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

Nadiya Stratton\textsuperscript{1}, Raghava Rao Mukkamala\textsuperscript{1,2}, Ravi Vatrapu\textsuperscript{1,2}

\textsuperscript{1}Centre for Business Data Analytics, Department of Digitalization, Copenhagen Business School, Denmark
\textsuperscript{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 healthcare posts on Facebook. Both supervised and unsupervised learning techniques are used to achieve this goal. This research aims to contribute to strategy of healthcare organizations to engage regular users and build preventive mechanisms in the long run through informative healthcare 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 \cite{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 \cite{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” \cite[p.501-502]{3}. 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 \cite{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 \cite{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 \cite{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 \cite{7,8}. Popularity expectation might depend on the platform such as Youtube or other social media sources. Authors in \cite{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 \cite{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 \cite{10}. Authors in \cite{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 \cite{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 \cite{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 \cite{13}
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 behavior. 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 1.

<table>
<thead>
<tr>
<th>Activity/Nr.of Actions</th>
<th>Unique Actor Stats of the Dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td>Facebook Page Likes</td>
<td>10476523</td>
</tr>
<tr>
<td>Shares Count</td>
<td>4225739</td>
</tr>
<tr>
<td>Post Likes Count</td>
<td>243831264</td>
</tr>
<tr>
<td>Comment Count</td>
<td>1734154</td>
</tr>
<tr>
<td>Comment Like Count</td>
<td>1507697</td>
</tr>
<tr>
<td>Comment Reply Count</td>
<td>208512</td>
</tr>
<tr>
<td>Total actions</td>
<td>42941330</td>
</tr>
</tbody>
</table>

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.

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:

$$\text{Gap}_n(k) = E_n^\star \{ \log(W_k) \} - \log(W_k)$$

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 estimates quality of the cluster results visually. 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) clusters/assignments. 'Low engagement' (Cluster 0) cluster constitutes 63% of data points from the total data set,'Medium' (Cluster 1) - 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 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 .

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

[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

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 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 (i.e., Holiday, 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 $K=3$.

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.
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, whereas prediction label classes might contain combination of the ‘True’ classes. Ideally each of the Prediction label classes should contain 100% of only one class.

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
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