**CHAPTER THREE**

**RESEARCH METHODOLOGY**

**3.1 Introduction**

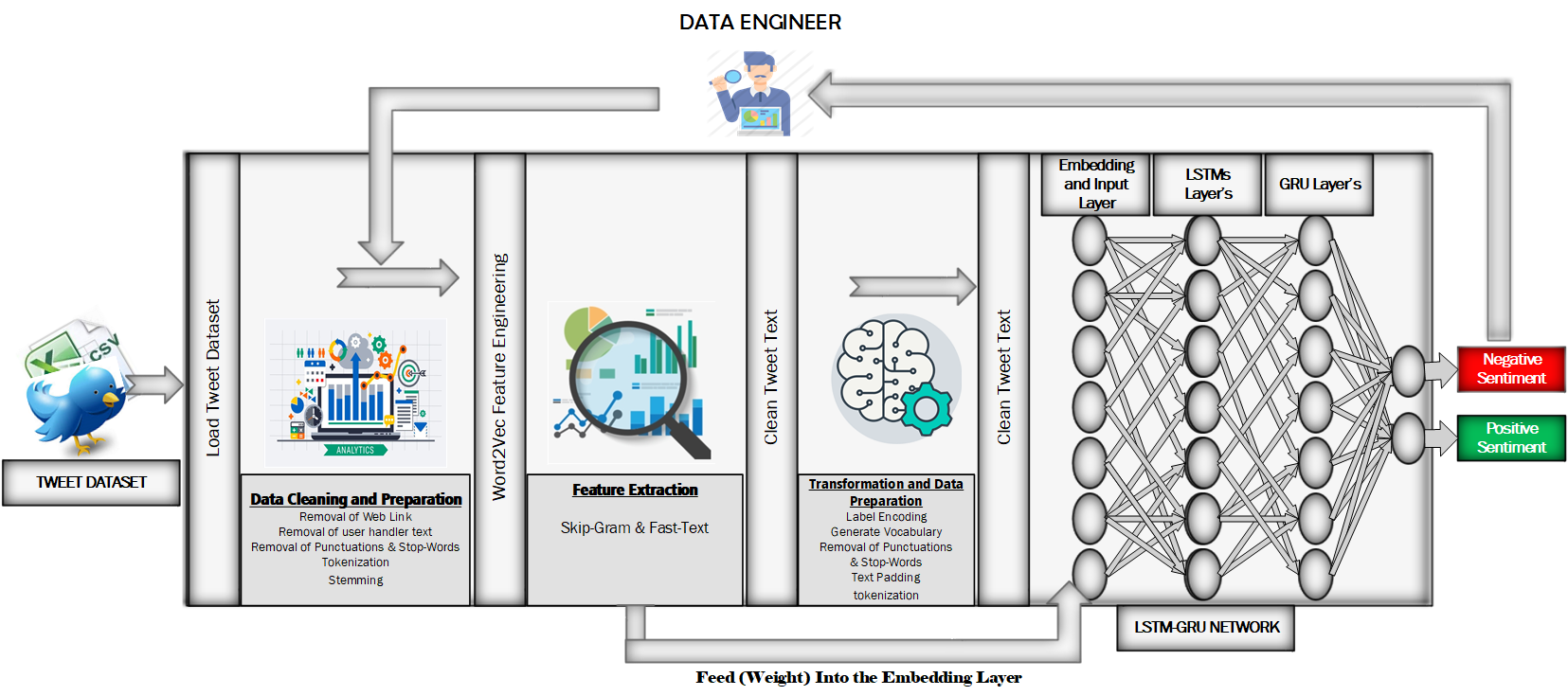
The Research methodology is considered as a very crucial chapter in any scientific writing, due to the disclosure of systematic or scientific step in actualizing the proposed ideas (model or system). detail description of how the propose model is actualize is described in this section using various concept diagram (system architecture, flowchart, use case and the likes). Additionally, the source of data collection, data preprocessing approach, evaluation metric and the proposed hybridize deep learning algorithm (LSTM-GRU) for sentiment analysis.

**3.2 LSTM-GRU Tweet Sentiment Concept Diagrams**

This subsection includes various diagram for describing the functionality, technology and methods used in developing the propose model. Diagram considered includes conceptual, data flow, and used diagrams.

**3.2.1 LSTM-GRU Conceptual Diagram**

Based on the figure 1, showing the logical concept behind designing the proposed method. This range from data importing, preprocessing mechanism, feature extraction techniques, the network layer (embedded, LSTM, and GRU layers), and finally the data engineer. Firstly, the tweet dataset is imported in a csv format and then perform feature extraction on the raw dataset using Word-2-Vec approaches (Skip-gram and FastText word embedding). The Word-2-Vec Model are shallow neural network architecture use for describing or encoding word in their contextual form in a text document. Words can be represented in various dimensional vectors using the Skip-gram and word level and at character level using FastText approach. with this feature we can have contextual information of words which is now feed to the embedding layer of the LSTM-GRU network.

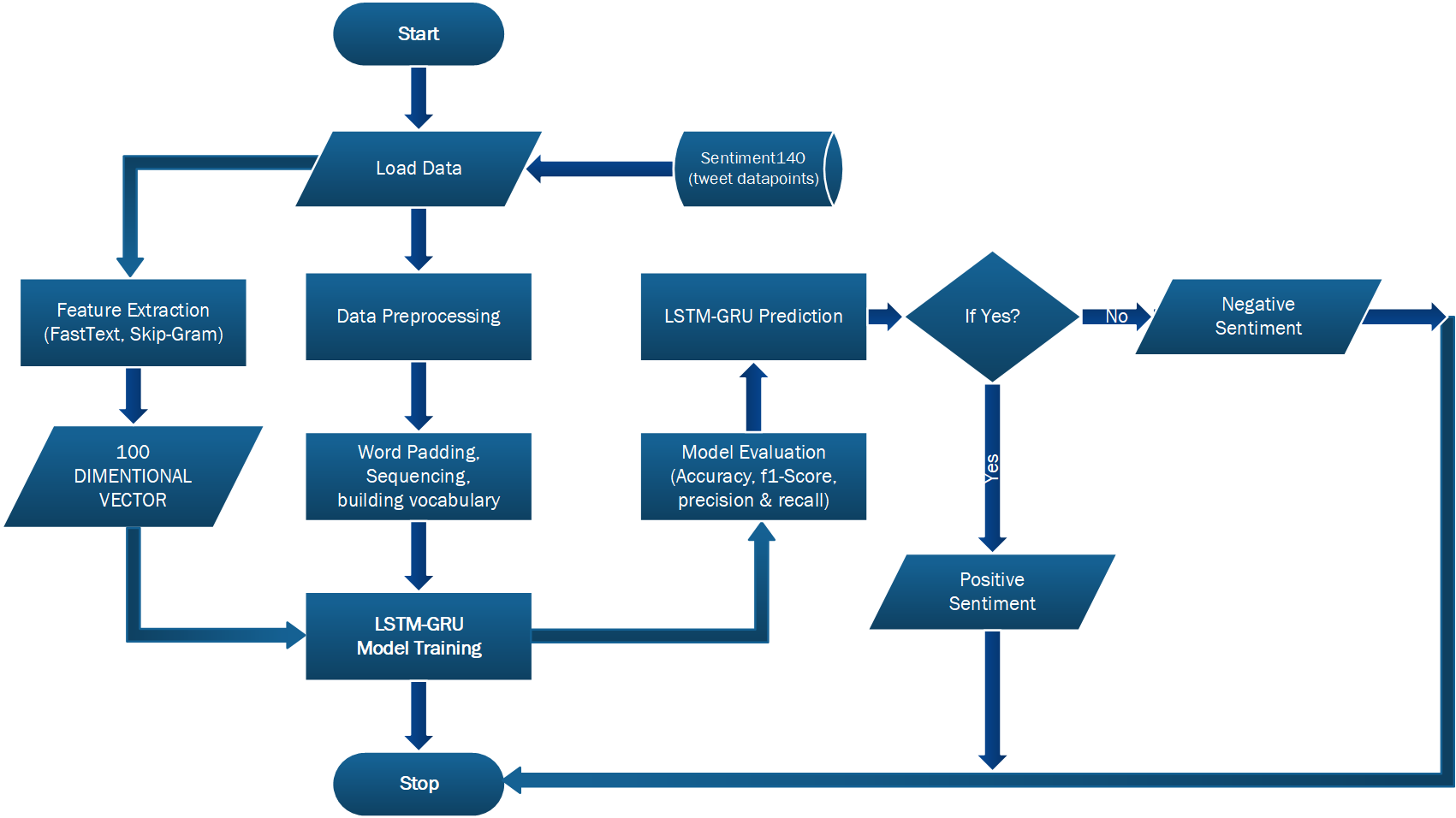


**Figure 1. Conceptual Diagram for the proposed Approach**

Furthermore, Data are clean and preprocessed using various regex matching pattern to remove, punctuation marks, stop-words, tweet user handlers and the like. Finally, the clean data are transform into sequence of number based on the vocabulary that are generated using TensorFlow modules before forwarding into the LSTM-GRU deep network. the Data engineer indicate the individual that monitors the performance of the proposed model and update certain functionality of the model if needs arise.

**3.2.1 LSTM-GRU Data flow Diagram.**

The dataflow diagram is describe using the figure 2. This show how data flows through various stage of developing the LSTM-GRU tweet sentiment model using deferent flow chart diagram, and arrow shape for denoting direction of data flow. The Figure includes start denoting the beginning of data flow, then loading of tweet dataset from local repository (HDD) for data preprocessing (stemming of words, tokenization, removal of stop words, handler, weblinks and the likes). At the point of preprocessing a copy of the clean data is taking to feature extraction using word embedding techniques (Skip-gram and Fast Text). Hundred-dimensional vector space word embedding is generated for describing the contextual information of each word. However, further preprocessing is essential before the data point features can be feed into the deep neural network input layer. Preprocessing such as creating vocabulary of words (thus, mapping of each unique word in a document to a unique number), padding which ensure each datapoint are of equal lengths, sequencing for generating numerical features of word using a vocabulary (dictionary of word to a unique number) in a sequential manner has word is originally arranged. Hence, after all the necessary feature extraction has been carryout then we can now feed the data into the LSTM-GRU network for training. Firstly, the features extracted using Word Embedding approach are

****

**Figure 2. Data flow chart diagram**

feed into the embedding layer of LSTM-GRU network, then secondly the further preprocessed, sequential, and padded datapoint are feed into the input layer of the proposed network for training. The model output is further evaluated using different performance metrics such has Accuracy, F1-score, recall and precision. After successful evaluation the LSTM-GRU network model can now take in external data for sentiment classification (positive or negative sentiment). Then the entire data flow terminated with a stop.

**3.3 Data Collections**

The Kaggle online data science repository is the sources platform considered in this sturdy for data collections and gathering. The tweet sentiment140 dataset is download on Kaggle to trained the proposed model. The file downloaded his of CSV format and it contain textural document of users’ tweet history. Kaggle repository is known to offer data scientist complete access to data for various problem and machine learning competition to facilitate learning. Compressively, Kaggle repository is a remote data science platform that is primarily established for solving academic and research problem, developing machine models, and solving difficult task using artificial intelligent (Casper et al., 2020).

**3.4 Feature Selection or Engineering**

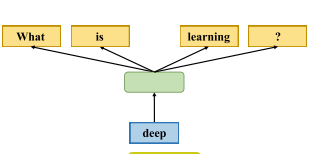
In natural language processing tasks, feature engineering includes the process of transforming textual information into numerical patterns that can help machine or deep learning model understand spoken or written words. The process of transforming word into numerical representation is called vectorization, Word2Vec models are neural network approach in extracting features by creating word embeddings (Priyadarshini & Cotton, 2021). The Word2Vec model consist of 2 layers of neural network, first layer is the input of word2vec text corpus while the output is a set of N-dimensional feature vectors. Word-2-vec ensure word with similar context generate similar embedding of various dimension. They are popular two approach of word embedding this includes Continues Bag of Words (CBOW) and Skip-gram genism model, both model is capable of learning vector representations of words.

**3.4.1 Fast Text Word Embedding**

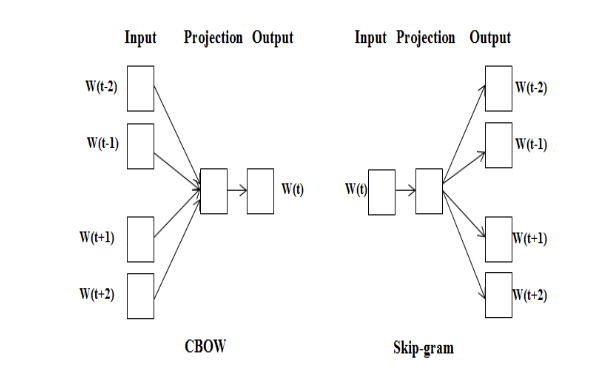
FastText is a deep learning framework developed by Facebook to generating contextual word embedding for words (Kulkarni & Shivananda, 2019). Considering situation where embedding of slang words or rare words with can poorly estimate. Hence, this prompt researcher in developing various approach to tackled this issue, Fast-Text approach is one of the methods. The Fast-text adopt sub-word level encoding of contextual information to embed rare words, but still used skip-gram model to generate feature in sub-word level. Its proven that this approach have significantly improved the performance of syntactic task and not too well in semantic problems (Wang et al., 2019). The fastText model is associated to a group of techniques, the uses combination of method to encode word representation and ordering. However, the approach considered word ordering while trying to attain efficient result. It uses similar approach as CBoW to calculate word embeddings, and it generate word embedding for general use without taking into account classes (Stein, 2018).

**3.4.2 Skip-Gram Word Embedding**

Skip-gram is another typical example of Word2Vec approach for generating word contextual embedding. The skip-gram make it practically possible to predict next word giving the surrounding or contextual words (Dang, 2020). Although word2vec uses two approaches for feature extraction in textual document and this includes Continuous back of Words (CBOW) and Skip-gram. this sturdy considered Skip-gram and the figure 3 illustrate how skip-gram generate datapoints and labels before the actual training using a shallow neural network architecture as described in figure four.



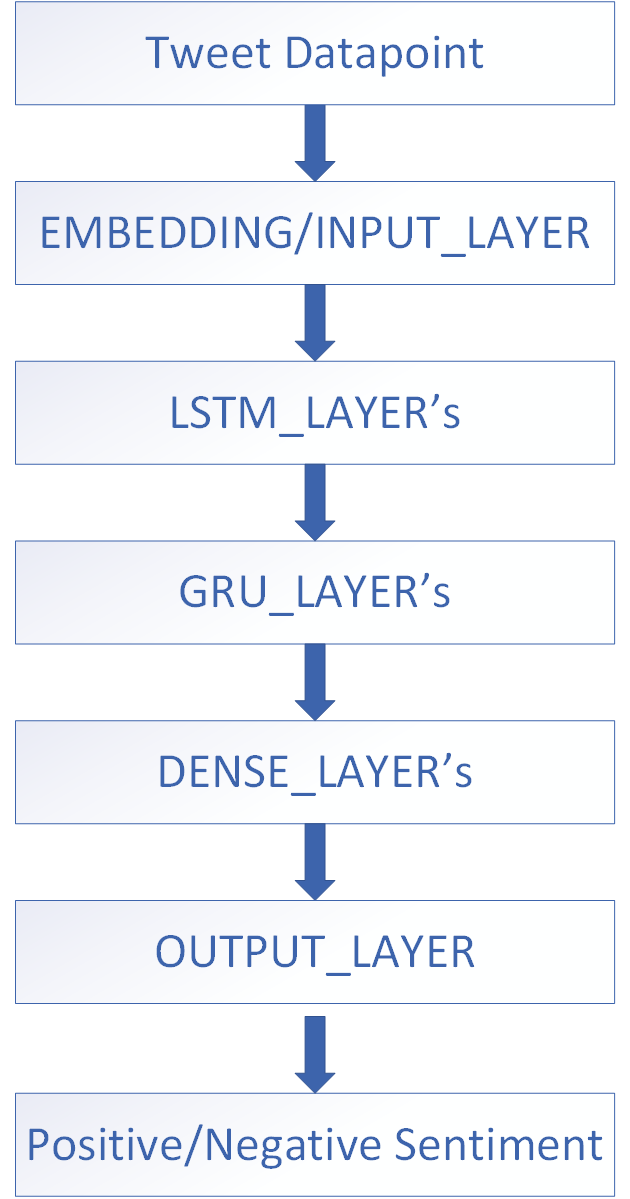
**Figure 3. Skip-gram Feature extraction (Li et al., 2020)**



**Figure 4. CBOW & Skip-gram Shallow Neural Network Architecture**

**(Sinan & Kayaalp, 2021)**

**3.5 LSTM-GRU Model Architecture**

****

**Figure 5. LSTM-GRU Model Architecture**

the research thesis considered hybridizing the Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) for training and evaluating the tweet sentiment analysis dataset. Both LSTM and GRU are deep neural network for processing time series data, this sturdy stack both network together in such a way that the weighted output of the LSTM are feed into the GRU for further weight update. LSTM perform more better in situation where the dataset is large and GRU is more efficient in terms of computational speed. However, we proposed utilizing the high-level accuracy of LSTM and process efficiency of GRU by stacking both network together instead of having only either layers of LSTM’s or GRU’s. Generally, the proposed network comprises of input layer, embedding layer, LSTM layers, GRU layers, fully connected dense layer and finally and output layer. The model will accept 85% training tweet preprocessed datapoint via the input layer and weighted features from Fast-Text and Skip-gram are supply to the embedding layer of the proposed model. furthermore, two LSTM and two GRU layer will be considered along with a fully connected dense layer. Finally, an output layer with two possible output (Positive or Negative sentiment class).

**Reference**

Casper, B., Bojer, S., & Meldgaard, J. P. (2020). *Learnings from Kaggle ’ s Forecasting Competitions*.

Kulkarni, A., & Shivananda, A. (2019). Natural Language Processing Recipes. In *Natural Language Processing Recipes*. https://doi.org/10.1007/978-1-4842-4267-4

Li, Q., Peng, H., Li, J., Xia, C., Yang, R., Sun, L., Yu, P. S., & He, L. (2020). A Survey on Text Classification: From Shallow to Deep Learning. *ACM Transactions on Intelligent Systems and Technology*, *37*(4). http://arxiv.org/abs/2008.00364

Priyadarshini, I., & Cotton, C. (2021). A novel LSTM – CNN – grid search ‑ based deep neural network for sentiment analysis. *The Journal of Supercomputing*, *0123456789*. https://doi.org/10.1007/s11227-021-03838-w

Sinan, M., & Kayaalp, F. (2021). *Sentiment Analysis on Social Media Reviews Datasets with Deep Learning Approach*. *4*(1). https://doi.org/10.35377/saucis.04.01.833026

Stein, R. A. (2018). *An Analysis of Hierarchical Text Classification Using Word Embeddings ✩*.

Study, A. C. (2020). *Sentiment Analysis Based on Deep Learning :*

Wang, B., Wang, A., Chen, F., Wang, Y., & Kuo, C. C. J. (2019). Evaluating word embedding models: Methods and experimental results. *APSIPA Transactions on Signal and Information Processing*, *8*(2019). https://doi.org/10.1017/ATSIP.2019.12