Classification Connection of Twitter Data using K-Means Clustering

Rashmi H Patil, Siddu P Algur

Abstract—The rise of social media platforms like Twitter and the increasing adoption by people in order to stay connected provide a large source of data to perform analysis based on the various trends, events and even various personalities. Such analysis also provides insight into a person's likes and inclinations in real time independent of the data size. Several techniques have been created to retrieve such data however the most efficient technique is clustering. This paper provides an overview of the algorithms of the various clustering methods as well as looking at their efficiency in determining trending information. The clustered data may be further classified by topics for real time analysis on a large dynamic data set. In this paper, data classification is performed and analyzed for flaws followed by another classification on the same data set.

Keywords— Classification of Twitter data, K-Means Clustering, Euclidian Distance, TF/IDF, Social Media.

INTRODUCTION

A large number of dataset sources may be found which may not be readily classified or labeled such as newspapers, magazines, blogs and various social media platforms. Such sources may provide raw unformatted texts with large amount of data where the time to obtain the data may range from seconds to years which may cause it to be stored in millions of racks. Efficient storage of such data would require an asserted way to sort and arrange the data to determine the type of data. This is critical for efficient and accurate query and retrieval of the required data since even though retrieval time is important, the retrieval of the right data type is equally crucial to avoid unnecessary overheads and loss of efficiency seen even with the quickest retrieval.

Feature based clustering method such as K-Means clustering [1] provide quick and reliable classification of streaming data into groups by the features of the available data. The most important aspect of K-Means would be that it may be applied to raw and unformatted data into smaller groups called clusters. The number of clusters needed initially is determined by the requirements of the problem definition. The number of clusters is equivalent to the type of data corresponding to the subject of interest. The input data would be raw and stream from the user source which would mean that the unsupervised learning would require no prior processing of the data with metadata tags like labels. The basic preprocessing of removing whitespaces and irrelevant patterns may be performed. A centroid [2] is a point of concentration in a data group. Such centroids are randomly place in the clusters. The input sentences are represented as data points associated with a value calculated using vectorization [3] techniques of term frequency (TF), inverse document frequency(IDF) and counts to obtain the vector of the text which is a dot product representation of the scalar quantities like the value associated with the text and various parameter values, if any, considered during the classification process. Each vector may be defined as a set of features. The distance between the vectors and the selected centroid of each cluster are calculated using various techniques. One such technique is the Euclidean algorithm [4] and the distance is termed as Euclidean distance [5] which determines the closest centroid to the vectors assigned to the clusters. A vector with the shortest Euclidean distance to a centroid would be a part of the cluster with that centroid. Thus the Euclidean distance and centroids play a major role in cluster formation of the vector data points and classification of datasets.

Twitter is considered as one of the world's largest social media platforms which helps people keep up to date with the world's current affairs, events and different perspectives and opinions in the form of tweets. Information updates and propagation are made in real time through a series of "re-tweets"of the original tweet post by the original handle. Twitter provides functionalities to the users to follow as well as unfollow people or accounts of interest such as famous personalities, friends, media and organizations through a twitter handle. The adoption and usage of Twitter was greatly increased due to the release of highly portable and affordable smart-phones. People can easily propagate news or updates quickly from the point of origin through tweets and is even said that news channels first break the news on twitter through their handle followed by the television broadcast. One of the advantages of such social media platforms to data scientists is that they offer applications to retrieve twitter data through different custom approaches based on keywords using Open Authentication (OAuth) mechanisms. The tweets matching the keywords are retrieved seamlessly independent of size and time. The number of tweets retrieved is dependent on the frequency of the searched keyword in the form of tweets and re-tweets. In general, topics are tweeted as hash-tags beginning with the '#' symbol or uppercase letters for personalities. A personality with a twitter profile would addressed beginning with the '@' symbol. Such conventions make retrieving data relatively easy for data miners to find the abundantly available worldwide data on a single platform.

Various methods exist to extract the data, process it in the required format to obtain the information and then analyze the information for various goals such as marketing, social

Revised Manuscript Received on April 12, 2019.

Rashmi H Patil, Department of Computer Science,Rani Chennamma University, Belagavi. (E-mail: patilrashmi139@gmail.com)
Siddu P Algur, Department of Computer Science, Rani Chennamma University, Belagavi.(E-mail: siddu_p_algur@hotmail.com)

trends, sentiment analysis etc. The most effective retrieval technique is through clustering and classification of the data to determine trending topics. The number of sentences used to perform the classification is randomly determined and depends on the performance of the computing machine. A good computing machine is required to process millions of data texts. In the project however, 8, 22 and 50 sentences are used to perform the classification and analysis of the results while the number of sentences may be extended to hundredsof sentences depending on the computing power of the machine.

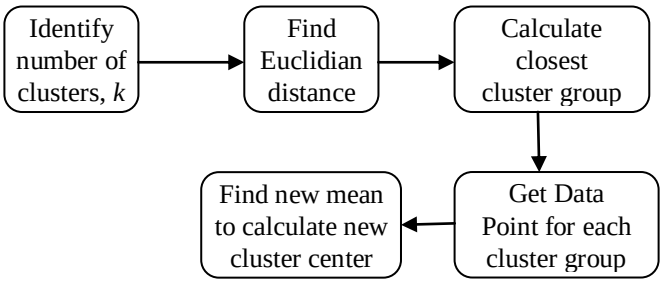

Figure 1: K-Means Clustering Flow Diagram

IMPLEMENTATION& RESULTS

The tweets retrieved would be unstructured initially due to all the metadata associated with it making it ineligible. It is therefore required to identify important words while removing unnecessary text from the tweet like the proper nouns starting with the '@' character followed by the twitter handle. Other unnecessary data are whitespaces, stop-words, digits and various emoticon texts. Spell checks are needed to correct any case spelling errors in the content. A good convention is to also convert all text to lower case letters to reduce the time for processing and reduce redundant checks. Normalization of text [6] is not mandatory during processing but is an internationally recognized and accepted approach before any data analysis. This also reduces the randomness coefficient related to the data and gradually reduces the error percentage with each iteration in the result.

After normalization, the keywords are separated into nodes where the nodes with similar meaning are grouped into clusters. This implies that words with different meanings are in different clusters. The relative distance of two words with different meanings would be higher than two words in the same cluster. The clustering would occur such that given any set of words from the different cluster groups, they would refer to different meanings or an entirely different context overall. The distance measured between the words within the same cluster is called the intra-cluster distance [7] whereas the distancemeasures between words from two different clusters is called the inter-cluster distance. The inter cluster distance is usually considered greater than the mean inter-cluster distance [8]. This property of clustering helps determine the overall efficiency of the process. There may be cases of improper assignment of a word into a wrong cluster. Over the various iterations, a word from each cluster would be nominated as the center such that the center would be surrounded by words representing similar meaning of the cluster. This helps reduce the error rate compared to normal keyword based search methods. The cluster homogeneity computation [9] helps to determine the proximity or the distance for interchangeability of the words which are part of a cluster. Higher values mean easier substitution of words while lower values would cause changes to the overall meaning. There are many techniques to calculate the distance between the nodes since it helps to determine the homogeneity computation of a cluster. Some of these methods are the Euclidean distance, Manhattan distance [10], Chebychev distance [11], Spearman distance [12], Euclidean square distance [13], Pearson square distance [14] and the Pearson correlation distance [15]. Among these the most frequently used method would be the Euclidean distance due to its ease of application to majority of the real world dataset objects as well as its relative simplicity. Clustering with inter-cluster distance may be done using methods such as single link, complete link and average link. Single link makes use of the distance between the closest adjoining data points throughout the cluster. Complete link uses the distance between two of the farthest data points in two clusters. Average link has the mean distance of the data points between two clusters.

From the various clustering methods available like K-Means clustering, Agglomerative Hierarchical clustering [16] and Jarvis-Patrick clustering [17] for clustering the data from twitter, a popular technique is the Naïve-Bayesian method [18] used for classification but this method provides less flexibility and extensibility for data points compared to clustering.

K-Means offers a numerical based grouping where the number of clusters is 'K'. This method requires at least one data point in each of the k clusters. The clusters do not converge since there is no data which belongs to more than one cluster at any given space and time. The advantage of K-Means is due to its use of numerical values which help in faster mathematical deductions making it comparatively fast and efficient. One drawback would be the ambiguity in deciding the number of 'K'clusters to initiate the process which in itself is hard to determine the features of the clusters.

Jarvis-Patrick clustering uses a natural algorithm without heavy numerical calculation like K-Means and makes use of the distance measure to determine the similarities between two data nodes. Like K-Means it also uses a deterministic approach to the number of clusters but it doesn't need to be provided with the number as input initially at the beginning.

Similar to Jarvis-Patrick clustering, the agglomerative hierarchical clustering uses a natural algorithm based on the gene merging process seen in nature. Each data node gets assigned to separate clusters followed by the cluster merging process when conducive data points are identified. The merging process results in the final number of clusters to be significantly less than the initial number. The main disadvantage would be that merged data points cannot be rearranged in case of wrong clustering once merged.

Test Case 1:

Consider the following eight sentences with k=3 for the word lied:

[1] You lied about the weather now everyone expecting a storm
[2] Should it be discovered that Ramaphosa lied under oath, will he resign?
[3] There actually isn't or it would be on the front door of every "Holocaust" Museum.
[4] You've been lied to. One of my confession
[5] I have money but I can't spend it because I already lied that I don't have money so idk what to do smh
[6] Yup. We can see how much we've LIED.
[7] i lied i hate every type of weather i can feel myself becoming a bitter old lady i can't wait to be cute
[8] @Nicole_Cliffe Yo I saw one video where the parents LIED to the kid and had the father call him to say Daddy couldnt…

Table 1: Clustering 8 sentences with k=3

Cluster #	Sentence #
1	3,4,5,6,7,8
2	1
3	2

Test Case 2:

Consider the following 22 sentences with k=4 for the word crane:

[1] "A major upgrade to one of the world's oldest deep underground railways, at Bank in London, relies on a single @ghcranes overhead crane nestled in a side street. Positioning the crane, and the logistics of its operation, posed a considerable challenge.",
[2] "Crane 2 operation class conducted by Mr.S.Majumder (15/11/18)",
[3] Whooping Crane #31-16 lingers on the midwest landscape with his Sandhill Crane entourage, but for how long? #migration #endangeredspecies #habitat #nativespecies @OperMigration @savingcranes",
[4] "OSHA Publishes Report on Cause of 3 Tower Crane Collapses During 2017's Hurricane Irma",
[5] "Specialized Crane Installed for 'The One" #Toronto",
[6] "Ar an daoine ag The Crane.",
[7] "Another new #Liebherr delivered#constructinhour! Berry Cranes has taken delivery of a 90 tonne Liebherr LTM 1090-4.2 All Terrain #crane http://ow.ly/Mhfg30mwIqWvia @vertikalnet",
[8] "How to make a paper Crane Dish | Fold an Origami Easy via @YouTube",
[9] "I added a video to a @YouTube playlist How to make a paper Crane Dish | Fold an Origami Easy",
[10] "@FullerFarmer I like my pie unadulterated, but my MA-raised father will go for the sharp cheddar every time.",
[11] "American model Carmen Dell'Orefice balances on a crane, in front of the Old Bailey in London, England, for the cover of Queen magazine, September, 1960. Photo by Norman Parkinson",
[12] "@Yellowyorkie @nevtownsend Ah the old game of you rubbish my info, and I rubbish yours? Can't be arsed with it mate. The point still stands though that Nick Clegg branded the idea of a European army 'a dangerous fantasy'. So he basically lied/talked utter bollocks",
[13] "Car shopping and my mom talking about "y'all can't fit no car seat and Crane in that backseat.She right but I was gonna make Crane ride in the front with Jared",
[14] "Visited the concrete plant and the area for the production of concrete products. they make a stock of products for the winter (according to the plan 5000-7000 pieces). These blocks are made yesterday, so they have not yet been transferred by crane.",
[15] "@aundrey_crane Exactly bro",
[16] "@NielsFonsboel @bo_elkjaer When shipping tanks from Vladivostok to Nicaragua they didn't use a RoRo b/c the port had an adequate crane to lift them out of a regular bulk carrier. That's what makes me wonder if they are dropping off heavy weapons in a port that doesn't have that kind of equipment.",
[17] "The crane is in for the new City Hall!Time to set some steel!",
[18] "A rooftop UNIC URW-506 spider crane uses a Hydraulica 2000 vacuum lifter to move large format glass into a building",
[19] "Looking to join Cameron Craig Group (@cameroncraig), Vanderbilt University Medical Center (@VUMCcareers), or Hoist & Crane Service Group (@HCSG)? View 3 openings from these companies and more.",
[20] "A-A crane's voice!? Th-That might be too hard to mimic... Sorry.",
[21] "@Darla_Crane I don't know how this little prick isn't in prison yet??!",
[22] "On the crane cloak"

Table 2: Clustering 22 sentences with k=4

Cluster #	Sentence #
1	22
2	6,
3	1,2,3,4,5,7,10,11,12,13,14 ,16,17,18,19,20,21
4	8,9,15,

Test Case 3:

Consider the following 50 sentences with k=5 for the word engage:

[1] "@sbmitche @sarahbauerle Seems ok for reviewers to ask to engage w/ directly related working papers & other unpublished work, provided it's available on line (and usually is). ",
[2] "US, Taiwan should engage in regular economic dialogue amid China threats: think tank",
[3] "When she refused the invite of BBC, I thought she shouldn't have done so (yes, while I as a person don't believe in any dialogue at all, as a brand I thought we can engage). She has been proven right. The invite was a sham. They had already pre-judged and decided the conclusions.",

[4] "Reluctant to bring #cybersecurity to your district or #school? We are learning how to get past that fear and engage in this amazing pathway here at #PLTWSummit #CSForAll #PLTW",
[5] "We often talk about how migrants in this country should assimilate and engage with our British way of life. I want to show you how proud I am of my German wife for doing her utmost to become a Briton. This is her response to seeing Nigel Farage appear on Sky News.",
[6] "And this last immediate previous tweet of mine on this thread path is just a scenario you can engage in to expose the loophole... I didn't say you said so",
[7] "This was a unique opportunity for the students to explore ways of using theatre and social media to reach and engage with different audiences, and to find creative ways of telling an important narrative and have their voice heard.",
[8] "Health and safety is about much more than red tape and bureaucracy, our UAE Networking Conference will hear. The event will explore how health and safety professionals can engage and motivate employees to help their business thrive ",
[9] "Michaels taps into experiential retail @MichaelsStores #innovation #education #community",
[10] "@Claire_M @yvonneseale @Wikimedia This doesn't disregard the very valid points you make, btw, I just think our responses need to be to engage and change things.",
[11] "You hand over the ransom to the West and engage your plasma jets would likely break if you land.",
[12] "@yvonneseale @Claire_M @Wikimedia Given all the fora in which I can write which engage with the public, why should I choose to invest my time and expertise in a project which is structurally hostile? You frame that as 'wash[ing] my hands' I think of it as picking my battles.",
[13] "@HskyToki Same happened to me when I went to an open house party in SoCal. Showed up, entered a house I've never been to, tried to engage people…owner of the house never said hello or left their pc gaming.",
[14] "I'm not wrong.Your agenda is to attack DSA and the Democratic party through means your own club wouldn't allow you to do.I find it very bad faith organizing and I won't engage with it any further.",
[15] "Help enhance the Penn State student-athlete experience on #GivingTuesday!",
[16] "Nice story by @navpersaud abt engaging patients in designing and conducting a trial of providing free medications to patients who can't afford them #SPORSRAP.Although it seems like the outcome would be obvious ability to engage policymakers and tell personal stories was key.",
[17] "Heading home for the holidays & need a ride to the airport? Check out the graphic below for info & sign into Engage to request your ride here",
[18] "Before cancer he was always so chipper and happy to engage other scientists, now it's a chore. 'The chemo has my brain half cooked' having some trouble with memory and fatigue wears",
[19] "Promotion of institutional integrity in partnership w/ @anticorruption is a key pillar of CRIMJUST: Glad to share our updates in West Africa where @GhanaIntegrity and @cislacnigeria actively engage and share findings of CRIMJUST report regarding Ghana police services",
[20] "NICE today unveiled the Journey Excellence Score (JES), a pioneering new metric designed to precisely measure the quality of customer experience across multiple channels.",
[21] "@dpurvis_dWo @mattkoonmusic But is this something that can't be "fixed" with a little bit of creativity and consistency of storytelling. Yes, times have changed, and there is so much more content, but is it impossible to engage the audience in 2018?",
[22] "@LTF_01 Capital C versus little c is general versus particular. Ethnographic work occurs within a little-c cultural context, that is nearly always true. Some cross-cultural work does engage with translation, and it may be unfair to say that isn't hypothesis-generating.",
[23] "Looking for a way to engage with more students who share similar acadmic interests? Visit http://myuniversityconnect.com@ Arizona State University ",
[24] "I'm accepting applications for a friend with whom I can crack irreverent Bible jokes and occasionally engage in deep philosophical discussions such as what was God's rationale in being born blue-eyed and blond in the Middle East. Gender/Religion/Age no bar. Must be irreverent AF",
[25] "@vicenewsThis is such an ignorant tweet. They are there to take over duties that CBP/ICE would normally do to prepare for this many people. CBP and ICE are the ONLY ones who can engage with illegal immigrants so it makes sense to bring in the Military to do all the boring day to day",
[26] "Their self-image is so grounded in their positive power for good that they seem unable to engage in constructive discussion about the social problems they are creating.",
[27] "Command over 70+ units at your disposal and engage in a heart-to-heart battle with your foes and friends",
[28] "@GarbodorianGray @BadAstronomerWe *could* engage with the concerns of the left wing of the party, but it's easier to just dismiss them as cranks until we need their votes again huh",
[29] "DO NOT engage with the negative comments",
[30] "London's AI sector saw a 200 percent venture capital funding increase between 2015 and 2017. #ArtificialIntelligence",

[31] "This video is very informative and eye-opening however, this fandom continues to bring up MX on wearing that shirt. I'm tired, I don't want to engage anymore. They're stupid and can't see beyond what they think it's fair for their faves",
[32] "When I was on the @MCATA_Tweets executive, our conferences also ended up heavily skewed towards Middle and High School. We tried to engage more elementary teachers. Our theory was the same. Most MS and HS teachers teach mostly one subject. Elementary has to attend to all.",
[33] "Lines at the mics to engage in conversation at the #NRC2018Forum- so far we explored the importance of community partners and experiential learning, and that seeing 'froot' doesn't mean that food contains any actual fruit. Excited for the next workshops! #UnpackingFoodLiteracy",
[34] "While others continue to engage in political gamesmanship about who gets what power, @ocasio and the team already getting to work...because climate change doesn't care.",
[35] "Inside your head tho bc some of y'all are ignorant and don't really have the range to engage the rest of us in conversation.",
[36] "It is very lamentable that the leader of the Opposition has chosen to engage in a Trumpian style of rhetoric, chastising the government through either deliberate distortion, or a more fundamental ignorance of existing law. - Neil Boyd, criminology prof at SFU in BC #cdnpoli",
[37] "yes we have to engage with the ministry of environment to put in place tangible policies on environmental conservation,and the ministry of education to introduce a subject that centres around conservation so as to plant seeds of responsibility in the younger generation",
[38] "Good questions/thoughts re: motivation & change. Fear can create urgency & action but it is NOT sustainable, nor r we encouraging creativity & lowering stress. Leaders must engage heads & hearts about the promise of a bright future",
[39] "More alcohol brands using AR to engage consumers (AR big this holiday season) @JagermeisterUSA Reveals Halloween Fortunes With AR Tarot Cardshttps://www.alistdaily.com/technology/jagermeister-halloween-tarot-ar … #BrandedEngagement",
[40] "Teaching Tip Thursday: Struggling to engage your student's in higher-level thinking skills? Utilize this Bloom's Tech Tools Wheel to help!",
[41] "Hire a @BRIDGEEnergyGrp expert #utility consultant to complete your next project. Tell us what you need!",
[42] "JW held out the American olive branch, and thats how we should always be. Most of the snark to conservatives are from young people. Rise above, dont engage w snark, just facts.",
[43] "It's not too late to register for tomorrow's 11:00 am MT marketing webinar.Come learn how to create brands that engage consumers. Join our free webinar with @mikesolo Sign up at https://www.eventinterface.com/en/everything-we-know-is-wrong-how-to-create-brands-that-engage-consumers-in-the-wild-new-world-of-marketing/ … #marketing #whywebuy #consumerbehavior @ingomu_learning @500SPKRS",
[44] "@BluestMagoo @SewellTim @jessphillips I don't engage with fake accounts.",
[45] "*You're, not *Your.Find someone else to debate.I have no desire to engage with a person who doesn't know the difference between a possessive adjective and a contraction.",
[46] "@lilaedd Hello Lila! We take pride in offering quality academic writing services to our clients. DM and engage us if interested.",
[47] "We've enjoyed taking part in @WWLondon18 fringe today - here's our CEO #LuisDeSouza speaking on how to engage #workers and deliver effective #changeprogrammes. Great feedback!",
[48] "@dr_simon I didn't engage him. Assumed he had better things to do! Sure we'll have him on #TheHearingPodcast soon though",
[49] "This guy thinks that Free Speech demands the right to call for its own end and that a private company is acting of their own will when censorship advocates engage in campaigns to ruin their public image. Genius right?",
[50] "Report the negative comments and move on. Don't engage with them."

Table 3: Clustering 50 sentences with k=5

Cluster #	Sentence #
1	9
2	15
3	29,50
4	27
5	1,2,3,4,5,6,7,8,10,11,12,13,14,16,17,18,19,20,21,22,23,24,25,26,28,30,31,32,33,34,35,36,37,38,39,40,41,42,43,44,45,46,47,48,49

Test Case 4:

Consider the following 50 sentences with k=5 for the word accident:

1. "It's with a heavy heart that we must confirm Paul Walker passed away today in a tragic car accident...MORE: http://t.coäó_",
2. "5 years ago I was involved in a roll over car accident that left me paralyzed from the shoulders down. This is my progress tod‰Û_",
3. "I just sharted myself. That's when u fart and u shit yourself on accident!",
4. "two souls don't just meet by accident",
5. "TIP OF THE DAY: Always pay attention. You don't meet people by accident and stuff doesn't happen by coincidence.",
6. "When am I gonna see you againFirst of all, you saw me by accident.",

7. "On June 11, Brandon Rogers tragically passed away in a car accident. At the request of his family, we share his audition with you.‰Û_",
8. "Still no Instagram it was an accident",
9. "On June 11, Brandon Rogers tragically passed away in a car accident. At the request of his family, we share his audition with you.‰Û_",
10. "i watched your snapchat story on accident fyi",
11. "My little sis had an accident today at kindergarten & this is how my dad left to pick her up so she wouldn't feel so saäó_",
12. "We don't meet people by accident. They are meant to cross our path.",
13. "When ur momma died from Alzheimer's, your sister died in plane crash & husband in a car accident but you still alive htt‰Û_",
14. "Horrific accident at the Iowa State Fair",
15. "Remember the universe doesn't ever put someone on your path by accident. Everything is a lesson, pay attention.",
16. "Finally, as if by accident, thebreaks down & admits the truth about where the violence is coming from htäó_",
17. "Horrific accident at the Iowa State Fair",
18. "This Tweet from @paulsaab6 has been withheld in: India.",
19. "Health Insurance stocks, which have gone through the roof during the ObamaCare years, plunged yesterday after I endedäó_",
20. "Sad 2 know about train accident in MP. God grant peace 2 souls of those who lost their lives & strength 2 their famili‰Û_",
21. "LMAOOOOO bro Migos done fucked around and made a hit on accident",
22. " Before Warriors vs. Mavs, Steph comforts Devin Harris' nephew who lost his dad in a car accident this past Thursday Õæ_Õ_ • äó_",
23. "LMAOOOOO bro Migos done fucked around and made a hit on accident Õæ_ÕŸ‰",
24. "Horrific accident at the Iowa State Fair",
25. "please donate to my gofundme. i was in a horrible accident--my birth",
26. "3 AN ACCIDENT CAR",
27. "Horrific accident at the Iowa State Fair",
28. "Liberals: Men are dangerous rapistsAlso liberals: Men should go in girl locker rooms and girls should go camping wiäó_",
29. "Ystrday night I survived 2 car accident in a row. I lost a lot of blood but I miraculously ended up with minor injuries.",
30. "íó deux doigts de l'accident mais son whooooo il me tue Õæ_ÕŸ‰Õæ_ÕŸ‰",
31. "It is unconscionable that millions of people in the world‰Ûªs richest country are one diagnosis or car accident away from fin‰Û_",
32. "íó deux doigts de l'accident mais son whooooo il me tue Õæ_ÕŸ‰Õæ_ÕŸ‰",
33. "When you've been in the U.S. for 17 years and you need a translator to call Trump a white supremacist in English:",
34. "How black moms be during a car accidentÕæ_ÕŸ‰Õæ_ÕŸ‰",
35. "Brilliant idea! They can use all the money they looted from Haiti",
36. "Term limits is the only thing that will tear down establishment and career politicians and stop the cover ups and out oäó_",
37. "These are not things they're doing by accident.",
38. "How black moms be during a car accident’Ê_’üâ’Ê_’üâ",
39. "bruh a fight AND a car accident",
40. "Teaching Tip Thursday: Struggling to engage your student's in higher-level thinking skills? Utilize this Bloom's Tech Tools Wheel to help!",
41. "his proposal plans got pushed back by a car accident but that didn't stop him from getting down on one knee ’Ê_’ü_‰□ _•ü",
42. "when your mom died from Alzheimer's, your sister died in plane crash & husband in a car accident but you still alive htt‰Û_",
43. "äóìMathematics operates with unearned privilege in society, 'just like whiteness.äó»äó • You, madame, are an imbecile. htäó_",
44. " I live in a country where waving an ISIS flag outside Parliament won't get you arrested, but mean comments about Islam onäó_",
45. "Les rebeu ils aiment bien snapper au volant íæ 112 km/h en ville apríÂs quand ils font un accident ils disent c'est l'3aí¿n fäó_",
46. "TIP OF THE DAY: ALWAYS ALWAYS PAY ATTENTION. YOU DONäó»T MEET PEOPLE BE ACCIDENT, AND THINGS DONT JUST HAPPEN BY COINCIDENäó_",
47. "rub ur ass against his dick by accident on purpose",
48. "Fan: What happened to Bizzle?Justin: Bizzle died.. In a fatal car accident.",
49. "Stephen Curry comforts Devin Harrisäó» nephew who recently lost his father in a car accident. (via @KDISAWARRIOR)",
50. "16 year old girl sexually assaulted on train between Preston and Blackburn.trying to trace this man. Do you k‰Û_"

Table 4: Clustering 50 sentences with k=13

Cluster#	Sentence#
1	18
2	10
3	14,17,24
4	3,4,5,12,15,33,46,28,29,33,46
5	47
6	25
7	35
8	8
9	21,23,30,32,34,38
10	39
11	6

12	1,2,7,9,11,13,16,19,20,22,31,36,40,41,42,43,44, 45,48,49,50
13	37

Optimization through Affinity Propagation Technique

K-Means clustering requires the number of clusters to be provided initially determining which requires good knowledge of the application domain and the problem definition without which there would either be too many or too few clusters which would not meet the requirements of the problem definition. Too many clusters would take more time iterating through the clusters formed degrading overall efficiency. Affinity Propagation [19] is an algorithm which is used to identify the exemplars [20] in the data points. The algorithm has the potential to determine the number clusters with the exemplar being the mean value of the cluster but not necessarily the centroid like in K-Means. The exemplar is not a space value but a data point. Each data point initially raises a claim to be the exemplar of the cluster. The data points exchange messages until a good set of exemplars and clusters are obtained. These messages can be of two types:

1. Responsibility messages $r(d,e)$ sent from the data points to the exemplars where d represents the data point and e the exemplar indicate how well the data points are would be suited to be a member of the exemplar's cluster.
2. Availability messages $a(d,e)$ sent from exemplar to the data point indicating how suitable exemplar e would be to the data point d.

Using the sum of all the messages provides the clustering information called the preference value $p(d,e)$ of exemplar e for d^{th} data point calculated as,

$$p(d,e) = r(d,e) + a(d,e) \quad (1)$$

The higher the preference value, the higher the chance for the data point to become the exemplar for the cluster. The number of clusters can be adjusted using the preference values where increasing the preference value increases the number of clusters and vice versa since lower preference values would make the data point to prefer to be part of another cluster with a higher preference value.

Affinity Propagation for 8 sentences with word lied:

Exemplars:
2 7
Clusters:
Cluster 1, exemplar 2:
2
Cluster 2, exemplar 7:
1 3 4 5 6 7 8

Table 5: Clustering 8 sentences using Affinity Propagation

Cluster #	Sentence #
1	2
2	1,3,4,5,6,7,8

Affinity Propagation for 22 sentences with word crane:

Exemplars:
5 6 8 15 16 22
Clusters:
Cluster 1, exemplar 5:
5
Cluster 2, exemplar 6:
6
Cluster 3, exemplar 8:
8 9
Cluster 4, exemplar 15:
15
Cluster 5, exemplar 16:
1 2 3 4 7 10 11 12 13 14 16 17 18 19 20 21
Cluster 6, exemplar 22:
22

Table 6: Clustering 22 sentences using Affinity Propagation

Cluster #	Sentence #
1	5
2	6
3	8,9
4	15
5	1,2,3,4,7,10,11,12,13,14,1 6,17,18,19,20,21
6	22

Affinity Propagation for 50 sentences with word engage:

Cluster 1, exemplar 9:
9
Cluster 2, exemplar 11:
11
Cluster 3, exemplar 15:
15
Cluster 4, exemplar 17:
17
Cluster 5, exemplar 27:
27
Cluster 6, exemplar 29:
29
Cluster 7, exemplar 30:
30
Cluster 8, exemplar 31:
1 2 3 4 5 6 7 8 10 12 13 14 16 18 19 20 21 22 23 24 25 26 28 31 32 33 34 35 36 37
38 39 40 42 43 45 46 47 48 49 50
Cluster 9, exemplar 41:
41
Cluster 10, exemplar 44:
44

Table 7: Clustering 50 sentences using Affinity Propagation for the word *engage*

Cluster #	Sentence #
1	9
2	11
3	15
4	17
5	27
6	29
7	30

8	1,2,3,4,5,6,7,8,10,12,13,14,15,18,19,20,21,22,23,24,25,26,28,31,32,33,34,35,36,37,38,39,40,42,43,45,46,47,48,49,50
9	41
10	44

Affinity Propagation for 50 sentences with word accident:

Clusters:
Cluster 1, exemplar 6:
6
Cluster 2, exemplar 7:
7
Cluster 3, exemplar 8:
8
Cluster 4, exemplar 9:
9
Cluster 5, exemplar 10:
10
Cluster 6, exemplar 18:
18
Cluster 7, exemplar 21:
21 23
Cluster 8, exemplar 25:
25
Cluster 9, exemplar 26:
1 2 3 4 5 11 12 13 14 15 16 17 19 20 22 24 26 27 28 29 31 33 36 38 39 40 41 42 43
44 45 46 48 49 50
Cluster 108, exemplar 32:
30 32 34
Cluster 11, exemplar 35:
35
Cluster 12, exemplar 37:
37
Cluster 13, exemplar 47:
47

Table 8: Clustering 50 sentences using Affinity Propagation for the word *accident*

Cluster#	Sentence#
1	6
2	7
3	8
4	9
5	10
6	18
7	21, 23
8	25
9	1, 2, 3, 4, 5, 11, 12, 13, 14, 15, 16, 17, 19, 20, 22, 24, 26, 27, 28, 29, 31, 33, 36, 38, 39, 40, 41, 42, 43, 44, 45, 46, 48, 49, 50
10	30, 32, 34
11	35
12	37
13	47

CONCLUSION

The centroids for K-Means clustering are selected and given. Centroids which are close to a good solution seem to work the best and having small clusters is suitable in terms of efficiency in time to iterate through all the data points of the cluster. The centroids are initialized randomly for each cluster and may be either space value or data value with the centroid scalar value which is out of any end user's control. There also exists the possibility of having too many clusters causing loss in efficiency, increase in processing overhead.

The affinity propagation method is used to identify exemplars among the data points in clusters which are surrounded by the data points. These exemplars are data values and not space values. The algorithm determines the number of clusters by using the preference values intrinsically to distinctly specify the outer values for a set of data points.

REFERENCES

1. [1] Quing Yang, Ye Liu, Dongxu Zhang, Chang Liu, Improved k-means algorithm to quickly locate optimum initial clustering number K, Proceedings of the 30th Chinese Control Conference, 2015.
2. [2] Md. Sohrab Mahmud, Md. Mostafizer Rahman, Md. Nasim Akhtar, Improvement of K-means clustering algorithm with better centroids based on weighted average, 7th International Conference on Electrical and Computer Engineering, 2016.
3. [3] Farid Bourananni, Mouhcine Guennoun, Ying Zhu, Clustering Relational Database Entities Using K-means, Second International Conference on Advances in Databases, Knowledge, and Data Applications, 2015.
4. [4] Shuang Chen, Junli Li, Xiuying Wang, A Fast Exact Euclidean Distance Transform Algorithm, Sixth International Conference on Image and Graphics, 2016.
5. [5] Abhay B. Rathod, Sanjay M. Gulhane, Shailesh R. Padalwar, A comparative study on distance measuring approaches for permutation representations, IEEE International Conference on Advances in Electronics, Communication and Computer Technology (ICAECCT), 2016.
6. [6] Shruti Gupta, Abha Thakral, Shilpi Sharma, Novel technique for prediction analysis using normalization for an improvement in K-means clustering, International Conference on Information Technology (InCITe) – The Next Generation IT Summit on the Theme – Internet of Things: Connect your Worlds, 2016.
7. [7] Sarra Ben Hariz, Zied Elouedi, IK-BKM: An incremental clustering approach based on intra-cluster distance, ACS/IEEE International Conference on Computer Systems and Applications – AICCSA, 2015.
8. [8] Chellamal Surianarayanan, Gopinath Ganapathy, An Approach to Computation of Similarity, Inter-Cluster Distance and Selection of Threshold for Service Discovery Using Clusters, IEEE Transactions on Services Computing, 2017.
9. [9] M. Sato-IIic , On evaluation of clustering using homogeneity analysis, Smc 2000 conference proceedings. 2000 ieee international conference on systems, man and cybernatics. 'cybernatics evolving to systems, humans, organizations and their complex interactions, 2015.
10. [10] Shruti Kapil, Meenu Chawla, Performance evaluation of K-means clustering algorithm with various distance metrics, IEEE 3rd International Conference on Power Electronics, Intelligent Control and Energy Systems (ICPEICES), 2017.

11. [11] Senem Kumova Metin, Bahar Karaoglan, Tarik Kisla, *Text similarity analysis using IR lists*, 21st Signal Processing and Communications Applications Conference (SIU), 2016.
12. [12] Yaqin Xie, Yan Wang, Armugam Nallanathan, Lina Wang, *An Improved K-Nearest-Neighbor Indoor Localization Method Based on Spearman Distance*, IEEE Signal Processing Letters, 2017.
13. [13] Lu Liu, Jianqin Zhou, *A route-like demand location problem based on squared-Euclidean distance*, International Conference on Logistics, Informatics and Service Sciences (LISS), 2016.
14. [14] Mehmet Yesilbudak, Ilhami Colak, Ramazan Bayindir, *k-Means Partition of Monthly Average Isolation Period Data for Turkey*, 15th IEEE International Conference on Machine Learning and Applications (ICMLA), 2016.
15. [15] G. Chen, Yang Dai, *A new distance measurement for clustering time-course gene expression data*, The 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2014.
16. [16] Yu Jiankun, Guo Jun, *An Improved Agglomerative Levels K-Means Clustering Algorithm*, International Conference on Management of e-Commerce and e-Government, 2014.
17. [17] Xi-xian Niu, Guo-bin Han, Li-li Zhao, *Clustering algorithm research and realization based on Local Gathering Features*, Eleventh International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), 2014.
18. [18] G. Santafe, J,A. Lozano, P.Larranaga, *Bayesian Model Averaging of Naïve Bayes for Clustering*, IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 2016.
19. [19] Preeti Arora, Deepali Virmani, Shipra Varshney, *Substantiation of K-means and Affinity Propagation algorithm*, 7th International Conference on Cloud Computing, Data Science & Engineering, 2017.
20. [20] Shenping Xia, Jianjun Liu, Edwin Hancock, *Mining Exemplars for Object Modelling Using Affinity Propagtion*, 26th International Conference on Pattern Recognition, 2016.