Sentimental Analysis Using MongoDB and MySQL under Hadoop for Suicidal Thoughts

**Chapter 1: Introduction**

Suicide is a significant social issue. The World Health Organization (WHO) estimates that 600 million individuals try suicide each year, and many more especially adults in their 20s and 30s—commit suicide. The second leading cause of mortality for people betwen the ages of 10 and 36 is suicide [2]. Suicidal ideantion, usually referred to as suicidal thoughts or suicidal thoughts in general, entails considering ending one's own life. For a number of reasons, including shock, , guilth, depression, and anxiety, suicidal thoughts can afflict people of all ages. Long term depresssion may lead to suicide if adequate therapy is not obtained, even though the majority of persons who have suicidal thoughts do not actually want to end their own lives [3]. With the aid of medical experts and drugs, thoughts of self harm could be controlled. But most people who experience suicidal thoughts chose not to seek treatment due to the stigma surounding medical interventions. Many people opt to publicize their plan to kill themselves on social media instead. Because mental ilness may be identified and treated, early warning signs or indications of danger may prove to be the most effective way to prevent suicide [4].In general, suicidal thoughts, a suicide scheme, and a suicide attempt are all attempts to commit suicide that have varying degrees of depth. Suicidal conception is the first unatempt strategy, a suicide plan is a technical aproach with a specific intent, and a self-harm attempt is an intended behaviour that can result in death. However, there are presently few reliable ways to spot persons who may be contemplating suicide as soon as possible so that prompt interventions can be given to keep them from acting suicidally [5]. Researchers have recently looked at two different viewpoints on people's mental health issues [6]. One is based on a conventional viewpoint that depends on professional contacts between medical personnel and patients and uses conventional scales and questions to evaluate suicidal ideation. The disadvantage of this strategy is that persons frequently hide their plans for suicide before committing themselves, suffer from willful concealing and misreporting, and are reluctant to visit a psychologist or counsellor [5,6]. In contrast, a new and developing subject is the use of social media data to access and analyze suicide screening strategies. Previous studies have shown - that due to the widespread use of the internet, young people who are considering suicide may announce - their plans or seek assistance on social networking sites [7,8,9]. Due to the growinng popularity of websites like Facebook and Twitter [10,11,12], young individuals with suicidal thoughts are posting more and more remarks on social media. Numerous research have demonstrated that internet-based vocalization of suicidal thoughts is linked to psychologically judged suicide risk [12,13], while it is unknown to which degree this online expression is analogous to physician-derived suicide risk [14].

Twitter has shown to be a useful resource for understanding the intricacy of a person's thoughts and ideas for most social media users. Most of the time, these people' tweets are unfiltered, spontaneous views that are shared with no regard for the potential implications. This can aid in determining the person's current, appropriate feelings. With such a big number of users actively using the site to express their emotions, there is a sizable amount of ongoing data to be analyzed.

To identify users who, require help, we try to build a suicidal ideation identification model based on their tweets and social media responses. It uses MongoDB and MySQL as two databases along with Apache Pyspark and ARIMA for the prediction of suicidal tendencies amongst Twitter users.

**Chapter 2: Literature Review**

Among of the fastest-growing research areas in the field of information technology is sentimennt analysis. [15] asserts that the origins of sentiment analysis can be found in public sentiment research surveys conducted at the turn of the 20th -century. In the 1990s, a group called computational linguisttics also pioneered test analysis. The accessibility of texts on the online has increased the prominence of computer-based sentiment analysis. Aside from that, the impact of text availability on the web has significantly enhanced several fields in terms of understanding the underlying emotions that accompany any text or voice message. The literature has several works on analysis of sentiment that use a variety of methodologies.

Researchers are increasingly turning to social media and NLP for their studies on mental health. Digital social media data have been a developing area of study in sentiment analysis due to the prevalence of forums dealing with mental health. [16] constructed a model that integrated LDA (Latent Dirichlet allocation), LIWCA (Linguistic Inquiry and Word Count Algorithm), bigram, and MLP ( Multilayer Perceptron) to achieve an accuracy of 90%.

Long-short-term memory, also known as LSTM, and convolutional neural network networks (CNN) are DL techniques that are significantly advancing the field of NLP as word embedding becomes more widespread. ML techniques cannot be employed for all applications because to several limitations, such as dimension explosion, limited data, and lengthy processing times. The adoption of deep learning (DL) methodologies, which enable the most important characteristics from input data, has the potential to considerably boost conventional machine learning methods. To obtain great accuracy, a model's layer count might be increased. As a result, the model will offer a categorization that is more reliable and accurate. Because DL models achieved a higher level of reliability in predicting the likelihood of suicidal thoughts than ML classifiers, it has been demonstrated in [17,18] that they are preferable to ML classifiers. With a 93.8% accuracy rate, [19] employed a CNN-LSTM model incorporating Word2Vec to forecast suicidal thoughts. This was rendered possible by the model's ability to extract both local semantic information and long-term global dependencies; nevertheless, the authors only employed a small dataset.

**Chapter 3: Data and Methodology**

It is crucial to clean and normalize the gathered data before moving on to the evaluation of the profiles and the extraction of features. Tweets that have been posted tend to be brief and noisy. For instance, the language is quite informal and contains acronyms, Unicode characters, and poor punctuation. Existing literature chooses to extract data through general methods of existing databases and read them through any IDE. But Big data technologies like Hadoop's parallel processing capacities, tolerance for failures, scalability, and affordability make suicide sentiment research using these platforms superior. These benefits enable effective analysis of big datasets, assisting in understanding and successfully resolving issues associated to suicide.

3.1 Data

Twitter provides a public API that allows for the automatic collecting of tweets as they are sent out and filtered according to predetermined parameters. Phrases such as "suicidal; suicide; kill myself; the suicide note; my self-harm letter; the end my life; never awaken up; cannot continue on; worth nothing to live; prepared to jump; rest forever; desire to die; become dead; happier with no me; better away dead; self-harm plan; suicide pact; worn out of living; not wanting to be here; die alone; go to sleep forever" indicated that the data was gathered to understand the further sentiment.

3.2 Methodology

Due to its capacity to process sizable datasets concurrently across a cluster of computers, allowing quicker analysis and adaptability to handle growing data volumes, the structure of MapReduce is advantageous for the prediction of suicidal ideation because it makes it easier to identify sequences and risk factors.

A picture containing screenshot, sketch, diagram, design

Description automatically generated

Figure : Framework

1. Hadoop Framework:

Hadoop's two main parts, MapReduce and HDFS, are what make it so strong and effective to utilize. A programming model called MapReduce is used for distributed, parallel processing of massive data collections. To create the final output, the data is separated first, then combined.

In this paper, only Hadoop MapReduce framework shall be discussed.



Figure 2: Hadoop MapReduce [20]

The Hadoop ecosystem, a framework for creating large-scale data processing, includes MapReduce. Apache Pig, Yarn, and Hadoop Distributed File System (HDFS) are additional parts of Apache Hadoop. In the Hadoop ecosystem, this component provides large-scale data processing utilizing a variety of compatible methods. Hadoop’s MapReduce architecture is used to create systems that can process enormous volumes of data. It is also known as an editing model since it enables us to process massive databases across all computer collections. This application makes it possible to store data in a distributed manner, which simplifies both a sizable machine and a sizable volume of data. MapReduce has two primary operations: map and trim. The prior task was completed before saving. The supplied data is divided into pieces and processed in accordance with the map function's instructions.

Phases of MapReduce: The phases as shown in the diagram above discusses the crux of MapReduce.

1.Mapping Phase:

This is the program's initial phase. This phase is divided into two steps: classification and mapping. In the division process, the database is split into equally sized "units" (also known as "input divisions"). A RecordReader in Hadoop transforms input variables through keyword pairs by means of TextInputFormat.

The map step then uses key-value pairs as input. A map editor is able to read or comprehend this data format. The reasoning behind the algorithm used to create these data blocks is contained in the map step. This stage involves the map analysing key pairs and producing output in the form of key-value pairs.

2. Shuffling Phase:

Following the final step of the mapping phase, this is the subsequent phase. The two main procedures are merging and filtering. Keys are used to filter keywords in the filter phase, and combining makes that key-value pairs are covered.

The eradication of duplicate data and the gathering of values are made easier during the shoplifting phase. The combination of various values with the same keys. As mentioned in the Map section, the output of this group includes keys and values.

3. Reducer Phase:

User input is the push phase's output in the reduction phase. These inputs are continually processed by the subtractor, which shrinks the median values summarizes the complete database and is provided. This category's output is kept in HDFS. In some instances, the data can be stored in a NoSQL or SQL databases as well [20].

1. **MongoDB**

An freely available document-oriented database called MongoDB allows you to work with data very effectively while storing a lot of it. Because MongoDB does not store or retrieve data in the form of tables, it is classified as a NoSQL (Not Only SQL) database [21].

The MongoDB database, initially made accessible by MongoDB.Inc. in February 2009, is created, maintained, and used in accordance with the Server-Side Public Licence (SSPL). The data is saved in the documents and collections due to its NoSQL database.Similar to how a MySQL database has tables, a MongoDB database has collections. You may make as many databases and collections as you like. We have documents inside the collection.

These files contain the data that we want to keep in the MongoDB database, and because they are schema-less, several documents may be found in a single collection without necessarily being related to one another. The fields are used to build the documents. Document fields are key-value pairs, just as relational database columns. The value of a field can be any BSON data type, such as double, string, boolean, and others. BSON documents are the format used for the data saved to MongoDB. BSON refers to the binary encoding of JSON documents in this context. Or, to put it another way, the MongoDB server transforms the JSON data into the more efficient BSON binary format in the backend, which is then stored and queried [22].

You are able to save nested data in MongoDB documents. In contrast to SQL, this data nesting lets you to construct complicated relationships among data and save it within the same document, which makes working with and obtaining data incredibly efficient. To obtain the data from tables 1 and 2, you must create intricate joins in SQL. The BSON document can be up to 16MB in size [23].

1. **MySQL**

RDBMSs like MySQL, which manages relational databases, are free to use.In one or more tables, a relational database organizes data so that connections may be made between the tables [24].With the help of the SQL programming language, which is also used to control user access to the database, data may be added to, changed, and extracted from relational databases. An RDBMS, like MySQL, interacts with the operating system to configure a relational database in the computer's memory system, manages users, provides network access, and makes it simpler to assess database integrity and build backups [25]. MySQL is open-source and free software according to the terms of the GNU General Public License. It is also accessible under a number of proprietary licences. The Swedish business MySQL AB, which sold itself to Sun Microsystems (now Oracle Corporation), was the owner and sponsor of MySQL.[8]

Due to its dependability, interoperability, and scalability, MySQL is crucial for data storage in a Hadoop cluster following data reading via Apache PySpark. MySQL, a popular relational database management system, ensures data integrity by offering robust consistency of data, transactional support, and ACID features. Because of its compatibility between Python and PySpark, data transfer is quick and easy. Sharding and replication are two of MySQL's scalability features that make it possible to handle big datasets. Furthermore, MySQL provides sophisticated querying options and optimisation strategies to improve retrieval of information and analysis. Its importance in storing data produced by PySpark within a Hadoop cluster is further bolstered by its developed ecosystem, which has a large amount of community support and tools.

1. **Apache Pyspark**

On top of Hadoop, Apache PySpark is a potent open-source platform for distributed data processing. By utilising the distributed computing power of Hadoop clusters, it enables users to quickly process and analyse massive datasets. Analysts and data scientists can utilise PySpark because of its easy Python API. PySpark supports smooth interaction with the Hadoop ecosystem and scalable data processing operations in a distributed environment by reading and processing data from a variety of file types stored in the Hadoop Distributed File System (HDFS).

1. **SARIMA (Seasonal Autoregressive Integrated Moving Average)**

The popular time series forecasting model SARIMA (Seasonal Autoregressive Integrated Moving Average) combines the ideas of ARIMA and seasonal components. It is intended to depict both the seasonal patterns and temporal dynamism found in a time series. When working with time series data that demonstrate seasonality, such as weekly, monthly, or quarterly trends, SARIMA models perform exceptionally well.

The moving average (MA), differencing (I), and autoregressive (AR) elements, along with the seasonality component, make up the SARIMA model's three primary parts. The differencing component deals with trend and stationarity difficulties, the AR component represents the dependence on prior error terms, and the MA component models the connection between the present observation and their previous values. The time series' periodic patterns that repeat themselves at regular intervals are captured by the seasonal component.

SARIMA models can successfully capture the intricate fluctuations and periodicity in data from time series by including these elements, enabling precise and trustworthy forecasting. Statistical methods like maximum likelihood estimation are used to estimate the variables of the SARIMA model.

1. **Python Dashboard**

Dash is an open-source, free Python framework for building web applications with analytical features. It is a useful package that helps developing data-driven apps simpler. For Python data analysts who aren't very experienced in web programming, it's especially helpful. With dash, users can build stunning dashboards right in their browser.

Dash connects cutting-edge UI components like dropdowns, sliders, and graphs straight to your analytical Python code. Dash is built upon top of Plotly.js, React, and Flask.

Dash apps consists of a Flask server that exchanges JSON packets over HTTP requests with front-end React components [26].Using only Python code, Dash makes it simple to create and share interactive dashboards for your data research. HTML, CSS, and advanced JavaScript frameworks like React.js are not necessary to learn.

**Chapter 4: Results and Discussion**

4.1 Results

A picture containing text, screenshot, menu, font

Description automatically generated

Figure : Hadoop Unzip

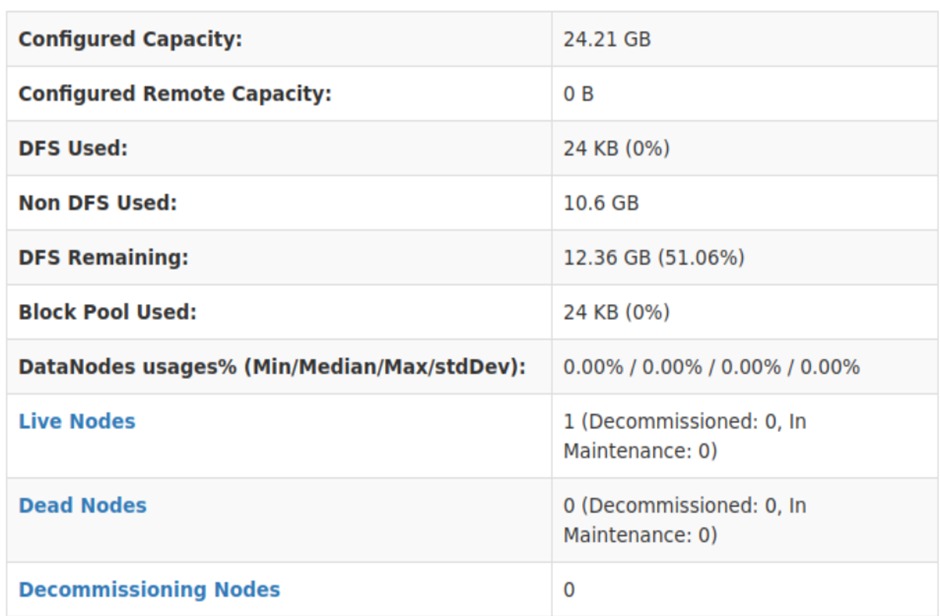


Figure : Hadoop Configuration via User Interface

A screenshot of a computer

Description automatically generated with medium confidence

Figure 5: Hadoop via UI

The following figures above showcase the installation phase of Hadoop

A picture containing text, screenshot, black and white, font

Description automatically generated

Figure : MongoDB Shell

A screenshot of a computer

Description automatically generated

Figure : SQL Schema

Figure 7 shows the SQL Scheme, whereas Figure 6 shows the MongoDB shell.

A screen shot of a computer program

Description automatically generated with low confidence

Figure 8: PySpark Preprocessing

Figure 8 showcases the preprocessing required for analysing the tweets for the sentiment of suicide ideation.

A screenshot of a computer program

Description automatically generated with medium confidence

Figure : Sample Results of PySpark

A screenshot of a computer program

Description automatically generated with medium confidence

Figure : PySpark data to be saved in MongoDB

A screenshot of a computer program

Description automatically generated with low confidence

Figure : PySpark data now saved in MongoDB

4.2 Discussion

A picture containing text, font, screenshot

Description automatically generated

Figure : Database Comparison

The paper devises a core CPU and memory utilisation test strategy to compare SQL and MongoDB quantitatively, however, given the figure 12, no distinguishable boundary could be drawn. But, from a qualitative standpoint, although raw data traversal in SQL requires rudimentary pre-processing but it is a fast-paced retrieval system for the Apache PySpark pre-processing pipelines to utilise. After which the results is stored in MongoDB as a NoSQL format, allowing fast bulk insertion which can be utilised by the dashboard for processing 7, 30 and 90 days predictions.

A screenshot of a computer

Description automatically generated with low confidence

Figure : Database Performance

**Chapter 5: Conclusion**

In conclusion, a strong framework for comprehending and forecasting sentiment patterns associated with suicidal thoughts is provided by the sentiment analysis for suicidal thoughts using a Hadoop MapReduce cluster, storing data in MongoDB, utilising Apache PySpark for processing and archiving information in MySQL, and incorporating SARIMA models for time series analysis.

Using a Hadoop MapReduce cluster makes it possible to process enormous amounts of data efficiently, which speeds up the sentiment analysis of thoughts about suicide. The flexibility and scalability of MongoDB storage makes it possible to store and retrieve unstructured or semi-structured data.

By offering robust data processing abilities and seamless Hadoop connection, Apache PySpark improves the framework by facilitating effective data reading and storage in MySQL. Combining these two elements guarantees dependable and organised database administration for sentiment analysis.

The examination of temporal patterns and trends in thoughts of suicide is made possible by the incorporation of SARIMA models for time series analysis. By capturing dependencies and autocorrelations underlying the data, these models make it possible to identify underlying patterns and predict the future using data from the past.

Overall, this combined strategy makes it easier to analyse suicidal thoughts from a sentiment perspective, adding to our knowledge of this important subject and assisting in the forecasting of sentiment trends. To assist with efforts to address and prevent suicide, it combines the scalability of Hadoop, the adaptability of MongoDB, the dependability of MySQL, and the temporal analysis abilities of SARIMA models.

**Chapter 6: References**

1.     World Health Organization. (n.d.). "Suicide Ideation Detection System." WHO Fact Sheet, Available at: https:/[www.who.int/news-room/fact-sheets/detail/suicide](http://www.who.int/news-room/fact-sheets/detail/suicide) (Accessed: 18 July 2021).

2.     Ivey-Stephenson, Z., Crosby, A.E., Stone, D.M., Gaylor, E., Wilkins, N., Lowry, R., Brown, M. (2021). "Suicidal Ideation and Behaviors Among High School Students—Youth Risk Behavior Survey, United States, 2018." MMWR Suppl, 69, 47–55. doi: 10.15585/mmwr.su6901a6.

3.     Gliatto, M.F., Rai, A.K. (1998). "Evaluation and Treatment of Patients with Suicidal Ideation." American Family Physician, 59(1500-1506).

4.     Aldhyani, T.H., Alsubari, S.N., Alshebami, A.S., Alkahtani, H. and Ahmed, Z.A., 2022. Detecting and analyzing suicidal ideation on social media using deep learning and machine learning models. International journal of environmental research and public health, 19(19), p.12635.

5.     Giachanou, A., and Crestani, F. (2016). "Like it or not: A survey of Twitter sentiment analysis methods." ACM Computing Surveys, 49(1-41). doi:10.1145/2938640

6.     Pompili, M., Innamrati, M., Di Vittorio, C., Sher, L., Girardi, P., and Amore, M. (2013). "Sociodemographic and clinical differences between suicide - ideators and attempters: A study of mood disordered patients 50 years and older." .Suicide and Life-Threatening Behavior, 44, 34–44… doi:10.1112/sltb.12050.

7.     Stankevich, M., Latyshev, A., Kuminskaya, E., Smirnov, I., and Grigoriev, O. (2019). "Depression detection from social media texts." CEUR Proceedings, 6/2523.

8.     Abdulsalam, A., and Alhothali, A. (2022). "Suicidal Ideation Detection on Social Media: A Review of Machine Learning Methods." [Title of Journal/Conference Proceedings], [Year]. [(Accessed on Day Month Year)].

9.     De Choudhury, M., Kicman, E., Dredze, M., Copersmith, G., and Kumar, M. (2016). "Discovering Shifts to Suicidal Ideation from Mental Health Content in Social Media." In Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, San José, CA, USA, 9–12 December 2016. New York, NY, USA: ACM, pp. 2098–2110

10.  Coppersmith, G., Ngo, K., Leary, R., and Wood, A. (2016). "Exploratory analysis of social media prior to a suicide attempt." In Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology, San Diego, USA, 16 June 2016, pp. 106–117.

11.  Lumontod, R.Z., III (2020). "Seeing the invisible: Extracting signs of depression and suicidal ideation from college students' writing using LIWC, a computerized text analysis." International Journal of Research Studies in Education, 9, 31–44. doi:10.5861/ijrse.2020.5007.

12.  Masuda, N., Kurahashi, I., and Onari, H. (2013). "Suicide Ideation of Individuals in Online Social Networks." PLoS ONE, 8, e62262. doi:10.1371/journal.pone.62262.

13.  Pestian, J., Nasrllah, H., Matykixwicz, P., Bennett, A. and Leenaars, A., 2011. Suicide note classification using natural language processing: a content analysis. Biomed. Inf. Insights 3.

14.  Liu, J., Shi, M. and Jiang, H., 2022. Detecting suicidal ideation in social media: An ensemble method based on feature fusion. International journal of environmental research and public health, 19(13), p.8197.

15.  Mäntylä, M.V., Graziotin, D. and Kuutila, M., 2018. The evolution of sentiment analysis—A review of research topics, venues, and top cited papers. Computer Science Review, 27, pp.16-32.

16.  Tadesse, M.M., Lin, H., Xu, B. and Yang, L., 2019. Detection of depression-related posts in reddit social media forum. IEEE Access, 7, pp.44883-44893.

17.  Aldhyani, T.H., Alshebami, A.S.A. and Alzahrani, M.Y., 2020. Soft Computing Model to Predict Chronic Diseases. J. Inf. Sci. Eng., 36(2), pp.365-376.

18.  Singh, C., Imam, T., Wibowo, S. and Grandhi, S., 2022. A deep learning approach for sentiment analysis of COVID-19 reviews. Applied Sciences, 12(8), p.3709.

19.  Tadesse, M.M., Lin, H., Xu, B. and Yang, L., 2019. Detection of suicide ideation in social media forums using deep learning. Algorithms, 13(1), p.7.

20.  Dekate, C. (2022) Learn everything about mapreduce architecture & its components, Analytics Vidhya. Available at: https://www.analyticsvidhya.com/blog/2022/07/learn-everything-about-mapreduce-architecture-and-its-components/ (Accessed: 07 May 2023).

21.  Hassan, M.K., Shakthi, S.P. and Sasikala, R., 2017, November. Sentimental analysis of Amazon reviews using naïve bayes on laptop products with MongoDB and R. In IOP Conference Series: Materials Science and Engineering (Vol. 263, No. 4, p. 042090). IOP Publishing.

22.  Trupthi, M., Pabboju, S. and Narasimha, G., 2017, January. Sentiment analysis on twitter using streaming API. In 2017 IEEE 7th international advance computing conference (IACC) (pp. 915-919). IEEE.

23.  Saini, A. (2021) What is mongodb - working and features, GeeksforGeeks. Available at: https://www.geeksforgeeks.org/what-is-mongodb-working-and-features/ (Accessed: 08 May 2023).

24.  MySQL 8.0 Reference Manual :: 1.2.1 what is mysql? (no date) MySQL. Available at: https://dev.mysql.com/doc/refman/8.0/en/what-is-mysql.html (Accessed: 25 May 2023).

25.  Engines ranking (no date) DB. Available at: https://db-engines.com/en/ranking/relational+dbms (Accessed: 03 May 2023).

26.  Dash framework (no date) Tutorials Point. Available at: https://www.tutorialspoint.com/python\_web\_development\_libraries/python\_web\_development\_libraries\_dash\_framework.htm (Accessed: 25 May 2023).