Nepal Journal of Multidisciplinary Research (NJMR)
Vol. 7, No. 2, June 2024. Pages: 1-11
ISSN: 2645-8470 (Print), ISSN: 2705-4691 (Online)
DOI: https://doi.org/10.3126/njmr.v7i2.68189

Comparative Study among Term Frequency-Inverse Document Frequency and Count Vectorizer towards K Nearest Neighbor and Decision Tree Classifiers for Text Dataset

Tula Kanta Deo*
Department of Computer Science and Engineering, Kalinga University, Raipur(CG), India
Email: tuladeo@gmail.com

Rajesh Keshavrao Deshmukh
Department of Computer Science and Engineering, Kalinga University, Raipur(CG), India

Gajendra Sharma
[3] Department of Computer Science and Engineering, Kathmandu University, Bagmati, Nepal

Corresponding author*
Types of Research: Original Research

Received: May 11, 2024; Revised & Accepted: July 17, 2024

Abstract

Background: Text classification techniques are increasingly important with the exponential growth of textual data on the internet. Term Frequency-Inverse Document Frequency (TF-IDF) and Count Vectorizer(CV) are commonly used methods for feature extraction. TF-IDF assigning weights to terms based on their frequency. CV simply counts the occurrences of terms. The performance of CV as well as TF-IDF are evaluated and compared with KNN and DT classifiers across text datasets.

Methodology: The investigation begins with preprocessing. The feature vectors are created using both TF-IDF and CV. Feature vectors are passed into the KNN and DT classifiers at in training stage. Experiments are executed the usage of Kaggle's public database Ukraine 10K tweets sentiment_analysis dataset and the Womens ecommerce clothing reviews dataset.

Findings: The average of precision, recall, f1 score and accuracy of KNN with TF-IDF were 84.5%, 87%, 83%, 87% respectively and KNN with CV were 83.5%, 87%, 83.5%, 87% respectively. Similarly, average of precision, recall, f1 score and accuracy of DT with TF-IDF

Nepal Journal of Multidisciplinary Research (NJMR)
Vol. 7, No. 2, June 2024. Pages: 1-11
ISSN: 2645-8470 (Print), ISSN: 2705-4691 (Online)
DOI: https://doi.org/10.3126/njmr.v7i2.68189

were 89%, 89%, 89%, 89% respectively and DT with CV were 89%, 89.5%, 89.5%, 89.5% respectively. The results obtained in this research is consistent with previous similar research result.

Conclusions: The performance of TF-IDF is almost similar as CV for a particular dataset and a particular classifier in this study.

Novelty: The experiment performed using these classifiers and feature extraction methods on the datasets is a novelty and contribution of this research.

Keywords: Count Vectorizer , Decision Tree, K Nearest Neighbor, Term Frequency and Inverse Document Frequency

Introduction

The volume of textual data that is available on digital platforms demands efficient textual classification techniques. Natural language processing (NLP) like text classification are vital for numerous applications, including sentiment analysis and data categorization (Qaiser & Ali, 2018). Feature extraction plays a big role in text classification. TF-IDF and CV are widely used. TF-IDF weights terms primarily based on their importance inside the documents, while CV creates a frequency representations. KNN and DT classifiers are recognized for their simplicity and performance in text classification. This study examines how nicely two famous feature extraction strategies TF-IDF and CV perform whilst used with KNN and DT classifiers. This take a look at objectives to compare overall performance evaluation of KNN and DT classifiers which carry out with TF-IDF and CV in terms of classifying exclusive text datasets. A popular method for changing textual inputs into numerical representations in NLP is TF-IDF. The metric referred to as TF-IDF examines the significance of term in a document with regards to collections of documents. It is an important for strategies which include sentiment analysis and document classification. The TF-IDF is calculated by means of combining TF and IDF scores. TF-IDF is used in information retrieval systems, sentiment classification, search engines (Suryaningrum, 2023). Based at the importance of the terms, TF-IDF can be extracted features from texts and classify texts in sentiment analysis may be labeled as positive, negative. TF-IDF is a widely utilized method. It uses data in numerical format. Its capacity to capture the relevance of term in documents makes it unmatched in several mining tasks(Addiga & Bagui, 2022).

CV is a crucial method in NLP to transform data right into number representation. As the quantity of data generated remains to enhance the requirement for effective procedures together with analysis of these data comes to be a lot more crucial. CV is a regularly made use of strategy for feature extraction plus creates the basis of numerous text analysis algorithms. CV is a method to transform text right into a matrix of token counts where each row stands for a document plus each column stands for a certain term in the document collection. The value in the matrix represent the frequency of each term in particular documents. It can make use of machine learning(ML) algorithms for automated text handling by transforming text right into numerical format. It is a basis for text classification, text summarization, sentiment analysis(Peter the Great St.Petersburg Polytechnic University et al., 2020).

Nepal Journal of Multidisciplinary Research (NJMR)
Vol. 7, No. 2, June 2024. Pages: 1-11
ISSN: 2645-8470 (Print), ISSN: 2705-4691 (Online)
DOI: https://doi.org/10.3126/njmr.v7i2.68189

Literature Review

This section summarizes associated works of TF-IDF as well as CV in the domain name of SA. A research study was performed making use of the features of TF-IDF, CV plus support vector machine(SVM). The comparision metrics are precision, accuracy, recall as well as f1-score. The precision, accuracy, recall, and f1-score acquired making use of the TF-IDF has 88.77%, 87.45%, 88.77%, as well as 87.81% respectively. The precision, accuracy, recall along with f1-score got utilizing the CV has 87.14%, 85.89%, 87.63% and 86.72% respectively. The outcomes reveal that the percent of the TF-IDF feature is above making use of the CV feature (Suryaningrum, 2023).

In an additional research study, Turkish text were used with various ML algorithms to discover cyberbullying. Accuracy, precision, recall as well as F1-score were made use of to review the efficiency of the classifiers. Evaluation result of DT classifier utilizing CV for f1-score, precision, recall, precision were 99.96%, 100%, 99.92%, 87.19% respectively. Evaluation result of random forest making use of CV for f1-score, precision, recall, precision were 99.21%, 99.66%, 98.75%, 86.36% respectively. Evaluation result of DT classifier utilizing TF-IDF for f1-score, precision, recall, precision were 99.16%, 99.16%, 98.42%, 85.69% respectively. Evaluation result of random forest classifier utilizing TF-IDF for f1-score, precision, recall, precision were 99.16%, 99.66%, 98.67%, 86.86% respectively (Sadigzade & Nasibov, 2021).

A research study have actually compared three ML algorithms coupled with two vectorization strategies. The ML algorithms made use of are Naive Bayes(NB), logistic regression(LR) and also SVM, and also the vectorization strategies made use of are TFIDF as well as CV(bag of words (BoW)). Precision of NB, LR, and also SVM for CV were 97.34%, 95.69% 93.56% specifically. Precision of NB, LR, and SVM for TF-IDF were 81.33%, 98.45%, 93.08% specifically. Out of the 6 mix versions of classification as well as vectorization the model developed making use of LR and also TF-IDF showed to produce
greatest accuracy of 98.45%. NB with CV provided accuracy 97.34%. TF-IDF as well as NB classifier provided least accuracy 81.33% (Bhansali et al., 2022).

In a study BBC News has actually separated its newspaper group right into 5 teams. The objective was to categorize news right into their groups making use of classification strategies ML algorithms to assist users accessibility appropriate news rapidly and also conveniently without losing time. After extracting out feature from the data making use of CV and also TF-IDF vectorizer different classification algorithms SVM and also multinomial Naive Bayes(MNB) were used to the data. The SVM algorithms outperform MNB with 99.1% accuracy in the CV and also 98.2% accuracy in the TF-IDF vectorizer(Ogaib M.F. and Hashim K.M., 2022). A research study was done on OK cuisine reviews from the Amazon site compared the performance of NB, LR, and also SVM making use of two technique of feature selection including CV and also TF-IDF. Accuracy, precision, recall, f1-score for SVM with CV were 89%, 90%, 89%, 89% respectively, for NB with CV were 68%, 64%, 68%, 65% respectively, for LR with CV were 85%, 95%, 85%, 89% respectively, for SVM with TF-IDF were 91%, 93%, 91%, 91% respectively, for NB with TF-IDF were 81%, 97%, 81%, 87% respectively,

Nepal Journal of Multidisciplinary Research (NJMR)
Vol. 7, No. 2, June 2024. Pages: 1-11
ISSN: 2645-8470 (Print), ISSN: 2705-4691 (Online)
DOI: https://doi.org/10.3126/njmr.v7i2.68189

for LR with TF-IDF were 88%, 92%, 88%, 90% respectively. The findings showed that the SVM classifier had actually attained the highest possible accuracy of 91%, by CV (Fadel & Behadili, 2022).

A research study was carried out to figure out the effect of vectorization strategies on public opinion analysis in preparation for offline education and learning almost everywhere during the Covid-19 pandemic in Indonesia. Authors make use of two various techniques to identify sentiment: by manually and using NLP library Text Blob. CV, TF-IDF along with mix of both were compared.

Feature vectors were classified in three methods: LR, NB as well as KNN for both manual and automated labeling. Precision, recall, f1-score for KNN with TF-IDF were 76.48% 77.21% 75.11% respectively and for KNN with CV were 72.83%, 72.86%, 71.36% respectively(Kunang & Mentari, 2023).

Methodology

In this section, the general process of TF-IDF and CV in sentiment classification are explained. This research is an experimental study of TF-IDF and count CV using KNN and DT classifiers.

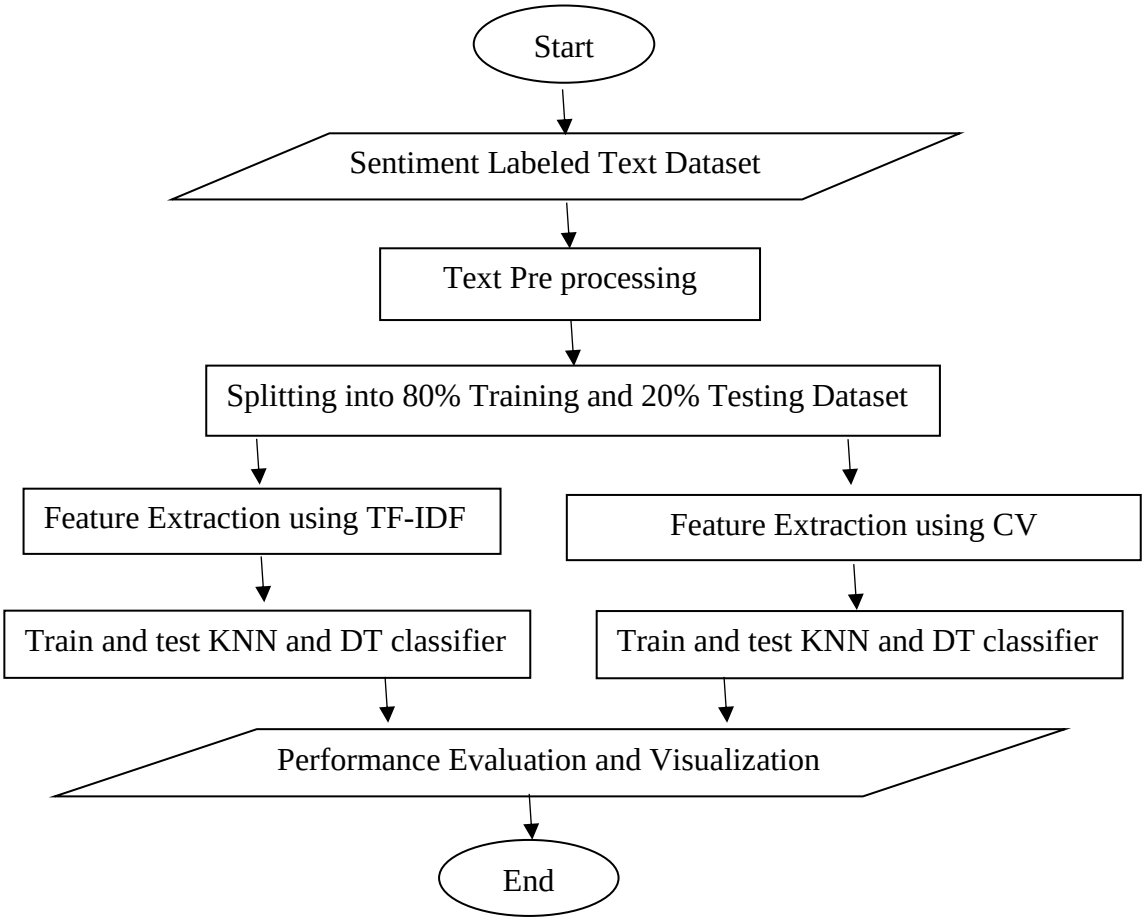

Figure 1: Schema for performance comparison of TF-IDF and CV

Nepal Journal of Multidisciplinary Research (NJMR)
Vol. 7, No. 2, June 2024. Pages: 1-11
ISSN: 2645-8470 (Print), ISSN: 2705-4691 (Online)
DOI: https://doi.org/10.3126/njmr.v7i2.68189

This research study initially gathered information from Kaggle's public database and afterwards preprocessed the information by getting rid of non-English words, getting rid of spelling marks, getting rid of numbers, getting rid of unneeded words. Tokenization, part of speech tagging and also lemmatization are executed. After utilizing the TF-IDF and also CV technique for feature extraction from the training dataset and also testing dataset are separated according to the proportion of 80% training dataset as well as 20% testing dataset. Training and also testing of KNN as well as DT classifiers starts after that running it for performance evaluation. The performance evaluation made use of, f1-score, accuracy, recall as well as precision. Visualization of the evaluations are likewise executed to present tables along with charts.

Experimental Setup

This section deals the experimental setup for the research. This section contains details about the dataset and matrices associated with the performance. The experiments have performed on python language with Scikit Learn, NLTK, Pandas, and other library packages.

Datasets

Two datasets have used in this research. The dataset "Ukraine_10K_tweets_sentiment_analysis" contains 10,000 tweets about the keyword "Ukraine" is downloaded from Kaggle public repository(*Tweets with Sentiments Dataset | Kaggle*, n.d.). The sentiment values in the dataset were visualized using the Textblob library. "Womens-ecommerce-clothing-reviews" dataset is a collection of women's clothing ecommerce dataset based on reviews written by customers. It includes 23486 rows and 10 feature variables. (*Women's E-Commerce Clothing Reviews | Kaggle*, n.d.).

Table 1: Descriptions of the datasets

Dataset	Data Set Description	Positive	Negative
Ukraine_10K_tweets_sentiment_analysis	It contains 10,000 tweets about the keyword "Ukraine"	2064	7935
womens-ecommerce-clothing-reviews	It is a collection of women's clothing ecommerce dataset	19314	4172

Performance Measurements

Numerous performance metrics are generally made use of in SA to evaluate the effectiveness of models in predicting sentiment. A few of the key performance metrics consist of:

Accuracy: It is calculated as the ratio of the number of correctly predicted samples to the total number of samples.

Precision: It is a proportion of real positives to the amount of real positives as well as incorrect positives.

Nepal Journal of Multidisciplinary Research (NJMR)
Vol. 7, No. 2, June 2024. Pages: 1-11
ISSN: 2645-8470 (Print), ISSN: 2705-4691 (Online)
DOI: https://doi.org/10.3126/njmr.v7i2.68189

Recall: A proportion of true positive predictions amongst all actual positive samples in the dataset.
F1-score: It is calculated as 2 x (precision x recall)/(precision + recall)(Tasnim et al., 2022).

Results and Discussions

Results

This section contains the details about the performance of TF-IDF and CV for KNN and DT on Ukraine_10K_tweets_sentiment_analysis and Womens-ecommerce-clothing-reviews datasets. This section contains the tables, charts obtained while performing the sentiment analysis using KNN and DT classifiers on the datasets.

Table 2: Performance metrics of KNN classifier with TF-IDF and CV

Metrics	Ukraine_10K_tweets_sentiment_analysis		Womens-ecommerce-clothing-reviews	
	TF-IDF	CV	TF-IDF	CV
Precision	0.92	0.92	0.77	0.75
Recall	0.92	0.92	0.82	0.82
F1 score	0.91	0.91	0.75	0.76
Accuracy	0.92	0.92	0.82	0.82

Source: Experimental results of this this research

Table 2 presents a tabular representation of the outputs. It shows the metrics for KNN classifier with TF-IDF and CV applied over the two datasets individually.

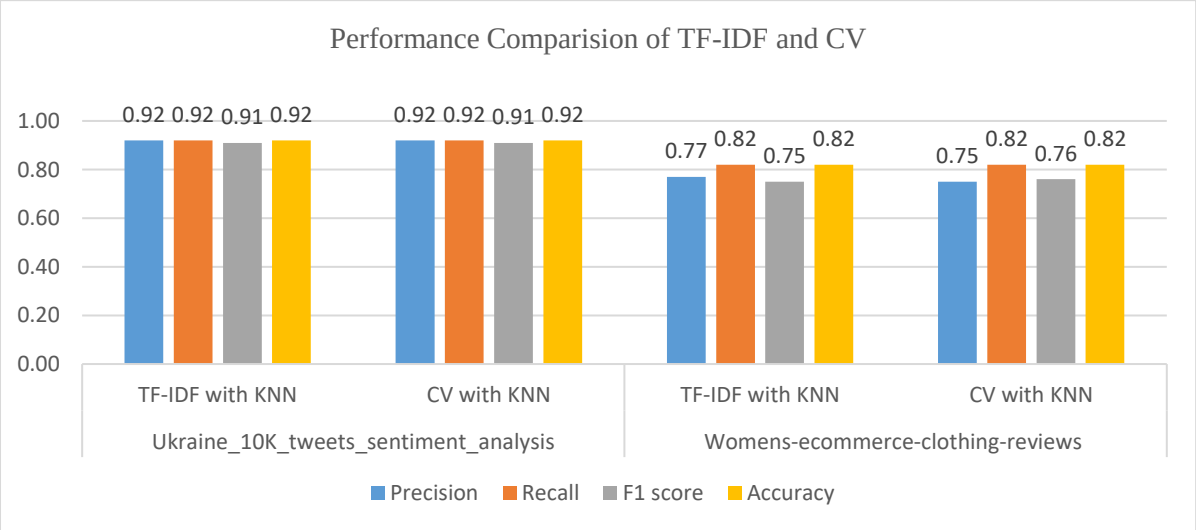

Figure 2: Performance comparison of TF-IDF and CV for KNN classifier

Nepal Journal of Multidisciplinary Research (NJMR)
Vol. 7, No. 2, June 2024. Pages: 1-11
ISSN: 2645-8470 (Print), ISSN: 2705-4691 (Online)
DOI: https://doi.org/10.3126/njmr.v7i2.68189

Figure 2 shows the metric results for the Ukraine_10K_tweets_sentiment_analysis and Womens-ecommerce-clothing-reviews datasets. This bar chart shows an overall comparison between TF-IDF and CV of KNN classifier on this particular data. TF-IDF performs almost same as CV in all parameters.

Table 3: Performance metrics of DT classifier with TF-IDF and CV

Metrics	Ukraine_10K_tweets_sentiment_analysis		Womens-ecommerce-clothing-reviews	
	TF-IDF	CV	TF-IDF	CV
Precision	0.97	0.97	0.81	0.81
Recall	0.97	0.97	0.81	0.82
F1 score	0.97	0.97	0.81	0.82
Accuracy	0.97	0.97	0.81	0.82

Source: Experimental results of this this research

Table 3 presents a tabular representation of the outputs. It shows the metrics for DT classifier with TF-IDF and CV applied over the two datasets individually.

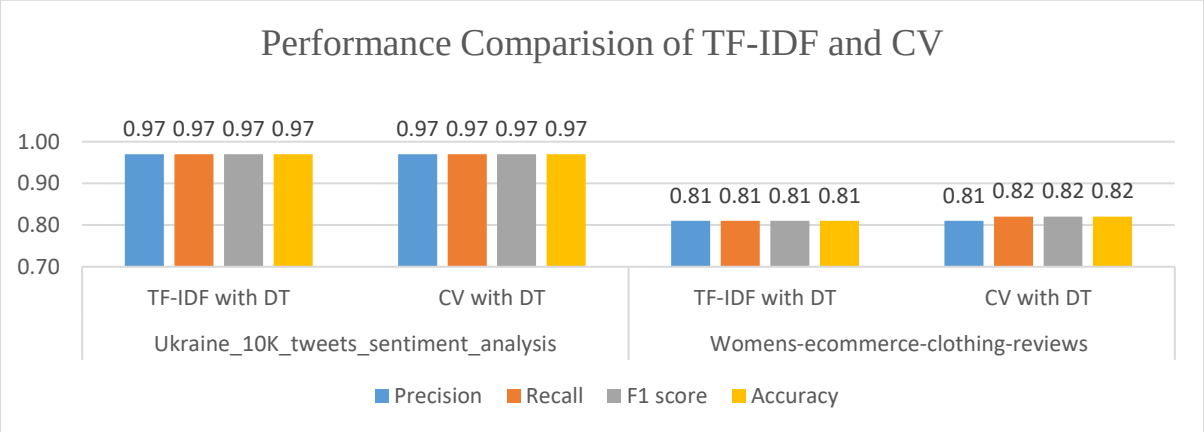

Figure 3: Performance comparison of TF-IDF and CV for DT classifier

Figure 3 shows the metric results for the Ukraine_10K_tweets_sentiment_analysis and Womens-ecommerce-clothing-reviews datasets. This bar chart shows an overall comparison between TF-IDF and CV of DT classifier on this particular data. TF-IDF performs almost same as CV in all parameters on a particular dataset.

Nepal Journal of Multidisciplinary Research (NJMR)
Vol. 7, No. 2, June 2024. Pages: 1-11
ISSN: 2645-8470 (Print), ISSN: 2705-4691 (Online)
DOI: https://doi.org/10.3126/njmr.v7i2.68189

Table 4: Average values of metrics for two datasets for KNN classifier

Metrics	Average of two datasets	
	TF-IDF	CV
Precision	0.845	0.835
Recall	0.87	0.87
F1 score	0.83	0.835
Accuracy	0.87	0.87

Source: Experimental results of this this research

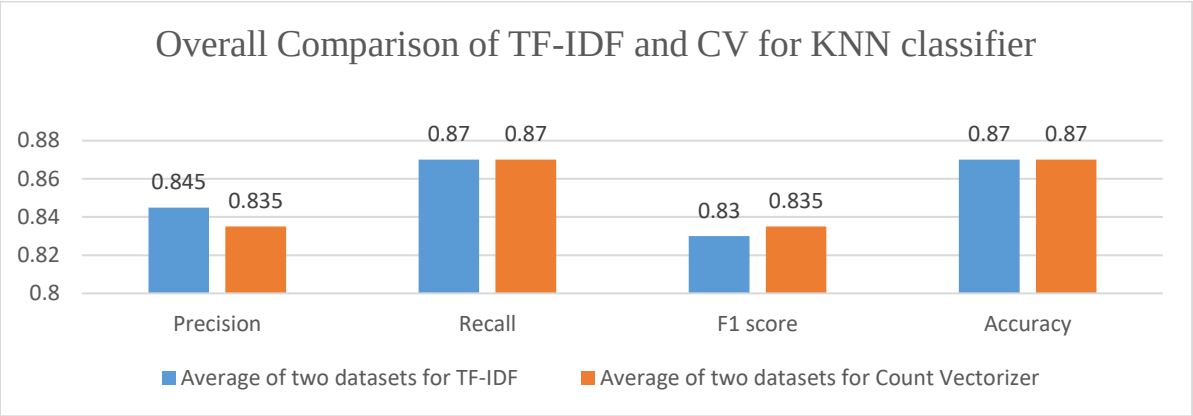

Figure 4: Average Comparison of TF-IDF and CV for KNN classifier

The average of all parameters in table 4 and figure 4 , TF-IDF is almost same as CV for KNN classifiers in the two datasets.

Table 5: Average values of metrics for two datasets for DT classifier

Metrics	Average of two datasets	
	TF-IDF	CV
Precision	0.89	0.89
Recall	0.89	0.895
F1 score	0.89	0.895
Accuracy	0.89	0.895

Source: Experimental results of this this research

Nepal Journal of Multidisciplinary Research (NJMR)
Vol. 7, No. 2, June 2024. Pages: 1-11
ISSN: 2645-8470 (Print), ISSN: 2705-4691 (Online)
DOI: https://doi.org/10.3126/njmr.v7i2.68189

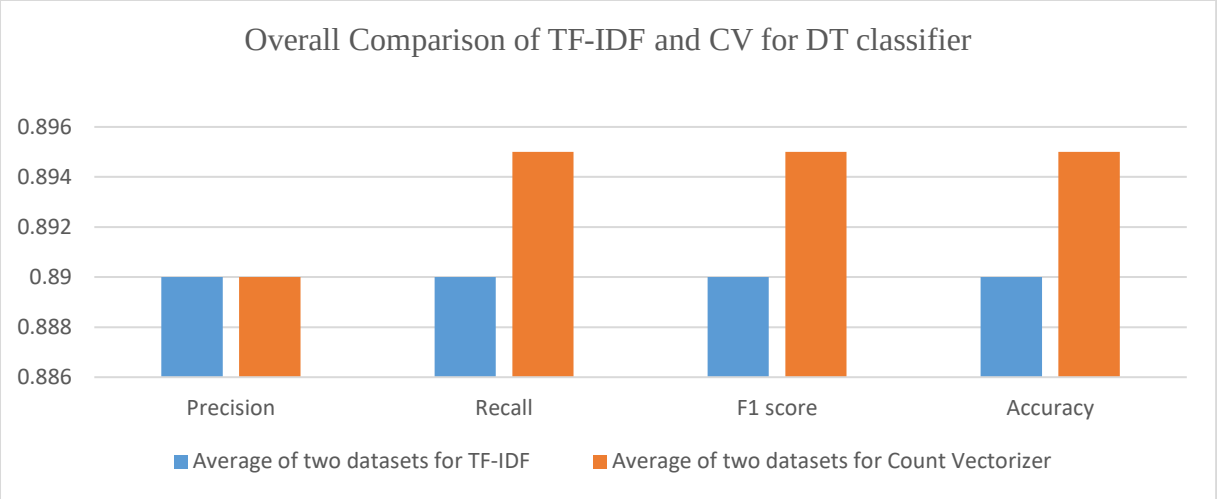

Figure 5: Average value Comparison of TF-IDF and CV for DT classifier

The average of all parameters in table 5 and figure 5, TF-IDF is almost same as CV for DT classifiers in the two datasets.

Discussion

The current study found that average of precision, recall, f1 score and accuracy for TF-IDF with KNN were 84.5%, 87%, 83%, 87% respectively, for CV with KNN were 83.5%, 87%, 83.5%, 87% respectively, for TF-IDF with DT were 89%, 89%, 89%, 89% respectively and for CV with DT were 89%, 89.5%, 89.5%, 89.5% respectively.

Multiple ML algorithms were used to detect cyberbullying in Turkish texts. Evaluation Results of DT model using CV for f1 score, precision, recall, accuracy were 99.96%, 100%, 99.92%, 87.19% respectively. Evaluation Results of DT model using TF-IDF for f1 score, precision, recall, accuracy were 99.16%, 99.16%, 98.42%, 85.69% respectively(Sadigzade & Nasibov, 2021). These results are consistent with current study.

A study was conducted to determine the impact CV, TF-IDF and combination of both. Precision, recall, f1score for KNN with TF-IDF were 76.48%, 77.21%, 75.11% respectively, for KNN with CV were 72.83%, 72.86%, 71.36% respectively(Kunang & Mentari, 2023), These results are consistent with current study.

Conclusion

This study presents the results of a comparative study to investigate the two well-known ML methods KNN and DT classifiers using TF-IDF and CV on Kaggle publicly available Ukraine 10K tweets sentiment analysis dataset and Womens ecommerce clothing reviews datasets. This research includes different metrics such as accuracy, precision, recall and f1 score, which help to clarify the comparison between the two methods. The results and discussions shows the performance of TF-IDF and CV for KNN and DT classifiers on Ukraine 10K tweets sentiment analysis dataset and Womens ecommerce clothing reviews datasets is consistent with most of

the other previous research. The effectiveness of TF-IDF and CV depending on the dataset characteristics and the classifier used. The performance of TF-IDF is almost same as CV for a particular dataset and a particular classifier in this research. This research provides insights into the comparative effectiveness of different combinations of feature extraction techniques and classifiers for text classification tasks.

In the future, an ensemble of KNN and DT could be developed that combines the results of the two methods and therefore improves performance.

References

Addiga, A., & Bagui, S. (2022). Sentiment Analysis on Twitter Data Using Term Frequency-Inverse Document Frequency. *Journal of Computer and Communications*, *10*(08), 117–128. https://doi.org/10.4236/jcc.2022.108008

Bhansali, A., Chandravadiya, A., Panchal, B. Y., Bohara, M. H., & Ganatra, A. (2022). Language Identification Using Combination of Machine Learning Algorithms and Vectorization Techniques. *2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)*, 1329–1334. https://doi.org/10.1109/ICACITE53722.2022.9823628

Fadel, F. H., & Behadili, S. F. (2022). A Comparative Study for Supervised Learning Algorithms to Analyze Sentiment Tweets. *Iraqi Journal of Science*, 2712–2724. https://doi.org/10.24996/ijs.2022.63.6.36

Kunang, Y. N., & Mentari, W. P. (2023). Analysis of the Impact of Vectorization Methods on Machine Learning-Based Sentiment Analysis of Tweets Regarding Readiness for Offline Learning. *JUITA : Jurnal Informatika*, *11*(2), 271. https://doi.org/10.30595/juita.v11i2.17568

Ogaib M.F. and Hashim K.M. (2022). *Comparative Study for News Categorization by Multinomial Naïve Bayes and Support Vector Machine*.

Peter the Great St.Petersburg Polytechnic University, Kozhevnikov, V. A., Pankratova, E. S., & Peter the Great St.Petersburg Polytechnic University. (2020). Research of the Text Data Vectorization And Classification Algorithms of Machine Learning. *Theoretical & Applied Science*, *85*(05), 574–585. https://doi.org/10.15863/TAS.2020.05.85.106

Qaiser, S., & Ali, R. (2018). Text Mining: Use of TF-IDF to Examine the Relevance of Words to Documents. *International Journal of Computer Applications*, *181*(1), 25–29. https://doi.org/10.5120/ijca2018917395

Sadigzade, M., & Nasibov, E. (2021). *Comparative Analysis of Count Vectorization vs TF-IDF Vectorization For Detecting Cyberbullying In Turkish Twitter Messages*.

Suryaningrum, K. M. (2023). Comparison of the TF-IDF Method with the Count Vectorizer to Classify Hate Speech. *Engineering, MAthematics and Computer Science (EMACS) Journal*, *5*(2), 79–83. https://doi.org/10.21512/emacsjournal.v5i2.9978

Tasnim, A., Saiduzzaman, Md., Rahman, M. A., Akhter, J., & Rahaman, A. S. Md. M. (2022). Performance Evaluation of Multiple Classifiers for Predicting Fake News.

Nepal Journal of Multidisciplinary Research (NJMR)
Vol. 7, No. 2, June 2024. Pages: 1-11
ISSN: 2645-8470 (Print), ISSN: 2705-4691 (Online)
DOI: https://doi.org/10.3126/njmr.v7i2.68189

Journal of Computer and Communications, *10*(09), 1–21. https://doi.org/10.4236/jcc.2022.109001

Tweets with Sentiments Dataset | Kaggle. (n.d.). Retrieved March 17, 2024, from https://www.kaggle.com/datasets/abhishek14398/10k-tweets-with-sentiments-dataset

Women's E-Commerce Clothing Reviews | Kaggle. (n.d.). Retrieved March 17, 2024, from https://www.kaggle.com/datasets/nicapotato/womens-ecommerce-clothing-reviews