STD_5	סטיית התקן של מחירי הסגירה ב-5 הימים האחרונים.	מספק מידע על רמת הסיכון והתנודתיות בתקופה האחרונה.	שיפור זיהוי מצבי קיצון או חוסר יציבות בשוק.
MACD_RSI_Ratio	יחס בין ערכי MACD ל RSI- לצורך לכידת מומנטום מחוזק.	מספק מדד יחסי על עוצמת השוק בהקשר טכני- מגביר הבחנה במומנטום.	התמודדות עם שונות גבוהה בין אינדיקטורים נפרדים.
Sentiment_Score	ציון סנטימנט ממוצע מספרי מחושב מניתוח טקסטואלי של החדשות לכל יום.	מכניס מימד רגשי למודל- מבטא את הלך הרוח הכלכלי.	התמודדות עם היעדר הקשר חד משמעי בין חדשות לתנועה במדד.
sentiment_label	קטגוריה מקודדת ממוצעת של סנטימנט כולל: חיובי, שלילי, ניטרלי של אותו יום חדשות.	מאפשר למודל להבין את סוג התגובה הציבורית לחדשות.	צמצום איבוד מידע סנטימנטלי גולמי על ידי קידוד חכם.
Is_News_Day	עמודה בינארית המצביעה אם באותו יום פורסמה חדשה כלכלית וכל יום לאחר פרסום החדשה ירד בצורה מתונה של 0.005 הערך המקודד של אותה חדשה.	מאפשר למודל להבדיל בין אירוע חדשותי חדש לבין השפעה נגררת.	התמודדות עם ריבוי ערכים חסרים עבור ימים ללא חדשות.
Volatility	שיעור התנודתיות המחושב ממדדים טכניים לאורך זמן.	מודד את התנודתיות ומשפר זיהוי סביבות שוק מסוכנות	התמודדות עם תנודתיות בלתי מוסברת במדד לאורך זמן
Change_Pct	מדד נוסף של אחוז שינוי בין ערכי סגירה ליום קודם	מוסיף מידע נוסף על שינוי יומי מעבר ל-Pct_Change	מתן מבט משלים על דינמיקת המחירים מעבר לשינויים כלליים



score_Day	*Sentiment_Score Is_News_Day	מכניס ממד אינטנסיביות של הסנטימנט בהתאם לפרסום חדשות	שילוב בין תוכן החדשה לבין מועד פרסומה – תוך חיזוק האות הסנטימנטלי
-----------	---------------------------------	---------------------------------------------------------	-------------------------------------------------------------------------