Forecasting stock market movement using sentiment analysis

By: Osama Mohamed Elawadi Elsaeed

(2051822264)

Part A Graduation Project

Department of Data Science

Arab Open University

Dr. Sara Nabil

May14, 2024

Abstract

Knowing human behavior has become an important trend in business. Especially with the increased profit from knowing these behaviors. Which was later called "psychology".

The field in which this knowledge seems most useful is finance, specifically the stock market, because it depends entirely on the dynamics of supply and demand in the market at the current moment.

But it’s hard to follow up on all aspects that influence supply and demand. So, we will focus on the analysis of the psychology of the market with sentiment analysis. Although we divide the big problem into small but sentiment analysis is not easy to measure manually cause of the resources a lot. Here the role of Machine learning and AI comes to help investors in that.

Acknowledgment

It’s my pleasure to thank all the professionals who helped me implement this project. Most importantly I extend my most heartfelt thanks to Sara Nabil (My supervisor) whose support and guidance were of immense help during this work.

Chapter 1 Introduction

1.1 Overview

1.2 Problem Statement

1.3 Aims and Objectives

1.4 Motivation

1.5 Project Scope

Chapter 2 Literature Overview

2.1 Attachment

2.2 Sentiment Analysis Type

2.3 Extracting Information

2.4 Similarity Technologies

Chapter 3 Requirements and Analysis

3.1 Project Requirements

3.2 Use Case with Diagram

3.3 Flow Chart

3.4 Critical Problem

3.5 Suggested Solutions

Chapter 1: Introduction

1.1 Overview

The factors that investors think about during day trading are many and following everything by themselves regardless of the money they will incur will take a lot of time. In the stock market, every moment equals a life.

Therefore, the investor must follow the market psychology represented in newspapers, social media, reports, and decisions of governments, organizations, and countless other resources.

We will help investors by machine learning and AI obtain this service and give them the flexibility to control the model.

1.2 Problem Statement

With the exponential growth of platforms like social media and online communications, the necessity of using automated sentiment analysis techniques has significantly increased. Organizations offering psychological analysis of the stock market with high fees and a lot of documentation, which is a burden for small investors. In addition, they show you what they want, not what you want in terms of analysis of news, decisions, and reports.

Here the role of AI has appeared to solve this problem by building models that do the same service that other organizations represent but cheapest and make investors control it.

1.3 Aims and Objectives

- The model helps investors predict the direction of the stock market based on the force of supply and demand laws of the market in the current moment.
- Investors can use the model in other businesses whatever type of data that related to this business.
- The model was created in a dynamic way that allows anyone to use it without having to rebuild its setup.

1.4 Motivation

Psychology is the scientific study of the mind and behavior and understanding mental processes, brain functions, and behavior by Developing, evaluating, and applying new quantitative methods for the analysis of psychological data using applications of statistical models to real-world problems and Bayesian models of human cognition.

The stock market is a set of exchanges and other Venus where shares of publicly held companies are bought and sold, regardless of the simple definition of the meaning of the stock market. The processes that take place in it are very complicated. One of these processes is the analysis of people's psychology.

1.5 Project Scope

We are using deep learning techniques to extract complex patterns and features from unstructured text data, which makes them a powerful tool for sentiment analysis. Display the process of sentiment analysis, including data preprocessing, feature extraction, model architectures, and evaluation metrics. And explore the use of recurrent neural networks (RNNs), Convolutional neural networks (CNNS), and transformer models in sentiment analysis tasks. We examine the utilization of RNNs, incorporating long short-term memory (LSTM) and gated recurrent unit (GRU), to model sequential dependencies in text data.

Chapter 2 Literature Overview

2.1 Attachment

- With the growth of the internet, there have been a larger number of online factors. Finance' decision-making is influenced by the deals experiences of other individuals that are often form, news, reports, …etc. Therefore, the analysis of online data is essential. It can help investors reflect results in deals and raise percent success of it. in addition to, it may provide a practical recommendation system that helps investors have a better decision experience.
- And to do that the modern internet trying to find the best way to develop tools that help use specially with increasing of valuable resources for political science and business. Until appear subfield of natural language processing (NLP) called sentiment analysis or "Opinion mining" that focus on learn how to use machine learning algorithms to classify documents based on their sentiment and build predictor that can distinguish between positive and negative behaviors.

2.1.1 Support financial sector:

i) **Public opinion**: By monitoring news articles, financial reports, and social media posts, sentiment analysis can provide insights into public opinion and sentiment towards specific companies or sectors. This information can be used to inform investment decisions and predict market movements.

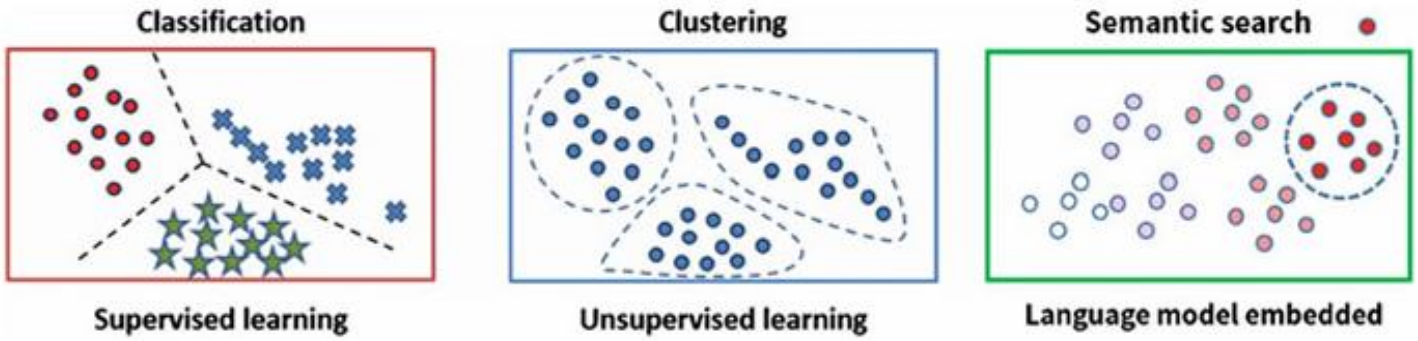
ii) **Consumer sentiment**: By monitoring customer opinions on products, services, or brands, sentiment analysis can help businesses predict future sales and identify potential risks and opportunities. For example, if sentiment analysis reveals a negative sentiment towards a particular product or service, businesses can address the issue and improve customer satisfaction.

iii) **Economic policies and events**: By understanding public opinion on issues such as taxes, regulations, and economic indicators, businesses can prepare for potential market shifts and adjust their strategies accordingly.

2.2 Sentiment Analysis Types

- **Opinion sentiment analysis**, for example, focuses on understanding the opinions expressed in text data, such as positive, negative, or neutral sentiments. This type of sentiment analysis is commonly used in customer service and marketing to gauge customer satisfaction and identify areas for improvement. It helps to understand clients' attitudes toward the product, service, or company.
- **Emotion sentiment analysis**, on the other hand, identifies specific emotions, such as joy, anger, or sadness, from text data. This type of sentiment analysis is useful to understand emotional reactions to your advertisement, product, and services.
- **Intent sentiment analysis** is another type of sentiment analysis that is used to determine the intent behind the expressed sentiment, such as purchase intent or customer service inquiries. This type of sentiment analysis is commonly used in e-commerce and customer service to identify potential customers and address their needs.
- **Polarity sentiment analysis:** it's numerical values that range from -1 to 1, where -1 indicates a very negative sentiment, 0 indicates a neutral sentiment, and 1 indicates a very positive sentiment. Polarity scores can help you quickly identify the overall mood of a text, whether it is a product review, a social media post, or customer feedback. For example, you can use polarity scores to analyze how satisfied your customers are with your service, or how your brand reputation is perceived online.
- **Classification sentiment analysis:** it's the automated process of identifying and classifying emotions in the text as positive sentiment, negative sentiment, or neutral sentiment based on the opinions expressed within. It helps determine the nature and extent of feelings conveyed using Natural Language Processing (NLP) to understand what customers say or feel about your brand, products, and services.
- **Brand monitoring**: it's the process of tracking different channels to identify where your brand is mentioned. Knowing where and how people are talking about your brand will help you better understand how people perceive it, and lets you collect valuable feedback from your audience. You can also keep an eye on potential crises and respond to questions or criticism before they get out of control and protect your brand's reputation.
- **Multilingual sentiment analysis**: it's the AI-driven process of extracting sentiment from data containing several languages. It is achieved through native language machine learning (ML) models built individually for different languages. A highly varied corpus of manually tagged data is gathered for every language to develop these models.
- **Competitive analysis**: it's the process of comparing your competitors against your brand to understand their core differentiators, strengths, and weaknesses. It's an in-depth breakdown of each competitor's market position, sales & marketing tactics, growth strategy, and other business-critical aspects to see what they're doing right and find opportunities for your business.
- **Customer insight**: it's a way the used to collection of trends in consumer behavior, data, and feedback that helps businesses deeply understand their customers and their purchasing decisions.
- **Market research:** is the process of evaluating the viability of a new service or product through research conducted directly with potential customers. It allows a company to define its target market and get opinions and other feedback from consumers about their interest in a product or service.

2.3 Extracting information



Techniques for extracting meaningful information from big data already exist and it divided to three topics:

2.3.1 Classification:

- For classifications with these techniques, computers are employed to automatically classify input objects according to their content in predefined classes. This requires the use of labeled data to be used for training in supervised learning. Because it is difficult to obtain large datasets including labels, this method is both demanding and expensive.

2.3.2 Clustering:

- Clustering is a technique that identifies similarities between objects and groups them according to their characteristics. However, because clusters are ambiguous, the accuracy of these clustering results may be poor.

2.3.3 Semantic search:

- To avoid the shortcomings of traditional techniques, we use a semantic search to effectively extract the necessary information. When a query is entered, the semantic search returns a score for the semantic similarity between the query and corpus. By filtering data at a certain score, it is possible to coarsely define a boundary that contains data related to a common query. This study uses semantic search with this property to design a ranking system.

2.4 Similar Technologies

2.4.1 Bloomberg terminal:

2.4.1.1 Intro:

- The Bloomberg Terminal revolutionized the industry by bringing transparency to financial markets. More than four decades on, it remains at the cutting edge of innovation and information delivery — with fast access to news, data, unique insight, and trading tools helping leading decision-makers turn knowledge into action.

2.4.1.2 Unparalleled Coverage:

- The terminal provides coverage of markets, industries, companies & securities across all asset classes.
- Drive innovation and productivity with connected data and technology built for how it works.

2.4.1.3 Bloomberg VS My Project:

2.4.1.3.1 Similarities:

- Both of them can depend on financial decisions.
- Both of them can't reach for source code of models.
- Both of them offer reach for statistics indicators.
- Both of them tell you about the story of the news analysis.

2.4.1.3.2 Differences:

- My application doesn't have a consultant to help investors.
- Bloomberg focuses on the finance sector only.
- My application gives data take collect to your hand.
- Bloomberg is very expensive.

2.4.2 Reuters power of APIs:

2.4.2.1 Intro:

- It's a service that delivers critical information to leading decision-makers in the legal, tax and accounting, global trade, and media markets. APIs serve as connectors and integrate Thomson Reuters (TR) solutions with existing business systems to automate data manipulation and help complete tasks more efficiently. They also serve as a bridge to connect multiple TR solutions to get the most value.

2.4.2.2 Unparalleled coverage:

- Transform data into intelligent action. Extract select content from Thomson Reuters legal platforms to integrate into your systems, workflows, and processes.
- Unlock endless ways to increase efficiency and better serve clients with a powerful tax compliance ecosystem enabled by APIs that deliver automation in each step of the workflow.

2.4.2.3 Reuters VS My Project:

2.4.2.3.1 Similarities:

- They are both link by APIs.
- They both have a recommendation system for investors.
- The cost for both is low.
- They are both used in other businesses like marketing and things related to political science.

2.4.2.3.2 Differences:

- My application needs a server to save data.
- Reuters needs to understand how it can interact with Python code.
- My application doesn't have an interface.
- Reuters doesn't support finance sectors.

2.4.3 FACTSET:

2.4.3.1 Intro:

FactSet creates flexible, open data and software solutions for investment professionals worldwide, providing instant access to financial data and analytics that investors use to make crucial decisions.

2.4.3.2 Unparalleled coverage:

- Provide exceptional client service by exploring and evaluating ideas faster with programmatic access to data exploration, research, and analysis.
- Gain a richer data experience with access to structured and unstructured data in an open flexible environment.
- Evaluate data extraction that represents, natural language understanding, and sentiment analysis with generative AI.

2.4.3.3 FACTSET VS My Project:

2.4.3.3.1 Similarities:

- They both work in multiple data not finance data only.
- They are both available for all.
- You are welcome to try both of them for free.

2.4.3.3.2 Differences:

- My application collects data.
- FACTSET represents multiple services like corporation and insurance.
- My application gives decisions not reports.
- FACTSET has a customer service.

Chapter 3 Requirements and Analysis

3.1 Project Requirements

Functional requirements	Non-Functional requirements
• The user should be able to use websites it's API is open and free or not. • The data used is in English. • The user must be familiar with the field he is searching for. • The result of the model does not have any correlation with the decisions of the user, model just only gives suggestions, not recommendations. • Used website satisfaction metrics.	• The application should have an interface that is user-friendly. • The user doesn't need to understand the code to interact with the model. • Service is not high so you can determine how you want to pay.

3.2 Use Case with diagram for the process

- **Chrome Extension:**
  - The Chrome extension built has two important modules; a background.js file and a manifest.Json, which are used for successful execution. The manifest.json file tells Chrome about the required crucial information about the extension out layer, its parts, and how to handle each one. The background script (background.js file) listens for key events or actions in the browser and reacts with specific code. To brief, the javascript file is injected, which detects the search bar changes and sends that to a service worker. After receiving these changes, it is further sent to a web socket server and is processed via a machine learning model. The final output predicted from our sentiment analysis model is sent to the foreground and a simple HTML pop-up is displayed with the predicted result (sentiment of the input text). We can simply set up the Chrome extension on our local machine by switching to developer mode on the extensions page and uploading our source code folder to apply our desired functionalities.
  - To summarize, the text entered in the search engine is taken as input to the Random Forest model via Chrome extension using service worker and WebSocket server. The trained Random Forest model and its weights are saved using a library called joblib and are further used in our Python script to establish a connection between our pre-trained model and extension. The predicted output from the model will be displayed on the web page immediately, or as per the requirement.

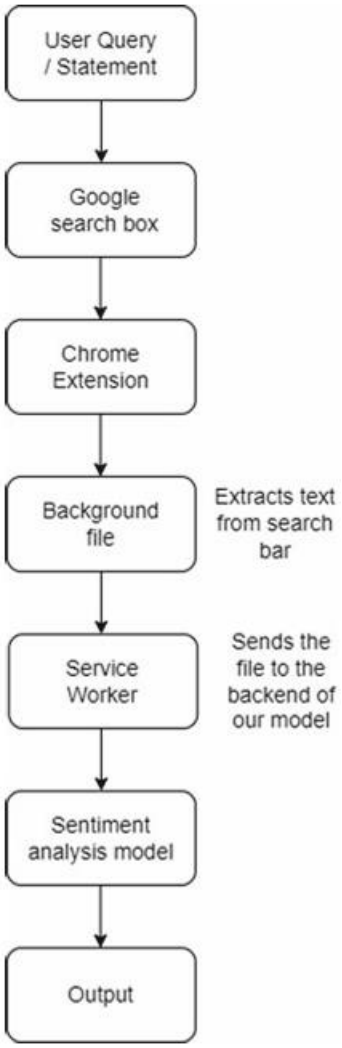
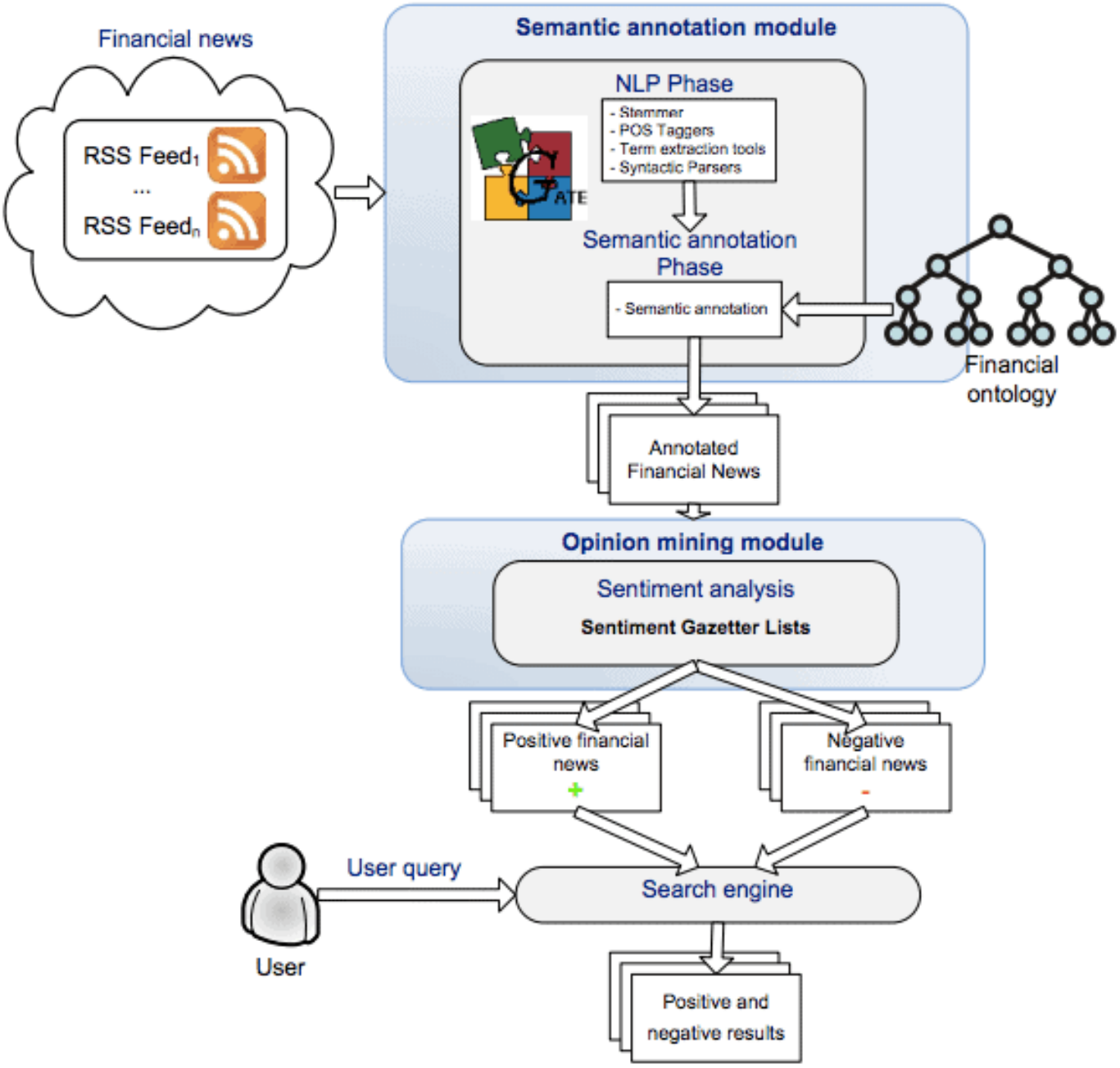


Fig.1 Flow chart of proposed model.

3.3 Flow Chart

- In the following figure, we can see the steps to analyze the sentiment tendencies or attitudes of features from data.



3.4 Critical Problems

- Most people run away that if they can predict the movement of the market, they can earn from it, it's an illusion.
- Used data not related to the topic that you are interested in.
- Relying on the model 100% is not good because the model gives a suggestion not a recommendation.
- User must be familiar with the field that data is related by, any decisions will be built on illusion not studying for the deal.
- Ensuring to comply with legal and ethical standards.
- we use good resources for scraping as we avoid evolving website structures and anti-scraping measures.
- Check your data such as its source and reliability.
- Understanding Complex human emotions.
- Language and cultural differences

3.5 Suggested Solutions

- The model will be offered as SAAS (Software as a service).
- The cost of service will be calculated according to the period you want it and your purpose for it.
- Users can use the model in multiple fields like (marketing, e-commerce, and political).

- References:

- Machine Learning with PyTorch and Scikit-Learn. By Sebastian Raschka, Yuxi (Hayden) Liu, Vahid Mirjalili. Published (2022).
- Sentiment Analysis and Deep Learning. Subarna Shakya, Ke-Lin Du, Klimis Ntalianis. Published (2023).
- Multi-Model Sentiment Analysis. Hua Xu. Published (2023).
- Create a Pipeline to Perform Sentiment Analysis using NLP. Vaibhav Haswani. Available at:https://www.analyticsvidhya.com/blog/2020/11/create-a-pipeline-to-performsentiment-analysis-using-nlp/ (Accessed: 09 November 2020)
- How can you use sentiment analysis techniques to analyze scraped text data from websites?. Available at: https://www.linkedin.com/advice/0/how-can-you-use-sentiment-analysis-techniques-analyze-n2kjc.