ELSEVIER Scopus doi

Journal of Vibration Engineering

ISSN:1004-4523

Registered

- SCOPUS
- GOOGLE SCHOLAR
- DIGITAL OBJECT IDENTIFIER (DOI)
- IMPACT FACTOR 6.1

Our Website
www.jove.science

Usingtextmining,analysetextmessagesandonl inehealth records

Dr.Savyasachi
[1]AssistantProfessor,DepartmentofInformationTechnology,L.N.MishraCollegeofBusinessManagement,M uzaffarpur,Bihar

Abstract-ExaminingOnlineHealthDataUsingTextMiningMethods.Onlineexchangeofhealthinformation is a different approach. To disseminate information on health-related topics, social networksiteslikeTwitter,Facebook,Reddit,andonlinesupportgroupsforspecificconditionsareincreasinglyuse d.Inordertoreceivecounsel,thismayneedgivingpersonalhealthinformation,oritmayentailrespondingtoquestio ns from other patients based on their medical histories. This increase in social media usage offers afresh perspective on how to use user-generated content from social networks to enhance the presentlandscape of health communication. These internet channels of contact allow health professionals to assistthose looking for guidance more quickly. Non-profit organisations and federal agencies can also dispersepreventativeinformationinsuchnetworksforimprovedoutcomes.Researchersstudyinghealthcommun ication might extract information into patient experiences that may be difficult to obtain throughconventional surveys by analysing user-generated content on social networks. Getting the signal from thenoise is the main challenge in mining social health data. The informal nature of social data, typos,emoticons, tonal changes (such as sarcasm), and ambiguities resulting from polysemous words make itchallengingtodevelopautomatedmethodsforextractinginsightsfromsuchsources.

Keywords—*Analysis,TextMessages,OnlineHealthRecords,TextMining,HealthCommunication*.

INTRODUCTION

Throughout the past ten years, social media platforms have expanded quickly, resulting in numerousimportant changes to people's daily lives. Online knowledge sharing and consumption have becomewidespreadthankstosocialmediasiteslikeFacebook,Twitter,andInstagram,aswellasdiscussionboards likeReddit,Quora,andStackOverflow.76%ofAmericansutiliseasocialnetworkingserviceonline.71%of online adults use Facebook, compared to 18% who use Twitter and 3% who use Instagram. One in fourUSteenagersand39%ofinternetBlackAmericankidsuseTwitter.Withmorethan100milliondailyactiveuser s creating more than 140 million tweets since its start in 2006, Twitter has become one of the top tenwebsitesontheInternet(Alexa,2016).withmorethan100milliondailyactiveuserscreatingmorethan500milli on tweets everyday (Twitter, Inc, 2013). Twitter users, often known as tweeters, can openly expresstheiropinionsthrough140-character-limittweets.Byreplyingorutilisingahashtagtojointheconversation,tweeters can also participate in real-time dialogues and discussions with other tweeters. Twitter is anasymmetric network; a user can only see the feeds of the users they follow, and they won't see the feeds oftheirfollowersuntiltheyfollowthemback.Redditisaforumontheinternetthatwasfoundedin2005andfocuses on issues in news, music, sports, gaming, health, and entertainment. Reddit users, also known asRedditors, can publish text or Web links and receive feedback in the form of comments, upvotes, anddownvotes. Reddit utilises an algorithm based on the Newton cooling method and up/down votes to rankposts. Reddit, as opposed to Twitter, allows for lengthier, more in-depth posts that can deliver moreinformation in a better structured fashion. This essay will concentrate on a few subreddits (also known assub-forums or topic-related discussion groups). Reddit manages the themes using subreddits while Twittertracks them using hashtags. As the content of social networks gets more in depth, it has changed the wayresearchers conduct their research and patients seek health information. Everyday social network userscreatemillionsofposts.Comparedwithtelephone/volunteerbasedsurveymethodstocollectdata,social network based data collection is an impossible mission for the human mind to analyze the posts manually.Identifying the different aspects of the themes that the posts describe, and evaluating the opinion of eachaspect of the posts is computationally challenging. Researchers are developing automated methods toanalyze these large, unstructured datasets. In this context machine learning, natural language processing(NLP),andstatisticalanalysespresentthepossibilityofutilizingthismassivedataforderivinginsightsfromsocialstreams.

ONLINEHEALTHRECORDS

As information is simply facts about anything or someone, health information is simply facts aboutsomethingorsomeone'shealth.Itincludestheclinicalcontextofthepatient,includingtheirmedicalhistory,diagnosis, allergies, current therapies, drug side effects, and lifestyle factors that may affect their health(e.g.,exercise,smoking,drinking).Personalhealthrecords,suchaselectronicmedicalrecords,aretypically storedsecurelyandareonlyaccessibletoindividualsandhealthcareprofessionals.Patientsnowhaveanewavenue for information-seeking thanks to the increased popularity of social media platforms and onlinehealth forums. By sharing their personal story, and viewing others' posts, patients can have a deeperunderstanding of what they may face in the future; how to find a good health provider; and where to findsupport groups. As social media is not as regulated and is available for everyone to post their thoughts, wetypically don't know who the posters are in real life or to what extentaretheyknowledgeable.Thefollowingconcernsneedtobeconsidered:

- Thetrustworthinessoftheseposts
- Privacyissues
- Misleadinginformation
- Incompletedata

Thesehealth-relateddatacanbeusedtoextractavarietyofhealth-relatedinformation,including:

- Publicbeliefs,perceptions,andattitudestowardsproducts,regulations.
- Latesttrendsinsubstanceabuseandaddiction.
- Factorsassociatedwithmentalhealthconcerns,suchassuicidality,depression,andanxiety.

RELATEDWORK

Peopleareencouragedtopublishtheirideasandpersonalinformationonsocialmedia.Asinglepostmightnotprovideinsightfulhealthinformation,butmillionsofpostingsonthesamesubjectshowanumericshiftthat results in a qualitative shift. According to numerous research, compiling millions of posts can revealinformation about public health. Some significant public health surveillance examples include influenzadetection (Aramaki et al., 2011) and infectious disease outbreaks (Choi et al., 2016). In (Brownstein et al.,2009) it is estimated 37%-52% of Americans seek health-related information on the Internet. Usually,inaccurate or irrelevant information is also available to the public, and it is crucial to identify whichinformationiscorrect,especiallywhenitcomestohealth-relatedinformation.Inaccurateinformationcouldpotentially have a negative impact on our well-being. Nowadays, people prefer to use the Internet as apriority option to seek advice regarding their illnesses, drug use, and self-treatment. Chung studied theaccuracyofonlineinformation regarding the safety of infants during sleep (Chung et al., 2012).TheAmericanAcademyofPediatricshaspublishedrecommendationsforreducingtheriskofsuddeninfant death syndrome (SIDS), suffocation, strangulation, entrapment, and other accidental sleep-related infantdeaths.However,theserecommendationsaregivenasguidelinesbyhealthprofessionalscontainingmedical jargon that cannot be easily understood by the general public without a related background in medicine.Therefore, people probably enter the keywords related to infant sleep safety into a search engine and mayfollowthesuggestionslistedinthesearchresults.Chuangetal.(Chungetal.,2012)analyzed1300websiteson infant safety sleep (13 keywords and first 100 websites for each). The overall proportion of accurateinformationis43.5%,inaccurateinformationis28.2%,and28.4%irrelevantinformation.Theyalsofoundthat different data sources have huge differences. Government websites (.gov or .state) and organizationalwebsites(.org)achievedthehighestlevelofaccuracy:80.9%and72.5%,respectively.Onthecontrary, blogsandpersonalweb-siteshadverylowaccuracyscore:25.7%and30.3%respectively.Anotherfindingwas that different keywords brought different outcomes from as high as 82% accuracy to as low as 18%accuracy. This study shows that the Internet does provide an opportunity for people to seek health-relatedinformation, and patients need skills to identify what information is most accurate. On the other hand, thestudy also shows the quality of keywords is important for health-related information. We can see on linehealth information still has a long way to go to improve the accuracy of health information. It requirescollaboration between health professionals, researchers, and Internet users. Besides static websites, socialnetworksitesalsocontributeamassiveamountofhealthinformation.ThestudybyChouetal.(Chouetal.,2009)showsthatin2007,about69%ofAmericanadultshadInternetaccess.AmongInternetusers,5% of them joined in an online support group. As health providers mainly focus on clinical outcomes,patients'mentalissuesusuallyaren'tgivenenoughattention.Throughafflictionofemotionalandphysicalpain,patientsmaydevelopdepressionandsuicidalthoughts.Thereareseveralstudiesonsuicideprevention(Kavuluru etal., 2016; Luxton et al., 2012), many identifying the most important posts during theconversion which led to a sentiment shift. By analyzing millions of posts, we may findsome patterns which can help health providers and patients' families help patients through difficult times.Monitoringsentimentchangeduringaconversationandmanualanalysisistime-consumingandunrealistic.WiththepowerofNLPandmachinelearning,sentimentanalysis(alsoknownasopinionmining)isusedtoclassify the polarity of a given context automatically. More details on sentiment analysis will be discussedin section 2.1. By classifying the sentiment of posts during online communication, we canusetopicmodeling/statisticalanalysistosummarizeandcategorizewhattypesofinformationpatientswillneed for support, what types of support are most helpful (sharing personal stories, general support,information support, etc.), and how to attract patients and keep them in the conversation.The benefits of online health support forums such as Cancer Survivors Network (CSN), Lungevity, andPatients Like Me are immense. Although the numbers of users are far below the numbers of more generalsocial networks such as Facebook and Twitter, online health support forums offer patients the chance tointeract with others who have been diagnosed with the same diseases such as lung or breast cancer.

Asonlinehealthsupportforummembersarepatients,healthproviders,patients'familiesandfriends,thepostsmadebytheseusersaremoreaccuratethanblogsandpersonalwebsites;whensomepostsarerecognizedasinaccurate, other users will quickly move to correct them. Online health-related social media offers anabundance of information for patients, health providers, and researchers. Wicks et al. (Wicks et al., 2010)show that over 70% of Patients Like Me users think the site is "moderately" or "very helpful"; over 50%ofpatientsfoundthesitehelpfulforunderstandingthesideeffectsoftheirtreatments;and42%of patientsagreed that site had helped them to find another patient who can help them understand a specific treatmentfor their symptoms. This shows an opportunity that online health information and communication canprovideacriticalmassofusefulinformationfordifferentparties.

TEXTMINING

Manyscientificfields,includingstatistics,computerscience,linguistics,andlibraryscience,havecontributed to the development of text mining. Text mining approaches deal with unstructured text andconcentrate on automated analysis of textual data as a kind of natural language. Despite the lack of auniversaldefinitionfortextmining,thegeneralmethodofanalysisisaccepted.TextminingisalsolinkedtoNatural Language Processing (NLP), which is concerned with the study of natural languages. Softwareoptions are accessible for analysing social media applications due to the requirement of using automatictechniques for textual data analysis and extracting pertinent information. Text mining tools are used toidentify and analyze posts, likes, followers in online social networks to explore people's reactions andbehavior.

Moreover,itshowsthevariationinviewsandopinionsregarddifferenttopics.Thefundamentalprocessoftextminingincludesdatacollection,preprocessing,contentanalysis,findingandintegration.

METHODFORTEXTCLASSIFICATION

(i) TraditionalMachineLearning-TraditionalMLmethodsincludenaiveBayes,decisiontrees,k-nearestneighbors, logistic regression, and support vector machines(SVM). These algorithms are based on featureengineeringwhereasetofdiscriminativehandcraftedfeaturesisconstructedtoimprovemodelperformance. Although deep learning has increased in popularity recently, in most Kaggle competitionscurrently,competitorsarestillwinningbytraditionalmethods.Especiallywhendataisstructured,ahumancanfindgoodfeaturerepresentationstotrainMLmodels.Onthecontrary,deeplearningisadeptatfindingfeatures from unstructured data such as images, audio, video. Such models can extract the features thathumanscannoteasilyunderstand,butarestillmeaningfultoamachine.Foradatasetwithafewhundredtoa couple of thousand training samples, traditional machine learning usually outperforms deep learningmethods.Sinceadeeplearningstructureismorecomplex than a traditional method, it haslarger parameter spaces it needs to search through and learn. A small dataset cannot fully tune parametersthat are generally representative for a domain leading to poor generalization. Usually, the challenge intraditionalmachinelearningisidentifyinganappropriatemodel andfeatureswhileindeep learningitistosearch forappropriatearchitectures.

(ii) Deep Neural Networks- In the sixties, a single layer neural network was introduced and was called aperceptron. The structure was one input layer, one hidden layer, and one output layer.Duringthattime,themachinesweresimpleandcouldnothandlecomplexoperations,untilthe1980s,whenthemultilayerperceptron(MLP)wascreatedbyRumelhart,Williams,andHinton.Inthefigure.wecanseethatallthelayersarefullyconnected,whereeachnodeinonelayerisconnectedtoallnodesinneighboringlayers. This brings an issue when the network gets larger: the number of parameters increases drastically.Forinstance,with1000cellsinthehiddenlayerconnected with1000×1000elementsintheinputmatrix,we need 109 weight parameters and 1000 bias parameters for a single hidden layer. Additional layers willfurtherincreasetheparametersetsize.Thislimitedthesizeofeachlayerandthedepthofthenetworkduetocomputingconstraints.Inaddition,thismayalsoleadtooverfittingandthenetworkgettingstuckinlocalmaxima.

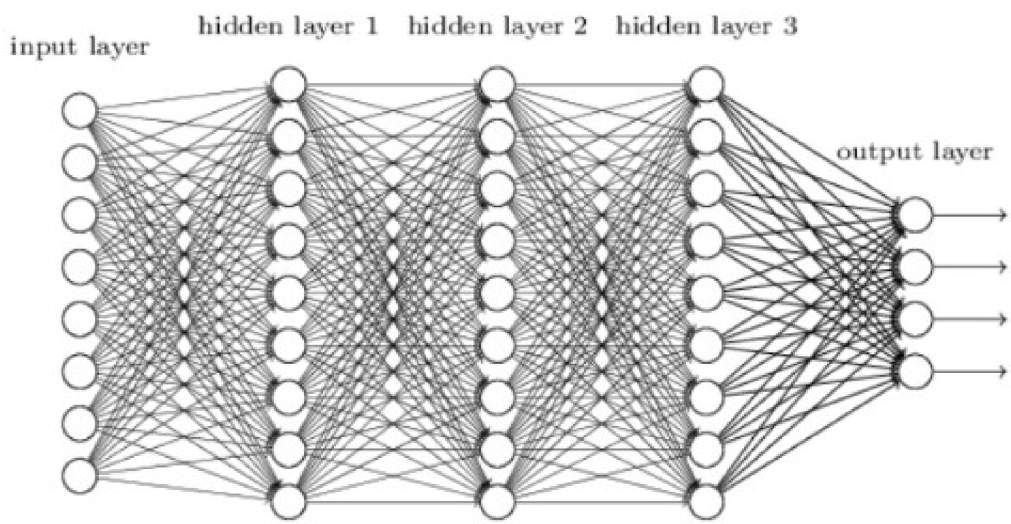

Figure1-Afeedforwarddeepneuralnetwork

(iii) ConvolutionNeuralNetwork-
AnexampleCNNisasshowninFigure2forthetaskoftextclassification.Thedifferenceinthisnetworkisthatconvolutioncellsonlyconnectwithapartofthe

input cells. It extracts local information from the connected cells in the previous layer.For instance, in the image classification task, the first hidden layer can extract somecurves by looking at a part of the pixel matrix, the second hidden layer might knowthe combination of curves and recognize them as part of the information of an object,and the third layer might know what the object could be and give a probability foreachpossibleclass.Thepurpose of anactivationfunctionismakingoutputresultfrom the convo-lutional layers compressed to a fixed real range so that the range of values that areinput to the next layer is controllable; it also introduces non-linearity to extend net-work abilities to capture more complex functions. The common activation functionsincludesigmoid,ReLU,LeakyReLU,Maxout,tanh.

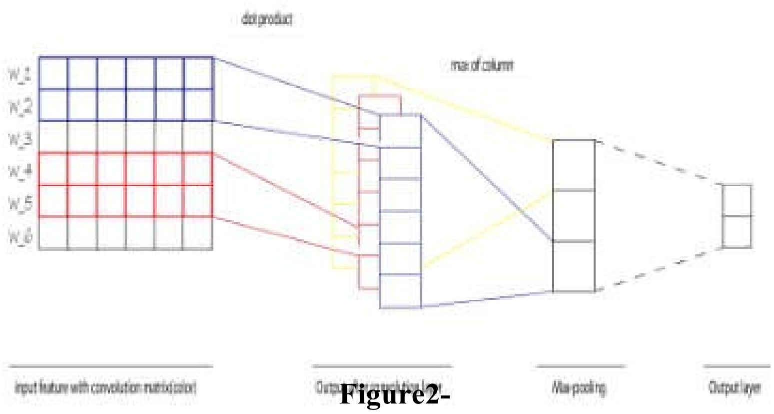

Figure2-

TheCNNmodelwithabinaryoutputlayerfortextclassificationTEXTMININ

GALGORITHMS

The most algorithms used in analyzing the text in social networking are classification and clustering.Classification is a supervised learning that learn from training process a set of rules. The classificationmethodcomprisesquantitativeapproachestoautomateNLPtoclassifyeachtexttoacertaincategory .

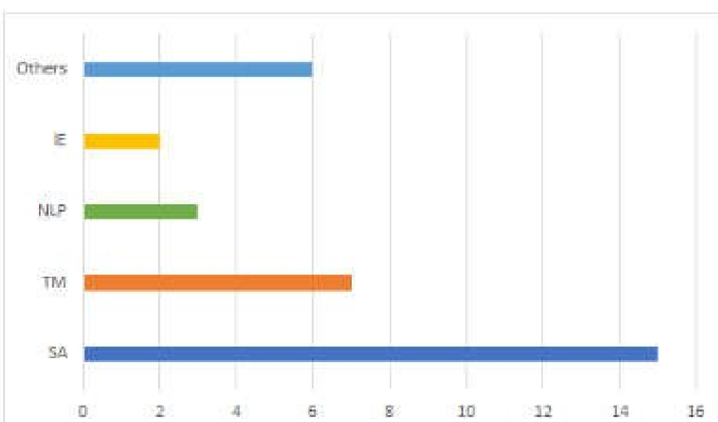

Figure3-Mostappliedtextminingtechniquesinonlinenetworking

The most common algorithms are K-Nearest Neighbour (KNN), Decision Trees (DT), Support VectorMachine(SVN),andArtificialneuralnetworks(ANN).Clusteringisanunsupervisedalgorithmthat groupedthetextinclusters.Differentclusteringtechniquesincludedifferentstrategiesthatcanbecategorized in three types, naming, partitional, hierarchical, and semantic-based clustering.The studies inthis paper cover a variety of text analysis algorithms. Most of the articles focus on classification orclustering. We noted that the number of articles in the dataset that employed clustering and classificationalgorithms increased recently. Clustering and classification are the most data mining techniques thatextensivelystudiedinthecontextoftext.

CONCLUSION

Text mining applications have altogether influenced the inquire about in social systems investigation.Through investigating theinquire about in online social systems, 32 inquire about thinks about wereanalyzed to supply significant bits of knowledge on the approachesapplied to progress decision-makingcompletely different regions utilizing social media platforms. The survey uncovers finding thatreplytheinquireaboutquestions,theforemostcommontextminingstrategieswereassumptionexaminationand subject modeling. Clustering and classification are the foremost information calculations that broadlyconsidered.Asinformationpreprocessingisbasicstepincontentminingthatmightinfluencetheprecisionoftheinvestigation,theanalystsaresuggestedtodepictthesestepsindetail.Thedifferentnatureofcontentinformationinsocialmediaposturesnumerouschallenges,countingthegiganticvolumeoftheinformation,commotionand etymologyissues.

REFERENCES

[1] L.Sorensen,"Usermanagedtrustinsocialnetworking-ComparingFacebook,MySpaceandLinkedin,"20091stInternationalConferenceonWirelessCommunication, VehicularTechnology,InformationTheoryandAerospace&ElectronicSystemsTechnology,2009,pp.427-431.

[2] M.Naaman,“SocialMultimedia:HighlightingOpportunitiesforSearchandMiningofMultimediaDatainSocialMediaApplications,”MultimediaToolsAppl,vol.56,no.1,pp.9-34,2012.

[3] S.RaniandP.Kumar,“Asentimentanalysissystemforsocialmediausingmachinelearningtechniques:Social enablement,”DigitalScholarshipintheHumanities,vol.34,no.4,2018.

[4] D. Sudarsa, S. K. Pathuri, and L. J. Rao, “Sentiment Analysis for Social Networks Using MachineLearning Techniques,” International Journal of Engineering and Technology(UAE), vol. 7, no. 2, pp. 473-476,2018.

[5] K. Jani, M. Chaudhuri, H. Patel, and M. Shah, “Machine learning in films: an approach towardsautomationinfilmcensoring,”JournalofData,InformationandManagement,vol.2,pp.55-64,2019.

[6] N. Naw, “Twitter sentiment analysis using support vector machine and K-NN classifiers,” Int. J. Sci.Res.Publ,vol.8,no.10,pp.407–411,2018.

[7] M. Grčar, D. Cherepnalkoski, I. Mozetič, and N. K. Petra, “Stance and influence of Twitter usersregardingtheBrexitreferendum,”ComputationalSocialNetworks,vol.4,no.1,2017.

[8] G. Piatetsky-Shapiro, “Data mining and knowledge discovery 1996 to 2005: overcoming the hype andmovingfrom“university”to“business”and“analytics”.”DataMinKnowlDisc,vol.15,no.1,pp.99–105,2007..

[9] J Cao, K. Basoglu, H. Sheng, and P. Lowry, “A Systematic Review of Social Networks Research inInformation Systems: Building a Foundation for Exciting Future Research,” Communications of theAssociationforInformationSystems,vol36,pp.227-758,2015.

[10] H. B. Haq, H. Kayani, S. K. Toor, A. Mansoor, and A. Raheem, “The Impact Of Social Media: ASurvey,”InternationalJournalofScientific&TechnologyResearch,vol.9,pp.341-348,2021.

[11] D. M. Boyd and N. B. Ellison, “Social network sites: Definition, history, and scholarship,” Journal ofComputer-MediatedCommunication,vol.13,no.1,pp.210-230,2007.

[12] R.Kumar,P.Raghavan,S.Rajagopalan,andA.Tomkins,“TrawlingtheWebforemergingcybercommunities,”ComputerNetworks:TheInternationalJournalofComputerandTelecommunicationsNetworking,vol.31,no.11-16,pp.1481-1493,1999.