Analysis of stock market recommendations using computer vision

Submitted in fulfilment of the requirements
of the degree of

Masters of Technology

by

Pronoy Mandal (Registration no. 20 - 27 - 09)

Supervisor:
Dr. (Mr.) Dasari Srikanth

Co - supervisor:
Mr. Himanshu Chaudhary

Department of Applied Mathematics
Defence Institute of Advanced Technology (Pune)
2021 - 22

CERTIFICATE

This is to certify that the project entitled "**Analysis of stock market recommendations using computer vision**" is a genuine work of **Pronoy Mandal (Registration no. 20 - 27 - 09)** submitted to the Defence Institute of Advanced Technology (Pune) in fulfilment of the requirement for the award of **Masters of Technology in Data Science**.

Certified: _____

Dr. (Mr.) Dasari Srikanth
Supervisor/guide

Certified: _____

Mr. Himanshu Chaudhary
Co - supervisor
Assistant General Manager, Integrated Surveillance-
Department, SEBI

Certified: _____

Dr. (Mr.) Somanchi VSSNVG Krishnamurthy
Head of the Department of Applied Mathematics

Dissertation Approval for M.Tech.

This project report entitled **Analysis of stock market recommendations using computer vision** by **Pronoy Mandal** is certified for the degree of **Masters of Technology in Data Science**.

Examiners

1. _____

2. _____

Date:

Place:

DECLARATION

I declare that this written submission represents my ideas in my own words, and where other ideas or words have been included, I have adequately cited and referenced the original sources. I also declare that I have adhered to all principles of academic honesty and integrity and that I have not misrepresented, fabricated, or falsified any idea, data, fact or source in my submission. I understand that any violation of the above will be cause for disciplinary action by the institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

Pronoy Mandal (Registration no. 20 - 27 - 09)

Date:

Abstract

There are a large number of business NEWS channels broadcasting a variety of topics throughout morning and evening primetime. These range from simple stock market recommendations (strategies to BUY/SELL/HOLD) to detailed fundamental and technical analyses of various companies to in-depth quarterly performance reviews of various parties in the market. With the availability of such a volume of information from public sources; it becomes difficult to assimilate all the information in a single place and analyse it for meaningful information.

This project encompasses the ability to analyse NEWS which gives stock market recommendations (a subset of business NEWS) and presents the results in a tabular format to the user for analysis. The tabular format includes basic information from the broadcast such as the telecast date of the show, the analyst presenting the same, the stock which is being discussed, the NSE listing symbol of the concerned stock, the recommendation etc. (totalling 8 parameters). The exact reason why the Securities and Exchange Board of India (SEBI) is attempting to analyse NEWS videos is something that can't be specified within the scope of this report due to its confidentiality. However, it can be said that this project's output forms the base as well as the input of many anomaly (in the context of violations in the securities market) detection models at a later stage.

This project goes a step ahead to ease as well as automate the job of SEBI officers from the surveillance department by deploying the same onto an MLOPs platform as well as by creating an equivalent free-running workflow in a virtual environment.

The aforementioned reason forms the principal motivation for doing this project i.e. it helps automate a large part of a task that is prone to human errors and mistakes such as collecting data manually by watching the NEWS shows and at the same time reduces the time required to carry out such tasks. Apart from that, since an MLOPs workflow is included in the project, this project helps establish various principles which need to be followed for any ML project in general so that it can be deployed into production.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Dissertation name and title

My dissertation is titled **Analysis of stock market recommendations using computer vision**.

1.2 Problem statement

There is a large no. of daily broadcasts which take place throughout primetime (morning and evening) as well as outside of it. This project aims to digitize as well as summarise the recommendations cumulatively and present them to the end-user at the same time this project and its associated machine learning system are input to more sophisticated models for anomaly detection (in the context of violations in the securities market) at a later stage.

1.3 Proposed solution

The proposed solution is a deep learning model aimed at solving two fundamental machine learning problems at hand:

(a) Object detection – Detect and classify important regions of interest in the frames of various NEWS broadcasts (or NEWS shows) i.e. regions having (but not restricted to) the name of the analyst/presenter, the recommendation being given, the company, commodity or the market segment being discussed and its corresponding listing symbol (if any), the various metrics of the recommendation such as stop loss, target price etc.

(b) Optical Character Recognition (OCR) – Bilingual (English and Hindi) text obtained from the above regions of interest is passed onto sophisticated OCR models which have a text detection model and an actual language model(s) (corresponding to

the concerned languages) working under the hood to obtain or extract the required information.

The final output consists of a CSV (or) an excel file as required which summarises the above-obtained information.

1.4 Scope and report contents

Chapter 2 goes into a detailed discussion of the literature which has been studied while preparing this report. It includes several details such as (but not restricted to) how **YoLo** models or in general object detection models are evaluated and which performance metrics and common datasets are referred to for their speed and accuracy. It has additional details regarding how advanced deep learning models like **YoLov4** can be optimised for special use cases wherein necessary. The complete work presented in this dissertation report is broadly divided into two parts which may be completed sequentially or parallelly and are described as follows. Additionally, the chapters to which they pertain to are also mentioned.

(a) Machine learning and computer vision: There are a total of 8 NEWS shows which are currently being targeted under this project. Machine learning used within this project is further divided into two parts: the one dealing with the recognition of various numerical parameters from the video frames of a broadcast (wherein active work is going on to increase efficiency as well as the accuracy of models) are random forest classifiers which have been trained on a large number of video frames which were earlier trained on **YoLov3** output (but needed additional training due to lesser accuracy) and templates followed by a much more robust and reliable framework i.e. YoLov4 which has been trained only for one NEWS show (i.e. *PehlaSauda* from CNBC Awaaz).

The second part of ML goes into OCR models wherein **no active work is being pursued** rather readymade models and OCR engines that are doing the job within tolerable error limits are being used. Both Tesseract – OCR and EasyOCR provided models have been used for this purpose and their performance metrics analysed. It should be noted that **no part of the report** goes into the details of the inner workings of the OCR models which are deep learning models in themselves.

As opposed to standard ML projects wherein a considerable amount of time is used for training a particular model; this project uses more time and space to carry out a variety of image processing operations on the captured video frames. This includes histogram equalization, simple grey level transformations as well as complex erosion and dilation processes. After carrying out sufficient thresholding operations, these are fed into the ML models described above.

The details for the above i.e. the process of collecting the data, preparing the data for training and testing and using appropriate deep learning models for the same are discussed in 3 in detail.

(b) Deployment into production: Rarely do data science projects in academia ever reach production and into a streamlined architecture. There are several reasons which

prevent it from happening so. Such details, as well as reasons, have been given in brief in chapter 4's opening sections.

Fortunately, for this project, the deployment of the data models falls within its current scope. All the ML models which would be trained and tested would be loaded onto the proprietary **Ezmeral MLOPs** platform by **HPE**. The servers for the same reside alongside the required hardware resources on SEBI's internal servers. Doing this (without going into the details) would make the entire process streamlined i.e. the process of uploading videos, and training models on new data as soon as it is available and at the same time would decrease the manual work involved while doing so. As soon as the deployment has been completed, whether for all the NEWS shows or even a few of them, the internal server would be able to serve requests from within the organization domain (*POST* methods with *JSON* body) and at the same time would be able to abstract all the complex code and scripts running in the background.

Apart from the MLOPs deployment, it is more desirable for SEBI to have a continuously running workflow that shall consist of processing a bulk or a batch of videos at once. A workflow about this has been deployed on a standalone virtual machine as a part of this project. Which projects require an MLOPs workflow and which don't, as a result, have also been discussed in chapter 4.

Chapter 5 would then go on to discuss the various results which have been obtained at the end in terms of speed and accuracy of the YoLo models involved which metrics are suitable for evaluating projects spanning multiple ML models. Chapter 6 of the report would then finally present important conclusions drawn from the entire project.

Chapter 2

Literature review

Chapter 3

Methodology and development

Chapter 4

Deployment in production

Chapter 5

Results and discussions

Chapter 6

Conclusions

Acknowledgements

I am thankful to my institute Defence Institute of Advanced Technology (Pune) for considering my dissertation and providing help at all stages necessary during my work collecting information on my dissertation.

It gives me immense pleasure to express my deep and sincere gratitude to Dr.(Mr.) Dasari Srikanth (Dissertation guide) for his kind help and valuable advice during the development of the dissertation synopsis and his guidance and suggestions. Apart from that, I would like to thank Mr Himanshu Chaudhary (co-guide), Assistant General Manager from the Integrated Surveillance Department (ISD) of the Securities and Exchange Board of India (SEBI) for guiding me throughout this project and giving his valuable suggestions wherein necessary. I would like to express my deepest sense of gratitude to the employees of the Integrated Surveillance Department (ISD) and Information Technology Department (ITD) of SEBI who have put in their time and efforts to give additional guidance and suggestions during the project.

I would also like to thank the employees of Hewlett Packard Enterprise (HPE), a contractor of SEBI, who have allowed me to access their development environments, sophisticated hardware and cloud platforms, and associated tools and have been an immense help throughout this project.

I am deeply indebted to the Head of the Applied Mathematics Department, Dr. (Mr.) Somanchi VSSNVG Krishnamurthy, for giving me this valuable opportunity to do this project. I express my hearty thanks to him for his assistance, without which it would have been difficult to finish this project and review the project successfully. I would like to convey my deep sense of gratitude to all the teaching and non-teaching staff for their constant encouragement, support, and selfless help throughout the dissertation work. It is my great pleasure to acknowledge the help and suggestions that I received from the Department of Applied Mathematics.