A Comprehensive Review of Arabic and English Sentiment Analysis in BBC and SANAD News

Hassan Al-Sukhni[1], Qusay Bsoul[2],Sharaf Alzoubi[3] , Fadi yassin Salem Al jawazneh[4], Dalia Ehab Abdelaziz[5], Hisham Mohamed Gamel[6], Diaa Salama AbdElminaam[7,8,*]

[1]Cybersecurity Department, Faculty of Science and Information Technology, Jadara University, Irbid, Jordan
[2]Cybersecurity Department, College of Computer Sciences and Informatics, Amman Arab University, Amman, Jordan
[3]Department of Software Engineering , College of Computer Sciences and Informatics, Amman Arab University, Amman, 11953, Jordan
[4]Faculty of Information Technology, Applied Science Private University, Amman, Jordan
[5]Misr International University, Cairo, Egypt
[6]BIS Department, Obour High Institute for Management and Informatics, Cairo, Egypt
[7]MEU Research Unit, Middle East University, Amman, Jordan
[8]Jadara Research Center, Jadara University, Irbid, Jordan

Emails: h.sukhni@jadara.edu.jo; q.bsoul@aau.edu.jo; skalzubi@aau.edu.jo; F_aljawazneh@asu.edu.jo; dalia.ehab@miuegypt.edu.eg; hishamg@oi.edu.eg; diaa.salama@miuegypt.edu.eg

Abstract

News agencies connect global events to local communities. It plays a pivotal role in influencing public opinion. Thus, the necessity arises to recognize news article's sentiment. The purpose of this paper is to analyze sentiment for English and Arabic news articles in terms of positivity, negativity, or neutrality. Analyzing the articles of Arabic and English news can be challenging from the perspective of morphology. In this paper, we introduce 4 Machine Learning methods, including Logistic Regression (LR), k Nearest Neighbors (K-NN), Random Forests (RF) and Naive Bayes (NB), with the TF-IDF as the feature extraction. The study was validated using 2 data sets (BBC, SANAD Arabic news), and two learning models (Hold out and 10-fold cross-validation). The evaluation was based on; Accuracy (ACC), Precision (PREC), Recall (REC), F1-score (F1), and The Matthews Correlation Coefficient (MCC) where it shows an outstanding performance for ML on a 10-fold strategy. The experiments provided in the paper indicated that the proposed ML models achieved the best results.

Keywords: Machine Learning; Arabic News; Sentiment Analysis; Supervised Learning

1 introduction

Media can be a two-edged weapon; it can affect sentiments and transfer facts. It can be an instrument of manipulation and can be a way to help spread the truth. News Agencies are considered a representation of the community thus it's essential to recognize the nature of sentiment represented in articles. The history of news agencies has been one of interest for many years. As early as the mid-19th century, these organizations were designed to provide international news to domestic markets, and later, through alliances with other news agencies, they became involved in providing news on a global scale.[?] BBC, with its Western ideology, is affected by economic and political forces. Its widespread name makes it ideal to use sentiment analysis.

DOI: https://doi.org/10.54216/FPA.180115

Received: July 08, 2024 Revised: October 06, 2024 Accepted: December 29, 2024

The English language is the most dominant in the world. More than 350 million people speak it as a first language and 430 million speak it as a second language.[2] Figure 1 shows the numbers of language speakers on 16th june, 2023 according to? It can be seen that English is the first language and Arabic is in the sixth place. The importance of researching the English language lies in it being the language of the internet and the most globalized language available. English is a communicational tool that allows people to reach out and unite. It is influential part makes it a tool that nearly the whole world can understand.[2]

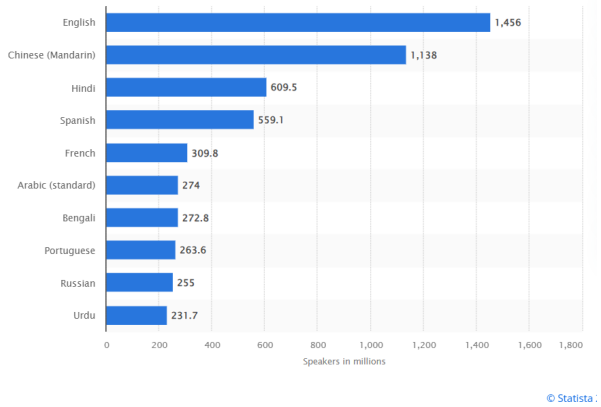

Figure 1: The most spoken languages worldwide in 2023 (by speakers in millions)

Machine learning has been in constant development. Giving researchers the space to discover more about the language and learn more about its features. Sentiment analysis is considered an optimal computational technique that can help understand and analyze text from the perspective of sentiments, opinions, and emotions. Many works have been presented on blogs and product reviews, however, when it comes to news articles it gives a space to be objective. Using news articles gives a space to evaluate the positivity and negativity of words and automatically identify where the negative spike lies.?

This research will analyze and evaluate BBC and SANAD news dataset articles using sentiment analysis. The analysis outlines some points in the text where the positive and negative news is focused on the BBC website in English and SANAD dataset in Arabic and recognizes the efficiency of sentiment used. Sentiment analysis can also help discover how positive and how negative the news is in the source language, giving the advantage that the results would be more accurate.

The models used for this research will be Random Forest, Naive Bayse, Logistic Regression and kNN and using methods like Vader and multilingual with sentiment analysis.

As a summary, this research paper shall include the following:

- Providing a valuable framework for English and Arabic news articles using sentiment analysis.
- Process Sentiment Analysis by selecting data sets that contain a large number of articles,
- Discussing the importance of pre-processing articles including removing hashtags removing stop words, tokenization, and stemming.
- Providing methods to analyze sentiment in lengthy text documents viewing the positive and negative texts with compound in English and Arabic and discussing sentiment analysis models such as the Vader method and multilingual.
- Discusses various semantic analysis models like Random Forest, Naive Bayes, Logistic Regression and kNN and extracts keywords to define positive and negative statements.
- Techniques to extract meaning and context from textual data and compare subgroups to the main group and document map.

DOI: https://doi.org/10.54216/FPA.180115

Received: July 08, 2024 Revised: October 06, 2024 Accepted: December 29, 2024

The rest of the paper is organized as follows:

section 2 presents the collected related work. Section 3 focuses on aspects of the English language's structure and orthography that are essential to the study. Section 4 explores the key challenges facing both English language and sentiment analysis. Section 5 introduces the methodology. The experimental results are shown in section 6. Section 7 shows a comparison of the collected results. Section 8 a detailed explanation of the proposed model. Section 9 concludes the paper and provides some future works.

2 related work

Many factors affect sentiment analysis in news. According to some researchers[?] the factors affecting news can be for example idiomatic expressions and quotes. In this research, the aim is to evaluate those challenges with the help of sentiment analysis. The researcher used different domains including JRC Tonality, SentiWordNet, and MicroWN. The metrics approached for such matters were accuracy, precision, recall, and F1-score. Moreover, the aim was to test the appropriateness and the challenges of sentiment analysis. However, the paper didn't put much focus on the usage of negation and valence shifters. The paper proved that using JRC Tonality and MicroWN on 6 words gave a much higher accuracy. It also shows that large lexicons don't mean that the performance of systems is increased.

On the other side, some researchers used a lexicon-based approach for sentiment analysis on BBC News between the years 2004 and 2005.[?] Whereas the results were drowned out as positive, negative, and neutral. The method used was by WordNet lexical dictionary and calculating the polarity scores. Most of the of news articles were negative, while the neutral articles had the least outcome. The researcher also discussed that business and sports had the most positive sentiments while entertainment, politics, and tech explicited the negative. One of the mentioned limitations was that the accuracy and performance depend on the type of data sets and constant change of the new semantics.

Another researcher tackled the lexicon-based approach for sentiment analysis on BBC datasets.[?] The paper aimed at providing insights into how large amounts of data can be analyzed. The limitations of the paper were the limited lexical database and the constant change in datasets. The methods adopted on this research were Dictionary-based methods and Corpus-based methods and all of them were based on a document level. WordNet lexical dictionary was used to measure the polarity of sentiment words and the total sentiment score. The results presented that the categories Entertainment and Tech were negative, while business and sports were positive. While politics were a mix of positive and negative.

Meanwhile, some researchers delved into the matter of sentiment analysis on fake news detection.[6] The technique used was with the current state-of-art while it discussed the challenges faced. The research elaborated further on the importance of fake news detection amid the spread of misinformation and political polarization. For this matter, the researcher discussed various approaches and models to implement fake news detection. It focused on providing a comprehensive review of the recent research paper presented nowadays. Furthermore, the research elaborated on the difficulty of detecting fake news and the availability of high-quality training data.[6]

Researchers have also tackled an overview of sentiment analysis to grasp the classification of sentences and opinion extraction as well as supervised and unsupervised learning techniques.[7] It also highlighted the sentiment polarity shift issue and the lack of structured data in sites like Twitter. The approaches discussed for such matter was machine learning approaches using Support Vector Machines (SVM), Naive Bayes, and neural networks. Another approach was lexicon-based methods by N-gram and Part-of-Speech (POS) tagging. When it comes to sentiment analysis many challenges could arise such as difficulty on detecting sarcasm and irony. Also, handling the intensifiers and negation can be tricky. Lastly, the lack of annotated datasets for training and evaluation is one the most common challenges for sentiment analysis.[7]

Moreover, researchers have tackled sentiment analysis from Twitter by using techniques such as K-means, Naive Bayes, and SVM to label basic emotions.[?] The use of it would be to recognize the emotions from large texts for providing box office predictions. The methodology for such a technique would be to use labeling for emotions and classifying them through supervised learning techniques. Then, evaluate the classifiers with

DOI: https://doi.org/10.54216/FPA.180115

Received: July 08, 2024 Revised: October 06, 2024 Accepted: December 29, 2024

accuracy, precision, recall, and F1-score. In this paper, the accuracy for SVM is 72.5 percent. However, some limitations were mentioned such as the quality of Twitter data, Syntactic Level Parsing and to use of a more exhaustive Bag of Words.

Moving further to emotion detection, researchers have presented some perspectives in that regard.[10] The paper presents some practical issues that happen during emotion detection and provides solutions such as e-learning virtual world emotion tagging and global opinion of TV viewers. One of the main problems discussed was domain dependency, where the ML algorithms often depend on the training model and the effectiveness of the data sets may be affected. Another problem raised was the emotion representation for interoperability and the difficulty of neutralizing emotions. To solve such issues the researcher used unsupervised or semi-supervised machine learning, using hybrid methods, and advocated the use of EmotionML for emotion representation. The results of the paper were represented in using a symbolic approach that got 56.3 percent accuracy on the Semeval-07 dataset and 65.86 percent accuracy on the Semeval-13 dataset. Moreover, the paper suggests that sentiment analysis was achieved and can be delegated during video conferences in a virtual world forum with emotions. However, the limitation lies in the text length, domain dependence, and evaluation techniques.

Another paper disscussed emotion detection from the perspective of their various levels, approaches, and challenges in the context of social media and big data.[11] A lot of problems could arise from tackling sentiment analysis on informal language such as spelling mistakes, new slang, and incorrect use of grammar. Also, the challenges of detecting sarcasm and multiple emotions expressed. Moreover, the language is in a consistent change due to the development of new trends. Also, detecting polarity from comparative sentences is a challenging matter. The approaches used for sentiment analysis was n-grams, word embeddings, and sentiment lexicons, rule-based methods, machine learning algorithms, and deep learning models. The paper presents a comprehensive review of the methods used with different data sets. Some limitations were presented, such as lack of resources for domain-specific corpora and the Web slang presented on social media.

On the other hand, some researchers maintained a three-fold technique, where they evaluated pre-trained word embeddings and classifiers with neural networks.[?] They applied it to various domains in different scenarios. The research aimed to evaluate the deep learning model in application with utility in Twitter messages. The following problems were highlighted (slang language, limited length of tweets and not leveraging the ordering of words). The models used were softmax, support vector machine (SVM), and a whole feed-forward neural network. The results provided effectiveness in deep learning systems for sentiment analysis in Twitter messages. The paper expresses some limitations such as the use of an automatically labeled dataset for emotion identification and the unavailability of fine-tuning of model embeddings in a semi-supervised manner.

Lastly, Some researchers tackled sentiment analysis from the perspective of appraisal theory of online news.[?] The paper analyses the content of the news text and automatic sentiment analysis. They also address sentiment polarity and various types of sentiment-biased behaviors. The paper aims to analyze the political news in particular while discussing problems such as (lack of tense treatment, emotion typology, and types of political behaviors). Interestingly, the data set chosen discusses two Presidents on two major issues, the economy and the Iraq war. The results show the bias of the appraisers and the author, type of attitude, and manner of expressing the sentiment with their prominent polarity and attitude type. To sum up, the paper presented some limitations, for example, the results can't be generalized due to the limited data set. Also, this framework may not be applied to other domains. Table 2 represents a summary of the work related.

3 English Language

In the news agency world, the English language is considered essential. Where all the well-established agencies look after their terminology and consistency. BBC and CNN are considered a dieverse platform yet it contains a wide range of topics. Opinion mining through sentiment analysis needs such consistency to elevate the efficiency of results. The direct grammar of the English language helps facilitate the workflow.

3.1 English morphology

The structure of words in English focuses on small units. Each unit is called a morpheme. There are two types of morphemes, free and bound morphemes. The free morphemes are direct words such as sing, dog, and sad.

DOI: https://doi.org/10.54216/FPA.180115

Received: July 08, 2024 Revised: October 06, 2024 Accepted: December 29, 2024

Table 1: Work Related Summary

Ref	Year	Algorithm	Dataset	ACC	Advantage (A) Disadvantage (D)
3	2010	JRC Tonality, SentiWordNet, and MicroWN	News	82%	D: Didn't focus on the usage of negation and valence shifters
4	2019	WordNet, A Lexicon-based Approach, TF-IDF	BBC	91%	A: Large dataset D: Most of the news articles were negative
5	2020	WordNet and Lexicon-based Approach	BBC	91%	D: Limited word coverage in lexical databases
6	2021	Support Vector Machines (SVM), Neural networks, and Lexicon-based approaches	LIAR, FakeNewsNet, BS Detector, and Fake vs. Real News	94.4%	A: Fake news detection includes its usefulness
7	2020	SVM, Naive Bayes, neural networks, N-gram, and Part-of-Speech (POS) tagging	Twitter, Amazon, and BBC	89.1%	D: Lack of annotated datasets for training and evaluation
8	2018	K-means, Naive Bayes, and SVM	Twitter	64%	A: Uses Plutchik's Wheel of Emotions D: Limited to explicit emotions in tweets
9	2013	Dictionary-based methods, Corpus-based methods, polarity, and WordNet	Semeval-07 for news Semeval-13 for Twitter	65.86%	A: Can be used for video conference D: Limited to a certain domain
10	2021	n-grams, word embeddings, sentiment lexicons, rule-based methods, and deep learning models	SemEval, (SST), (ISEAR)	83.4%	D: Lack of resources for domain-specific corpora
11	2018	softmax, support vector machine (SVM)	Twitter	62.45%	D: Unavailability of fine-tuning of model embeddings in a semi-supervised manner

The bound morpheme is a word that is connected with other units such as suffixes and prefixes. The prefixes are associated at the beginning of the word. Where the suffixes end the word like singing. Moreover, each morpheme gives the word another meaning or another state.[9]

3.1.1 Derivational morphology

Another type of morphology branch is the derivational where adding affixes can affect the word itself and create a whole new meaning. What matters about this branch is its contribution in many factors such as changing the class of words, creating new words, and changing the meaning.[9]

3.1.2 Inflectional morphology

The last factor to be discussed in morphology is the inflectional.[9] This branch is concerned with the grammar aspect. Informations are usually affected by tense, number, gender, or aspect. The main aspect of inflectional morphology lies in grammatical markers and grammatical categories.

Grammatical markers: The grammatical markers don't change the core meaning rather they modify it. For example, The suffix 's' in 'dogs' marks the plural form in 'dog'

Grammatical Categories: Grammatical categories are classes that share the same grammatical functions such as the tense, gender, and numbers in the sentence

4 Challenges for English language Sentiment analysis

The following section shall discuss the challenges researchers faced in sentiment analysis with the English language and how the proposed model will tackle them later in this paper.

DOI: https://doi.org/10.54216/FPA.180115

Received: July 08, 2024 Revised: October 06, 2024 Accepted: December 29, 2024

4.1 Contextual Understanding

One of the main issues with sentiment analysis is the context meaning and how a word can convey several outcomes.[?] According to Balhur, English problems with contextual understanding lie in identifying the target of sentiment as news articles cover a wide range of subjects and targets. In addition, The non-lexical expression of sentiment can hinder the sentiment analysis. Most importantly it can be challenging to detect irony and sarcasm in the text and acquire the polarity for such matters.

4.2 Data Imbalance

For the English language, Data imbalance can be challenging for training sentiment analysis models. As some researchers faced the outcome was either too positive or too negative, which affected the results outcome. The number of positive, negative, and neutral instances in the dataset was not balanced. These issues can result in biased results and impact the accuracy of results as they used only one data set which is BBC. This can also lead to poor generalization on the results and misclassification of minority class

4.3 Domain-specific Sentiments

Domain-specific sentiments occur when using text from specific domains.[11] The challenges from this perspective can be the limited training data, making it hard to train accurate models for a certain domain. Also, the dynamic nature of each domain with all the new terms and expressions that are always coming up. The model needs adaptation to such changes for classification accuracy. Thus, standardization needs to be implemented to avoid such obstacles.

5 Methodology

Figure 2shows the proposed framework for sentiment analysis, which is drown to six steps :

- Data collection
- Data preprocessing
- Features extraction techniques
- Data splitting
- Classification based on ML
- Prediction and evaluation metrics

5.1 Data collection

The analysis focuses on data from BBC written in English and Sanad News dataset written in Arabic. First, the data needed to be cleaned and prepare a huge amount of data for the sentiment analysis.

The data sets collected are 2, the first data set is from BBC[?] which contains articles from 2004 to 2005. The data are a set of documents with the following categories (business, entertainment, politics, sport, and tech). It consists of 2225 documents. The documents also have a summarized duplicate.

Sanad dataset[19] is collected by web scraping techniques from many famous news sites such as Al-Arabiya, Al-Youm Al-Sabea (Youm7), Al-Jazeera, the news published on the Google search engine and other various sources. The data contains more than 500 document at each category with original Arabic news texts, without pre-processing. It is categorized into Culture, Finance, Politics, Sports and Tech.

DOI: https://doi.org/10.54216/FPA.180115

Received: July 08, 2024 Revised: October 06, 2024 Accepted: December 29, 2024

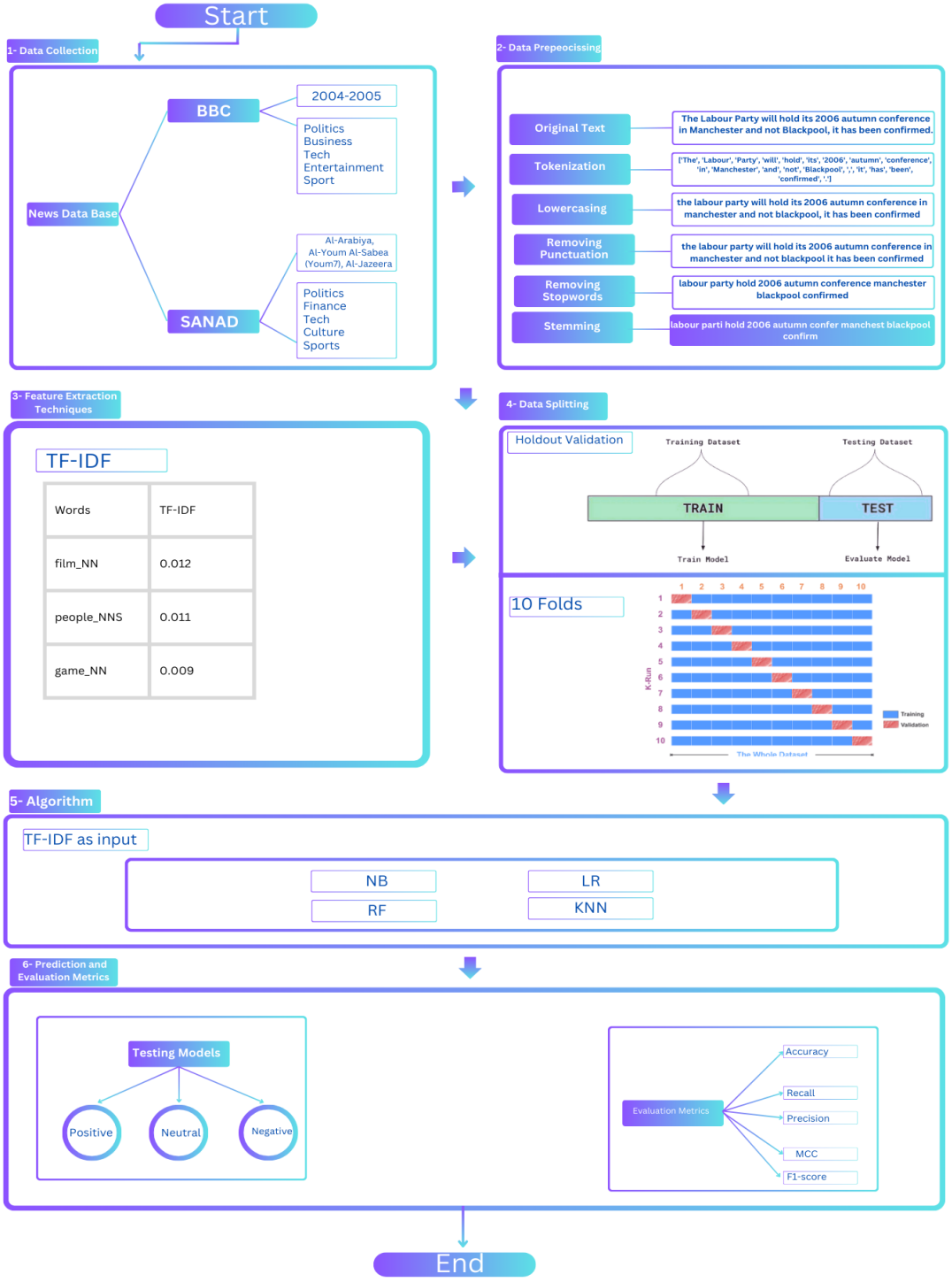

Figure 2: The Proposed System of Sentiment Analysis.

5.2 Data Preprocessing

The preprocessing phase is considered essential for the dataset to be deemed efficient for sentiment analysis. This phase goes through the following steps (tokenization, lowercasing, removing punctuation, eliminating stopwords, potentially stemming). These methods also include fitting the data into the ML models, which helps in eliminating information that isn't needed. Table 2 show the Phases of preprocessing. The preprocessing steps contain sub-phases that shall be mentioned:

DOI: https://doi.org/10.54216/FPA.180115

Received: July 08, 2024 Revised: October 06, 2024 Accepted: December 29, 2024

Original Text	The Labour Party will hold its 2006 autumn conference in Manchester and not Blackpool , it has been confirmed.
Tokenization	['The', 'Labour', 'Party', 'will', 'hold', 'its', '2006', 'autumn', 'conference', 'in', 'Manchester', 'and', 'not', 'Blackpool', ',', 'it', 'has', 'been', 'confirmed', '.']
Lowercasing	the labour party will hold its 2006 autumn conference in manchester and not blackpool, it has been confirmed
Removing Punctuation	the labour party will hold its 2006 autumn conference in manchester and not blackpool it has been confirmed
Removing Stopwords	labour party hold 2006 autumn conference manchester blackpool confirmed
Stemming	labour parti hold 2006 autumn confer manchest blackpool confirm

Table 2: Preprocessing Phases

- **Phase 1: Tokenization:**

  Original text:

  The Labour Party will hold its 2006 autumn conference in Manchester and not Blackpool, it has been confirmed.

  After Tokenization:

  Tokens: The, Labour, Party, will, hold, its, 2006, autumn, conference, in, Manchester, and, not, Blackpool, it, has, been, confirmed, .

- **Phase 2: Lowercasing:**

  Convert all words to lowercase to ensure consistency (unless case sensitivity is required for your analysis). Sentence after lowercasing: "the labour party will hold its 2006 autumn conference in manchester and not blackpool, it has been confirmed."

- **Phase 3: Removing Punctuation:**

  Remove all punctuation marks as they don't hold any significant in sentiment analysis.

  Sentence without punctuation: "the labour party will hold its 2006 autumn conference in manchester and not blackpool it has been confirmed"

- **Phase 4: Removing Stopwords:**

  Eliminate common words (like "the," "will," "its," "in," "and," "not," etc.) that don't carry significant sentiment or meaning.

  Sentence after removing stopwords: "labour party hold 2006 autumn conference manchester blackpool confirmed"

- **Phase 5: Stemming:**

  Reduce words to their root form to handle different variations of words.

  Sentence after stemming: "labour parti hold 2006 autumn confer manchest blackpool confirm"

5.3 Applying Feature Extraction Techniques for ML

In this step, we implemented Feature Extraction techniques for Standard ML models, as shown in the following subsection:

Machine Learning Feature Extraction Method

The following step is using TF-IDF with the documents, Table 3 shows a sample for the TF-IDF of BBC dataset. TF-IDF is Term Frequency-Inverse Document Frequency. It is used to retrieve information in a numerical matter to decide the importance of a word in the documents. The TF will calculate how many times a word was repeated and which one occurs the most. While the IDF measures the word across the whole corpus. The two measures provide a score for the term frequency and its weight. After preprocessing the data

DOI: https://doi.org/10.54216/FPA.180115

Received: July 08, 2024 Revised: October 06, 2024 Accepted: December 29, 2024

Words	TF-IDF
film_NN	0.012
people_NNS	0.011
game_NN	0.009

Table 3: TF-IDF

from news article, the TF-IDF will be calculated for each word. This will create a matrix with each word and it's TF-IDF score. Then the scores will be used as a feature to train the machine for the sentiment prediction (positive, negative, neutral) for each news article. In TF-IDF, the weight is determined by Eq.

(1), Eq.(2) and Eq. (3) :

$$TF(i,j) = \frac{\text{Frequency of term i in news } j}{\text{Total number of terms in news } j} \tag{1}$$

$$IDF(i,j) = \log\left(\frac{\text{Number of news in datasets}}{\text{Number of news include } i \text{ term}}\right) \tag{2}$$

$$W(i,j) = TF(i,j) \times \text{IDF}(i,j) \tag{3}$$

Where $TF(i,j)$ is the frequency of term i in review j, $IDF(i,j)$ is the frequency of feature concerning all reviews. Finally, the weight of feature i in review j , $W(i,j)$ is calculated by Eq.(3).

5.4 Data splitting

The next step is data splitting, Where the data set will be separated into portions for training and testing. The importance of this step lies in ensuring that the model learns from one set of data and evaluates it. The approach of data splitting goes as follows, hold out (80% training and 20% testing).

Another approach that shall be implemented is the 10-fold where both will be analysed and compared to the results.

5.5 Classification based on ML models

In this step, four regular ML algorithms have been used including machine learning algorithms (e.g., Naive Bayes, Logistic Regression, Random Forest, KNN) to classify CNN and BBC datasets.

- **Naive Bayes (NB)[?]** Uses Bayes' theorem for probabilistic machine learning. This algorithm is simple and fast. It is good to train and can work on a large number of features. More importantly, it is good with text classification, especially for news

- **K-Nearest Neighbor (KNN)[?]** This algorithm is a Supervised Classification one. The KNN is a straightforward algorithm that works on the proximity of feature space. It doesn't need a training phase, it uses the data for prediction.

- **Logistic Regression[?]** Logistic regression is great in High-Dimensional Spaces and is useful for differentiating between categories of news articles and using binary classification problems.
- **Random Forest (RF)[?]** Random Forest is a learning method used for classification and regression tasks. It works on decision trees and the bagging techniques of random selection reducing manual feature engineering.

DOI: https://doi.org/10.54216/FPA.180115

Received: July 08, 2024 Revised: October 06, 2024 Accepted: December 29, 2024

5.6 Performance metrics

Five standard performance metrics; Accuracy (ACC), Precision (PREC), Recall (REC), F1-score (F1) and The Matthews Correlation Coefficient (MCC) are used to evaluate the performance of the proposed models. They are calculated as follows:

$$\text{ACC} = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

$$\text{PREC} = \frac{TP}{TP + FP} \tag{5}$$

$$\text{REC} = \frac{TP}{TP + FN} \tag{6}$$

$$F1 = \frac{2.\text{PREC}}{\text{PREC} + \text{REC}} \tag{7}$$

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{8}$$

6 Experimental Results and Discussion

To examine the efficiency of Machine Learning (ML methods) in sentiment analysis, we have used two datasets: BBC news dataset and Sanad news dataset. The strategy being used is Holdout validation (80% in training and 20% in testing) and 10-Folds cross validation. Moreover the features used in ML methods are (KNN, LR, RF, NB) is used by TF-IDF.

6.1 Case Study I (BBC Dataset)

– Hold out Validation

This part explains the performance results of ML validation over BBC datasets.

Table 4 shows the values of six metrics including Accuracy (ACC), Recall (REC), Precision (PREC) F1-score (F1), (CA) and The Matthews Correlation Coefficient (MCC). The Holdout Validation (80-20) shows the sentiment analysis performance. Naïve Bayes and Logistic Regression models present consistency with a Classification Accuracy (CA) and F1 Score of 95.5%. Also, the model shows a balance for Precision and Recall at 95.5%, indicating that the model is equal in negative and positive and with no bias. However, KNN shows a moderate accuracy of 83.6% indicating its lower ability in detecting sentiment. Moreover, The Random Forest made an accuracy of 88.7%. Overall, Naïve Bayes and Logistic Regression show a higher accuracy in detecting sentiment.

– 10-Folds Cross Validation

This part shows the impact of splitting using 10-Folds cross validation over BBC dataset by employing ML as indicated in 4. Naïve Bayes and Logistic Regression models provide the highest percentage above 95% in AUC, CA, F1 Score, Precision, Recall, and MCC. Also, Random Forest performs well with around 90%, which is lower compared to Naïve Bayes and Logistic Regression in terms of CA and MCC. Moreover, KNN shows lower performance with around 50-53%, which is less accuracy and precision compared to the other models.

DOI: https://doi.org/10.54216/FPA.180115

Received: July 08, 2024 Revised: October 06, 2024 Accepted: December 29, 2024

Table 4: BBC and SANAD Performance Results

BBC Dataset	**Holdout Validation (80%-20%)**						
	Model	**AUC**	**CA**	**F1**	**PREC**	**Recall**	**MCC**
	kNN	0.836	0.524	0.526	0.763	0.524	0.473
	NB	0.996	0.955	0.955	0.955	0.955	0.943
	LR	0.995	0.955	0.955	0.955	0.955	0.943
	RF	0.982	0.887	0.887	0.89	0.887	0.859
	10 folds cross Validation						
	Model	**AUC**	**CA**	**F1**	**PREC**	**Recall**	**MCC**
	kNN	0.838	0.53	0.534	0.768	0.53	0.479
	NB	0.977	0.96	0.961	0.961	0.96	0.95
	LR	0.996	0.958	0.958	0.958	0.958	0.948
	RF	0.986	0.897	0.897	0.897	0.899	0.871
SANAD Dataset	**Holdout Validation (80%-20%)**						
	Model	**AUC**	**CA**	**F1**	**PREC**	**Recall**	**MCC**
	kNN	0.853	0.608	0.595	0.689	0.608	0.532
	NB	0.992	0.898	0.901	0.919	0.898	0.877
	LR	0.997	0.968	0.968	0.969	0.968	0.96
	RF	0.987	0.914	0.914	0.916	0.914	0.893
	10 folds cross Validation						
	Model	**AUC**	**CA**	**F1**	**PREC**	**Recall**	**MCC**
	kNN	0.852	0.607	0.594	0.683	0.607	0.53
	NB	0.993	0.903	0.906	0.922	0.903	0.883
	LR	0.998	0.972	0.972	0.972	0.972	0.965
	RF	0.988	0.917	0.917	0.92	0.917	0.897

DOI: https://doi.org/10.54216/FPA.180115

Received: July 08, 2024 Revised: October 06, 2024 Accepted: December 29, 2024

6.2 Case Study II (SANAD Dataset)

ML methods are tested on five categories Culture, Finance, Politics, Sports and Tech. Two splits are used including Hold out validation and 10-Folds Cross validation.

– Hold out Validation

The holdout validation showed that Naïve Bayes and Logistic Regression have higher performance percentages compared to kNN and Random Forest in table 4. Regarding the Accuracy Logistic Regression has shown the highest performance with 96.8%. However, Naïve Bayes is the highest performance in AUC followed by Logistic Regression. Meanwhile, KNN has been classified as the lowest percentage as it is focused on the K nearest neighbors by Euclidean distance which can be sensitive to Arabic language. To sum up, Naïve Bayes and Logistic Regression prove to be the best model for SANAD dataset for accurately classifying sentiments in Arabic news articles.

– 10-Folds Cross Validation

The 10 folds have also shown a high performance for Naïve Bayes and Logistic Regression across all metrics as indicated in table 4 . Naïve Bayes showed an AUC of 99.3%, with accuracy, F1 score, precision, recall, and MCC percentages nearing or exceeding 90%. While the Accuracy was 90.3%. As for Logistic Regression, it shows Accuracy (CA) and and F1 Score for 97.2%. As for the kNN and Random Forest, they were the lowest percentages. Naïve Bayes and Logistic Regression models work the best for Arabic sentiment analysis based on their performance metrics. To conclude, 10-fold cross-validation is better from the aspect of performance and consistency due to its comprehensive evaluation.

6.3 Graphical analysis

Figure 3 summarizes the best values of metrics in terms of ACC, PREC, REC, F1 and MCC, obtained by the ML methods for the two datasets BBC and SANAD, using both strategies of splitting (hold out and 10 folds cross-validation).

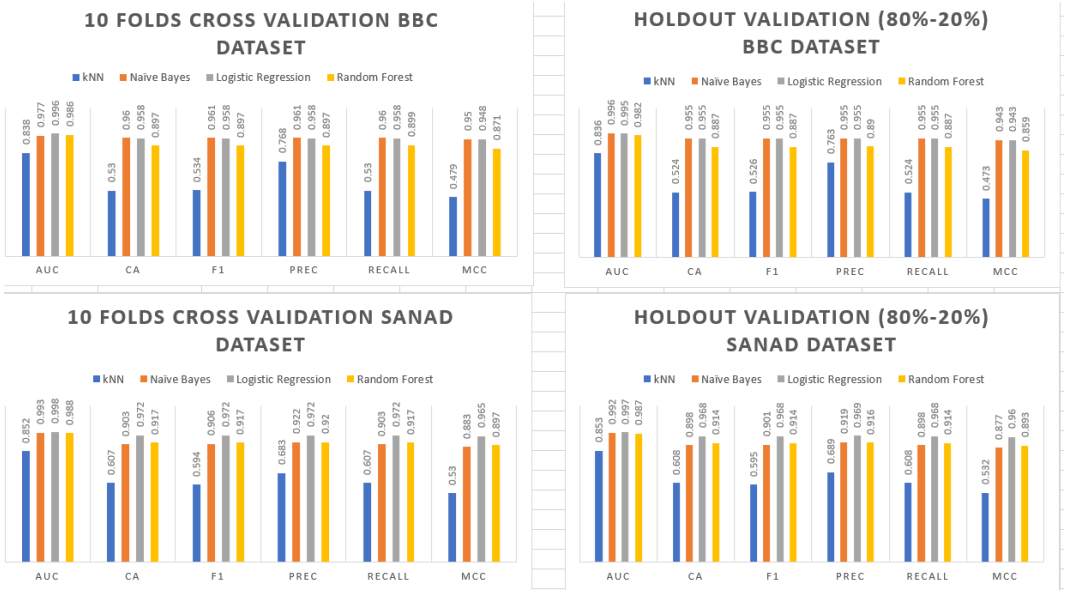

Figure 3: The Performance Results for the BBC and SANAD Dataset

DOI: https://doi.org/10.54216/FPA.180115
Received: July 08, 2024 Revised: October 06, 2024 Accepted: December 29, 2024

7 Data Comparison

BBC data set results has shown that Naïve Bayes and Logistic Regression works better for news dataset while kNN and Random Forest are lower in performance. As for SANAD Dataset, the 10 folds have shown a better performance and Naïve Bayes and Logistic Regression exhibit a higher performance than the rest. They provide high AUC values and consistently high accuracy, F1 score, precision, recall, and MCC. Based on both performances Logistic Regression is the most suitable model for such a field as it stands out for sentiment analysis in Arabic and English.

8 Proposed ML Model

The proposed system of Sentiment Analysis for news includes:

- **Collect data**: The data is collected from 2 datasets BBC and SANAD Arabic news data sets.
- **Features extraction techniques**: The features are extracted using TF-IDF for Machine learning models.
- **Data splitting**: The data is splited into (80% training and 20% testing) and 10-Folds cross validation.
- **Classification Based on Machine Learning**: Sentiment analysis is classified based on logistic regression (LR), K-nearest neighbor (KNN), random forest (RF) and Naive Bayes (NB).
- **Evaluation Parameters**: The parameters used for evaluation are Accuracy (ACC), Precision (PREC), Recall (REC), F1-score (F1) and The Matthews Correlation Coefficient (MCC).

9 Conclusion and future work

News articles gives a representative view of the sentiment analysis among the world events. In this paper, sentiment analysis in Arabic and English has been discussed based on 4 basic machine learning algorithms. In the study, we have selected two data sets with various categories included. The articles are listed as positive, negative, or neutral. Moreover, the extraction feature used was TF-IDF. Both data sets have accomplished great testing results for the 10-fold with logistic regression and naive Bayes. Where the performance measure results suppress the 90%. Furthermore, the best classifier for machine learning is the logistic regression. The lowest classifier performance was KNN as it uses Euclidian distance.

Our future work: We will seek to use Deep Learning methods using more classifiers to dig deep into the sentiment analysis of Arabic news articles and it's dialects.

Conflicts of interest

The authors have declared that there is no conflict of interest. Non-financial competing interests.

References

[1] Bielsa, Esperança, "Translation in Global News Agencies," *Target*, vol. 19, 2007, doi: 10.1075/target.19.1.08bie.

[2] Ilyosovna, Niyozova Aziza, "The importance of English language," *International Journal on Orange Technologies*, vol. 2, no. 1, pp. 22–24, 2020.

DOI: https://doi.org/10.54216/FPA.180115

Received: July 08, 2024 Revised: October 06, 2024 Accepted: December 29, 2024

[3] Balahur, Alexandra, Ralf Steinberger, Mijail Kabadjov, Vanni Zavarella, Erik van der Goot, Matina Halkia, Bruno Pouliquen, and Jenya Belyaeva, "Sentiment Analysis in the News," 2010.

[4] Taj, Soonh, Areej Meghji, and Baby Shaikh, "Sentiment Analysis of News Articles: A Lexicon-based Approach," 2019, doi: 10.1109/ICOMET.2019.8673428.

[5] Samuels, Antony and John Mcgonical, "News Sentiment Analysis," 2020.

[6] Alonso, Miguel A., David Vilares, Carlos Gómez-Rodríguez, and Jesús Vilares, "Sentiment analysis for fake news detection," *Electronics*, vol. 10, no. 11, pp. 1348, 2021.

[7] Mehta, Pooja and Sharnil Pandya, "A review on sentiment analysis methodologies, practices and applications," *International Journal of Scientific and Technology Research*, vol. 9, no. 2, pp. 601–609, 2020.

[8] Salam, Shaikh and Rajkumar Gupta, "Emotion Detection and Recognition from Text using Machine Learning," *International Journal of Computer Sciences and Engineering*, vol. 6, pp. 341–345, 2018, doi: 10.26438/ijcse/v6i6.341345.

[9] Carstairs-McCarthy, Andrew, *Introduction to English Morphology: Words and Their Structure*, Edinburgh University Press, 2017.

[10] Denis, Alexandre, Samuel Cruz-Lara, and Nadia Bellalem, "General purpose textual sentiment analysis and emotion detection tools," *arXiv preprint arXiv:1309.2853*, 2013.

[11] Nandwani, Pansy and Rupali Verma, "A review on sentiment analysis and emotion detection from text," *Social Network Analysis and Mining*, vol. 11, no. 1, pp. 81, 2021.

[12] Stojanovski, Dario, Gjorgji Strezoski, Gjorgji Madjarov, Ivica Dimitrovski, and Ivan Chorbev, "Deep neural network architecture for sentiment analysis and emotion identification of Twitter messages," *Multimedia Tools and Applications*, vol. 77, 2018, doi: 10.1007/s11042-018-6168-1.

[13] Soo-Guan Khoo, Christopher, Armineh Nourbakhsh, and Jin-Cheon Na, "Sentiment analysis of online news text: A case study of appraisal theory," *Online Information Review*, vol. 36, no. 6, pp. 858–878, 2012.

[14] Zhang, Zijun, "Improved adam optimizer for deep neural networks," in *2018 IEEE/ACM 26th International Symposium on Quality of Service (IWQoS)*, IEEE, 2018, pp. 1–2.

[15] Cover, Thomas and Peter Hart, "Nearest neighbor pattern classification," *IEEE Transactions on Information Theory*, vol. 13, no. 1, pp. 21–27, 1967.

[16] Peng, Joanne, Kuk Lee, and Gary Ingersoll, "An Introduction to Logistic Regression Analysis and Reporting," *Journal of Educational Research*, vol. 96, pp. 3–14, 2002, doi: 10.1080/00220670209598786.

[17] Pal, Mahesh, "Random forest classifier for remote sensing classification," *International Journal of Remote Sensing*, vol. 26, no. 1, pp. 217–222, 2005.

[18] Greene, Derek and Pádraig Cunningham, "Practical Solutions to the Problem of Diagonal Dominance in Kernel Document Clustering," in *Proc. 23rd International Conference on Machine Learning (ICML'06)*, ACM Press, 2006, pp. 377–384.

[19] Einea, Omar, Ashraf Elnagar, and Ridhwan Al Debsi, "Sanad: Single-label arabic news articles dataset for automatic text categorization," *Data in Brief*, vol. 25, pp. 104076, 2019.

[20] Schonlau, Matthias, "The Naive Bayes Classifier," in *The Naive Bayes Classifier*, 2023, pp. 143–160, doi: 10.1007/978-3-031-33390-3_8.

[21] Statista, "The Most Spoken Languages Worldwide," 2023, [Online]. Available: https://www.statista.com/statistics/266808/the-most-spoken-languages-worldwide/, Accessed: December 12, 2023.

DOI: https://doi.org/10.54216/FPA.180115

Received: July 08, 2024 Revised: October 06, 2024 Accepted: December 29, 2024