**NATURAL LANGUAGE PROCESSING PROJECT**

**SENTIMENT ANALYSIS ON NEWS JOURNALS**

**Github URL :** [**https://github.com/udayk09/NLP**](https://github.com/udayk09/NLP)

**Team Structure:**

|  |  |
| --- | --- |
| Team member | Email |
| Konduru Uday Kiran | [udaykirankonduru@my.unt.edu](mailto:udaykirankonduru@my.unt.edu) |
| Veerla Nagi Reddy Aanabolu | [veerlanagireddyaanabolu@my.unt.edu](mailto:veerlanagireddyaanabolu@my.unt.edu) |
| Likith Guduru | [likithguduru@my.unt.edu](mailto:likithguduru@my.unt.edu) |
| Raja Srinivas Nelluri | [rajanelluri@my.unt.edu](mailto:rajanelluri@my.unt.edu) |

**Goals and Objectives:**

This project helps in correcting the words that uses in the newspapers and electronic media which helps in giving the right information to the public.

# Motivation:

Through print and electronic media, journalists provide breaking news about events occurring all around the world. We frequently encounter grammatical errors and profanity that is either offensive or changes the sense of the story. We may correct those errors by utilizing our program.

To learn more about the data for our project, we are employing a verbal dataset and sentiment analysis. This method can be further customized in order to acquire insights from various data.

# Significance:

Now-a-days, almost everyone follows the news through Print, News channels, and Social media which create a lot of influence on the public in a day to day life. This kind of initiative is crucial to getting the message out to the public.

# Objectives:

The main intention of this project is to modify the text by removing the abusive words in the content and providing disciplinary meaning to the public who uses print and media without violating the content. To conduct sentiment analysis on the dataset, we will use Python programming and tools like Pandas, NLTK, e.t.c.

# Features:

Important features used in the project:

* Data collection: As a data source for this project, we are using news journal data that has been saved in a CSV file.
* Data processing: Different levels of data processing, such as punctuation removal and lemmatization, are performed here. The analytical part can be made simple by following these procedures.
* Data Analysis: As part of data analysis, we will use sentiment analysis to group the messages into various categories.

# Workflow:

Below are steps to be followed in the project.

1. Data Preparation
2. Data Pre-Processing
3. Analysis
4. Machine Learning Model applying
5. Evaluation of model

**Table of Contents**

[Introduction 3](#_Toc118124402)

[Related Work (Background) 3](#_Toc118124403)

[Dataset 4](#_Toc118124404)

[Detail design of Features 5](#_Toc118124405)

[Analysis 6](#_Toc118124406)

[Implementation 8](#_Toc118124407)

[Preliminary Results 9](#_Toc118124408)

[Project Management 10](#_Toc118124409)

[Implementation status report 10](#_Toc118124410)

[Work completed 11](#_Toc118124411)

[Issues/Concerns 13](#_Toc118124412)

[Conclusion 13](#_Toc118124413)

[References 14](#_Toc118124414)

# Introduction

***“Natural Language Processing”*** has improved better than any additional segment of Artificial Intelligence. With substances like supporting chatbots evolving better functional and correct and permitting businesses to acquire beneficial discernment from any client deliberations fast, Natural Language Processing is helping establishments in uncovering discounts within personal and important information with ***“sentiment analysis”***. The attribute is that users can create assessments, and forecasts and provide the knowledge for devices to understand by themselves. This sentiment analysis is very much influential and endless in the duration of application opportunities.

# Related Work (Background)

The selected dataset that was just concerned with retains tweet message reviews. Each consideration is either marked as favorable or unfavorable. The selected dataset includes the ***‘review text message’*** and ***‘review sentiment’*** domains. The area of Natural Language Processing dates to a periodic decade and has evolved rather greatly overall these particular years. Originally contained to collecting all information from a narrow collection of digitized records. A substantial portion of the standing was accomplished in the ***“Information Retrieval”*** domain which is deemed an application of all different types of Natural Language Processing specialization. Before examining the information retrieval techniques, delve into the academic and functional elements of Natural Language Processing after a period.

**Figure 1: Steps of Sentiment analysis**

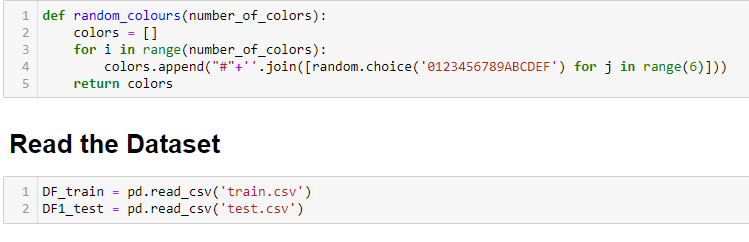
(Source: Qi *et al.* 2020)

The foremost problem is to prepare a provided record into comments and corrections. The comment token was originally determined to upload programming terminologies idea, this part is currently interchangeable with segmenting the reader into comments. Most speeches operate in white areas like all delimiters while this part can be a time problematic in specific vocabularies. Confusing the issue additionally would be the vocabulary of the composition. The Unicode normal benefits like individual texture are transferred a distinctive Significance and consproduce produces it applicable to determine the accurate language.

In extra to utilizing text messages in tickets or different types of words, the Natural processing Language field identifies a prominent focus on discovering the perimeter of every statement correction. While multiple speeches will utilize punctuation spots to represent correction edges, other terminologies. Proven to exist considerably additional challenges in this particular consideration. Confusing the issue additionally are straightforward arrangements that utilize the duration symbol. While utilized for completing a correction, a permit strength transmit the immoral movement.

# Dataset

The first step of this project is to import all libraries and then load the required dataset. The selected dataset data utilizing with ***panda’s*** *package to read\_csv()* process. The first pieces of information of laden information are displayed in the output section. Firstly insert all required libraries needed for reading the dataset, create a visualization graph, and analyze part. After that read the dataset with the proper library (Neumann *et al.* 2019). After loading this dataset checked the shape of the dataset. They mainly used two different datasets for text sentiment analysis. The desire to comprehend more about the application of machine learning in sentiment classification, which enables a method of natural language processing a subject that is now very hot led to the creation of this research.



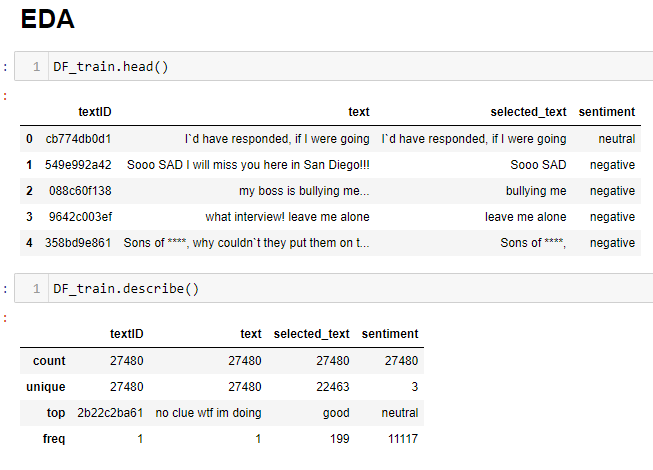
**Figure 2: Read the Dataset**

(Source: Acquired from Python)

Here I adopted the very same concept effort determine whether they are bad, building on my prior expertise with short tracks and categorizing it in terms of sentiment. After that checked all null values in the dataset, when finding any null value in this dataset so removed all null values from the dataset. The dataset is the training dataset there checked the null values and then removed these null values after that check the test dataset was for null value find. After completing this checking process now going EDA portion with these two fresh datasets.

# Detail design of Features

Now, this part solved the ***EDA part*** with some important features. So here firstly called some values and then check all sentiment categories and count the number of texts in each statement category. So after completing this total process there have found three different types of statements in the dataset. Identifying the prevailing feeling of a document collection is the goal of sentiment classification, also known as collaborative filtering, which is a subtask of learning algorithms. Researchers can retrieve the emotional data contained within a categorize it depending on its polarity, including pleasant, neutral, or negative, utilizing machine learning methods learning using natural language processing algorithms.



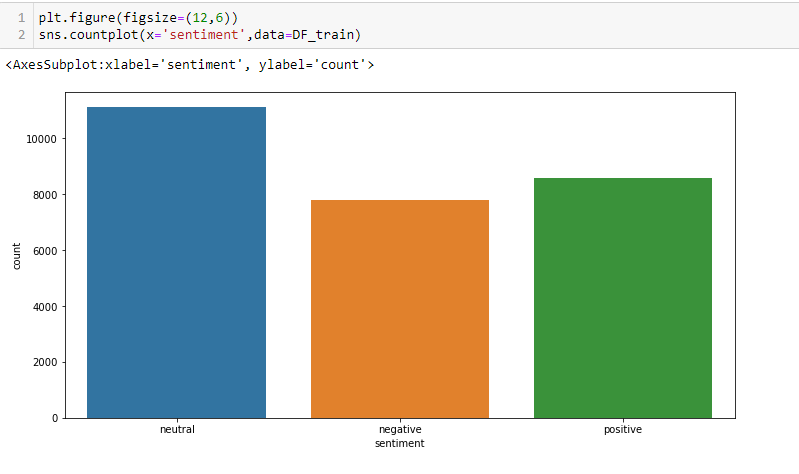
**Figure 3: EDA data Analysis**

(Source: Acquired from Python)

It is very helpful research since we might maybe ascertain the general perception of a selling item or forecast stock prices for a specific company, for example, if so many people have a favorable opinion of it, perhaps its stock prices would rise, and so on. In this project, researchers have opted to use a prediction algorithm to try and categorize tweets from Twitter as having "good" or "bad" sentiments (Lauriola *et al.* 2020). Twitter is a social media site that allows users to write short text tweets too rapidly and impulsively express their including the target symbol or the hashtag "#" in your twee tweet, you can join in a topic or address anyone specifically in a tweet. Twitter is a wonderful source of information to ascertain the current general sentiment about almost anything due to its popularity.

# Analysis

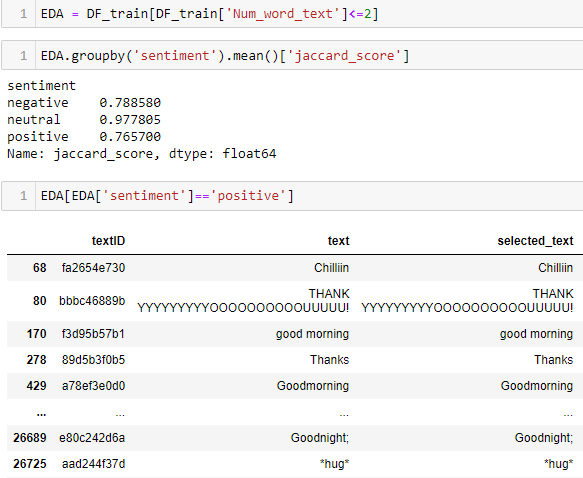
Now can convert the raw tweets provided that possess the library of comments as well as any websites that might be helpful (Shah *et al.* 2019). It is crucial since the performance of the classification algorithm will be directly impacted by all the changes we make during this process. Cleaning, normalizing, transformations, extraction, classification, and selecting are all included in the pretreatment. Preprocessing will produce homogeneous prediction and detection that can be used to optimize the classification performance of the classifier.



**Figure 4: Sentiment types Analysis**

(Source: Acquired from Python)

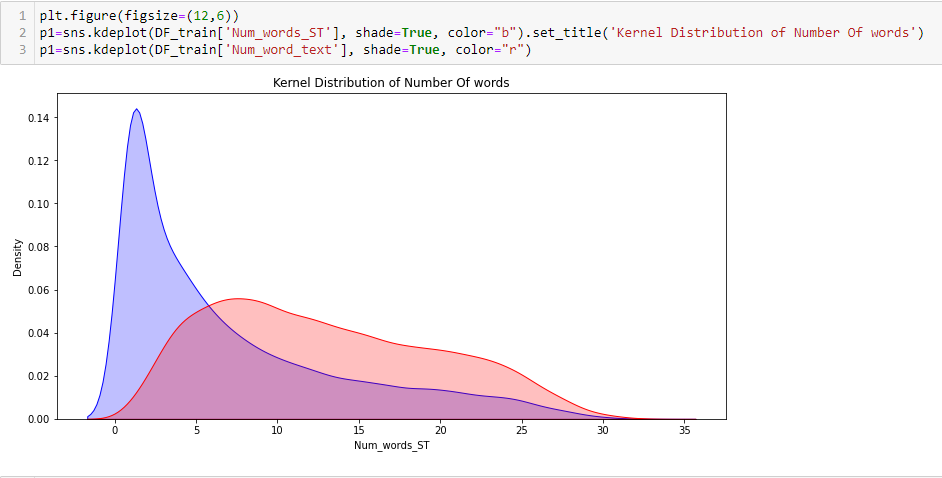
Utilizing the emoticon dictionary, we swap out all emoticons for their positive and negative sentiment polarities. To substitute, we iterate through each tweet and use a regex to determine whether it includes symbols and if so, replace them with the appropriate polarity. All abbreviations are changed to their interpretation. An acronym is an abbreviated created from a quote's or words basic building blocks.



**Figure 5: Sentiment Value Analysis**

(Source: Acquired from Python)

Typically, these pieces include phrases or names, particular letters (such as North Atlantic treaty organization or light), or both. As readers will see from the accompanying bar diagram, our data collection of tweets contains a lot of abbreviations (Sun *et al.* 2019). At this stage, tweets will be converted into binary by removing whitespace and utilizing split to finish the procedure quickly. Designers might need parts, but it would be considerably slower.



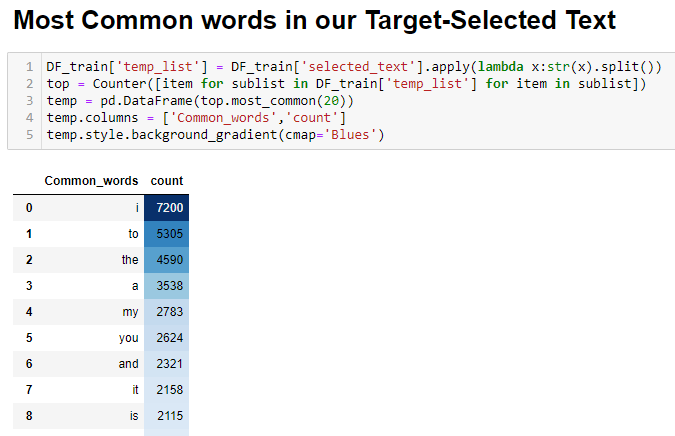
**Figure 6: Kernel Distribution of statements**

(Source: Acquired from Python)

To remove a smaller portion of statements such as "I don't enjoy it," designers use the negativity vocabulary to substitute all negativity terms like "never," "absolutely," and "hardly ever" with the label ||not||. Given that there was a "don't" previously, like ought not to be interpreted positively here.

# Implementation

Classifier, Multilayer Perceptron, and Communicative Language teaching are the three principal models used throughout sentiment classification to categorize a text as either favorable or negative. SVM is renowned for producing the greatest results, however, in this research, designers solely concentrate on generative models, specifically Multilayer Perceptron and Communicative Language teaching, which are frequently employed in this area (Guo *et al.* 2020). Let's start by introducing the Naïve Bayesian model, which is renowned for being both straightforward and effective at classifying texts. Ignorant Bayes classifier is a simple straightforward probabilistic classifier used in learning algorithms.



**Figure 7: Common words found**

(Source: Acquired from Python)

To achieve this, "don't" will be changed to "not," and the word "like" will not be considered a compliment. Designers should note that whenever a negativity is met, the words in the both positive and negative phrase databases that precede the negativity word will be switched around. Words of encouragement will have become negligible, and vice versa, as we look for both positive and negative terms (Neumann *et al.* 2019). After putting the take-place stages into practice, designers can concentrate on that supervised learning portion.

# Preliminary Results

Designers investigate the practice of an individual group, the process by which individuals react to separate parties, and the process thought around detail commanders. Utilizing live tweets as all different information appointed. Users obtained a precision of approximately 82% for all Multinomial Classifiers. After using this classifier means all classifier is functioning perfectly. Other effects, users can discover every user’s conviction and accuracy. The variety which all users have accomplished is essentially like a franchise.



**Figure 8: Tree of unique words**

(Source: Acquired from Python)

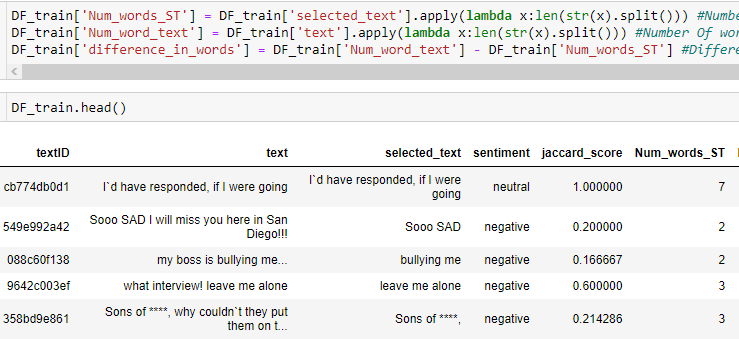
Therefore, designers can involve a franchise method to estimate the highest numeral of franchises for quality by operating method and try to make it as enthusiastic. It existed discovered that all tweets category of pessimistic conviction is trying to make 100% accuracy, whereas the approving category is around 85% and this is dedicated. The accuracy of every measure was discovered at around 70%, so that represents most of all derivatives replaced by all classifiers.

# Project Management

All algorithms are based on the implementation of the Bayesian hypothesis with significant (naive) partial information across the characteristics. The amount of parameters required for naive Bayes classifiers is proportional to the number of variables (characteristics) in a machine learning model, making them extremely scalable. Evaluation of a shuttered statement like a quantitative statement capable of being examined in a limited amount of operations, which requires linear time, can be used for optimum training (Wolf *et al.* 2020). It is always beneficial to establish a baseline for any task involving machine learning. For the initial categorization, a basic model is frequently implemented in a "fast and dirty" manner, and based on the results, an attempt is made to enhance it. First, the learning algorithm and testing dataset are separated from the complete data set.

# Implementation status report

To achieve this, first, initially reshuffle the set of statistics to remove any order imposed on the information. Next, we select tweets from each pair of either positive or negative comments and combine them to create the learning algorithm. The sample set is created using the remainder. Finally, there are 1183958 retweets in the training data set and 394654 inside the testing dataset. They are symmetrical and share the same distributions as the preliminary data set, as can be seen. Their learning technique is the Multilayer Perceptron Bayes, while the traditional method of classifying texts is Laplace smoothness.



**Figure 9: Implement the dataset**

(Source: Acquired from Python)

The designer chooses the bags of words paradigm to describe the current large dataset of posts on Twitter due to the need to extract characteristics from it. The handbag of words and words is a condensed description of a publication that ignores the morphology or sentence order and simply represents the content as a handbag of its words (Qi *et al.* 2020). The frequency (number of times that each reader can understand a manuscript is utilized as a characteristic for training the classification algorithm in text summarization.

# Work completed

**1. Data Preparation**

The data preparation part is the main part of the project because here controlled the dataset. Here also import all libraries for reading the dataset and also solved the analysis part of this project.

**2. Data Pre-Processing**

After importing all libraries and the dataset, check the dataset and find if there is any null value present or not. After completing this, all things arrange this dataset with a new style and then start the next new parts of this project.

**3. Analysis**

This analysis part checked every text id from the dataset and arrange this dataset with their required process. Training all different types of classifiers related to already complete or negatively organized penalties and removing attributes that exist in comments, users subsequently try these all different classifiers on some corrections.

|  |  |  |  |
| --- | --- | --- | --- |
| **S.No** | **Task** | **Person** | **Percentage** |
| 1 | Data Prepartion | Likith & Raja | 36% |
| 2 | Data Pre-processing | Uday & Nagireddy | 39% |
| 3 | Analysis | Nagireddy, Raja Srinivas | 25% |

# Work to be done

**4. Machine Learning Model applying**

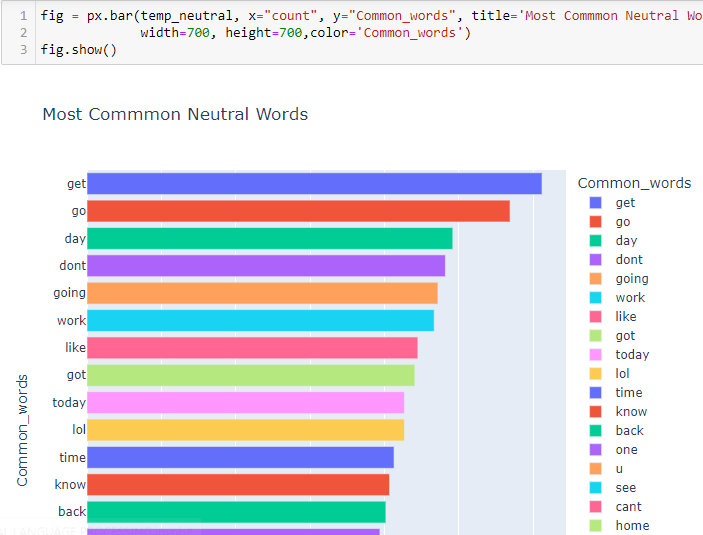
Support Vector Appliance classifiers are used for understanding ideals with different associated understandings all different algorithms that interpret data information to determine designs to segregate into different or similar classes. All classes are exploited for category and deterioration estimation.

**5. Evaluation of model**

Sentiment Investigation employs the evaluation metrics of Accuracy, rearrangement, Future score, and Precision. Also, intermediate calculations with weighted F1 values are helpful for multi-class problems. Depending on the balance of classes of the dataset the most appropriate metric should be used.

**Description**

Nowadays, this text sentiment analysis or thought analysis is a superheated subject in apparatus knowledge. The period of accumulating significant understandings from sociable media information has currently reached with the advancement in this sentiment analysis technology (Otter *et al.* 2020). The Twitter text sentiment analysis case analysis offers everyone a peek at the control of sentiment analysis Search. It’s a period for all user communities to push outside general opinion and countable metrics.



**Figure 10: Tree of Common words**

(Source: Acquired from Python)

Organizations have existed to leverage and control the intensity of information recently, but to reach the most resonant of communication, designer include control of the influence of all different types of AI, Deep knowledge, and intellectual classifiers like Emotion Analysis. The designer can reach all users to empower their machine learning technology or obtain a properly customized dashboard used for breeding expressive understandings from all different digital platforms.

# Issues/Concerns

Techniques can be challenging to diagnose verbally, and after challenging to gather release in all registered expressions (Maulud *et al.* 2021). Something brings actor confusion when a single type attempts to investigate the tremendous importance of information that exists and contains both personal and accurate replies. Labels can encounter complications in locating personal opinions and correctly investigating these problems for their planned technique.

* Individuals utilize satire and mockery in everyday discussions and different types of memes on social platforms.
* The effort of expressing differing opinions by employing backhanded commendations can terminate it challenging for emotional research instruments to catch the proper context of what this specific reaction is renting to.
* This process usually results in more heightened importance of “optimistic” outcomes than is reaagonistic.
* The issue with a sociable platform range that is valuable information, like this famous Twitter, is attacked with others things.

NLP studies are prepared to language characteristics. While users can remove textbooks from actual photos, a speech in itself (Qiu *et al.* 2020). Most sentiment research explanations minister things like unique qualities that easily withdraw from all information during all different approaches of emotion mining. But accomplishing so suggests that the characters' choice does not accept holistic understandings from all types of information.

# References

Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J. and Poon, H., 2021. Domain-specific language model pretraining for biomedical natural language processing. ACM Transactions on Computing for Healthcare (HEALTH), 3(1), pp.1-23.

Guo, J., He, H., He, T., Lausen, L., Li, M., Lin, H., Shi, X., Wang, C., Xie, J., Zha, S. and Zhang, A., 2020. GluonCV and GluonNLP: Deep Learning in Computer Vision and Natural Language Processing. J. Mach. Learn. Res., 21(23), pp.1-7.

Lauriola, I., Lavelli, A. and Aiolli, F., 2022. An introduction to deep learning in natural language processing: Models, techniques, and tools. Neurocomputing, 470, pp.443-456.

Maulud, D.H., Zeebaree, S.R., Jacksi, K., Sadeeq, M.A.M. and Sharif, K.H., 2021. State of art for semantic analysis of natural language processing. QubCubanademic Journal, 1(2), pp.21-28.

Neumann, M., King, D., Beltagy, I. and Ammar, W., 2019. ScispaCy: fast and robust models for biomedical natural language processing. arXiv preprint arXiv:1902.07669.

Otter, D.W., Medina, J.R. and Kalita, J.K., 2020. A survey of the usages of deep learning for natural language processing. IEEE transactions on neural networks and learning systems, 32(2), pp.604-624.