שם: ענבר שמיר 09/04/23

ת.ז: 318571643

נושא הפרויקט: ניתוח השפעת שביעות רצון עובדים על מניות החברה

פירוט:

הצעתי לפרויקט היא חיזוי מחירי המניות לפי מידע עבר על מחירי המניות יחד עם ציוני sentiment analysis על חוות דעת עובדים על מקום עבודתם מאתר glassdoor. עבור המידע על מניות החברה אשתמש בנתוני המניות מאתר yahoo finance. אבצע sentiment analysis על ביקורות מאתר glassdoor.

אשתמש ברשת LSTM למודל רגרסיה, עם מידע מביקורות ומניות בחלון זמן של 10 ימים ופלט של חיזוי ערך המניה ביום הבא. אשתמש ברשת LSTM לרגרסיה כיוון שהיא יכולה לשמור מידע מהעבר הרחוק ולהשתמש בו, יחד עם מידע מהעבר הקרוב. מה שמתאים לכך שהביקורות אינן עולות באופן יומי, אך מחירי המניות כן משתנים כל יום.

ה- feature vector של הרשת יהיה בנוי ממידע על המניה: open, low, high, previous close, percentage change, volume, ומהציון שהתקבל מ- sentiment analysis על ביקורות מאותם ימים, כאשר האינדקסים הם התאריכים. ה- target vector יהיה מחירי המניה בסוף כל יום (close).

אבצע sentiment analysis על הביקורות באנגלית באמצעות שימוש בספריית TextBlob, שמחזירה ערך polarity בין [-1,1] כאשר 1- אומר שה-sentiment שלילי, וערך 1 אומר חיובי. בנוסף, היא מחזריה ערך subjectivity בין [0,1] כך שככל שהערך גבוה יותר, הטקסט מכיל יותר דעה אישית של הכותב מאשר עובדות. אשתמש רק בציון שיתקבל מה-polarity כ-feature למודל, אוסיף לו 1 ואחלק את הסכום ב-2 כדי לקבל ציון בין [0,1]. בנוסף, קיימים מספר כלים של hugging face לביצוע sentiment analysis, כשהמתאים ביותר שמצאתי ע"י הכנסת ביקורות שונות באתר שלהם למודלים שונים הוא bert-base-multilingual-uncased-sentiment, שאומן על 150k ביקורות באנגלית למוצרים באמזון. המודל מחזיר ציון כמספר כוכבים מ-1 עד 5, ולכל אחד מהם הסתברות. הציון שאקח ממודל זה הוא חישוב הסכום של מכפלות מספר הכוכבים בהסתברות שלו, אחסיר 1 ואחלק הכול ב- 4, כדי לקבל מספר בין [0,1] המשקף כמה הביקורת חיובית.

אשווה את שני המודלים השונים שאקבל (עם ציונים ממקורות שונים ל- sentiment analysis) על ידי מטריקות MAE, MAPE מספריית sklearn, בין וקטור מחירי המניות שקיבלתי כתוצאה לווקטור מחירי המניות האמיתיים כדי לראות את ההבדלים בין המודלים השונים. מטריקות אלה מחשבת את השגיאה הממוצעת (באחוזים או בערך מוחלט בהתאמה) בין שני וקטורים שונים. כך אוכל להעריך עד כמה תוצאות המודל קרובות למציאות וגם להשוות את תוצאותיי לתוצאות המופיעות במאמר [4].

מחברי מאמרים [1],[3] בנו מודלים דומים עם שימוש בכתבות מאתר חדשות על החברות השונות. במאמר [1] הוא העריך את תוצאותיו על ידי חישוב אחוז הימים בהם המודל חזה נכון את תנועת המניה של כל חברה מתוך כלל הימים שחושבו. אשתמש גם במטריקה זו כדי להעריך עד כמה המודל צדק בחיזוי עליה/ירידה של מניה בימים מסוימים ולהשוות את עצמי לתוצאות המאמר הנ"ל.

אם זמני החישוב יאפשרו, אנסה להריץ את המודל עם ובלי המידע מהביקורות ואשווה את מחירי המניות ששני הגרסאות חזו עם המחירים האמיתיים, ואבדוק האם הוספת המידע מהביקורות אכן שיפר את החיזוי.

מסדי נתונים:

1. מאתר glass door בטווח שנים מסוים. הדרכה איך לייבא dataset מהאתר:
https://www.lexalytics.com/blog/voice-of-employee-analytics-/guide

2. מידע על מניות מאתר Yahoo Finance:
https://www.kaggle.com/datasets/achintyatripathi/yahoo-finance-apple-inc-aapl?select=AAPL_daily_update.csv
Yahoo Finance Apple Inc. (AAPL) – מידע על מניות חברת Apple בין השנים 2019-2020

References:

1. Kumar, D. (2020). Stock Forecasting Using Natural Language and Recurrent Network.

2. Akita, R., Yoshihara, A., Matsubara, T., & Uehara, K. (2016). Deep learning for stock prediction using numerical and textual information.

3. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7959635/

4. Shayan Halder. (2022). FinBERT-LSTM: Deep Learning based stock price prediction using News Sentiment Analysis.