SentimentalAnalysisOfCovid-19VaccinationTweets

Dr.SarikaZaware[1],SanjanaBhosale[2],RajeshreeKalburgi[3],ShrutiModale[4],andPratikKadam[5]

[1]HOD-AISSMSInstituteofInformationTechnology
[2]Student-AISSMSInstituteofInformationTechnology
[3]Student-AISSMSInstituteofInformationTechnology
[4]Student-AISSMSInstituteofInformationTechnology
[5]Student-AISSMSInstituteofInformationTechnology

Abstract.

In March 2020, Coronavirus disease was officially announcedas a pandemic all over the world by the World Health Organization(WHO) Since then, the whole pharmaceutical world is in a state of warwith COVID-19 and has a responsibility to provide its vaccine for theentire world as soon as possible. The coronavirus outbreak has broughtunprecedented measures, which forced the authorities to make decisionsrelated to the installation of lockdown in the areas most hit by the pan-demic. Social media has been an important support for people whilepassing through this difficult period. Tweets collected, analyzed, and in-cluded in the media reports. Based on the analysis, it can be seen thatMost tweets are neutral, while the number of compatible tweets exceedsthenumberoftweetsagainsttweets.[5, 15, 17]Intermsofnews,itisconsidered that the occurrence of tweets follows the practice of events.Moreover, the proposed method can be used in a long-term monitoringcampaign that can help governments to establish appropriate commu-nication systems and to evaluate them in order to provide clear andadequate information to the general public, which can increase publicconfidenceinthevaccinecampaign.Thedatasetistrainedonamachinelearningmodeltoclassifytheopinionsof peopleonthevaccinationpro-
cess.ThealgorithmsusedareBERT,SVMandNaiveBayes.[1,7]

Keywords:BERT, Multinomial Naive bayes, SVM, Covid-19, Vaccina-tion,Sentimentalanalysis

DOI Number: 10.14704/nq.2022.20.11.NQ66336 **NeuroQuantology 2022; 20(11): 3274-3258**

1 INTRODUCTION

In this proposed system, we are implementing a sentiment analysis model ontwitter dataset using Machine Learning algorithms like BERT, SVM and NaiveBayes which will classify the tweets into positive, negative and neutral classesand compare them. It was found that BERT gives the highest accuracy[3, 4]. Theproposed approach can be used for a longer monitoring campaign that can helpthegovernmentstocreateappropriatemeansofcommunicationandtoevaluatetheminordertoprovideclearandade quateinformationtothegeneralpublic,

eISSN 1303-5150

which could increase the public trust in a vaccination campaign.[8] Performingsentimental analysis on vaccination tweets dataset.This study emphasizes onexaminingandevaluatingthekeytopicsandthemescorrespondingtoCOVID-19vaccinerelatedtweets,whichwerepostedbymanyindividualandcommunityprofiles,andexploringtheopinionslaidbythemonTwitter.[11,14]

3250

2 LITERATURESURVEY

Opinions Dynamics From Tweets in the Month Following the First Vaccine An-nouncement Opinion mining is a growing area of the Natural Language Pro-cessing field commonly used to determine viewpoints towards targets of interestusingcomputationalmethods.[1, 6]Welookatonesuchpopularmicroblogcalled Twitter and build models for classifying "tweets" into positive, negativeandneutralsentiment.Webuildmodelsfortwoclassificationtasks:abinarytask of classifying sentiment into positive and negative classes and a 3-way taskofclassifyingsentimentintopositive,negativeandneutralclasses.[2, 5, 9]ASurveyofTechniquesSentimentAnalysisisatermthatincludesmanytaskssuch as sentiment extraction, sentiment classification, subjectivity classification,summarization of opinions or opinion spam detection, among others. [4, 11] Vastusage of social media by people for expressing their opinions in all aspects of lifeproduces a lot of information on the Web. Analyzers and analysts are constantlydealingwiththis,howtheycantransformthismassiveinformationavailableon the web to useful information. [6, 9] The COVID-19 outbreak has broughtsignificant attention to the healthcare sector in recent times, and it has changedthe concept of safety in every aspect of our lives.[6] Sentiments Analysis is a taskwhich mainly focuses on textual data and we expect there to be a huge amountoftextdata.[5]

- DataGathering
- Pre-Processing
- FeatureExtracting
- FeatureSelection
- ClassifyingMethods

Humans are defined as social beings that make us relate to each other by emo-tion and thoughts. Indeed, the development of the child's knowledge is throughsocialinteraction.[3, 13]Socialinteractionincreasedinthe20thcenturytobea technological platform to interact with humans around the world to get theacceptanceofsociety.Differentapproachesarealsoconsideredwhileperformingand classifying the sentiments such as Deep Learning, Neural Network and alsohybridapproach.[19,20]

eISSN 1303-5150

3 SYSTEMARCHITECTURE

The first step is to collect a COVID-19 vaccination dataset.[1] A monthly combineddatasethasbeenmanuallyannotatedasPOSITIVE,NEUTRALorNEG-ATIVE vaccination, in order to be used in the training phase of the classificationalgorithms.[5]

3251

COVID-19 Vaccine Dataset: In this stage, A machine learning approach ischosentodetectthestanceoftweetsinthedatasetwhichisobtainedbylabelingthedatasettotrainthemodels.[3]

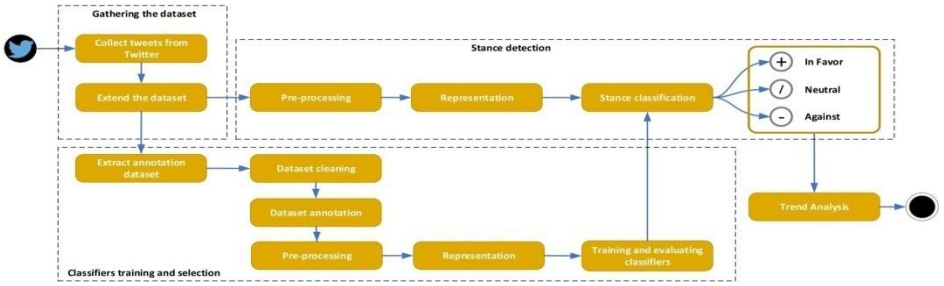


Fig.1. Systemarchitecture

3.1 DataCollection

– Severallargedatasetsincludingcollectionsoftweetsrelatedtothe COVIDhasbeentakenfortheanalysis.

– InordertocollectadatasetofCOVID-19vaccination,alargedatasetfromkagglewastaken.

3.2 DataAnnotation

To maintain the quality of the annotated dataset,which will be used for trainingthemachinelearningalgorithms,duplicatedtweetsareremovedandtheretweetshavebeeneasilyidentifiedduetothepresenceofthe“RT”symbolanddeleted.[2,4,5]

3.3 COVID-19VaccineSentimentDetection

Themaincomponentsofthestancedetectionprocessarethepre-processing,thefeatureextractionandthemachinelearningclassification.[3]

– Pre-processing:Asseenthesocialmediamessagesaremanyatimeswrit-ten using casual language, a preprocessing phase is used to filter out theunwantedtweetsfromthedatasetcontainingnonenglishlanguage.

eISSN 1303-5150

– Features: In order to use machine learning algorithms for text classification,thetextcontenthastobefirstconvertedintonumericalfeaturevectors.
– Learning Algorithms: Machine learning algorithms are used to predict thesentiments from the tweets. Some popular machine learning algorithms areusedlike:MultinomialNaiveBayes(MNB),SupportVectorMachine(SVM),BidirectionalEncoderRepresentationsfromTransformers(BERT). 3252

4 LEARNINGALGORITHMS

4.1 MNB(MultinomialNaiveBayes)

MultinomialNaiveBayesismostpopularsupervisedlearningclassificationthatis used for the analysis of categorical text data. We have selected the first algo-rithm as MNB because our model is predicting more than one class and MultinomialNaiveBayesisusedtosolveMultiClassclassificationproblems.[6]ItisalearningmethodthatismostlyusedinNaturalLanguageProcessing(NLP).ThealgorithmisbasedontheBayestheorem.Itcalculatestheprobabilityofeach tag for a given sample and then gives the tag with the highest probability asoutput. Mathematical Expression Bayes theorem, formulated by Thomas Bayes,calculates the probability of an event occurring based on the prior knowledge ofconditionsrelatedtoanevent.[4]Itisbasedonthefollowingformula:

$$P(A|B)=P(A)*P(B|A)/P(B) \quad (1)$$

WherewearecalculatingtheprobabilityofclassAwhenpredictorBisalreadyprovided.
P(B)=priorprobabilityofBP(A)=priorprobabilityofclassA
P($B|A$)=OccurrenceofpredictorBgivenclassAprobability.

Thisformulahelpsincalculatingtheprobabilityofthetagsinthetext.InMNB the dataset is divided into 80% for training purposes and 20% for testing.Multinomial Naive Bayes gives an accuracy of 61.62% which is very less so it isnotsuitableforregression.NaiveBayesalgorithmisonlysuitablefortextualdataclassificationandcannotbeusedtopredict numericvalues.Asthisalgorithmhasmanydisadvantages,weshiftedtowardstheSVMalgorithm.

4.2 SVM(SupportVectorMachine)

SVM is a supervised machine learning algorithm used for both classification andregressionchallenges.SVMperformsclassificationbyfindingthe hyperplanethatdifferentiatetheclasseweplottedinn-dimensionalspace.[10]Itdrawsthe hyperplane with the help of kernels. They are less prone to overfitting andcan perform multiclass classification. Mathematical Explanation The aim of theSVMistofindthehyperplanewhichmaximizesthemargin.Herethequestion

eISSN 1303-5150

ariseshowtofindtheoptimalhyperplane.

Theequationofanhyperplaneis:

$$wTx=0 \quad (2)$$

Let'slookatthestepstofindtheequationofthehyperplaneStep1:Youhaveadatas etDandyouwanttoclassifyit

$$D=\{(x_i,y_i)|x_i\epsilon R^p,y_i\epsilon\{+1,-1\}\}^{n}_{i=a} \quad (3)$$

Step2:Youneedtoselecttwohyperplanesseparatingthedatawithnopointsbetweenthem

$$y_i(w.x_i+b)\geq 1 forall 1\leq i\leq n \quad (4)$$

Step3:Maximizethedistancebetweenthetwohyperplanes

$$m=2\div||w|| \quad (5)$$

Step4:Finalequationofhyperplane

$$w^Tx=0 \quad (6)$$

In SVM the dataset is divided into 80% for training purposes and 20% fortesting.SupportVectorMachinegivesanaccuracyof83.49%whichislittleless.It is not suitable for large datasets. It does not execute very well when the datahas more noise. Due to these disadvantages, we are shifted towards the BERTalgorithm.

4.3 BERT

BERT(BidirectionalEncoderRepresentationsfromTransformers)isatransformer-basedmachinelearningtechniqueforNaturalLanguageProcessing(NLP)pre-trainingdevelopedbyGoogle.Historically,languagemodelscouldonlyreadtextinputsequentially(eitherleft-to-rightorright-to-left)butcouldn'tdobothatthesametime.[1, 12]BERTisdifferent becauseitisdesignedtoreadinbothdirectionsatonce.[7, 16]UsingthisbidirectionalcapabilityofBERT,itis pre-trainedontwodifferent,butrelated,NaturalLanguageProcessingtasks.ThespecialtokensusedbyBERTforfine-tuningandspecifictasktraining.Thesearethefollowing:

– CLS:Thefirsttokenofeverysequence.Aclassificationtokenwhichisnormallyusedinconjunctionwith asoftmaxlayerforclassificationtasks.Foranythingelse,itcanbesafelyignored.
– SEP:Asequencedelimitertokenwhichwasusedatpre-trainingforsequence-pair tasks (i.e. Next sentence prediction). Must be used when sequence pairtasksarerequired.
– MASK:Tokenusedformaskedwords.Onlyusedforpre-training.

eISSN 1303-5150

In BERT the dataset is divided into 80% for training purposes and 20%(halffor testing and other half for evaluation). After applying the BERT algorithmon the vaccination tweets, we can observe that BERT gives the highest accuracyrateamongallthealgorithmsasithastheabilitytoprocesslargeramountsoftext and language. It gives an accuracy of 91.24%. While calculating the accuracyofthemodel,accuracyformaskedwordsisconsidered.

3254

5 RESULTS

By applying MNB (Multinomial Naive Bayes), SVM (Support Vector Machine)and BERT (Bidirectional Encoder Representations from Transformers) on tweetsregardingCOVID-19vaccinationwecanobservethatMultinomialNaiveBayesgives61.62%,SupportVectorMachine83.49%andBERTgives91.24%.

TheresultsachievedbytheBERTarebetterthantheonesobtainedinthecaseof the other two algorithms in terms of classification report (i.e accuracy, recall,precisionandF1-score.)

Table1. Result

ier	acy	on		re
	% % %			

[5] df_train.head()

	date	tweet	replies_count	retweets_count	likes_count	hashtags
0	07-05-2021	@im_ankit_14 @DC_Hazaribag The whole vaccinati...	0	0	2	[]
1	10-05-2021	@cmohry @mlkhattar Sir Mera naam Rahul Gupta ...	0	0	0	[]
2	07-05-2021	@NewsNationTV Good governance, follow covid gu...	0	0	0	[]
3	07-05-2021	@WeAreVasai @VasaiVirarMcorp vaccination cent...	1	0	1	[]
4	07-05-2021	@htTweets @narendramodi @drharshvardhan Vaccin...	0	0	1	[]

Fig.2. InitialData

eISSN 1303-5150

```
[12] df_train.head()
```

	date	tweet	replies_count	retweets_count	likes_count	hashtags	clean_tweet
0	07-05-2021	@im_ankit_14 @DC_Hazaribag The whole vaccinati...	0	0	2	[]	vaccination drive given 3rd Year MBBS students...
1	10-05-2021	@cmohry @mlkhattar Sir Mera naam Rahul Gupta ...	0	0	0	[]	Sir Mera naam Rahul Gupta hai contact 78279736...
2	07-05-2021	@NewsNationTV Good governance, follow covid gu...	0	0	0	[]	Good governance follow covid guidelines vaccin...
3	07-05-2021	@WeAreVasai @VasaiVirarMcorp vaccination cent...	1	0	1	[]	vaccination center social distance super sprea...
4	07-05-2021	@htTweets @narendramodi @drharshvardhan Vaccin...	0	0	1	[]	Vaccines controlled PM fight corona shouldnt f...

Fig.3. Pre-processedData

	polarity	subjectivity	sentiment
0	0.000000	0.500000	1
1	0.000000	0.000000	0
2	0.550000	0.500000	1
3	0.088889	0.277778	1
4	0.133333	0.350000	1
...	...	...	...
6274	0.000000	0.000000	0
6275	0.000000	0.000000	0
6276	0.000000	0.000000	0
6277	0.266667	0.812963	1
6278	0.000000	0.625000	1

6279 rows × 3 columns

Fig.4. SentimentsAdded

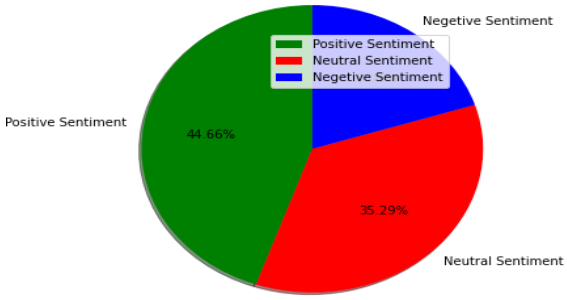
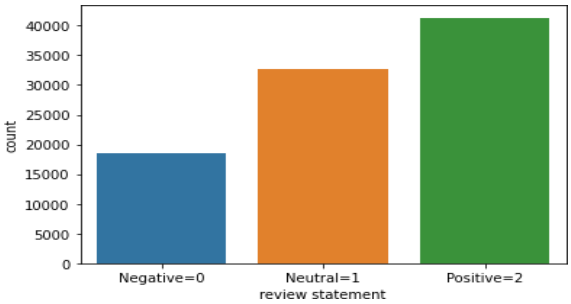


Fig.5. PieChart

eISSN 1303-5150



3256

Fig.6. BarGraph

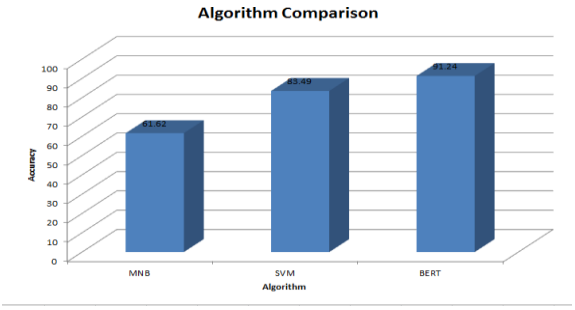


Fig.7. ComparisonGraph

6 CONCLUSION

Nowadays social media sites have become a strong tool for influencing peopleanddisseminatinginformationtothegeneralpublic.However,duetothelackofcontextualinformationinthetexts,sentimentanalysisforbrieftextssuchas Twitter is particularly difficult. As a result, numerous algorithms are alwaysbeingcreatedtoobtainthemostaccuratealgorithmofall.Inthecurrentpaper,a five-month period from January to May 2021 tweets has been analyzed usingmachine learning.The dataset used in this model is from kaggle. In our model wehave compared three Algorithms MNB,SVM and BERT.Based on the results it isconcluded that the majority of tweets i.e 44.66% were positive,20.05% Negativeand35.29%NeutralTweets.Multinomialnaivebayesgivesaccuracyof61.62% ,SVMgives83.49%andBertgives91.24%.SoitconcludesthatBertismostsuitable algorithm for performing sentiment analysis on large dataset.We haveclassifiedthetweetsintothreemainclassespositive,negativeandneutral.Countof sentiment and their percentage is calculated and final graph is displayed. Thisstudycanbeusefulforallgovernmentstopredictthesentimentofthepublicin order to make an effective strategy for better managing the situation. Futuretwitter data can also be considered for experimentation. The impact of COVID-19onthefinancialsector,employability,andpersonallifeoftheindividuals

eISSN 1303-5150

may be analyzed and prediction will be performed using this machine learningtechniques.

Acknowledgment

3257

It gives us immense pleasure in presenting the project report on “SENTIMEN-TALANALYSISOFCOVID-19VACCINATIONTWEETSUSINGMACHINE
LEARNING”. The success and the outcome of this report required a lot of guid-ance.WeareverygratefultoourguideDr.S.N.Zawarewhohasprovidedexpertise and encouragement. We thank sir who provided vision and knowledgethat was very helpful throughout the research. All that we have done is only duetothegreatguidance.WealsoexpressourgratitudetoDr.S.N.ZawareHeadof Computer Engineering Department, AISSMS’s Institute of Information Tech-nology, for the valuable support. We would also like to extend our sincere thankstoPrincipalDr.P.B.Mane,forhisdynamicandvaluableguidancethroughoutthe project and providing the necessary facilities that helped us to complete ourdissertationwork.

References

1. LIVIU-ADRIAN COTFAS, CAMELIA DELCEA, IOAN ROXIN, CO-RINAIOANAS , DANA SIMONA GHERAI AND FEDERICO TAJARIOL“TheLongestMonth:AnalyzingCOVID-19VaccinationOpinionsDynamicsFromTweets in the Month Following the First Vaccine Announcement” February 2021IEEEAccessPP(99):1-1doi:10.1109/ACCESS.2021.3059821
2. KumarRahul,BhanuRajJindal,KulvinderSinghandPriyankaMeel“Analyzing PublicSentimentsRegardingCOVID-19VaccineonTwitter”20217thIn-ternationalConferenceonAdvancedComputingandCommunicationSystems(ICACCS),Coimbatore,India,doi:10.1109/ICACCS51430.2021.9441693
3. Zainab Tariq Soomro, Sardar Haider Waseem Ilyas and Ussama Yaqub“Sentiment,Countand Cases: Analysis of Twitter discus-sionsduringCOVID-19Pandemic”,Bournemouth,UnitedKing-dom,doi:10.1109/BESC51023.2020.9348291
4. MahaA.AlaneziandNabilM.Hewahi“TweetsSentimentAnalysisDuringCOVID-19Pandemic”2020InternationalConferenceonDataAnalytics forBusi-nessandIndustry:WayTowardsaSustainableEconomy(ICDABI),Sakheer,Bahrain,doi:10.1109/ICDABI51230.2020.9325679
5. LokeshMandloi and Ruchi Patel “Twitter Sentiments Analysis Using Machine Learning Methods” 2020 International Conference for Emerging Technology (IN-CET)Belgaum,India.Jun5-7,2020doi:10.1109/INCET49848.2020.9154183
6. SupriyaRaheja and Anjani Asthana “Sentimental Analysis of Twitter CommentsonCovid-19”, 202111th International Conference on CloudComputing,DataScienceandEngineering(Confluence2021),Noida,In-dia,doi:10.1109/Confluence51648.2021.9377048

7. ApoorvAgarwal, BoyiXie, Ilia Vovsha, Owen Rambowand Rebecca Passonneau“Sentiment Analysis of Twitter Data” Columbia University New York,NY10027USA
8. VishalA.KhardeandS.S.Sonawane“Sentiment Analysis of Twitter Data:ASurveyofTechniques”InternationalJournalofComputerApplications(0975–8887)Volume139–No.11,April2016[9]AmanKhakharia,VruddhiShahandPragyaGupta“SentimentanalysisofCOVID-19vaccinetweetsusingMachineLearning”
9. NawSafrin Sattar and Shaikh Arifuzzaman“COVID-19 Vaccination Awareness andAftermath: Public Sentiment Analysis on Twitter Data and Vaccinated PopulationPredictionintheUSA”
10. Nijhum Paul and Swapna S. Gokhale“Analysis and Classification of Vaccine DialogueintheCoronavirusEra”2020IEEEInternationalConferenceonBigData(BigData)doi:10.1109/BigData50022.2020.9377888
11. Aman Khakharia, Vruddhi Shah and Pragya Gupta“Sentiment analysis of COVID-19vaccinetweetsusingMachineLearning”
12. A. H. Alamoodi et al., “Sentiment analysis and its applications in fighting COVID-19 and infectious diseases: A systematic review,” Expert Systems with Applications,p.114155,Oct.2020,doi:10.1016/j.eswa.2020.114155.
13. N.ÖztürkandS.Ayvaz,“SentimentanalysisonTwitter:Atextminingapproach to the Syrian refugee crisis,” Telematics and Informatics, vol. 35, no. 1, pp. 136–147,Apr.2018,doi:10.1016/j.tele.2017.10.006.
14. G.A.Ruz,P.A.Henríquez,andA.Mascareño,“SentimentanalysisofTwit- ter data during critical events through Bayesian networks classifiers,” Fu-tureGenerationComputerSystems,vol.106,pp.92–104,May2020,doi:10.1016/j.future.2020.01.005.
15. P.Tiwarietal.,“SentimentAnalysisforAirlinesServicesBasedonTwitterDataset,”inSocialNetworkAnalytics,N.Dey,S.Borah,R.Babo, and A. S.Ashour,Eds.AcademicPress,2019,pp.149–162.
16. Himanshu Batra,Narinder Singh Punn,Sanjay Kumar Sonbhadra and Sonali Agar-wal“BERT-BasedSentimentAnalysis:ASoftwareEngineeringPerspective”arXiv:2106.02581v3,2July2021
17. Muhammad Abbas, Kamran Ali,Saleem Memon and Abdul Jamali “MultinominalNaiveBayesClassificationModelforSentimentAnalysis”IJCSNSInternationalJournalOfComputerScienceandNetworkSecurity,VOL.19No.3,March2019
18. Kai-Xu Han , Wei Chien , Chien-Ching Chiu and Yu-Ting Cheng “Application ofSupport Vector Machine (SVM) in the Sentiment Analysis of Twitter DataSet”Appl.Sci.2020,10,1125;doi:10.3390/app10031125,7February2020
19. JatlaSrikanth,AvulaDamodaram,YuvarajaTeekaraman,RamyaKuppusamyandAmruth Ramesh Thelkar “Sentiment Analysis on COVID-19 Twitter Data StreamsUsingDeepBeliefNeuralNetworks”HindawiComputationalIntelligenceandNeuroscienceVolume2022,ArticleID8898100,6May2022
20. ChetanpalSingh ,Tasadduq Imam , Santoso Wibowo and SrimannarayanaGrandhi “A Deep Learning Approach for Sentiment Analysis of COVID-19 Re-views ” Appl. Sci. 2022, 12, 3709. https://doi.org/10.3390/app12083709 ,7 April2022

eISSN 1303-5150