Big Social Data Analytics for Public Health: Comparative Methods Study and Performance Indicators of Health Care Content on Facebook

Nadiya Straton[1], Raghava Rao Mukkamala[1,2], Ravi Vatrapu[1,2]
[1]Centre for Business Data Analytics, Department of Digitalization, Copenhagen Business School, Denmark
[2]Westerdals Oslo School of Arts, Comm & Tech, Norway
{ns.digi, rrm.digi, rv.digi}@cbs.dk

***Abstract*—This paper presents a novel approach that evaluates the right model for post engagement and predictions on Facebook. Moreover, paper provides insight into relevant indicators that lead to higher engagement with health care posts on Facebook. Both supervised and unsupervised learning techniques are used to achieve this goal. This research aims to contribute to strategy of health-care organizations to engage regular users and build preventive mechanisms in the long run through informative health-care content posted on Facebook.**

***Index Terms*—Gaussian mixture model, K nearest neighbors (KNN), BIC (Bayes Information criterion), AIC (Akaike information criterion), CV (Cross Validation).**

I. Introduction

Innovative advances in participatory Internet make social media sites such as Facebook an inescapable platform for health care promotion and education according to [1]. Social media provides new opportunities for interaction and distribution of information within and across organisations, which results in new kinds of socially mediated organisations [2]. Within the public health paradigm, the field of health informatics deals with: "the structures and processes, as well as the outcomes involved in the use of information and communication technologies (ICTs) within health" [3, p.501-502]. Situated within new public health informatics field, this paper compares different methods to study attributes of healthcare posts. Furthermore, the suitability of different algorithms is solely dependent on the data characteristics [4], therefore, there is a need for further in-depth analysis to find the suitable unsupervised and supervised machine learning algorithms to derive meaningful facts and actionable insights from social data.

A. Research Questions

The objective of this paper is to establish a post engagement frame and find the right classification model that can help potential stakeholders - public health care organizations with their social media strategy. Specifically what type of content to post and when. As part of this research work, two clustering algorithms: Gaussian mixture model (GMM) and K-means, two classification algorithms: K nearest neighbours (KNN) and Artificial Neural Networks (ANN) from [5] are applied and evaluated on the dataset to reach this goal. Therefore, we would like to explore the following research questions.

1) Which supervised and unsupervised algorithms will perform better on the public health social data?
2) Which Facebook post features do consumers of public health information find most engaging?

II. Related Work

Similar to current research, authors in [6] recognize the importance of rapid information dissemination on social media platforms. Even though, research is build on the specific case of MERS (Middle East Respiratory Syndrome), it touched upon the relevant topic in relation to health information sharing on social media platforms. Such as the differentiation between reliable information and false news, especially in cases of infectious outbreaks and the consequent response from users.

A. Unsupervised learning in social media and health-care contexts

Number of sources suggest that growth in popularity over time depends on the early measurements that define popularity [7][8]. Popularity expectation might depend on the platform such as Youtube or other social media sources. Authors in [7] use K-means clustering (with $k = 2$) to separate stories from Digg website into upper cluster stories (more popular) and the rest. Furthermore, the strength of a correlation between popularity cluster at different correlation times and reference time is calculated. Authors in [9] suggest clustering has a potential to extract actionable patterns that will be beneficial for businesses and private users. However, there are challenges posed by the vastness of the data, which is often noisy, unstructured and dynamic in nature. Therefore, novel approaches to data pre-processing and clustering might arise such as proposed in [10]. Authors in [11] reason that GMM performs better than cluster algorithms with hard assignments, such as K-means. The latter is more sensitive to outliers than probabilistic algorithms [11]. In contrast, cluster assignment with K-means perform equally if not slightly better than GMM in the current study. Moreover, authors do not mention data prepossessing and outlier detection prior to clustering in contrast to current research. More approaches have to be considered in case of different social media sites or a specific health care data set of a problem, as expressed previously in [12], an all purpose clustering algorithm will be very difficult to design. Conclusively, clustering algorithms should be benchmarked and tested with performance measures to find the most suited model for current research. Authors in [13]

978-1-5386-2715-0/17/$31.00 ©2017 IEEE

apply Gaussian Mixture Model to cluster user locations on Twitter. Geographical location and favorite topics per certain locations might infer personalized preferences of users in the region, infer user interests and predict user behaviour. Current research applies clustering techniques to national health care organization world wide, while taking into account similar features in their behavior. Both approaches are potentially invaluable for many reasons, such as patient engagement with social media content, behavior targeted to certain companies or differentiated on the disease and region.

B. Supervised learning in social media and health-care contexts

Authors in [7] suggests that prediction accuracy might depend on the choice of the model. Their hypothesis supports the previous assumption made by[14]: there is no one perfect algorithm or a model that performs with high accuracy in every single case. Lack of inherent superiority of any classifier: 'No Free Lunch theorem' and problem of over-fit are guiding principles in designing a prediction model in this study. Furthermore, two different classification algorithms are applied with various structures of the network and neighbors assignment, to find the most fitted model for the current data set. Similar, to current study [15] defines the number of features to predict the popularity of news articles on Twitter. Features are then used as inputs for classification algorithms: linear regression, k-nearest neighbors (KNN) and support vector machine (SVM) regression. Authors measure Euclidean distance between two articles based on their location in the feature space with the KNN algorithm. They do not give any background on why they use seven and three nearest neighbors. However, they state that KNN performs poorly on their data set, even more so when the data sample grows. In contrast, current study presents detailed reason on why and how number of neighbors is selected. Furthermore, authors use a similar approach to current research by defining popularity classes based on historical data. However, in contrast they do not take into account '0' posts and shares. The reason for one of the weakest predictors in "The Pulse of the News on Social Media: Forecasting Popularity" article is an overlap in the 'category score', as they are not disjointed [15]. Similarly, in current research 'Season' and 'Holiday/Not Holiday' attributes have the highest overlap in their features across three engagement classes/clusters and therefore lead to poor classification results.

III. Data Set Description

Data from Facebook is represented by 153 public health care organizations / health care walls and collected using Social Data Analytic Tool (SODATO) [16]. They represent the biggest health care organizations in the volume of posted content on Facebook, include private persons (bloggers), national as well as international organizations and can be distinguished though Facebook Wall Category. The total dataset contains information about approximately 43 million Facebook actions during the time period of 10 years: from the beginning of 2006 to the end of 2015.

Number of post performance indicators are presented in Table I .

Start date: **2006-01-01** End date: **2015-12-30**			
Number of Facebook Walls: **153**			
Activity/Nr.of Actions		**Unique Actor Stats of the Dataset**	
Facebook Page Likes	10476523		
Number Posts	280534	Post actors	101351
Shares Count	4225739	-	-
Post Likes Count	24331261	Post Like actors	7129957
Comment Count	1734154	Comment actors	788297
Comment Like Count	1507687	Comment Like actors	493266
Comment Reply Count	208512	Comment Reply actors	100379
Comment Reply Like Count	176920	Comment Reply Like actors	88202
Total actions	42941330	**Total Unique Actors**	7531865

Table I: Data description

Table I shows the time span, number of walls, unique actors and activity per each of the post performance measures. Dataset includes: 24 million Post Likes (approximately 55% of the total actions), 10 million Page Likes and 4 million Post Shares, 7,5 million unique actors that liked, commented or shared health-care content.

IV. Research Methodology and Process Flow

Initially, we used descriptive statistics to visualize data and perform reduction in the data attributes. Then, we applied unsupervised clustering techniques to segregate data into different clusters and to filter the post engagement attributes. Finally, supervised learning techniques were applied to understand other categories and attributes of the posts and to build a classifier to make predictions on the new data.

A. Engagement Frame

To find an association if any between engagement attributes coefficient of determination r^2 was applied. Page Like shows weak correlation with Post Like and Post Share attributes, therefore if companies pay for engagement with their Page, it does not necessary lead to engagement with the content. Post Share shows the highest correlation with Post Like (0.25) and Comment (0.18). Post Like has second highest correlation with Comment Like (0.30). Even though the strength of the linear relation between the attributes is below 0.50 and can be described as weak, evaluation of strength depends of the data type being examined.

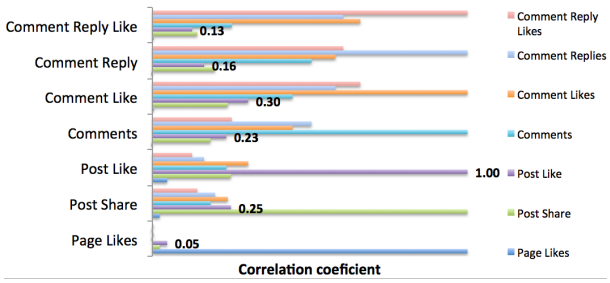

Figure 1: Correlation between engagement attributes.

Figure 1 shows seven engagement measures and their positive linear correlation with itself and another six attributes.

Since, Post Like is the most direct way to measure post performance, three other most correlated attributes will be included in the engagement frame: Post Like, Post Share, Comment and Comment Like to measure post performance and popularity with clustering algorithms.

1) Data Dimensionality Reduction: After attributes reduction to four performance measures - data pruning might be necessary. Attributes are dispersed in general between minimum and maximum values, even after removal of outliers. Therefore, further reduction, normalization or standardization is necessary.

Figures 2 and 3 show distribution of post engagement and their mean values before and after log-normalization.

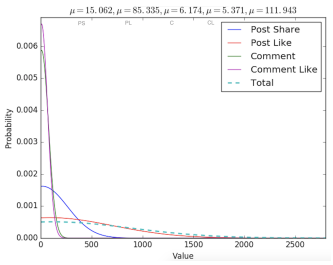

Figure 2: Distribution of Engagement Attributes

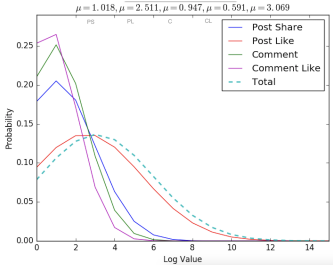

Figure 3: Distribution, after reduction with log values

Log value reduction/normalization allows to visualize data better and reduce risk of high-dimensional spaces while clustering. In general, distribution of values is skewed to the right (not normally distributed).

B. K-means

Selection of K - number of clusters can influence on the performance of K-means algorithm, however studies do not contain any explanation or justification for setting K to a specific value. Finding the optimal number of clusters for a given data set is a challenging problem in clustering [17]. However, number of studies looked into selection of K (clusters) and one method iterates algorithm several times and estimates quality of the cluster results visually. The research work in [17] suggests that optimal number of clusters K is when the value k, for which $log(W_k)$ falls the furthest below the reference curve:

$$Gap_n(k) = E_n^* \{log(W_k)\} - log(W_k)$$

The smaller the average within cluster sum of squares and the smaller the gap between two points on the elbow curve the more optimal is K that corresponds to the number of clusters on x coordinate. 'Elbow' points on Figure 4 show percentage of variance explained with different number of clusters and the most optimal are three and five clusters. Visually the most optimal and distinct distribution of the data points in the cluster in comparison to the manually assigned bounded classes is shown in *three* (figure 5) and *five* (figure 6) cluster assignments. 'Low engagement' (Cluster 0) cluster constitutes 63% of data points from the total data set,'Medium'(Cluster 1)

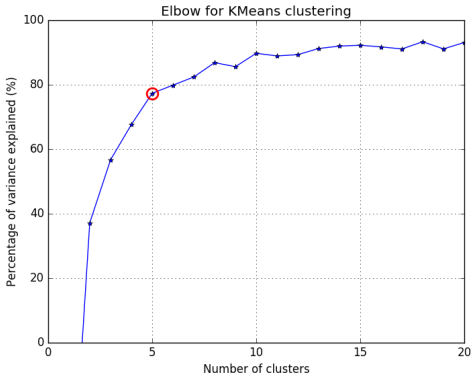

Figure 4: Percentage of variance explained

- 27% and 'High'(Cluster 2) - 10%. Each cluster with centroid is represented by a single color and manually assigned classes are triangularly shaped. Figure 6 shows the distribution of 5

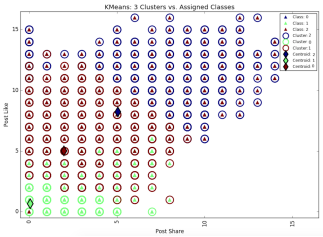

Figure 5: 3 Clusters/Classes K-means

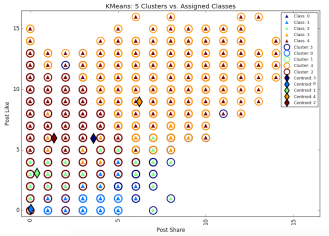

Figure 6: 5 Clusters/Classes K-means

clusters/classes From low cluster: (Cluster 0) with 49% of the total data set to high cluster: (Cluster 4) that constitutes 5% and 6% of the data.

C. Gaussian Mixture Model

Gaussian mixture model algorithm also requires selection of clusters, which might not be necessary a downside, as one has control over how many class labels are more suitable for the model and especially if model aims for smaller amount of clusters such as in the current research. Optimal number of clusters for GMM is determined with BIC (Bayesian Information Criteria), AIC (Akaike Information Criterion) and CV (Cross Validation). However, mixture of components is not always optimal number of clusters. In the case of BIC, if the number of mixture components are selected as clusters, the result can lead to overestimation as suggested in [18]. We have applied uni-variate Gaussian mixture model with G components, where observations are sampled from probability distribution with density. Figures 7 show BIC, AIC and CV estimates. BIC is based on high data count and strives for less complexity. Simulation results in [19], show that BIC tends to underestimate mixture model order .

$$BIC(G) = -2p(y|\hat{\tau}, G) - d * log(n)$$

[20] [21]. AIC tends to overestimate the correct number of components, is based on the low data count and also strives

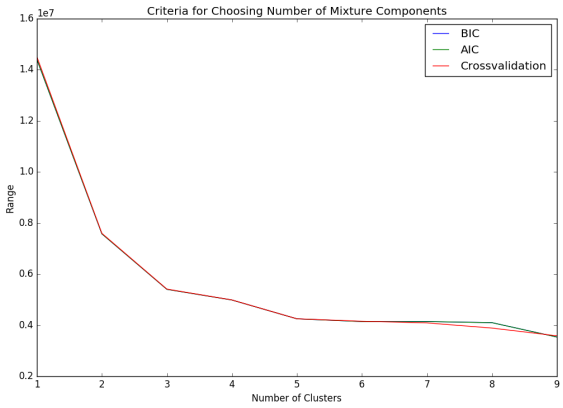

Figure 7: BIC, AIC and CV development

for less complexity

$$AIC(G) = -2logp(y|\hat{\tau}, G) + 2d$$

In practical situations, BIC criterion can help with over-parameterization of AIC [19]. Judging by the results from the data, BIC and CV show similar development, while AIC deviates slightly. The work in [22] conclude that CV based on averaging outperformed the other validation methods. Combination of AIC, BIC and CV helps to reach more balanced and optimal solution. All three models suggest five, as the most optimal number clusters for GMM, followed by three. Visual representation of classes versus clusters suggests more clear split between low and high engagement clusters in five cluster set up than three cluster set up. Figures 8 and 9 display scatter plot with combination of five and three cluster-class assignments. Visuals shows much better results for K-means

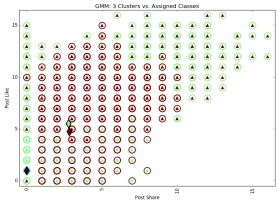

Figure 8: 3 Clusters/Classes GMM

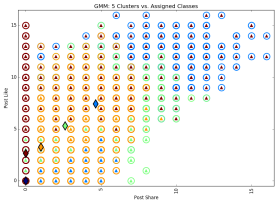

Figure 9: 5 Clusters/Classes GMM

cluster assignment with three clusters, as opposed to GMM. Moreover, evaluation coefficients in [5] show slightly better assignment with 3 clusters for K-means when compared to bounded, manually assigned class labels and therefore three cluster assignment will be chosen.

D. Post Engagement with K-Nearest Neighbors

In previous section: Post Like, Post Share, Comment and Comment Like were clustered on three engagement clusters with K-means. However, final goal of this research is to combine analyses from three engagement clusters ('Low', 'Medium' and 'High engagement') and eight other independent attributes (isHoliday, Season, Created Year, Month, Day of Week, Time of Day, Hour Span between Create and Update date and Post Type) to predict post popularity. KNN algorithm classifies the outcome for a new query point (marked as a star) surrounded by instances in figure 10. Depending on the majority of nearest neighbours, query point will be classified as class 0, 1 or 2. From figure 10, when *10 nearest neighbours* are selected, *KNN* will assign Class 0 - low engagement cluster to the outcome of the query point using majority voting rule. The choice of K - number of neighbors is important, as can influence on the quality of the prediction. Small value of K can lead to large variance in predictions where as a very large K can lead to model bias. Thus, there is an optimal value for K in the current data set: large enough to minimize probability of classification and small enough in relation to the number of cases in the sample, where query point is relatively close to K nearest points [23]. Figure 10 shows classification process with K=10 and K=3.

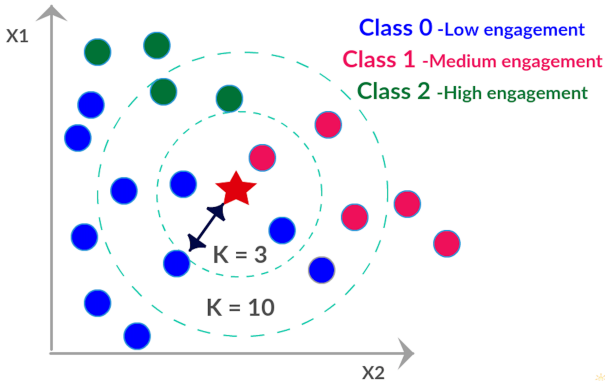

Figure 10: K-nearest neighbors

In order to decide on the optimal number of neighbors, training sample size was reduced from 80% to 18% (to keep the running time of algorithm low) to try out different values for K ($K = 1 \cdots 2000$). [24] suggests if k is too small, then nearest neighbor classifier can be susceptible to over fitting, because of the noise in the training data and lead to higher error rate predictions on the new data. Figure 11 shows automatic process of error rate estimation with varying number of K-nearest neighbors . Classification error rate decreases, as number of neighbors increases from 0 to 10, and then gradually increases. If k is too large the nearest neighbor algorithm may miss-classify *test instance*, because nearest neighbors are located too far from its neighborhood region [24]. To make predictions with KNN, one needs to decide on a metric for measuring the distance between the query point and points assigned to classes. After decision on K neighbors has been made, prediction is done based on KNN examples in the neighbor region. The prediction outcome of the query point is based on a voting scheme in which the winner is used to label the test query point. [25] Therefore in current work, number of neighbors was set to ten.

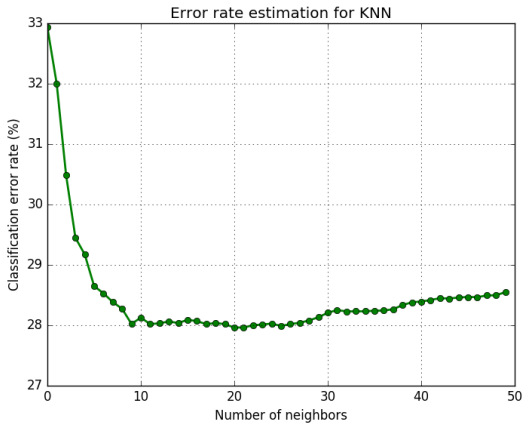

Figure 11: KNN error rate estimation with K neighbors selection.

V. Results

A. Post Engagement Prediction with KNN

K-nearest neighbors algorithm showed better accuracy results (72%) than ANN (69% accuracy rate) with 2 hidden layers, 2 hidden units and trained sample size of 5% (15000 lines), as seen in [5]. Evidence in [5] shows that right neural network architecture can be important for achieving more accurate results. Classification in figure 12 shows KNN classification results with 10 nearest neighbors (normalized and non normalized values), based on 0-low, 1-medium, 2-high engagement clusters. True label contains 100% of each class, where as prediction label classes might contain combination of the *'True'* classes. Ideally each of the *Prediction label* classes should contain 100% of only one class.

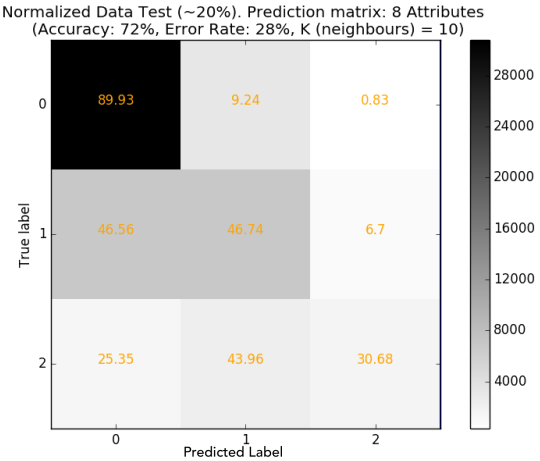

Figure 12: Total Data Matrix, Normalised

Classification matrix shows 'true' clusters and their spread at the Prediction label with reduced trained sample size that constitutes around 20% of the total data set (56000 data points) and test sample: 80% (224000). Since, the highest accuracy of 72% is achieved with 10 Nearest Neighbors, 10 KNN will be used to classify data for each of the attributes separately, to investigate which attributes are predicted with the highest accuracy and therefore affect post engagement the most.

Low engagement cluster is predicted with the highest accuracy of 90%, as it is the biggest class in the data set. Features that belong to medium engagement clusters are predicted with 47% accuracy. Moreover, features in the high engagement cluster, the smallest set of values are predicted with 31% accuracy rate. 72% accuracy rate was achieved as a result of prediction based on the total of 8 attributes: *Post Type, Hour Span, Time Of Day, Day Of Week, Month, Season, IsHoliday, Country Code*. Some attributes lead to higher engagement with the posted content than others, as shown in the table below. *Hour span* and *Post Type* have more expressive features in each of the clusters, therefore achieve best prediction results of 72% and 71% accuracy. KNN predicted label for 3 clusters: with around 90% accuracy for the low engagement cluster, 50% for the medium engagement cluster and around 30% in the high engagement cluster. *Season* and *isHoliday* show inconclusive results, where algorithm classifies labels into one prediction cluster. 'Time Of Day', 'Day Of Week', 'Month' and 'Country' are predicted with accuracy of less than 62%. Classification matrix in Figure 15 shows classification for 'Post Type' and 'Hour Span' attributes with 10 nearest neighbors. These attributes have the most distinct features in each of the clusters and conclusively can have the highest influence on the post engagement values. Visual content such as picture and video are of the most interest to the people who consume health care content on Facebook, followed by content that contains links. Even though, 'status' post types (short message) are posted by the health-care companies the most, this type of message does not engage users, results are build on previous work and are elaborated in greater detail in [26]. Features of the other attributes do not stand out strongly as part of any single cluster, have rather low accuracy prediction rates and can be disregarded from the classification and prediction for public health-care data set.

VI. Conclusion and Future Work

Hypothesis about lack of superiority of any model is supported in this research, where different methods are applied to find the most engaging features of the health-care posts. Problem of over-fit or large data sparsity are partial reason for accuracy decline with additional hidden units in ANN. In the case of KNN, weighing approach proved to be an important factor in assigning query point or right number of nearest neighbors to avoid over-fitting. In-spite of the number of advantages with GMM algorithm and adaptability to outliers, K-means had faster convergence was supported with better estimates and showed better visual assignment/separation of the data points in each of the clusters. Complex models such as DNN do not always perform well and depend on the type of the problem and data set in question, as in the current research where KNN achieved better prediction result. Moreover, time between between post creation - post update and visual content

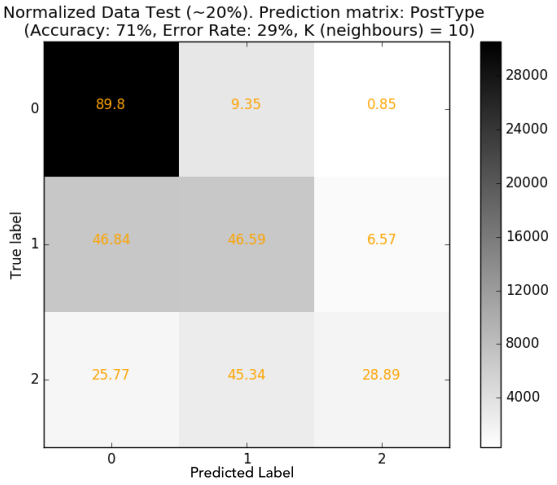

Figure 13: Post Type Matrix

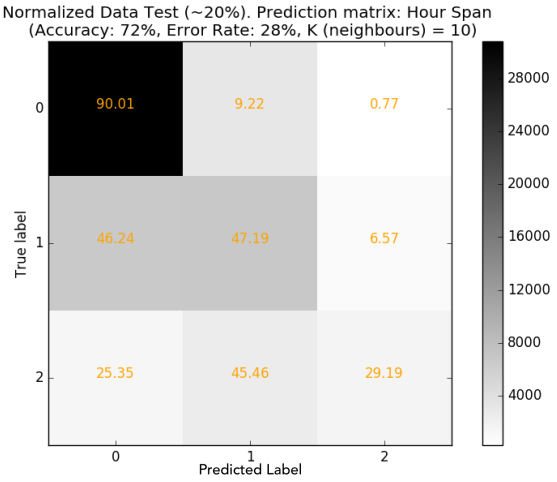

Figure 14: Hour Span Matrix

Figure 15: KNN classification results, based on 2 attributes: 'Post Type', 'Hour Span'

lead to greater engagement with health-care content and should be explored further.

ACKNOWLEDGMENT

We thank members of the Center for Business Data Analytics (bda.cbs.dk) for their feedback on the paper. The authors were partially supported by ReVus project, Big Data Analytics for Public Health funded by Copenhagen Health Innovation Fund. Any opinions, findings, interpretations, conclusions or recommendations expressed in this paper are those of its authors and do not represent the views of the Fund.

REFERENCES

[1] H. Korda and Z. Itani, "Harnessing social media for health promotion and behavior change," *Health promotion practice*, vol. 14, no. 1, pp. 15–23, 2013.

[2] R. Vatrapu, "Understanding social business." in *Emerging Dimensions of Technology Management*. Springer, 2013, pp. 147–158.

[3] P. A. Bath, "Health informatics: current issues and challenges," *Journal of Information Science*, 2008.

[4] K. A. Smith, F. Woo, V. Ciesielski, and R. Ibrahim, *Matching Data Mining Algorithm Suitability to Data Characteristics Using a Self-Organizing Map*. Heidelberg: Physica-Verlag HD, 2002, pp. 169–179.

[5] N. Straton, R. R. Mukkamala, and R. Vatrapu, "Big social data analytics for public health: Predicting facebook post performance using artificial neural networks and deep learning," in *Big Data (BigData Congress), 2017 IEEE International Congress on*. IEEE, 2017, pp. 89–96.

[6] J. Song, T. M. Song, D.-C. Seo, D.-L. Jin, and J. S. Kim, "Social big data analysis of information spread and perceived infection risk during the 2015 middle east respiratory syndrome outbreak in south korea," *Cyberpsychology, Behavior, and Social Networking*, 2017.

[7] G. Szabo and B. A. Huberman, "Predicting the popularity of online content," *Communications of the ACM*, vol. 53, no. 8, pp. 80–88, 2010.

[8] K. Lerman and T. Hogg, "Using a model of social dynamics to predict popularity of news," in *Proceedings of the 19th international conference on World wide web*. ACM, 2010, pp. 621–630.

[9] P. Gundecha and H. Liu, "Mining social media: a brief introduction," *Tutorials in Operations Research*, vol. 1, no. 4, pp. 1–17, 2012.

[10] J. Tang, X. Hu, H. Gao, and H. Liu, "Unsupervised feature selection for multi-view data in social media." in *SDM*. SIAM, 2013, pp. 270–278.

[11] A. Pal and S. Counts, "Identifying topical authorities in microblogs," in *Proceedings of the fourth ACM international conference on Web search and data mining*. ACM, 2011, pp. 45–54.

[12] A. K. Jain, "Data clustering: 50 years beyond k-means," *Pattern recognition letters*, vol. 31, no. 8, pp. 651–666, 2010.

[13] A. Ahmed, L. Hong, and A. J. Smola, "Hierarchical geographical modeling of user locations from social media posts," in *Proceedings of the 22nd international conference on World Wide Web*. ACM, 2013, pp. 25–36.

[14] D. H. Wolpert, "The supervised learning no-free-lunch theorems," in *Soft computing and industry*. Springer, 2002, pp. 25–42.

[15] R. Bandari, S. Asur, and B. A. Huberman, "The pulse of news in social media: Forecasting popularity," *arXiv preprint arXiv:1202.0332*, 2012.

[16] A. Hussain and R. Vatrapu, "Social data analytics tool (sodato)," in *DESRIST-2014 Conference (in press)*, ser. Lecture Notes in Computer Science (LNCS). Springer, 2014.

[17] R. Tibshirani, G. Walther, and T. Hastie, "Estimating the number of clusters in a data set via the gap statistic," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 63, no. 2, pp. 411–423, 2001.

[18] J.-P. Baudry, A. E. Raftery, G. Celeux, K. Lo, and R. Gottardo, "Combining mixture components for clustering," *Journal of Computational and Graphical Statistics*, 2012.

[19] G. Celeux and G. Soromenho, "An entropy criterion for assessing the number of clusters in a mixture model," *Journal of classification*, vol. 13, no. 2, pp. 195–212, 1996.

[20] L. Tierney and J. B. Kadane, "Accurate approximations for posterior moments and marginal densities," *Journal of the american statistical association*, vol. 81, no. 393, pp. 82–86, 1986.

[21] R. J. Steele and A. E. Raftery, "Performance of bayesian model selection criteria for gaussian mixture models," *Frontiers of statistical decision making and bayesian analysis*, pp. 113–130, 2010.

[22] Y. Fang and J. Wang, "Selection of the number of clusters via the bootstrap method," *Computational Statistics & Data Analysis*, vol. 56, no. 3, pp. 468–477, 2012.

[23] I. StatSoft, "Electronic statistics textbook. tulsa, ok: Statsoft," 2007.

[24] T. Pang-Ning, M. Steinbach, V. Kumar *et al.*, "Introduction to data mining," in *Library of congress*, vol. 74, 2006.

[25] statsoft. [Online]. Available: http://www.statsoft.com/textbook/k-nearest-neighbors

[26] N. Straton, K. Hansen, R. R. Mukkamala, A. Hussain, T.-M. Gronli, H. Langberg, and R. Vatrapu, "Big social data analytics for public health: Facebook engagement and performance," in *e-Health Networking, Applications and Services (Healthcom), 2016 IEEE 18th International Conference on*. IEEE, 2016, pp. 1–6.