Correlation Discovery between Student Web Queries and their GPA

Jigar Jadav, Greg Goldberg, Dawn Lindelin, Andrew Preciado, and Raviteja Ganti

Seidenberg School of CSIS, Pace University, Pleasantville, NY 10570

Abstract — In recent years, the K-12 learning space has been utilizing mobile devices to supplement student learning. The question is now begged, if these devices are actually enhancing a child’s capacity to learn? Going deeper, the area that needs inspecting is whether or not the students are using school issued mobile devices to search for school-related information. This study examines high school student web queries performed on school issued iPads using anonymized data from web filter logs. These web queries are first binary classified as either school-related or non-school related using Student Web Query Classifier (SWQC) algorithm created by the authors. Upon completing the classification, this study examines whether a correlation exists between these web queries and student GPA. Based on the regression analysis performed in this study, the null hypothesis, student web queries performed on school issued iPads has no impact on student GPA, is rejected.

Keywords — web query classification, mobile learning, data analytics, big data privacy and security, education

I. INTRODUCTION

The Internet is the primary source of information daily. With the rise of the Internet age, search engines have become a vital tool for information retrieval. Various proprietary search algorithms are used to retrieve relevant information to users. Not only has the use of search engines grown, but the number of studies related to it has also grown due to its demand [1]. This new information availability has led to an increase in the number of mobile devices used in the K-12 learning space [18]. Apple iPads and other tablet devices are contributing to schools having one-to-one mobile initiatives to facilitate the integration of technology in the classroom [20].

Mobile Learning can be defined as learning that is supported through mobile devices [19]. Ubiquitous Learning refers to mobile learning that allows a learner access to information anywhere over wireless networks [37]. Mobile devices, such as iPads, permit students to learn wherever they are whether they care connected to the Internet or not. No longer are software and Internet resource restricted to a computer laboratory with bulking, space monopolizing desktop computers [20]. This is an advantage our educational system can utilize to educate children. However, since schools are only recently starting one-on-one mobile device initiatives, we have little understanding of how students are using these devices, meaning are these devices being used to their full potential.

The use of big data analytics is common in advertising, finance, medicine, marketing, etc. due to its monetary potential. We are now starting to see the same trend in higher education through an emerging field of educational data mining [17]. However, such quantitative research may also be beneficial K-12 education in potentially improving the efficacy of school issued mobile devices.

As the use of mobile devices becomes more prevalent in K-12 education worldwide, the analysis from this study may reveal useful information to teachers, administrators and parents. This study analyzes real data collected from web filter logs of high school issued mobile devices, in particular student web queries. Previous studies done by authors of [3] explored data through exploratory text analysis to generate term frequencies and word clouds from student performed web queries. Later, an in-depth analysis of student web queries is performed similar to study performed by Spink, et. al.[13]. Data mining techniques and machine learning algorithms were used on web filter logs of these device, in order to classify student web queries as school related or non-school related to better understand the efficacy of these devices [4].

Data are initially explored through exploratory text analysis to generate term frequencies and word clouds from student performed web queries. Later, these web queries are classified as either school related or non-school related through machine learning algorithms and SWQC classifier created by the authors of [4]. A proposed procedure is presented to build a corpus of school-related terms and to compare it against students’ online activity [3]. This approach produced more accurate results than other that implemented supervised learning algorithms such as Naïve Bayes and Support Vector Machines [3].

This study attempts to find a correlation between student web queries and their Grade Point Average (GPA). To extrapolate
further, a student that consistently enters into a browser “school” related search queries may have a higher GPA than a student who consistently does not search for school related material. In order to investigate this correlation, the first step in this task, is the data collection and analysis of these students’ search queries. Most school-issued devices are required to have web filters installed on them to filter inappropriate content. These web filters logs collect student activities on the internet. These logs contain the web queries of the students, key to the analysis in question. By mining the data, categorizing it, and using regression analysis this study establishes that students using their devices to search for school-related terms typically have higher GPAs; confirming a positive correlation. This information can now potentially be used by teachers, guidance counselors and administrators as a way to help academically struggling students, allowing for early intervention. This study also highlights the ethical concerns of using data generated by minors and the laws that govern their data privacy and security.

The reminder of the paper follows as such: background of past work, literature review of related work, Big Data privacy and security concerns; methodology, methodology, result, conclusion, and finally, future work.

II. BACKGROUND OF PAST WORK

Authors of [13] performed an in-depth analysis of student web queries. For example, on average, how many queries did the students perform in a session? What was the average number of terms used in a search query? What was the frequency of terms related to school work? Authors in [13] extracted, raw data from web filters logs deployed by a local school district. Over 10,000 entries were collected from a two-hour time period in a school day. The following attributes were extracted: Suspicious, IP Address, User, User OU, User Groups, Computer Device ID, Search Query (SQ), Category Domain, Action, Rule Set (RS), Origin, Time. Of these attributes, SQ was used to create a corpus for text analysis [13]. It was concluded that a significant portion of the student web queries collected, over a specific time span, were, in fact, school related.

Authors of [4] performed binary classification of student web queries as either school related or non-school related. The study was conducted in three stages: data collection, model specification, and model evaluation. The raw data consisted of 6,477 student queries without any preprocessing. This data was then sorted by attribute RS and then segmented into four groups based on staff, administrator, teacher, and student. Only queries performed by students were used in the study and did not contain any student identifiers.

A new algorithm, Student Web Query Classifier (SWQC) was developed after the traditional classifiers Support Vector Machine (SVM) and Naive Bayes yielded poor initial results. SWQC took a two phased approach, 1) an unsupervised learning algorithm was used to classify the data, and 2) the classified data was used to train the SVM model, which was then used to classify the queries.

The results of the new algorithm were significantly better than those of either SVM or Naive Bayes alone, 90.68% accuracy compared to 1.29% and 46.03% respectively. These results can be built upon to delve into further questions that the study’s results imply, such as, is student time well-spent on school provided devices?

All in all, these works attempt to make sense of new paradigms in the K-12 learning space. Electronic devices are on the rise in classrooms, but the use of such devices is in its infancy. Technology is becoming more important toward a successful educational path for students. In this study, new techniques are explored and created.

III. LITERATURE REVIEW OF RELATED WORK

Classroom use of technology has progressed rapidly during the twenty-first century. While research has delved into the online habits of students, the effect those habits have on a student’s grade point average (GPA) has not been widely studied.

A 2014 study conducted by Doan and Bloomfield attempted to answer, “Whether or not using the Internet as a study resource increase the students’ ability to write an essay?” [22]. Participants in this study where comprised of 49 4th and 5th grade students grouped randomly by the school’s principal into three sets. These groups were divided as follows: Group I had 90 minutes to write their essay after being given a writing topic, Group II was given 30 minutes of Internet research time before their hour long writing session, Group III was given Internet training, three periods lasting 45 minutes each, before the day of the assignment. On the day of the assignment they had 30 minutes of Internet research time before their hour long writing session [22]. Authors of [22] analyzed their data using Analysis of Variance (ANOVA). During their analysis, a fourth group was formed, Group W. This group combined the results of Group II and III. The new group was used to compare results against the control group, Group I. They concluded that there was a statistical significance between the scores of Group W who were given access to the Internet, and Group I when comparing total essay score. Furthermore, their analysis also uncovered statistical significance when Group I was compared to Group III in the overall essay scores. There was a four-point
difference between Group I’s total score (11.69) and Group III’s (15.69). This result then suggests a subsequent question, “Did the Internet lessons alone improve the students’ writing scores?” [22].

Also in 2014, a study by Zhang was conducted to explore correlation between fourth grade students’ Internet searches for a particular Internet math game site, coolmath-games.com, and their performance in math and reading [21]. The study focused only on the students’ search for the site and not any actual use of the site contents. Publicly available data of two varieties was utilized by the author to conduct that research; the first was Internet Search, utilizing Google Trends, the second, Web analytics through Compete. The author of [21] attempted to answer the following question, “In what ways are the interest of search for a specific math game correlated to the state-wide academic performance?” Their study focused specifically on the students’ performance utilizing 2011 and 2013 results from the National Assessment of Educational Progress (NAEP). Data analysis found negative correlation for both years when looking at the exact amount of searches and the students’ academic performance. The analysis of Web and Internet Search reveal coolmath-games.com is an extremely popular gaming sites for education, the single factor of searching for that site however, is not a predictor of improved math or reading performance in fourth grade students nationally [21].

The question of one factor alone, such as school related web queries, being the predictor of student success has been captured in studies spanning grades K through 12 as well as university graduate and undergraduate students. One such study used data mining analytics to capture university library usage and discover if usage correlated to the students’ grade point average (GPA) [24]. The researchers used Pearson correlation analysis, a statistical tool that measures the exact correlation of two variables, in this case GPA and library use. The range -1 to +1 indicates the strength of the correlation, with zero being no correlation between the two variables. Using just one semester’s results indicated there was no correlation between library book checkout and GPA with a result of 0.08. Library usage, captured by turnstile count, showed a weak correlation with a result of 0.16. The results improved slightly when comparing the cumulative GPA to both book checkout and turnstile counts, with respective results of 0.12 and 0.21.

The authors concluded that further analysis was required due to factors such as the use of libraries matter more in some disciplines than others, suggesting a breakdown by major is required. They also concluded, with the data available to them, they were not able to attribute student success to library usage alone.

Gaudreau, Miranda, and Gareau found diametrically opposed results when they studied laptop usage during class time [23]. This study uncovered statistical significance between school-unrelated laptop behavior and academic success. When students participated in non-school related searches during class time their academic performance suffered. However, of interest to study done in [23], school related laptop behaviors showed no correlation to student success.

Educational data mining (EDM) is just one domain in the overarching data mining process which grew out of the generation of big data. Data mining in general is the process of taking large amounts of data and identifying patterns within it in order to glean an understanding of various domains, such as government, retail, security, marketing, or in this case education [29,32]. As an analytical tool EDM has proved beneficial in such areas as improved learning results [30], student behavior, and improved technology learning approaches [31]. Researchers strive to gain an understanding of the effectiveness of places of learning as well as how to improve student scholarship [32]. Even though EDM is an emerging discipline there has already been further refinement of this broader category. Academic analytics is one such sub category that looks specifically at how compiled data can make discoveries about student retention, for example. This discipline concentrates more on a narrow focus such as a specific department or institution, rather than looking at the bigger picture [32].

There are a number of approaches to EDM, quantitative analysis utilizes the tools of EDM such as machine learning or statistics. A qualitative approach draws on information gathered from surveys or interviews, and can also be effective although more subjective than the former [32]. Classification is the most prevalent means to analyze the data gathered during EDM, with clustering a close second; modeling and regression analysis are also methods used to study the data [33]. Various algorithms are used to draw comparison between two or more variables, these algorithms have different levels of success, however there may not be a great deal of statistical significance. A case in point, Strecht et al. [34] utilized six classification algorithms, including SVM and Naïve Bayes, to assess the effectiveness of these algorithms in predicting the success or failure of students. The results of their study found no statistical significance between the six algorithms trialed. This leads one to question the effectiveness of these tools or to search for ways to improve them.

As valuable as EDM is, it has come under scrutiny after cyber-breaching exposed hundreds of thousands students’ personal information [35]. There are arguments that the risk associated with EDM is no greater than what the student expose him or herself to in the normal course of daily social media activity [36]. This is obviously an area that requires more research and begs the question, how much information is too much?

IV. DEEP DIVE INTO NAÏVE BAYES ALGORITHM

One of the most popular and oldest classifiers is the Naïve Bayes algorithm. The implementation of Naïve Bayes is simple but, it requires minimal amounts of training data. Compared to other classifiers dealing with text classification, Naïve Bayes performs very poorly. Because of this, it makes the Naïve Bayes classifier unusable in spite of the simplicity and intuitiveness of the model [5].

Naïve Bayes is known as “the punching bag of classifiers” and has placed last in numerous head-to-head classification papers. Despite this title, it is frequently used for text classification due to its fast and easy implementation.
Algorithms that are less erroneous, tend to be slower and more complex [6]. Due to being fast and being easy to implement, Naïve Bayes text classifiers that contain the strong attribute conditional independence assumption have been studied well and numerous Naïve Bayes models are proposed [11].

Because of its simplicity, the Naïve Bayes text classifier has become widely used in both the classifying and training stage. It is less accurate than other discriminative methods, but many researchers have proved that it is effective enough to classify the text within many domains. Naïve Bayes models allow each of the attributes to contribute equally and independently towards the final decision from other attributes, in which it is more efficient when it is compared to other text classifiers. This focuses on employing Naïve Bayes’ approach as the text classifier for evaluating its classification performance against other classifiers as well as for document classification. [10].

A Naïve Bayes classifier can be defined as follows; variables are denoted using capital letters such as Xi, and their values will be denoted by lower-case letters such as xi, and sets of variables are denoted by boldface letters such as X.

Let X = {X1, ..., Xn} be a set of observed random variables, which are called features, where each feature takes a set of values from its Di domain. The set of all feature sets is denoted by \( \Omega = D_1 \times \cdots \times D_n \). Let c, such that \( c \in \{0, \ldots, u-1\} \), be a random variable which denotes the class of a set of features.

A hypothesis \( h : \Omega \rightarrow \{0, \ldots, u-1\} \), that assigns a class to a given set of variables is defined as a classifier. Each class c is assigned a function \( f_c(x) \), \( c = 0, \ldots, u-1 \). The classifier selects the class with the maximum discriminant function on a given set of variables, written as \( h(x) = \arg \max_c \epsilon \{0, \ldots, u-1\} f_c(x) \).

The Bayes classifier \( h^*(x) \) uses the following probabilities given a number of variables as the discriminant function, i.e.

\[
f^*(x) = (C = c \mid X = x). \]

Applying Bayes’ theorem from Eq. 1 to this function gives

\[
P(C = c \mid X = x) = \frac{P(X = x \mid C = c)P(C = c)}{P(X = x)}.
\]

Since \( P(X = x) \) is the same for all classes it can be ignored. Hence, the Bayes' discriminant function can be written as

\[
f^*(x) = P(X = x \mid C = c)P(C = c), \]

where \( P(x = x \mid C = c)P(C = c) \) is called the class-conditional probability distribution (CPD) [12]. Thus the Bayes' classifier written as in Eq. 3 finds the maximum posterior probability hypothesis given x.

Next equation 3: \( h^*(x) = \arg \max_c P(X = x \mid C = c)P(C = c) \).

Applying the notion that features are independent given the class, we have the Naïve Bayes classifier, which is now the fourth equation [12].

\[
h^*_N(x) = \prod_{j=1}^{N} P(X_j = x_j \mid C = c)P(C = c)
\]

V. DEEP DIVE INTO SUPPORT VECTOR MACHINES

Classification of text data is classic example of a utilization of Support Vector Machines (SVM). The preceding papers on this topic focused on the discovery, analysis, and classification on what is known as "term frequency". Term frequency analysis is used to discover the relationship between effective and inappropriate usage of mobile devices in school [3]. Analogous to this, text categorization techniques are used to classify a bevy of everyday items used in today’s society: Internet search engine results, news cycles, daily email alerts, and to making text auto-complete suggestions [2]. The reason SVM is extremely well suited for text categorization is that it a self-taught mechanism for binary classification problems [14]. Meaning, SVM is able to accomplish this concept: “input vectors are non-linearly mapped to a very high-dimension feature space. In this feature space a linear decision surface is constructed. Special properties of the decision surface ensure high generalization ability of the learning machine. The idea behind the support-vector-network was previously implemented for the restricted case where the training data can be separated without errors” [14]. SVM is a definite candidate to be able to efficiently classify whether or not a student entered term is considered "school related" or not. However, in order to train an SVM classifier, it requires training dataset because it is a supervised learning algorithm. Both Naïve Bayes and SVM are linear, efficient, and scalable on massive data sets, but the research in [7, 8] shows that SVM outperforms Naïve Bayes in large data sets when accuracy is considered for text categorization.

An SVM algorithm is constructed with the objective of discovering a separation between dimensional planes, which are demarcated by classes of data. So, SVM allows for the measurement of the variance of data as opposed to the finding the correspondence among types. Whereas Naïve Bayes classification algorithms are based on an assumption that the terms used in documents are independent.

In SVM, to get accuracy in large data sets, classifiers are selected. The classifier selection demands the tuning of several parameters which affect the generalization error. The process of fine tuning the parameters is termed as model selection [9]. With regards to machine learning, SVM becomes more accurate as it processes trained with more data; our correlation between query searches shows that non-linear SVM exceeds Naïve Bayes and linear SVM. So for our comparison SVM acts as a binary classifier to do multi-class classification, pair wise classification is used [16].

A student search query can either be school-related or non-school-related. Since SVM is well suited for binary classification, it is ideal for binary classification of student web queries as either school related or non-school related. Once each query is classified, a correlation test between student web queries and their GPA can be performed.

VI. BIG DATA PRIVACY & SECURITY CONCERNS

The Family Educational Rights and Privacy Act (FERPA) provides students with the ability to share info in certain cases without permission: student officials have an interest in the data determining financial aid, developing or administering tests that are predictive, administering student aid programs, local, state, and federal representatives conduct an audit of a school’s education program [25]. Under FERPA, a school may not unveil from educational records any of that student’s personal records to a third party. In order for the school to
disclose personal information, the student must deliver written approval for the school to do so.

Educational data mining is one of the major problems for students because of how the industry handles student data. “About 95% of school districts send student records to companies that manage student services such as Microsoft, Google, and others” [27]. The Protecting Student Privacy Act prohibits the mining of data by companies that includes student personal info in the interest of marketing. Student data should not be used and collected for commercial purposes, but students have no say with the companies that makes contracts with different school districts. Data mining can be useful for obtaining useful information, but companies need to provide students with a safe space so students can learn efficiently and not a way for corporations to learn about them. Private corporations are getting onboard with helping protect student data in accordance to the Student Privacy Pledge. This pledge is a solution for big data security for students; introduced by the Future of Privacy Forum and The Software & Information Industry Association [27]. These provisions are intended to protect student privacy in regard to the maintenance, collection, and the use of personal information of students. This pledge is intended to detail existing federal law regarding the handling and guarding of student data.

The Privacy Technical Assistance Center (PTAC), established by the United States Department of Education, further helps push the protection of student data. This center is a resource for education stakeholders to educate themselves on how they can enforce privacy, confidentiality, and security of data [26]. The PTAC offers information as well as assistance in regards to security practices by utilizing a menagerie of resources including: receiving data systems support, training materials, and providing observance with FERPA for fortifying the security of personal info.

The Privacy tool kit provides an up to date info guide for stakeholders who want to improve the privacy and security of educational data. Some of these resources included are checklists, online presentations, technical briefs, and a variety of other related materials. Topics covered within these resources include data sharing, data security, data governance, legal references, and disclosure avoidance. Local meeting forums and a support center are also provided [26]. A toolkit is also offered to Student Privacy & Data Security members (SIIA) to school providers and the sector with tools that allow the notification of their internal practices in addition to assist policy leaders, families of students, and teachers [28].

The Guidance and Technical Assistance workshop is a part of the SIIA tool kit. This workshop provides resources that inspect as well as explore new practices that improve data security and privacy. Also provided are tools to review and improve upon data security practices [28]. The laws and regulatory requirements, which include the State Student Privacy Law Reports, contains a section, which features a group of reports on security and privacy laws passed by their respective state legislatures. The reports generated each give a summary which guide educators as they interact with their local educational agencies [28].

VII. METHODOLOGY

This study includes web queries of 887 students that was collected over a two week period. Web filter logs of each day were extracted from a local school district web filter in the form of a Comma Separated Value (CSV) file. At the end of the two weeks, these CSV files were then merged together. The raw data pulled from this database comprised of approximately 148,502 SQ. Data contained in this merged file anonymized student identifiers for privacy and security reasons, see Table 1. Each student was given a unique unidentifiable id. A significant portion of this dataset contained duplicate search terms. This is caused from students clicking various links the were populated from the search query. An example would be: if a student typed into a search engine “in the heights” and then clicked on four links of the search results, then that search query would appear four times in the web filter log and eventually in the CSV file. The duplicate web queries were then removed leaving 40,404 unique SQ.

Next, a pivot table was created from this data. Using each student’s unique school log-in ID, the amount of that person’s SQ was mapped to them. In addition to this, from the school database, each student’s GPA was matched. A second data clean up is preformed, since not all of the original 887 students were able to be mapped due to an irregularity in the naming convention of the student log-in ID. For the most part student school log-in ID is created from the capitalized, first initial of the first name followed by the entire last name. In the event two students had the same first initial and last name, their ID did not follow that naming convention. These students were excluded from the dataset leaving the total sample size in this study to 455 students.

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>ID</td>
<td>SQ</td>
<td>Classification</td>
<td>Unique ID</td>
<td>Unique SQ</td>
<td>school related %</td>
</tr>
<tr>
<td>2</td>
<td>001</td>
<td>verbs</td>
<td>School</td>
<td>s001</td>
<td>47</td>
<td>14</td>
</tr>
<tr>
<td>3</td>
<td>001</td>
<td>large buy</td>
<td>Non School</td>
<td>s002</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>001</td>
<td>anthony robb</td>
<td>Non School</td>
<td>s003</td>
<td>51</td>
<td>24</td>
</tr>
<tr>
<td>5</td>
<td>001</td>
<td>rent movie</td>
<td>Non School</td>
<td>s004</td>
<td>58</td>
<td>12</td>
</tr>
<tr>
<td>6</td>
<td>001</td>
<td>illnesses</td>
<td>School</td>
<td>s005</td>
<td>19</td>
<td>15</td>
</tr>
<tr>
<td>7</td>
<td>001</td>
<td>aids</td>
<td>School</td>
<td>s006</td>
<td>21</td>
<td>16</td>
</tr>
<tr>
<td>8</td>
<td>001</td>
<td>mercy shawn menes</td>
<td>Non School</td>
<td>s007</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>9</td>
<td>001</td>
<td>justin breed</td>
<td>Non School</td>
<td>s008</td>
<td>44</td>
<td>11</td>
</tr>
<tr>
<td>10</td>
<td>001</td>
<td>a little fall of non lyrics</td>
<td>Non School</td>
<td>s009</td>
<td>19</td>
<td>7</td>
</tr>
<tr>
<td>11</td>
<td>001</td>
<td>on my own lyrics</td>
<td>Non School</td>
<td>s10</td>
<td>63</td>
<td>45</td>
</tr>
<tr>
<td>12</td>
<td>001</td>
<td>feel pretty lyrics</td>
<td>Non School</td>
<td>s11</td>
<td>49</td>
<td>33</td>
</tr>
<tr>
<td>13</td>
<td>001</td>
<td>take me or leave me lyrics</td>
<td>Non School</td>
<td>s12</td>
<td>55</td>
<td>11</td>
</tr>
<tr>
<td>14</td>
<td>001</td>
<td>take me or leave me rent</td>
<td>Non School</td>
<td>s13</td>
<td>55</td>
<td>30</td>
</tr>
<tr>
<td>15</td>
<td>002</td>
<td>translate</td>
<td>School</td>
<td>s14</td>
<td>75</td>
<td>11</td>
</tr>
<tr>
<td>16</td>
<td>002</td>
<td>davidson college test score</td>
<td>School</td>
<td>s15</td>
<td>91</td>
<td>42</td>
</tr>
<tr>
<td>17</td>
<td>002</td>
<td>davidson college</td>
<td>School</td>
<td>s16</td>
<td>43</td>
<td>23</td>
</tr>
<tr>
<td>18</td>
<td>002</td>
<td>maine schools</td>
<td>School</td>
<td>s17</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>19</td>
<td>002</td>
<td>democratic socialist of america</td>
<td>School</td>
<td>s18</td>
<td>56</td>
<td>40</td>
</tr>
<tr>
<td>20</td>
<td>002</td>
<td>john</td>
<td>Non School</td>
<td>s19</td>
<td>30</td>
<td>13</td>
</tr>
</tbody>
</table>

Table 2. Student-SQ Pivot Table
Each student web query in this dataset was then classified using the SWQC algorithm, as either school-related or non-school-related. A percentage of the amount of school-related SQ was formulated from the total amount of SQ and then mapped to that student’s GPA. With this percentage now known, a regression analysis is performed. A scatter–plot was created with GPA as the dependent variable on the y-axis and their percentage of school-related SQ as the independent variable on the x-axis.

VIII. RESULTS

At first glance, Fig. 1., shows a slight positive correlation between school-related web queries and higher GPA. The p-value from the regression analysis is significantly low rejecting the null hypothesis with $H_0$ the percentage of school-related web queries performed by high school students, originating from their iPads, has no impact on their GPA.

[Image: Initial scatter-plot graph showing slight positive correlation]

**Fig. 1.** Initial scatter-plot graph showing slight positive correlation

SQ was formulated from the total amount of SQ and then mapped to that student’s GPA. With this percentage now known, a regression analysis is performed. A scatter–plot was created with GPA as the dependent variable on the y-axis and their percentage of school-related SQ as the independent variable on the x-axis.

VIII. RESULTS

At first glance, Fig. 1., shows a slight positive correlation between school-related web queries and higher GPA. The p-value from the regression analysis is significantly low rejecting the null hypothesis with $H_0$ the percentage of school-related web queries performed by high school students, originating from their iPads, has no impact on their GPA.

![Initial scatter-plot graph showing slight positive correlation](image)

**Fig. 1.** Initial scatter-plot graph showing slight positive correlation

SQ was formulated from the total amount of SQ and then mapped to that student’s GPA. With this percentage now known, a regression analysis is performed. A scatter–plot was created with GPA as the dependent variable on the y-axis and their percentage of school-related SQ as the independent variable on the x-axis.

VIII. RESULTS

At first glance, Fig. 1., shows a slight positive correlation between school-related web queries and higher GPA. The p-value from the regression analysis is significantly low rejecting the null hypothesis with $H_0$ the percentage of school-related web queries performed by high school students, originating from their iPads, has no impact on their GPA.

With the threshold value now included in the regression model, a more defined positive correlation is observed in table 3. GPA (y-intercept) decreased from ~86 to ~79 and average amount of student SQ (slope) dramatically increased from ~5 to ~20.8. The new linear function now is $y = 19.793x + 79.584$. For approximately every 10% increase of student SQ being school-related, student’s GPA increases by 2 point. For example, if student SQ goes from 0% to 20%, $y = 19.793 \times 0.2 + 79.584 = 83.542$ i.e. ~4 points higher then would be expected if no SQ were performed while studying. Furthermore, the p-value dropped significantly and there was a significant increase in the R-squared value. Albeit, the R-squared value is roughly 25%, it can be concluded that percentage of school related search queries cannot be used as a determining factor for higher GPA. There are other discriminatory variables from performance data such as test scores, attendance, homework completion which play a greater role in GPA. However, based on the analysis presented here, it can be further concluded that percentage of school related school queries originating from school provided mobile devices can play a role significant role in student learning.

IX. CONCLUSION

The regression model was tweaked to include a threshold value which would implement a cutoff of how few of SQ a student can have in order to be included in the model. The number initially chosen was 10 SQ, resulted in slightly better results. Tests were continually performed at thresholds of increasing intervals of 10 SQ, until finally an optimal threshold of 70 SQ was reached. This returned the best results while maintaining a relatively high amount of observations, ~13% of the 455 students used in the initial regression analysis.

### SUMMARY OUTPUT

<table>
<thead>
<tr>
<th>Regression Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multiple R 0.498863795</td>
</tr>
<tr>
<td>R Square 0.031992308</td>
</tr>
<tr>
<td>Adjusted R Square 0.02985416</td>
</tr>
<tr>
<td>Standard Error 6.395862175</td>
</tr>
<tr>
<td>Observations 455</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ANOVA</th>
</tr>
</thead>
<tbody>
<tr>
<td>SS 78.871672</td>
</tr>
<tr>
<td>F 0.00125167</td>
</tr>
<tr>
<td>Significance F 4.108111-05</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>REGRESSION ANALYSIS WITH STRONG POSITIVE CORRELATION</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coefficients</td>
</tr>
<tr>
<td>Intercept 79.584446372</td>
</tr>
<tr>
<td>xvariable1 19.790872079</td>
</tr>
</tbody>
</table>

Table 4. Fixed Regression Analysis with Strong Positive Correlation

However, digging into the regression analysis, the R-Square value at ~0.03 is closer to 0 than 1, see table 3 The R-Square value is calculated by explaining the percentage of variation in the dependant variable (GPA) that is part of the regression line with the independent variable as the explanatory variable. Simply put: how well can you explain the variance in GPA given the predictor variable which is the percentage of SQ? Thus, the conclusion to be drawn from this initial result is that there is no real impact on GPA. However, this model included all students whether they performed one search query or 100+ search queries within this two-week period. An example would be a student with a 97 GPA having only 1 SQ. These outliers threw off the resulting regression analysis.

The regression model was tweaked to include a threshold value which would implement a cutoff of how few of SQ a student can have in order to be included in the model. The number initially chosen was 10 SQ, resulted in slightly better results. Tests were continually performed at thresholds of increasing intervals of 10 SQ, until finally an optimal threshold of 70 SQ was reached. This returned the best results while maintaining a relatively high amount of observations, ~13% of the 455 students used in the initial regression analysis.
IX. Future Work

The research presented here only consist data from first two weeks of school. With school year just underway, the total number of search queries performed by students is quite small. This model should be tested further with larger student data as the school year progresses.

Only 455 students were mapped to their userID, leaving other half of the data unused. Future study should include the remaining students once userID mapping is complete. This should provide more data to see whether this model holds true.

Finally, the question of the safety and ethical ramifications of having access to student’s (of which are minor) online data. What measures are in place to prevent student data from being exploited or stolen? What standards are instructors held to concerning their access to student information? Can parents opt out of mobile device learning, and will doing so put their children at a disadvantage? This work provides ground breaking research to under and improve efficacy of mobile devices as the standard in the educational system. Information presented in this study provides a glimpse into student learning that performance data does not provide. Additionally, as the use of mobile devices continue to grow, better monetary decisions can be made by school administrators. For example, teacher professional development may include how to maximize the utility of mobile devices in a classroom setting, restructuring curriculum to teach students how to effectively use school provided mobile devices thereby addressing the need of digital citizenship.

REFERENCES


[22] Doan, Kim, Bloomfield, Aaron, (2014) “The effects of browse time on the Internet on students’ essay scores”, West Chester University, University of Virginia

[31] Mehrnoosh Vahdat et al., ESANN 2015 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, Bruges (Belgium), 22-24 April 2015.