Leveraging Natural Language Processing Algorithms to Understand the Impact of the COVID

Syed Shoeb Ahmed [1], Mohammed Sohaib Zahoor [2], Syed Shadab [3], M. Shilpa [4]

[1, 2, 3] B.E Student, [4] Assistant Professor,
Department of IT, Lords Institute of Engineering and Technology,
Hyderabad, Telangana, India.

Published in IJIRMPS (E-ISSN: 2349-7300), Volume 11, Issue 1, (January-February 2023)

Abstract

Understanding the effects of a pandemic on the public sentiment is an important challenge in the study of social dynamics during a global pandemic. This paper puts forward a case study that throws light on the psychological impact of the COVID-19 pandemic on the people living in the Indian subcontinent. The study is based on a pipeline that involves pre-processing, sentiment analysis, topic modelling, natural language processing and statistical analysis of Twitter data extracted in the form of tweets. The results demonstrate the effectiveness of this pipeline in understanding the temporal impact of the different lockdowns implemented in the span of the pandemic on the public sentiment, which can be useful for healthcare workers, authorities, and researchers.

Keywords: Twitter, Python, SQLite Database

Introduction

The ongoing pandemic of novel coronavirus disease, COVID-19 pandemic has left a devastating impact on the global economy, healthcare and many more sectors. With countries trying to keep up with the soaring cases, the implementation of lockdowns and containment measures is being carried out. In the face of this rapidly changing situation, people all over the world are witnessing major impacts of the pandemic on their personal lives and mental health. Recording these impacts is essential not only for understanding the depth of the current problem, but also to document sentiment changes over time for creating robust support plans in the post pandemic world. The continuous analysis of trends in sentiment corresponding to the implementation of new policies can help the authorities understand the impact of policies on the public sentiment as well. To analyse the sentiment we require time series data, either obtained directly or indirectly from the people affected by the pandemic. In this study, we propose to use Twitter data which is in the form of tweets written in the English language from India for our preliminary analysis, as Twitter is widely used by people in India and contains sufficient textual content to conduct statistical analyses. We have employed a natural language processing pipeline that converts these tweets into interpretive insights for the authorities and healthcare workers.

Requirement Analysis

The project involved analysing the design of few applications so as to make the application more users friendly. To do so, it was really important to keep the navigations from one screen to the other well ordered and at the same time reducing the amount of typing the user needs to do. In order to make the application more accessible, the browser version had to be chosen so that it is compatible with most of the Browsers.

Requirement Specification

Functional Requirements

- Graphical User Interface for the users.

Software Requirements

- Python
- Django

Operating Systems Supported

- Windows 10 (64-bit)

Technologies and Languages used to Develop

- Python

Debugger and Emulator

- Any Browser (Particularly Chrome)

Hardware Requirements

- Processor: Intel Core i9
- RAM: 32 GB
- Space on Hard Disk: Minimum 1 TB

Input and Output Design

Input Design

The input design is the link between the information system and the user. It comprises the developing specification and procedures for data preparation and those steps are necessary to put transaction data in to a usable form for processing can be achieved by inspecting the computer to read data from a written or printed document or it can occur by having people keying the data directly into the system. The design of input focuses on controlling the amount of input required, controlling the errors, avoiding delay, avoiding extra steps and keeping the process simple. The input is designed in such a way so that it provides security and ease of use with retaining the privacy. Input Design considered the following things:

- What data should be given as input?
- How the data should be arranged or coded?
- The dialog to guide the operating personnel in providing input.
- Methods for preparing input validations and steps to follow when error occur.

Objectives

1. Input Design is the process of converting a user-oriented description of the input into a computer-based system. This design is important to avoid errors in the data input process and show the correct direction to the management for getting correct information from the computerized system.
2. It is achieved by creating user-friendly screens for the data entry to handle large volume of data. The goal of designing input is to make data entry easier and to be free from errors. The data entry screen is designed in such a way that all the data manipulates can be performed. It also provides record viewing facilities.
3. When the data is entered it will check for its validity. Data can be entered with the help of screens. Appropriate messages are provided as when needed so that the user will not be in maize of instant. Thus the objective of input design is to create an input layout that is easy to follow.

Output Design

A quality output is one, which meets the requirements of the end user and presents the information clearly. In any system results of processing are communicated to the users and to other system through outputs. In output design it is determined how the information is to be displaced for immediate need and also the hard copy output. It is the most important and direct source information to the user. Efficient and intelligent output design improves the system's relationship to help user decision-making.

1. Designing computer output should proceed in an organized, well thought out manner; the right output must be developed while ensuring that each output element is designed so that people will find the system can use easily and effectively. When analysis design computer output, they should Identify the specific output that is needed to meet the requirements.
2. Select methods for presenting information.
3. Create document, report, or other formats that contain information produced by the system.

The output form of an information system should accomplish one or more of the following objectives.

- Convey information about past activities, current status or projections of the
- Future.
- Signal important events, opportunities, problems, or warnings.
- Trigger an action.
- Confirm an action.

Implementation

Modules

- User
- Admin
- Data Pre-processing
- Machine Learning

Description of the Modules

User

The User can register first. While registering he required a valid user email and mobile for further communications. Once the user register then admin can activate the user. Once admin activated the user then user can login into our system. User can upload the dataset based on our dataset column matched. For algorithm execution data must be in Integer or Float format. Here we took Adacel Technologies

Limited dataset for testing purpose. User can also add the new data for existing dataset based on our Django application. User can click the Data Preparations in the web page so that the data cleaning process will be starts. The cleaned data and its required graph will be displayed.

Admin

Admin can login with his login details. Admin can activate the registered users. Once he activate then only the user can login into our system. Admin can view Users and he can view overall data in the browser and he load the data. Admin can view the training data list and test data list. Admin can load the data and view forecast results.

Data Preprocessing

A dataset can be viewed as a collection of data objects, which are often also called as a Sentiment tweets like positive, negative and neutral. Data objects are described by a number of features that capture the basic characteristics of an object, such as the mass of a physical object or the time at which an event occurred, etc. Features are often called as variables, characteristics, fields, attributes, or dimensions. The study is based on a pipeline that involves pre-processing, sentiment analysis, topic modelling, natural language processing and statistical analysis of Twitter data extracted in the form of tweets. We use Tweets and Sentiment amount of data.

Machine Learning

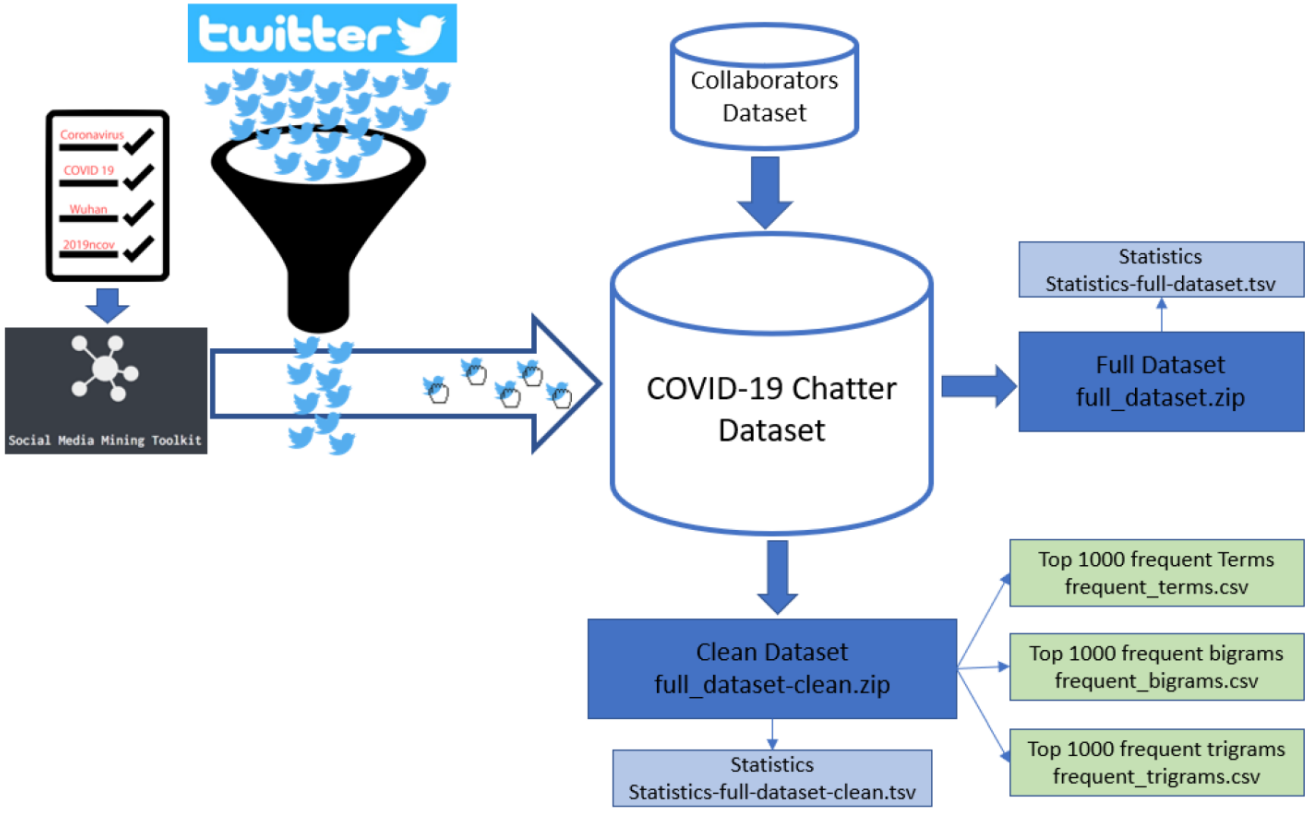
Based on the split criterion, the cleaned data is split into 80% training and 20% test, then the dataset is subjected to one machine learning classifier such as Natural Language Process (NLP). Sentiment analysis by fine tuning auto encoding models like BERT and ALBERT to achieve a comprehensive understanding of public sentiment. Thus, we have analysed the results of our experiment and methodology using the contextual information and verified the insights.

System Test

The purpose of testing is to discover errors. Testing is the process of trying to discover every conceivable fault or weakness in a work product. It provides a way to check the functionality of components, sub assemblies, assemblies and/or a finished product. It is the process of exercising software with the intent of ensuring that the Software system meets its requirements and user expectations and does not fail in an unacceptable manner. There are various types of test. Each test type addresses a specific testing requirement.

System Design

System Architecture



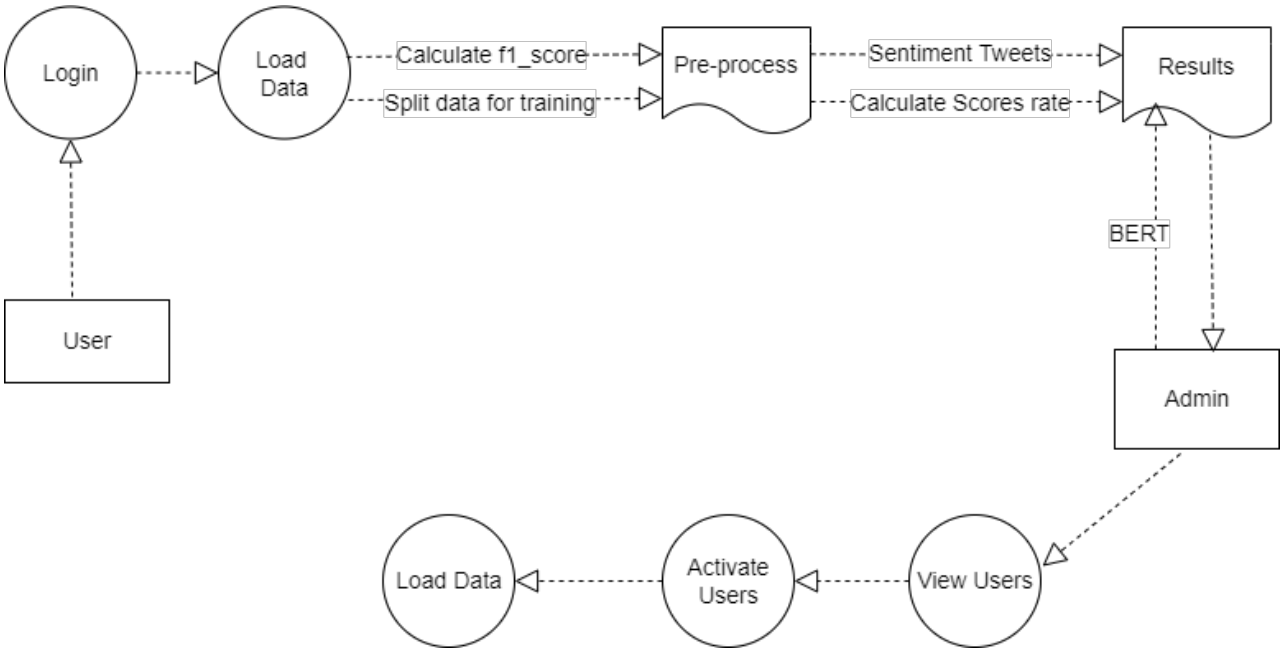
Sample Test Cases

S. No.	Test Case	Excepted Result	Result	Remarks (If Fails)
1.	User Register	User registration is successful.	Pass	If a user is already registered with the given email address then it fails.
2.	User Login	If User name and password are correct then it will get a valid page.	Pass	Non-registered Users will not be able to log in.
3.	User Add the Data	A new record will added to the dataset.	Pass	According to UCI repository, the data must be of Integer data-type, otherwise it failed.
4.	Data Cleaning	Data will be cleaned.	Pass	The data will be in Int or Float data-type, otherwise algorithm will not work.
5.	Tweet and Sentiment attribute plot a box.	Plot based graph is generated.	Pass	Target class is positive or neutral or negative then label class, else it will fail.
6.	User can add extra records for testing.	User added data will be consider for testing purpose.	Pass	Data added to test data for model.

7.	Calculate f1_positve, f1_neutral, f1_negitive Scores	For all models f1_positve, f1_neutral, f1_negitive are calculated	Pass	Data is considered for testing.
8.	Training and Testing score will be calculated.	For the four models, the training and testing data will be calculated.	Pass	Accuracy will be considered, the case fails if data is in binary format.
9.	Admin log-in	Admin can log-in with his log-in credential. If success, he get his home page.	Pass	Invalid log-in details will not be allowed.
10.	Admin can activate the registered users.	Admin can activate the registered user.	Pass	If user id not found then won't activate.

Data Flow Diagram

1. The DFD is also called as bubble chart. It is a simple graphical formalism that can be used to represent a system in terms of input data to the system, various processing carried out on this data, and the output data is generated by this system.
2. The data flow diagram (DFD) is one of the most important modelling tools. It is used to model the system components. These components are the system process, the data used by the process, an external entity that interacts with the system and the information flows in the system.
3. DFD shows how the information moves through the system and how it is modified by a series of transformations. It is a graphical technique that depicts information flow and the transformations that are applied as data moves from input to output.
4. DFD is also known as bubble chart. A DFD may be used to represent a system at any level of abstraction. DFD may be partitioned into levels that represent increasing information flow and functional detail.



System Study

Feasibility Study

The feasibility of the project is analysed in this phase and business proposal is put forth with a very general plan for the project and some cost estimates. During system analysis the feasibility study of the proposed system is to be carried out. This is to ensure that the proposed system is not a burden to the company. For feasibility analysis, some understanding of the major requirements for the system is essential.

Three key considerations involved in the feasibility analysis are:

- Economical Feasibility
- Technical Feasibility
- Social Feasibility

Economical Feasibility

This study is carried out to check the economic impact that the system will have on the organization. The amount of fund that the company can pour into the research and development of the system is limited. The expenditures must be justified. Thus the developed system as well within the budget and this was achieved because most of the technologies used are freely available. Only the customized products had to be purchased.

Technical Feasibility

This study is carried out to check the technical feasibility, that is, the technical requirements of the system. Any system developed must not have a high demand on the available technical resources. This will lead to high demands on the available technical resources. This will lead to high demands being placed on the client. The developed system must have a modest requirement, as only minimal or null changes are required for implementing this system.

Social Feasibility

The aspect of study is to check the level of acceptance of the system by the user. This includes the process of training the user to use the system efficiently. The user must not feel threatened by the system, instead must accept it as a necessity. The level of acceptance by the users solely depends on the methods that are employed to educate the user about the system and to make him familiar with it. His level of confidence must be raised so that he is also able to make some constructive criticism, which is welcomed, as he is the final user of the system.

UML Diagrams

UML stands for Unified Modelling Language. UML is a standardized general-purpose modelling language in the field of object-oriented software engineering. The standard is managed, and was created by, the Object Management Group.

The goal is for UML to become a common language for creating models of object oriented computer software. In its current form UML is comprised of two major components: a Meta-model and a notation. In the future, some form of method or process may also be added to; or associated with, UML.

The Unified Modelling Language is a standard language for specifying, Visualization, Constructing and documenting the artifacts of software system, as well as for business modelling and other non-software systems.

The UML represents a collection of best engineering practices that have proven successful in the modelling of large and complex systems.

The UML is a very important part of developing objects oriented software and the software development process. The UML uses mostly graphical notations to express the design of software projects.

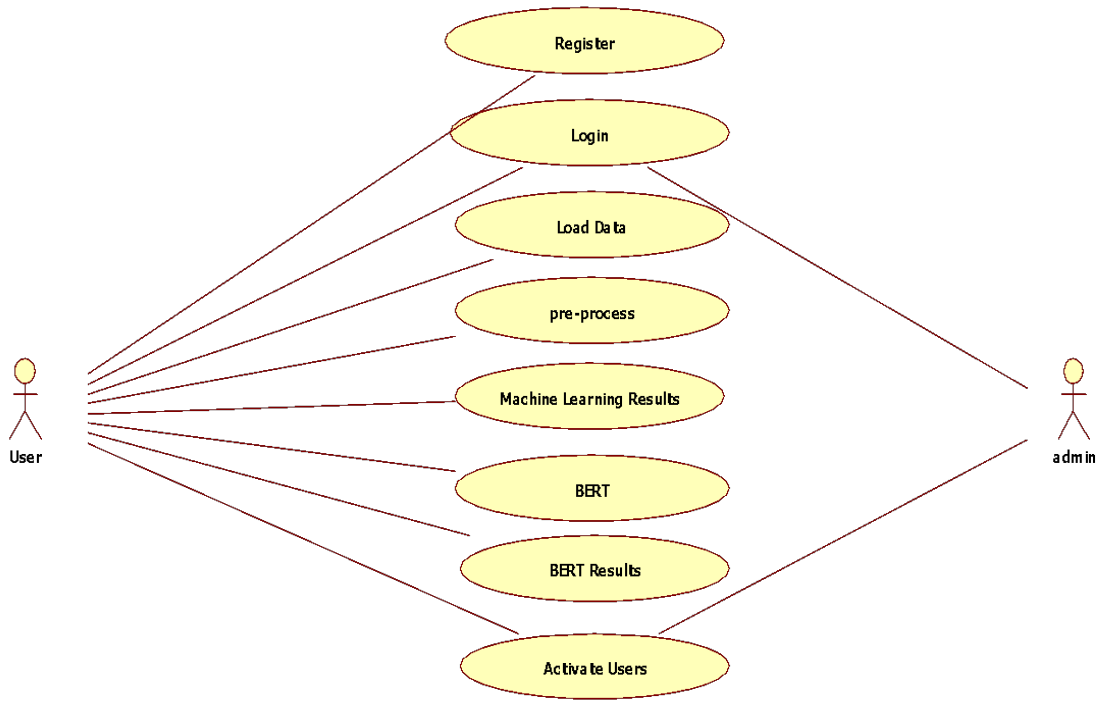
Goals

The Primary goals in the design of the UML are as follows:

1. Provide users a ready-to-use, expressive visual modelling Language so that they can develop and exchange meaningful models.
2. Provide extendibility and specialization mechanisms to extend the core concepts.
3. Be independent of particular programming languages and development process.
4. Provide a formal basis for understanding the modelling language.
5. Encourage the growth of OO tools market.
6. Support higher level development concepts such as collaborations, frameworks, patterns and components.
7. Integrate best practices.

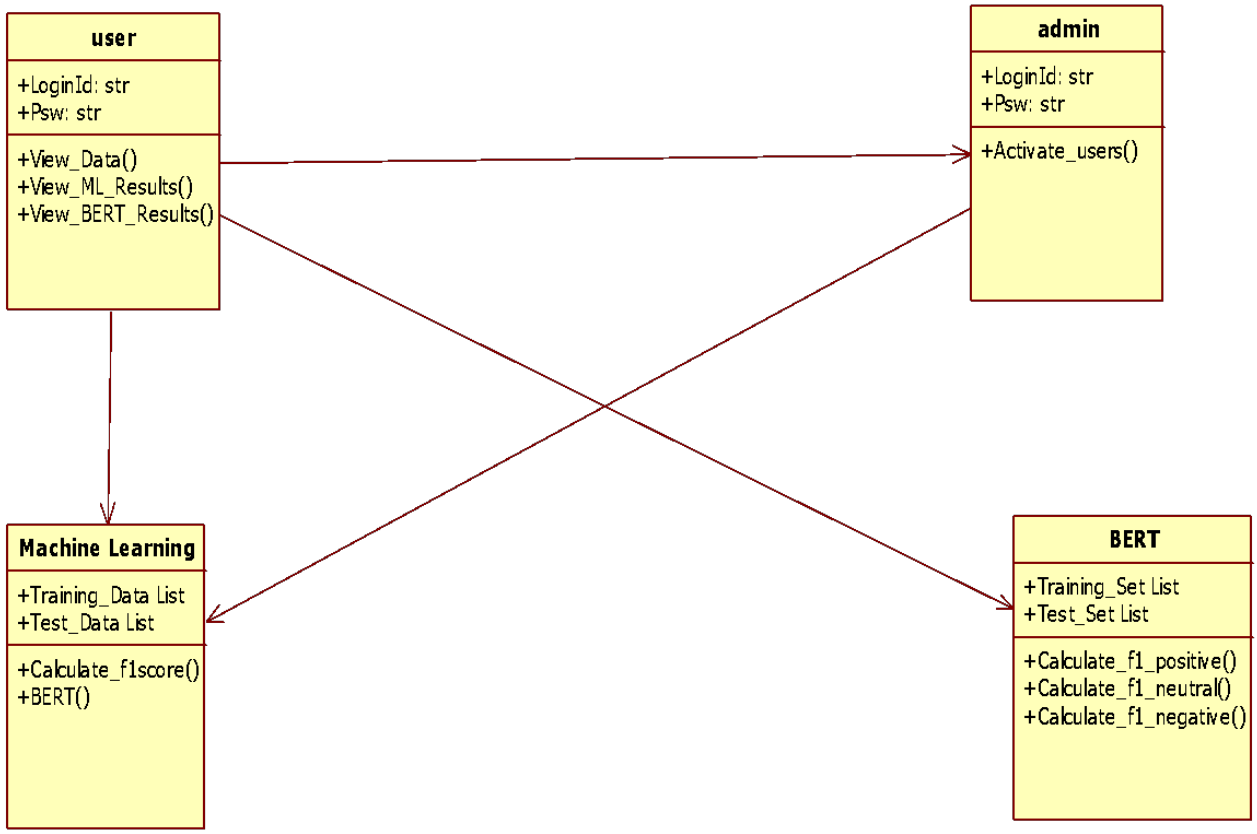
Use Case Diagram

A use case diagram in the Unified Modelling Language (UML) is a type of behavioural diagram defined by and created from a Use-case analysis. Its purpose is to present a graphical overview of the functionality provided by a system in terms of actors, their goals (represented as use cases), and any dependencies between those use cases. The main purpose of a use case diagram is to show what system functions are performed for which actor. Roles of the actors in the system can be depicted.



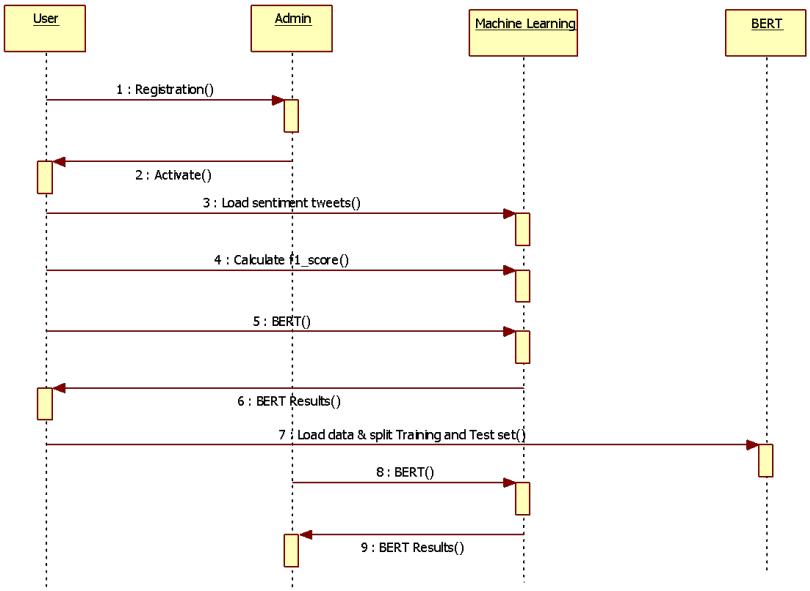
Class Diagram

In software engineering, a class diagram in the Unified Modelling Language (UML) is a type of static structure diagram that describes the structure of a system by showing the system's classes, their attributes, operations (or methods), and the relationships among the classes. It explains which class contains information.



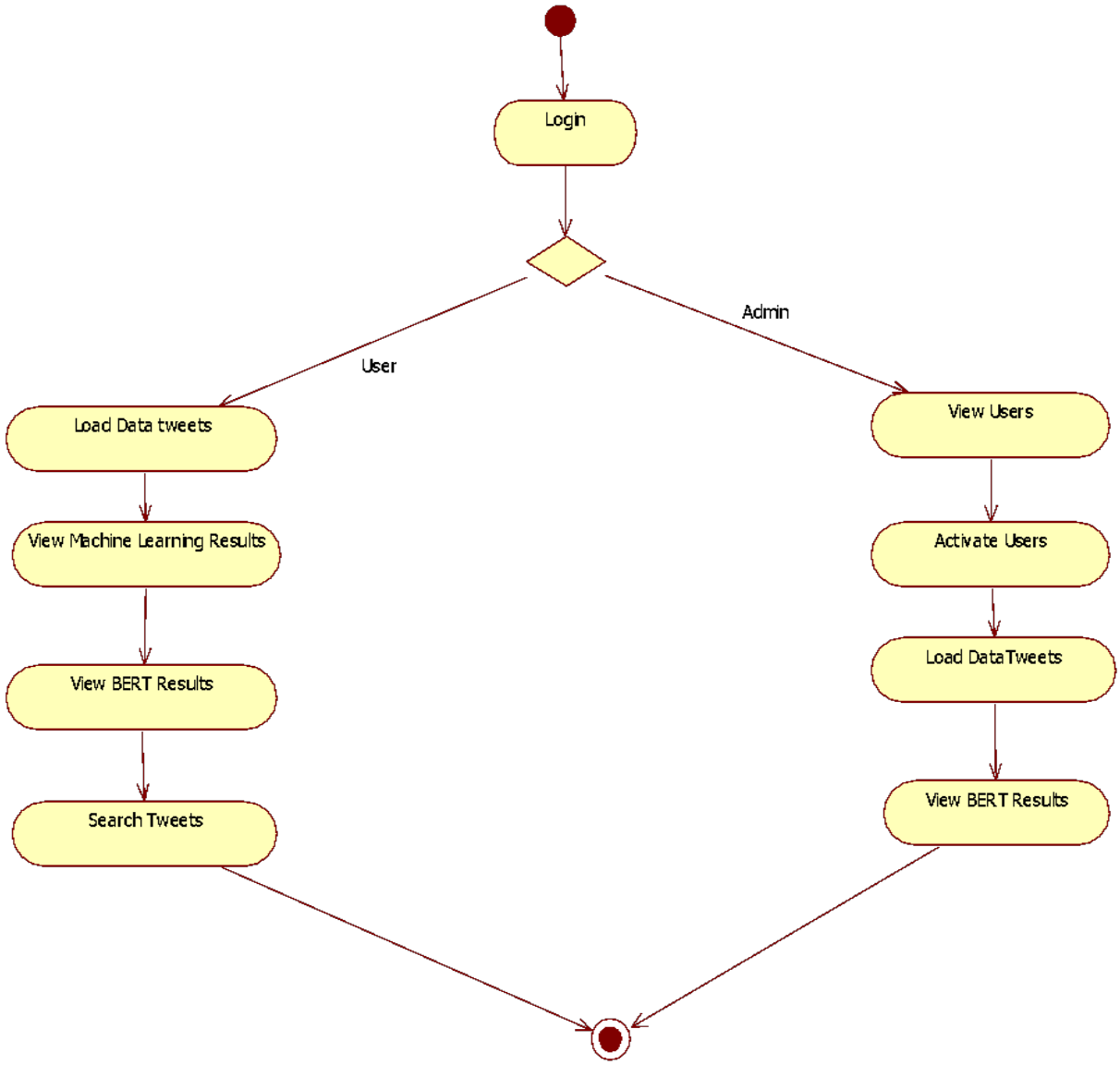
Sequence Diagram

A sequence diagram in Unified Modelling Language (UML) is a kind of interaction diagram that shows how processes operate with one another and in what order. It is a construct of a Message Sequence Chart. Sequence diagrams are sometimes called event diagrams, event scenarios, and timing diagrams.



Activity Diagram

Activity diagrams are graphical representations of workflows of stepwise activities and actions with support for choice, iteration and concurrency. In the Unified Modelling Language, activity diagrams can be used to describe the business and operational step-by-step workflows of components in a system. An activity diagram shows the overall flow of control.



References

[1] M.S. Satu, M.I. Khan, M. Mahmud, S. Uddin, M.A. Summers, J.M. Quinn, and M.A. Moni, "TClustVID: A novel machine learning classification model to investigate topics and sentiment in COVID-19 tweets", medRxiv, 2020. https://www.medrxiv.org/content/early/2020/08/04/2020.08.04.20167973

[2] M. Silva, F. Ceschin, P. Shrestha, C. Brant, J. Fernandes, C.S. Silva, A. Gregio, D. Oliveira, and L. Giovanini, "Predicting misinformation and engagement in COVID-19 Twitter discourse in the first months of the outbreak", arXiv preprint, arXiv:2012.02164, 2020.

[3] T.J.P. Luu and R. Follmann, "The relationship between sentiment score and COVID-19 cases in the USA", 2020.

[4] R. Chandrasekaran, V. Mehta, T. Valkunde, and E. Moustakas, “Twitter talk on COVID-19: A temporal examination of topics, trends and sentiments”, Journal of Medical Internet Research, 2020.
[5] S. Loria, “Textblob documentation”, Release 0.15, vol. 2, 2018.
[6] J.D. Hunter, “Matplotlib: A 2d graphics environment”, Computing in Science & Engineering, vol. 9, no. 3, pp. 90–95, 2007.
[7] M.L. Waskom, “Seaborn: Statistical data visualization”, Journal of Open Source Software, vol. 6, no. 60, p. 3021, 2021. https://doi.org/10.21105/joss.03021
[8] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding”, arXiv preprint, arXiv:1810.04805, 2018.
[9] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, “ALBERT: A lite BERT for self-supervised learning of language representations”, arXiv preprint, arXiv:1909.11942, 2019.
[10] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T.L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A.M. Rush, “Transformers: State-of-the-Art Natural Language Processing”, in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. Online: Association for Computational Linguistics, October 2020, pp. 38–45. https://www.aclweb.org/anthology/2020.emnlp-demos.6
[11] D. Lee and H. Seung, “Algorithms for non-negative matrix factorization”, in Advances in Neural Information Processing Systems 13 - Proceedings of the 2000 Conference, NIPS 2000, Neural Information Processing Systems Foundation, January 2001.
[12] Y. Zhang, R. Jin, and Z.-H. Zhou, “Understanding bag-of-words model: A statistical framework”, International Journal of Machine Learning and Cybernetics, vol. 1, no. 1-4, pp. 43–52, 2010.
[13] V. Kharde, P. Sonawane, et al., “Sentiment analysis of Twitter data: A survey of techniques”, arXiv preprint, arXiv:1601.06971, 2016.
[14] R. Wagh and P. Punde, “Survey on sentiment analysis using Twitter dataset”, in 2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA). IEEE, 2018, pp. 208–211.
[15] Imran, P. Mitra, and C. Castillo, “Twitter as a lifeline: Human annotated Twitter corpora for NLP of crisis-related messages”, arXiv preprint, arXiv:1605.05894, 2016.
[16] W.J. Corvey, S. Vieweg, T. Rood, and M. Palmer, “Twitter in mass emergency: What NLP can contribute”, in Proceedings of the NAACLHLT 2010 Workshop on Computational Linguistics in a World of Social Media, 2010, pp. 23–24.
[17] M. Kanakaraj and R.M.R. Guddeti, “Performance analysis of ensemble methods on Twitter sentiment analysis using NLP techniques”, in Proceedings of the 2015 IEEE 9th International Conference on Semantic Computing (IEEE ICSC 2015), IEEE, 2015, pp. 169–170.
[18] “NLP based sentiment analysis on Twitter data using ensemble classifiers”, in 2015 3rd international conference on signal processing, communication and networking (ICSCN), IEEE, 2015, pp. 1–5.
[19] L. Derczynski, A. Ritter, S. Clark, and K. Bontcheva, “Twitter part of speech tagging for all: Overcoming sparse and noisy data”, in Proceedings of the international conference on Recent Advances in Natural Language Processing (RANLP) 2013, 2013, pp. 198–206.