iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 8 **Issue:** XI **Month of publication:** November 2020

DOI: https://doi.org/10.22214/ijraset.2020.32112

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue XI Nov 2020- Available at www.ijraset.com

Event Detection from the News Headlines

Kaalishwar R[1], Gunanithi A[2], Dheveshwarann R[3], Dr. M. Sujithra[4], Dr. P. Velvadivu[5]
[1, 2, 3] *3rd Year, MSC. Data Science (Integrated), Coimbatore Institute of Technology, Coimbatore.*
[4, 5] *M.C.A., M.Phil., Ph.D, Assistant Professor, Coimbatore Institute of Technology, Coimbatore.*

Abstract: Internet has become the main source of news in the world. There are thousands of website which constantly publish and update the news stories around the world. Not every news items is relevant for everyone but some news items are very critical for some people or businesses. The event detection system is a big data based natural language processing system. The natural language processing system brings the intelligence to detect the events in the random headline sentences from the news items. The time series and Pyspark NLP, Naïve Bayes algorithm has been used.
Keywords: Website, news, big data, hadoop, Pysapark NLP, time series, Naïve Bayes machine learning algorithm, keywords

I. DATASET DESCRIPTION

The dataset has been taken from the Inshorts website where there is an instance of two lakh news events dating from 2012 to 2019 having a 41 categories which include education, politics, sports, entertainment and other categories.The dataset has been released which contains the authors, category, date, headline, link, short_description.

A. *Attribute Information*

1) *Authors:* The person who published the article
2) *Category:* The domain or the field which the news is based on or described.
3) *Date:* The date when it was published.
4) *Headline:* The headline of the news or the article.
5) *Link:* The website link(url) for the particular news.
6) *Short_description:* The shorts of the news which gives brief information about the news.

II. EXPLORATORY ANALYSIS

The exploratory analysis is done for the dataset. The data types of each column, luckily there are no missing values. The total number of counts under each category, having maximum, minimum count. Authors who published more articles .TThe number of ner=ws per day. Dicky fuller test has been used and the value of p is 0.984733

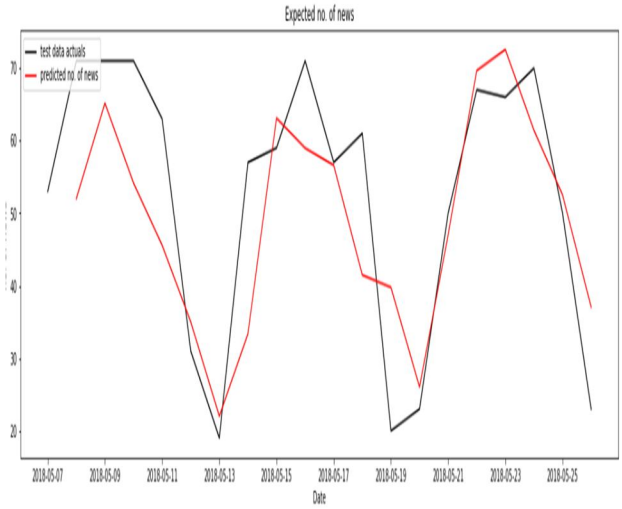
```
Results of Dickey-Fuller Test:
Test Statistic                   0.496717
p-value                          0.984773
#Lags Used                      27.000000
Number of Observations Used    2281.000000
Critical Value (1%)              -3.433220
Critical Value (5%)              -2.862808
Critical Value (10%)             -2.567445
dtype: float64
```

Now the ARIMA model has been constructed which gives the probability, minimum, maximum, percentiles, quartiles which is based on the date.

```
                              ARMA Model Results
==============================================================================
Dep. Variable:            count(date)   No. Observations:                 2309
Model:                     ARMA(2, 1)   Log Likelihood               -9136.242
Method:                       css-mle   S.D. of innovations             12.651
Date:                Tue, 27 Oct 2020   AIC                          18282.484
Time:                        09:10:02   BIC                          18311.207
Sample:                             0   HQIC                         18292.954

================================================================================
                   coef    std err          z      P>|z|      [0.025      0.975]
--------------------------------------------------------------------------------
const            86.9631      0.515    168.770      0.000      85.953      87.973
ar.L1.count(date)   0.1937      0.074      2.621      0.009       0.049       0.339
ar.L2.count(date)   0.0295      0.051      0.580      0.562      -0.070       0.129
ma.L1.count(date)   0.5208      0.070      7.414      0.000       0.383       0.658
                                    Roots
=============================================================================
                  Real          Imaginary           Modulus         Frequency
-----------------------------------------------------------------------------
AR.1            3.4022           +0.0000j            3.4022            0.0000
AR.2           -9.9787           +0.0000j            9.9787            0.5000
MA.1           -1.9202           +0.0000j            1.9202            0.5000
-----------------------------------------------------------------------------
                     0
count  2309.000000
mean      0.000791
std      12.656766
min     -66.902240
25%      -4.782593
50%       2.527452
75%       7.083868
max      39.070108
```

The linear graph has been built to show the comparison of the predicted with the original and to evaluate whether the procedure is correct or not. The black denotes the test and red denotes the predicted output.The two lines are similar.



III. CATEGORY DETECTION MODEL

The model has been built using Pyspark natural language processing to detect the category. The headline of the event has been converted to only string in which the words which are of less importance are omitted, and strings are converted to words,and they are encoded to some integer value. The filtered word has been used for the identification of the event, by assigning the category index. The Naïve Bayes model has been built for the event detection.

```
Accuracy of NaiveBayes is = 0.527288
Test Error of NaiveBayes = 0.472712
```

A. Word Search (KEY)

The word helps to find the headline containing the word like the word 'Football'. It lists out all the headlines.

authors	category	date	headline	link	short_description
Mary Papenfuss	CRIME	2018-02-18	Report: 2 Baylor ...	https://www.huffi...	The school has be...
Ron Dicker	ENTERTAINMENT	2018-02-02	Alex Trebek Mocks...	https://www.huffi...	"If you guys ring...
Ron Dicker	ENTERTAINMENT	2018-02-02	Justin Timberlake...	https://www.huffi...	The singer spoke ...
Harrison Wilkerso...	QUEER VOICES	2018-01-19	Gay Former Footba...	https://www.huffi...	Harrison Wilkerso...
	SPORTS	2018-01-09	Alabama Rallies T...	https://www.huffi...	The Crimson Tide ...
Ron Dicker	COMEDY	2018-01-03	'The Opposition' ...	https://www.huffi...	"While the NFL is...
Jim Buzinski, Out...	QUEER VOICES	2017-12-06	High School Footb...	https://www.huffi...	'My teammates and...
Andy McDonald	SPORTS	2017-12-03	We Now Know The 4...	https://www.huffi...	As always, debate...
Ron Dicker	SPORTS	2017-11-28	Video Surfaces Of...	https://www.huffi...	"Discipline was h...
Rebecca Shapiro	SPORTS	2017-10-27	A Kitten's Scramb...	https://www.huffi...	This tailback mus...
Mary Papenfuss	BLACK VOICES	2017-10-24	Middle School Foo...	https://www.huffi...	A Virginia town i...
Ed Mazza	SPORTS	2017-10-23	'Sunday Night Foo...	https://www.huffi...	The long-awaited ...
Alan Singer, Cont...	POLITICS	2017-10-09	Offended By Prote...	https://www.huffi...	On Sunday, Vice-P...
Lee Moran	ENTERTAINMENT	2017-10-06	John Cleese's Bum...	https://www.huffi...	"Why do you call ...
Bill Bradley	SPORTS	2017-10-02	NFL To Hold Momen...	https://www.huffi...	ESPN has announce...
Doha Madani	COMEDY	2017-10-01	'SNL' Does Its Be...	https://www.huffi...	Football, Puerto ...
Ron Dicker	CRIME	2017-09-29	High School Footb...	https://www.huffi...	Several players a...
Taryn Finley	BLACK VOICES	2017-09-20	Entire Third Grad...	https://www.huffi...	The team of 8-yea...
Center for Commun...	POLITICS	2017-09-09	Dreamers Are Peop...	https://www.huffi...	People should not...
Dr. Sudip Bose, M...	SPORTS	2017-09-08	Medics On Footbal...	https://www.huffi...	These professiona...

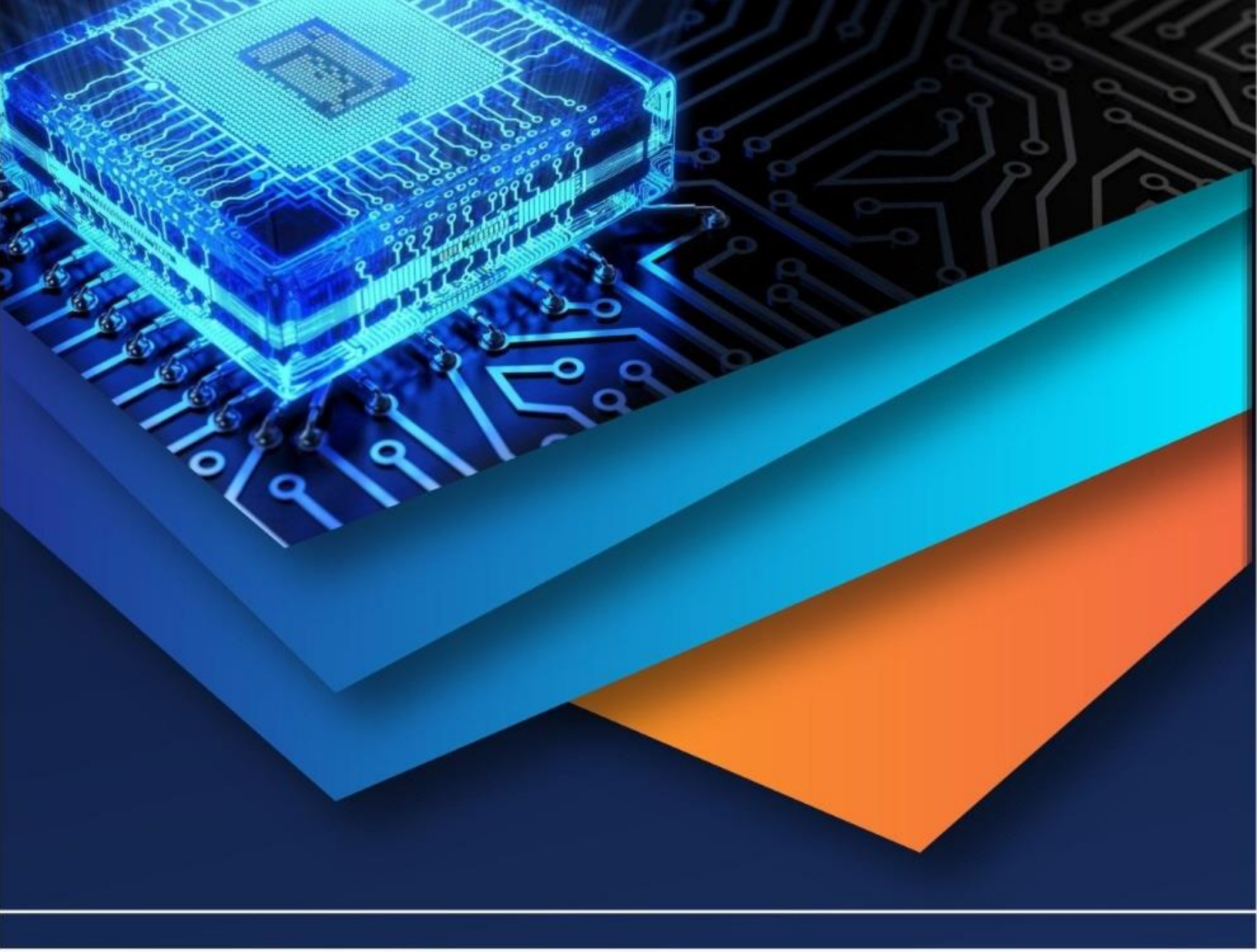
only showing top 20 rows

IV. CONCLUSION

The Naïve Bayes model has given an accuracy of 0.527288 and the test error is 0.472712.The accuracy is low since we need more data to be processed. Since the data is inadequate the accuracy of the model is low. The future implementation of this project is that all the scanned copies of old news headlines, converting them to string file and predicting the event and implementing for some more websites other than Inshorts website.

REFERENCE

[1] https://data-flair.training/blogs/nlp-natural-language-processing
[2] https://spark.apache.org/
[3] https://www.machinelearningplus.com/time-series/arima-model-time-series-forecasting-python/#:~:text=So%20what%20exactly%20is%20an,used%20to%20forecast%20future%20values.
[4] https://www.kdnuggets.com/datasets/index.html



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY