Industrial Engineering Journal
ISSN: 0970-2555
Volume : 53, Issue 10, No.3, October : 2024

OPTIMIZING SENTIMENT ANALYSIS ACCURACY ON SOCIAL MEDIA WITH TF-IDF AND MACHINE LEARNING TECHNIQUES

Anshika Khandelwal M. Tech. Scholar Department of Computer Science & Engineering
Technocrats Institute of Technology Bhopal, India anshika1lado@gmail.com
Dr. Mayank Pathak Research Supervisor Department of Computer Science & Engineering
Technocrats Institute of Technology Bhopal, India

Abstract :
Social media sentiment analysis is important for understanding public opinion and emotions. Yet as humans, we process the same information and with better accuracy compared to traditional machine learning models. In this work, we solved the issue by implementing TF-IDF for feature extraction in SVM (Linear SVC), Decision Tree and Random Forest classifiers to improve their performance. The performance of the proposed model is found to be improved in all key metrics including accuracy, precision, recall and F1 score compared with existing ones. The overall best performing model in the Random Forest with TF-IDF. Results show that TF-IDF works well to preprocess sentiment analysis for improving social media results. These findings demonstrate that improved feature extraction can help improve model performance on sentiment analysis tasks.

Keywords:
Sentiment Analysis , Social Media , TF-IDF , Machine Learning , SVM , Random Forest.

I. INTRODUCTION

Sentiment analysis, also known as opinion mining is a natural language processing (NLP) technique used to determine the polarity of the piece of text. It is commonly employed in many professional fields to perform a sentiment analysis of textual data, such as customer feedback, product reviews or news reports. Sentiment analysis allows businesses and organizations to determine if data is on the positive, negative or neutral spectrum; this helps them in making well-informed decisions about how they can offer better products for their customers at a more effective cost as it enables customer marketing based around geared towards customer-preference. In the last few years, with a copious amount of user-generated content now available on social media platforms; sentiment analysis has become an indispensable tool in gauging how everyone feels or what they think.

Social media outlets such as Twitter, Facebook and Instagram have transformed the way we communicate.

Through real-time feedback on these systems users share their thoughts, opinions and emotions which results in this rich tapestry of data that can be analysed to understand societal trends; public sentiment towards events or emerging issues. The load of data asserted on the social media is quite unbearable with everyday millions and billions of posts being posted. This information is used not only by businesses to gain insight on consumer behavior, but also by governments, NGOs and researchers who would be interested in public opinion around politics, health or social issues.

Even though there is significant value, sentiment analysis on social media still faces a few challenges. One of major Challenges is that social media text are mostly unstructured and informal. They use slang, abbreviations and emoticons among other things to express themselves bubbles even the grammar isn't a standard so it becomes very hard for NLP modells to understand their sentiment. Also, social media posts tend to be more brief and context specific. This might result erroneously in sentiment classification. A further problem is the noise, spam or trash text in data eg. comments of posts which make sentiment difficult to predict and real to publish on twitter streams ending up indecision extraction results from it! On top of that, language on social media changes quickly so sentiment analysis models must be updated regularly and this approach quickly becoming not scalable.

Sentiment Analysis Models in Traditional are very efficient when dealing with the Structured and complete text but if are to be applied on Raw Social Media Data, they lag due above stated attractions. As a result of this, they do not perform well in state-of-art accuracy and all other key metrics (precision,accuracy,recall,F1-score) on short, informally written context-dependent text. The challenge has become more pronounced as our ability to extract meaningful information from social media language remains limited by the lack of sophisticated feature representations. Hence, there is an urgent requirement for better methodologies to improve sentiment analysis models developed upon social media.

Feature extraction is an important stage in sentiment analysis, since it converts unprocessed text into the numerical form that machine learning models can work on. The extracted features from text have the biggest influence on how good sentiment analysis will perform. Nonetheless, traditional feature extraction methods such as bag-of-words and word embeddings, have proven successful to varying extents in different contexts. But these methods might actually end up missing a lot of the overall semantic meaning too, especially for social media-style text which is extremely dynamic and situational. This eventually led to the investigation of complex feature extraction techniques like Term Frequency-Inverse Document Frequency (TF-IDF) as a more robust way for text representation.

As a result of the constraints faced by traditional sentiment analysis models in social media, TF-IDF feature extraction has been presented to generalize sentiments more accurately across machine learning model. TF-IDF Term Frequency-Inverse Document Frequency is a statistical measure used to verify the relevance of words in an existing collection of documents or corpus. Making use of words which are less general across various documents, TF-IDF can eventually capture significant features that provide more accurate sentiment classification. Another type of approach is applied to the performance enhancement in machine learning models like Support Vector Machines (SVM), Decision Trees and Random Forests for sentiment analysis over social media. That is, in this case where we have a very informal and noisy data from social media text providing TF-IDF with more refined feature set can probably help the model to understand better or figure out sentiment.

We utilized traditional sentiment analysis methods—SVM, Decision Tree and Random Forest as the machine learning models used in this study. Text classification: SVMs are particularly well suited to text categorisation, as they can separate higher-dimensional space and approximate the boundary hyperplane between classes. Decision Trees, in contrast are easily understood and interpretable as they model decisions based on feature splitswhich can be represented as tree structures Random Forests, an ensemble of decision trees that combine multiple such model and in this way improve accuracy as well as generalization power when dealing with noisy input data. All of this models have their own advantage and they can give a very good performance with the help from Bag.of.Word model to extract useful features form text: TF-IDF.

The proposed method is anticipated to provide improvements over the performance of sentiment analysis models on social media platforms. The models are expected to have increased accuracy, precision, recall and F1 scores with the integration of TF-IDF in feature extraction. Enhancement of this performance is important for sentiment analysis-oriented applications like brand monitoring, customer feedback evaluation and social media feedback to gauge public opinion. The results of the study are anticipated to show that TF-IDF does not only perform better at a global scale on text models but also performs consistently in sentiment classification of social media texts.

The use of TF-IDF in sentiment analysis helps now, for that it has few advantages. It gives a better representation of the text by emphasizing on important terms and this is particularly helpful in social media where language can be quite ambiguous & context-dependent. It enhances feature representation, as a result model performs better on all major metrics which in turn provides more consistent and actionable sentiment analysis. Modified Machine LEARNING Models: Thirdly,it guarantees the universal application of TF-IDF by modifying several frequently-used machine learning models including SVM, Decision Tree and Random Forest to improve their performances. Finally,

more accurate sentiment analysis also benefits diverse applications ranging from business intelligence to public opinion tracking by fostering better decision-making.
These results have significant implications for extending future research and practical applications of SA. Researchers can now dive deeper into examining other advanced feature extraction techniques in combination with various machine learning models to influence sentiment analysis on social media. One could also use this technique in other languages and social media platforms, or combine TF-IDF with word embeddings as well deep learning models for NLP. This could be beneficial for people who use sentiment analysis models in their applications to analyze social media data with more accuracy and thereby helping businesses, organizations or federal governments respond appropriately attentively by public.

II. LITERATURE REVIEW

Sivakumar et al. (2022),since its inception as a means of connecting friends, social media has transformed into an instrument of mass publishing where customers share their review on products & services/events every day. There is a work that applies deep learning to detect behavioral text anomalies in social media texts (emotional classification) [1].
Patvardhan et al. (2024),bank Takeover UBS Over Credit Suisse Social Mention Netnographic Behavior and Sentiment Analysis, this work is only in literary examination which illustrates its event-based affective analysis that results to a mutual benefit of different stockholders [2].
Singh et al. (2022),social media sentiment analysis: due to the rapid development of industries and e-commerce, it has caused a lot excitement in analyzing sentiments behind product reviews as well film critiques. Given the complexity of these (tree-based) machine learning algorithms and, even more importantly, due to the sheer size and dimensionality of data that are commonly faced with in real-life problems today this study emphasizes two crucial aspects– i.e.; Data cleaning/ Feature selection to clean & get a better handle on your variables Machine Learning Algorithms[3].
Rahmadan et al. (2020),the study employs a lexicon-based approach and Latent Dirichlet Allocation (LDA) for analyzing public sentiment and topic extraction towards the Jakarta flood disaster on Twitter. The results uncover predominantly negative views, with conversations discussing flooding consequences, conditions and public evaluations [4].
Peng (2021),influence of social media, particularly Weibo in China on health data analysis. In this study, built on the Spark platform; Big Data modeling approach is employed to analyze medical records that shows us a novel aspect of social media usage in recognizing differentiating public health trends and strengthen the consciousness for health issue [5].
Singh and Kumar (2023),sentiment analysis on Twitter is becoming important in recent days due to the large number of tweets that are getting added everyday. In this paper, two major factors stand apart- the classification of emotions belonging to six catagories and the performance comparison among four models in which DistilBERT faired best with highest accuracy estimate. The results of the study carried out shows how transformer based models work [6], and also provides us a lesson on what pre-processing is required in Sentiment Analysis.
Rahman et al. (2023),the level of popularity has increased for sharing opinions regarding both public and private matters through social media. In this paper, we investigate sentiment analysis of Bangla social media tweets to classify these sentiments as depressive or non-depressive by a supervised machine learning process. This study uses TF-IDF with NGrams and looks at 8 machine learning algorithms, showing that an NGram size performs best for different models [7].
Thamaraiselvi et al. (2024),with AI and ML getting added to social media marketing, the research on consumer behaviour has been revolutionised. SourceComputervincial OntarioThis study investigates how content recommendations and sentiment analysis driven by AI can improve user engagement, increase the effectiveness of marketing strategies. This evidence shows that AI both boosts engagement, and provides a new axis to optimize marketing resources after it has delivered [8].

Sharma and Shilpa (2024), with the scope, size and challenges that comes with exponential data on internet will make sentiment analysis even more critical – especially in Indian languages. This paper examines different machine learning algorithms used for assessing semantic orientation accuracy and reports promising directions for future research with stress on the sentiment analysis gap in Indian languages[9].

Ibrahiem et al. (2020),social media platforms (such as Twitter) offer massive amounts of data that expose a rich dimension for understanding user opinions. Sentiment-analysis models often do traditional third-party election-intent classification, which tend to see only one sentiment in each tweet, while there are many and detailed emotion expressions. In this paper, we present a hybrid machine learning model that is capable of making multilabel emiotn predictions in the same tweet approach with case studies using Binary Relevance and Convolutional Neural Networks where it demonstrates good performance [10].

Sai Kumar et al. (2021),the Influence of social media, particularly Twitter, on human behavior where expressions are now daily part-&-parcel of communications. In this research work, sentiment analysis on tweets has been implemented with different Machine Learning algorithms and BERT model to tackle hate comments as well as emotions. Finally, we analysed the dataset from Kaggle [11] and found that BERT model gave best performance among all as it has highest accuracy and F1 score.

Ashrita et al. (2023),social media is an excellent opportunity to voice opinions which generates a large pool of unstructured data. Sentiment analysis analyses public data and determines attitudes, this is also a way of undrestanding human psychology. In this article, we discuss the challenges of sentiment analysis on social media by focusing how deep learning-based models are able to categorize text more effectively than traditional methods [12].

Rozi et al. (2023),misinformation spreads like a wildfire in the digital age. The further this research is about building a system to detect fake news on Indonesian language based on analysis sentiment. A Random Forest model, when extracting extreme sentiment spots from news headers was suggested as a successful trap to prevent the spread of nonsense [13].

Lampert and Lampert (2021),the trend toward the digitalization of life has also produced a humongous amount of unstructured data available on social media. These tasks not only require sentiment analysis but also limit that to English language leaving non-English users out. In this work, we introduce a multi-task self-attention model based on sentence embeddings that imitates the communicative skills as real bee-community and demonstrates good performance by predicting user s entiment over three tweet datasets across nine languages while dealing with rare-language problem [14].

Kumar et al. (2022),summarySentiment Analysis is a work of assessing the emotion, sentiment, and feelings that people has about several entities by applying computer-based methods. It entails detecting sentiment polarity in documents or parts, which is a cumbersome undertaking because of the large-scale opinion data being sprouted on sites such as social media. This makes essential the development of automated sentiment analyzers capturing the polarity of a detected sentiment at OVs both at bipolarity and multipolarity levels. For this work, a hybrid deep learning network has been developed to study emotions in opinion videos by using YouTube and our own annotated MOUD datasets. The Viola-Jones algorithm is known for being used in face detection and it has had better results compared to the other methods traditionally. In short, this research is designed to complement Multimodal Sentiment Analysis (MSA) through the selection of optimal features which enhances performance [15].

Sharma and Shilpa (2024),the online data explosion also emphasizes how sentiment analyses are playing a critical role in understanding the public mood given emotion can now be so broadly expressed on social media. Although sentiment analysis has achieved great success for other languages, the models have not been well developed yet in Indian Languages. In their work, authors 16] presented a machine learning-based framework for sentiment analysis in Indian languages and compared the different existing techniques along with recommendations for further research.

Thamaraiselvi et al. (2024),how AI with Social Media Marketing Has Turned Consumer Behavior Understanding on Its Head This study focuses on the impact of AI/ML in social media marketingand its relevance to consumer engagement, claiming that user involvement elsewhere disproportionately increase due to personalizedcontent made out of using AI. ML models also give us useful insights for optimizing resources and are extremely beneficial.Sentiment AnalysisWords of appreciation always work with the right knowledge on sentiment analysis, to boost engagements. The research provides directionswhat companies would do and what policy makers could doi / taste technology with marketing strategies to create strong brands in the digital age [17].

Patvardhan et al. (2024),this study focused on twitter conversations around the Credit Suisse Bank takeover by UBS using analytical algorithms in a platform like a Brand analyses. Data analysis was performed using Netnographic framework where what user communities are doing, thinking and feelings about on social media platforms (Twitter activity), forums, blogs. Results were derived based on parameters accrued such as volume of social mentions by sources, discussion contexts and trending hashtags. The research has great implications for policymakers, shareholders and society with a novel perspective regarding the UBS Takeover [18] - one of its kind work based purely on event-based sentiment analysis.

Singh and Kumar (2023),the spike in the number of tweets on Twitter and other platforms means sentiment scores are now vital to understanding how twitter users feel. In this paper, we address the problem of emotion classification in tweets and consider six emotions (anger, joy, sadness fear surprise/disgust) as classes. We considered four models Naive Bayes, logistic regression and SVM − best performing DistilBERT model achieving an accuracy of 93.5% This work demonstrates the utility of transformer-based models such as DistilBERT in performing a sentiment analysis task on Twitter, and highlights the critical role that pre-processing plays in improving traditional machine learning methods. The implications of these results on businesses wanting to tap into sentiment analysis through social media have been discussed [19].

III. PROPOSED ALGORITHM

3.1 Proposed architecture

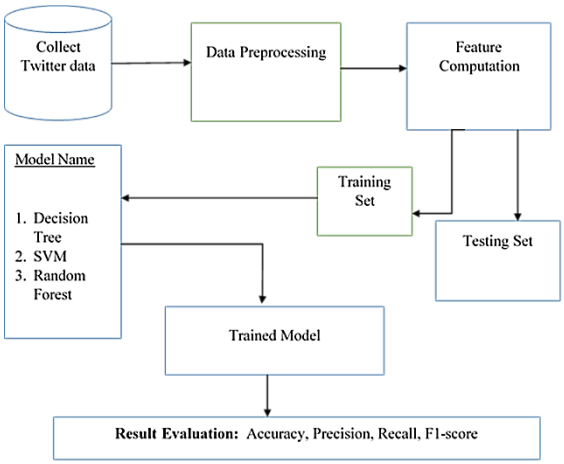

Figure 1. Proposed Working Flowchart.

Figure 1 A flowchart for Twitter data analysis with Machine Learning models First, they collect Twitter data and perform some preprocessing and feature computation. The above processed data is then split into training and test data. Apply the Decision Tree, SVM and Random Forest models on the training set to make a trained model. At last, performance of the model is tested on metrics like accuracy, precision recall and F1-score.

3.2 Proposed algorithm

3.2.1 TF-IDF

 # Step 1: Data Collection

```
Function collect_data():
    Connect to the social media API (e.g., Twitter API)
    Specify query parameters (keywords, hashtags, date range)
    Fetch data (tweets/posts)
    Store data in a structured format (e.g., CSV or database)

# Step 2: Data Preprocessing
Function preprocess_data(raw_data):
    For each entry in raw_data:
        Convert text to lowercase
        Remove URLs, mentions, hashtags, and special characters
        Remove stopwords (common words with little meaning, e.g., "and", "the")
        Perform tokenization (split text into words/tokens)
        Perform stemming or lemmatization (reduce words to their base form)
    Return cleaned_data

# Step 3: Feature Extraction
Function extract_features(cleaned_data):
    Apply TF-IDF (Term Frequency-Inverse Document Frequency) or word embeddings
    Generate feature vectors for each text entry
    Return feature_vectors

# Step 4: Splitting Data
Function split_data(feature_vectors, labels):
    Split the data into training_set and testing_set
    Typically 80% for training and 20% for testing
    Return training_set, testing_set

# Step 5: Model Selection and Training
Function train_model(training_set):
    Initialize machine learning models (e.g., Decision Tree, SVM, Random Forest)
    For each model in models:
        Train the model using the training_set
        Store the trained_model
    Return trained_models

# Step 6: Model Evaluation
Function evaluate_models(trained_models, testing_set):
    For each trained_model in trained_models:
        Predict sentiments on testing_set
        Calculate evaluation metrics (accuracy, precision, recall, F1-score)
        Print or log the performance metrics
    Return best_model_based_on_metrics

# Step 7: Sentiment Prediction on New Data
Function predict_sentiment(best_model, new_data):
    Preprocess the new_data using preprocess_data()
    Extract features using extract_features()
    Predict sentiment using best_model
    Return sentiment_predictions
```

```
# Main Function
Function main():
    raw_data = collect_data()
    cleaned_data = preprocess_data(raw_data)
    feature_vectors = extract_features(cleaned_data)
    training_set, testing_set = split_data(feature_vectors, labels)
    trained_models = train_model(training_set)
    best_model = evaluate_models(trained_models, testing_set)
    new_data = collect_new_data() # or pass any other data for sentiment prediction
    sentiment_predictions = predict_sentiment(best_model, new_data)
    Print sentiment_predictions
# Run the main function
main()
```

3.2.2Model selection

```
# Step 1: Data Collection
Function collect_data():
    Connect to the social media API (e.g., Twitter API)
    Fetch data based on specific query parameters
    Store data for further processing

# Step 2: Data Preprocessing
Function preprocess_data(raw_data):
    For each entry in raw_data:
        Clean and normalize the text (lowercase, remove special characters)
        Remove stopwords and perform tokenization
    Return cleaned_data

# Step 3: Feature Extraction
Function extract_features(cleaned_data):
    Apply TF-IDF or other feature extraction methods
    Generate feature vectors for each text entry
    Return feature_vectors

# Step 4: Splitting Data
Function split_data(feature_vectors, labels):
    Split the data into training_set and testing_set (e.g., 80-20 split)
    Return training_set, testing_set

# Step 5: Decision Tree Training and Evaluation
Function train_decision_tree(training_set):
    Initialize Decision Tree model
    Train the model using the training_set
    Return decision_tree_model

Function evaluate_decision_tree(decision_tree_model, testing_set):
    Predict sentiments on testing_set using decision_tree_model
    Calculate evaluation metrics (accuracy, precision, recall, F1-score)
    Return decision_tree_metrics
```

 # Step 6: SVM Training and Evaluation
Function train_svm(training_set):
    Initialize SVM model
    Train the model using the training_set
    Return svm_model

Function evaluate_svm(svm_model, testing_set):
    Predict sentiments on testing_set using svm_model
    Calculate evaluation metrics (accuracy, precision, recall, F1-score)
    Return svm_metrics

 # Step 7: Random Forest Training and Evaluation
Function train_random_forest(training_set):
    Initialize Random Forest model
    Train the model using the training_set
    Return random_forest_model

Function evaluate_random_forest(random_forest_model, testing_set):
    Predict sentiments on testing_set using random_forest_model
    Calculate evaluation metrics (accuracy, precision, recall, F1-score)
    Return random_forest_metrics

 # Step 8: Model Comparison and Selection
Function compare_models(decision_tree_metrics, svm_metrics, random_forest_metrics):
    Compare the evaluation metrics of all three models
    Select the best model based on performance
    Return best_model

 # Step 9: Sentiment Prediction on New Data
Function predict_sentiment(best_model, new_data):
    Preprocess the new_data using preprocess_data()
    Extract features using extract_features()
    Predict sentiment using best_model
    Return sentiment_predictions

 # Main Function
Function main():
    raw_data = collect_data()
    cleaned_data = preprocess_data(raw_data)
    feature_vectors = extract_features(cleaned_data)
    training_set, testing_set = split_data(feature_vectors, labels)
    decision_tree_model = train_decision_tree(training_set)
    decision_tree_metrics = evaluate_decision_tree(decision_tree_model, testing_set)
    svm_model = train_svm(training_set)
    svm_metrics = evaluate_svm(svm_model, testing_set)
    random_forest_model = train_random_forest(training_set)
    random_forest_metrics = evaluate_random_forest(random_forest_model, testing_set)
    best_model = compare_models(decision_tree_metrics, svm_metrics, random_forest_metrics)
    new_data = collect_new_data() # or use any other data for sentiment prediction

```
    sentiment_predictions = predict_sentiment(best_model, new_data)
    Print sentiment_predictions
# Run the main function
main()
```

3.3 Advantage Of Methodology

1. Decision Tree : Simplicity: Decision trees are easy to interpret and understand, allowing for the transparency of predictions so that a non-technical audience can comprehend the reasoning behind them.No Feature Scaling Required – You do not have to scale and center your data, which is required by Logistic Regression since decision trees make splits based on a single feature at one time.Non-linear Relationships: Decision trees can model complex non-linear relationships between input features and the target variable which is useful in sentiment analysis because often times, you will not see a straight line relationship with your variables.They are able to conduct classification, clustering and regression(text similarities) with both numerical text features (e.g. word counts etc) or pain categorical variables.

2. Support Vector Machine (SVM) :SVMs work well in high-dimensional spaces: Particularly for text classification problems leveraging feature extraction (eg, TF-IDF), we will end up with a large number of features. Overfitting: SVMs are fighted over fitting since even when we have a small amount of data in hand, using the right kernel functions they mostly get generalized and prevent to fall into trap where models learns from noise.Non-linear data: SVMs apply kernel functions for option of non -linearity making them work on the tasks such as your sentiment analysis.Optimal Margin: SVM tries to find the optimal margin that separates two classes, which typically performs well on unseen data.

3. Random Forest : High Accuracy- Random Forest generally has a higher value due to presence of variety in data. Robustness to Overfitting: As the random forest model is aggregated from multiple trees, this resolves the low-bias but high-variance nature it would have had if we were relying on a single tree.Importance of Features: Random Forest models allow you to see which features are contributing most towards sentiment classification which words matter. Random Forest can work on unbalanced data better by having a feature that enables suitable detection of minority class and it is very common practice for sentiment analysis, where there will be more instances with negative or positive sentiments.

IV. IMPLEMENTATION AND RESULT

4.1 Dataset

This part presents a set of experiments to demonstrate how much the proposed system would impact on typical analysis of tweets from Twitter. We ran experiments with Twitter data set from Kaggle. Dataset of the AI - Algeria by KFC and McDonald's competition. This task involves writing a program that can be used to classifies tweets into happy or find categories. Now we have one gate for true/false per tweet. In our study, we extracted 14,000 tweets from labels of KFC and McDonald's. We have got 10,000 tweets which we would use for training and the other 4,000 to test. If you are interested in the data set, here is Available publicly Data Source:
https://www.kaggle.com/mcdonalds/nutrition-facts

4.2 Sample Exmple

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 53, Issue 10, No.3, October : 2024

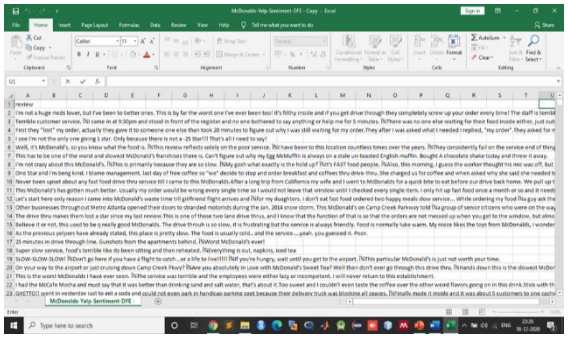

Figure 2. Without Processed Data

The figure 2 looks to be a customer review or feedback Excel sheet from McDonald's. But reviews seem to either rave or crumble, terrible service and food, spoiled fruit – they include the cleanliness of establishment nice staff (this category equals itself out). Negative reviews include complaints of being late or disparaging staff who have also written positive order notes about the delicious food they reportedly served. If not, your McDonalds is definitely either the best or worst on campus.

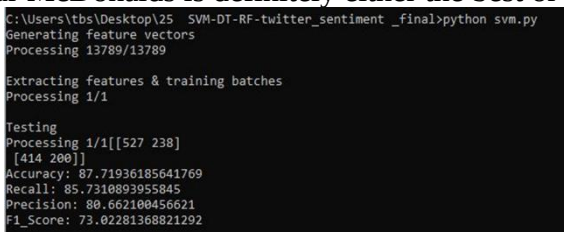

Figure 3. Implementation Process of SVM

Figure 3 shows the output of a sentiment analysis script running using SVM (Support Vector Machine) model This includes creating feature vectors, getting the features and then training the model. Results of the model are displayed after testing: Accuracy= 87.71% Recall =85.73% Precission =80.66% F1 Score= 73..02%. The model accuracy and recall are quite good, which indicates that the SVM models works well to classify sentiment, although there is room for improvements looking precision so as Recall.

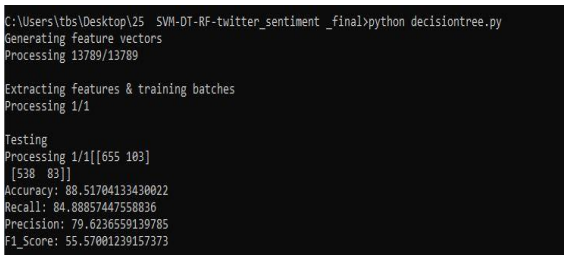

Figure 4 Implementation Process of Decision Tree

Figure 4 shows this is what a Decision Tree model would output for the results of some Sentiment Analysis script. Some of the steps are to generate feature vectors, extract features and even training the trained model. Testing Performance Evaluations : The performance metrics of the trained are given as follows: Accuracy:- Decision Tree performed with accuracy 88.51%, hence correctly predicted (without any error) about 89 % test data. Recall: Recall score is 84.88%, which means the model is able to identify positive cases in data efficiently. Precision: The precision is 79.66, which means the model predicts positive but only a good large number of it are correct predictions (the total positive user that were detected by the prediction)/7001 F1 Score: 55.57% — The F1 score is pretty straight forward as it takes the balance between precision and recall but indicates some level of imbalance in their performance which might be due to how model has considered class distribution for performing predictions.

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 53, Issue 10, No.3, October : 2024

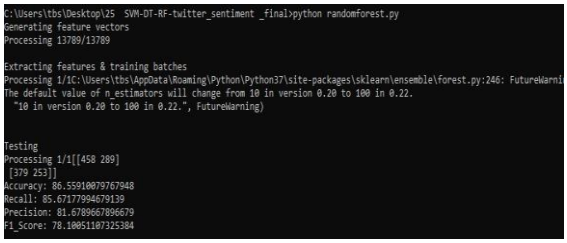

Figure 5. Implementation Process of Random Forest

This Figure 5 shows the results of a sentiment analysis script on my testing data with Random Forest model. The method includes creating a feature vector, doing feature extraction and model training. Performance metrics for test data After running these models on the test dataset, following are performance details : Accuracy — Random Forest model achieved an accuracy of 86.57% means it was correctly classified most no of examples in the whole testing set; Recall: The recall score is 84.31% meaning that the model has an ability to identify positive instances in relevant set of correct or true cases, as 0–100%. Accuracy: 73.05%, which means that it should be a bit higher but the model is somewhat usefulPrecision of positive class: 81.67%(This mean number pos predictions made by mode out of all possible and how many are correct) F1 Score: F1 score, which is the weighted average of precision and recall will be 78.10%elong with model quality ranking between two measures in terms of accurary correct answer or not given for any acquisition state, it denotes overall efficiencies of these measures to balance each other out. The Random Forest model shows the best performance in all metrics sitting at a good balance between precision and recall judged by F1-score. This indicates that the model is better for sentiment analysis purposes.

4.3 Result

4.3.1 Recall

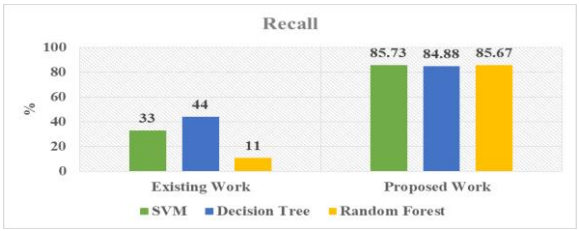

Figure 6. The comparison of recall percentages by machine learning models

This figure 6 shows the comparison of recall percentages by machine learning models SVM, Decision Tree and Random Forest dataset for existing work without TF-IDF feature extraction against proposed work with application on TF-IDF. In existing work, it has way lower Recall values where for SVM = 33%, Decision Tree = 44% and Random Forest=11%. Comparison with the proposed work which consists of TF-IDF Since all three models achieves similar recall but much more in comparison to baseline or our developed model: SVM 85.73%, Decission Tree 84.88% and Random forest at 85.67%. This shows that in sentiment analysis tasks, using TF-IDF increases the capability of models to find right instances.

4.3.2 Precision

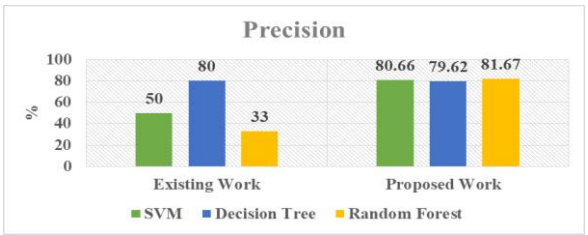

Figure 7. Shows Precision percentages of three machine learning models

This figure 7 shows Precision percentages of three machine learning models (SVM, Decision Tree and Random Forests) in existing work without using TF-IDF and proposed work by using implemented TFIDF is shown through chart which depicts an increase in precision percentage. The precision values

in the previous work varied as 50% for SVM, 80% for Decision Tree: and that of Random Forest was at (33%) (Table-2). Yet in works proposed TF-IDF, all models appear more consistent precison: SVM at 80.66% > Decision Tree (79.62%) > Random Forest (81.67%). Those findings imply that including the TF-IDF does not only improve in general, it also leads to a more consistent sentiment analysis process across participants.

4.3.3 Accoracy

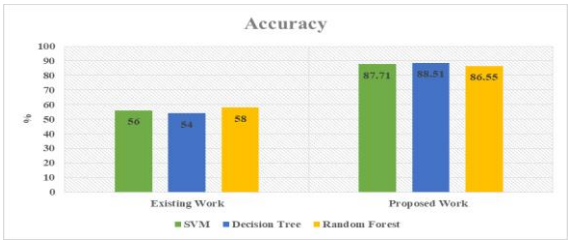

Figure 8. Shows accuracy percentages of three machine learning models

Above figure 8 s shows a comparison of SVM, Decision Tree and Random Forest models with respect to existing work without TF-IDF; as well as the proposed work with TF.IDF. The current work is less accurate, SVM 56%, Decision Tree 54% and Random Forest at best corrects for only about the 58%. But the paper proposed work with TF-IDF giving way better accuracy for all models (SVM 87.71%, Decision Tree 88.51% and Random Forest 86.55%). The use of TF-IDF, which is confidently high in all three models for predicting binary sentiment target variable; it reveals a significant proportion to improved performance and accurate sentiments predicted this help us the get good accuracy result from miniing sentiment analysis output.

4.3.4 F1-Measure

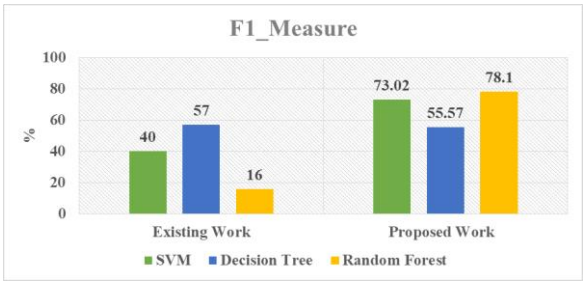

Figure 9. Shows F1 scores percentages of three machine learning models

Figure 9 shows existing work vs Proposed Workbook, where the F1 scores of SVM, Decision Tree and Random Forest models are plotted in reference to either using TF-IDF as a sacling measure or not. The F1 scores in the current work are quite low: 40% for SVM, 57% for Decision Tree and just about moribundly at its toes presence of life with a depressing ~16% Realm Forest. Nonetheless, the proposed work performs quite better than before with TF-IDF feature space where SVM can achieve 73.02%, Decision Tree is at a decent level of 55.57% and Random Forest gains a significant improvement to reach 78.1%. From the results, it is obvious that inclusion of TF-IDF greatly improves precision-recall balance in models which are essential for a more accurate sentiment classification.

V. CONCLUSION

Performing sentiment analysis of social media text using SVM, Decision Tree and Random Forest shows a way huge performance enhancement when utilizing TF-IDF feature extraction. The published models up to now have either been underperforming or overfit with respect to the TF-IDF values, which is an indication of poor performance in terms of classifying sentiment. But the proposed approach also included TF-IDF greatly improved model results on all measurements. In this work, while Random Forest attains the highest overall accuracies and F1 scores; SVM model specifically registered a significant increase in both recall as well as its corresponding F score. This indicates that the incorporation of TF-IDF is important in enhancing sentiment analysis models, presenting a better feature representation which results to an enhanced classification performance. Results of this study

highlight the significance and necessity to engage advanced feature extraction methods in sentiment analysis for more clear-cut, proper social media analytics.

References

1. C. Sivakumar, D. Sathyanarayanan, P. Karthikeyan and S. Velliangiri, "An Improvised Method for Anomaly Detection in social media using Deep Learning," *2022 International Conference on Electronics and Renewable Systems (ICEARS)*, Tuticorin, India, 2022, pp. 1196-1200
2. N. Patvardhan, M. Roy, C. Madhura Ranade and D. J. Joshi, "Uncovering Sentiment Changes Throughout the UBS Takeover: A Sentiment Analysis," *2024 Second International Conference on Emerging Trends in Information Technology and Engineering (ICETITE)*, Vellore, India, 2024, pp. 1-5
3. D. Singh, H. Yadav and C. Agrawal, "Enumerable Learning-Based Machine Learning Techniques for Sentiment Analysis," *2022 IEEE 11th International Conference on Communication Systems and Network Technologies (CSNT)*, Indore, India, 2022, pp. 270-275
4. M. Choirul Rahmadan, A. Nizar Hidayanto, D. Swadani Ekasari, B. Purwandari and Theresiawati, "Sentiment Analysis and Topic Modelling Using the LDA Method related to the Flood Disaster in Jakarta on Twitter," *2020 International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS)*, Jakarta, Indonesia, 2020, pp. 126-130
5. S. Peng, "Medical Analysis of Social Media Data Based on Spark and Machine Learning in China," *2021 2nd International Conference on Computer Science and Management Technology (ICCSMT)*, Shanghai, China, 2021, pp. 425-428
6. A. Singh and S. Kumar, "A Comparison of Machine Learning Algorithms and Transformer-based Methods for Multiclass Sentiment Analysis on Twitter," *2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, Delhi, India, 2023, pp. 1-9
7. H. Rahman *et al*., "An Analysis of Bangla Tweets on Social Media Platform for Polarity Detection Using Machine Learning Algorithms," *2023 4th International Conference on Big Data Analytics and Practices (IBDAP)*, Bangkok, Thailand, 2023, pp. 1-6
8. P. Thamaraiselvi, J. Masih, P. Giri, J. Sridevi, I. A. Karim Shaikh and M. V. R. Prasad, "Analysis of Social Media Marketing Impact on Customer Behaviour using AI & Machine Learning," *2024 Ninth International Conference on Science Technology Engineering and Mathematics (ICONSTEM)*, Chennai, India, 2024, pp. 1-6
9. D. Sharma and Shilpa, "Evaluation of Different Machine Learning Methods for Sentiment Analysis of Indian Languages," *2024 11th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)*, Noida, India, 2024, pp. 1-6
10. S. S. Ibrahiem, S. S. Ismail, K. A. Bahnasy and M. M. Aref, "A Case Study in Multi-Emotion Classification via Twitter," *2020 12th International Conference on Electrical Engineering (ICEENG)*, Cairo, Egypt, 2020, pp. 115-120
11. T. S. Sai Kumar, K. Arunaggiri Pandian, S. Thabasum Aara and K. Nagendra Pandian, "A Reliable Technique for Sentiment Analysis on Tweets via Machine Learning and BERT," *2021 Asian Conference on Innovation in Technology (ASIANCON)*, PUNE, India, 2021, pp. 1-5
12. Y. Ashrita, S. Abhiram, V. Hemanth, A. Srinivas and P. R. Vemula, "Deep Learning Techniques for Sentiment Analysis on Social Media Text," *2023 6th International Conference on Contemporary Computing and Informatics (IC3I)*, Gautam Buddha Nagar, India, 2023, pp. 2294-2300
13. I. F. Rozi, R. Arianto and H. H. Mahdyan, "Fake News Detection Using Sentiment Analysis Approach in Indonesian Language," *2023 International Conference on Advanced Mechatronics, Intelligent Manufacture and Industrial Automation (ICAMIMIA)*, Surabaya, Indonesia, 2023, pp. 206-211

14. J. Lampert and C. H. Lampert, "Overcoming Rare-Language Discrimination in Multi-Lingual Sentiment Analysis," *2021 IEEE International Conference on Big Data (Big Data)*, Orlando, FL, USA, 2021, pp. 5185-5192
15. V. S. Kumar, P. K. Pareek, V. H. Costa de Albuquerque, A. Khanna, D. Gupta and D. R. S, "Multimodal Sentiment Analysis using Speech Signals with Machine Learning Techniques," *2022 IEEE 2nd Mysore Sub Section International Conference (MysuruCon)*, Mysuru, India, 2022, pp. 1-8
16. D. Sharma and Shilpa, "Evaluation of Different Machine Learning Methods for Sentiment Analysis of Indian Languages," *2024 11th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)*, Noida, India, 2024, pp. 1-6
17. P. Thamaraiselvi, J. Masih, P. Giri, J. Sridevi, I. A. Karim Shaikh and M. V. R. Prasad, "Analysis of Social Media Marketing Impact on Customer Behaviour using AI & Machine Learning," *2024 Ninth International Conference on Science Technology Engineering and Mathematics (ICONSTEM)*, Chennai, India, 2024, pp. 1-6
18. N. Patvardhan, M. Roy, C. Madhura Ranade and D. J. Joshi, "Uncovering Sentiment Changes Throughout the UBS Takeover: A Sentiment Analysis," *2024 Second International Conference on Emerging Trends in Information Technology and Engineering (ICETITE)*, Vellore, India, 2024, pp. 1-5
19. A. Singh and S. Kumar, "A Comparison of Machine Learning Algorithms and Transformer-based Methods for Multiclass Sentiment Analysis on Twitter," *2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, Delhi, India, 2023, pp. 1-9