# Identification of Focused Influential Groups for Social Media Marketing using Big Data Analysis Methodologies

Riva Malik
*MS(DS)*
*FAST NUCES*
Islamabad , Pakistan
i201208@nu.edu.pk

Zarah Amer
*MS(DS)*
*FAST NUCES*
Islamabad, Pakistan
i202201@nu.edu.pk

***Abstract*—This research paper aims to apply Big Data Analytics to find insights from data and use them for effective marketing on social media platforms. These insights prove to be an important asset in understanding the audience and approaching them by personalised campaigns, targeted marketing and by making entertaining content according to users' interests. The approach is to obtain social media data and apply Big Data Analysis methodologies to discover insights from it. Using this approach we are able to identify numerous influential user groups within data based on age, area, education, interests, spoken languages, user activity on platform and many more. These focused groups can be treated as individual niches for promotional and marketing strategies used by different businesses.**

***Index Terms*—Social Media Marketing, Consumer Exploratory Insights, Big Data Analytic Methodologies, Big Data, Frequent Item Sets, Rare Item Sets, Data Analysis.**

## I. Introduction

THE rapid and increased adoption of social media platforms in our daily lives has opened doors to many innovative applications of them. One of such application is the use of Social Media as a tool for marketing, making it a source of multiple opportunities in online setting. It includes the use to social media platforms like Facebook, Twitter, Instagram, Linked-In to connect with your audience, spread awareness about some issue, build up you brand, drive traffic to your profile,publish interesting content to engage followers, increase sales and grow your business.

Numerous devices and platforms have drastically increased the data production as well. Thus data management and data analysis require smart approaches in order to support the big data revolution. Methodologies for Big Data Analytics are being introduced to aid in dramatic improvement in veracity and velocity of solving the problems.

With the emergence of Big Data and the revolutionary havoc in internet usage, Social Media Marketing with combination of Big data can reach an altogether new level. Companies and businesses can utilize this data in multiple ways. They can learn about customers and approach them personally according to their choices, likes and dislikes. The analysis can give in-depth insights of different user groups as well as their emotional, physical and social well-being. It can become convenient for marketers to understand their audience and strengthen their relationship, by creating personalized content for them. This will lead to elevated consumer trust and gain in business. [10]

In this research paper we have used different algorithms for better understanding of data and collecting insights from the data. The algorithm used are as follows:

- Exploratory Data Analysis for understanding variable and their relationships. [1]
- Apriori for extracting frequent pairs, items occurring in combination and their frequency. [4]
- PCY is used to mine frequent patterns using hash functions for memory and time saving. [6]
- FP Growth is used for even faster computation, better memory usage and optimal solutions. [7]
- For dealing with large number of streaming data we have used Apache Kafka which is a framework popularly applied in processing streaming data.
- Apriori Inverse is used to extract irregular rules by ignoring all candidate itemsets above threshold. [14]

Like Apriori Algorithm that extracts frequent patterns, we propose another algorithm to extract rare patterns. In this algorithm the minimum support is set very low which produces a large number of trivial frequent item sets. With the help of which we discover rare patterns by ignoring all candidate item sets above maximum support threshold. Rare patterns are the ones that appear in the data below the user defined threshold, which are said to be of no interest. However, we have seen that using that rare group of people, we can gain more insights and more knowledge. [15]

Sporadic Rules are the ones that fall in the range of user-defined maximum and minimum support. Further there are two types of sporadic rules;

- Perfectly Sporadic Rules **-** They have no subset above maximum support.They are the ones that rarely occur.

- Imperfectly Sporadic Rules **-** They have subset above maximum support.

Apriori Inverse finds all Perfectly Sporadic Rules since we have totally inverted the concept of Apriori by excluding all frequent items and extracting all rare items using maximum support and minimum confidence. [14]

To illustrate Rare Pattern Mining and Frequent Pattern Mining we took a dummy data of first 15 records to see the difference between the algorithms.

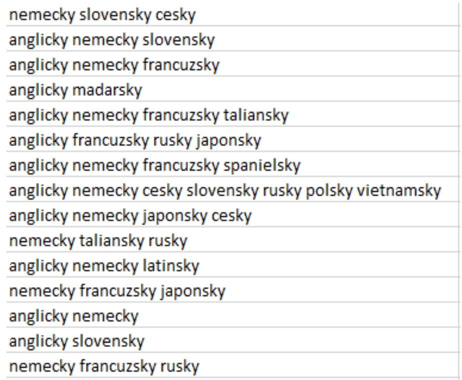

nemecky slovensky cesky
anglicky nemecky slovensky
anglicky nemecky francuzsky
anglicky madarsky
anglicky nemecky francuzsky taliansky
anglicky francuzsky rusky japonsky
anglicky nemecky francuzsky spanielsky
anglicky nemecky cesky slovensky rusky polsky vietnamsky
anglicky nemecky japonsky cesky
nemecky taliansky rusky
anglicky nemecky latinsky
nemecky francuzsky japonsky
anglicky nemecky
anglicky slovensky
nemecky francuzsky rusky

Fig. 1. Dummy Data for testing Apriori and Apriori inverse Algorithms

Now when we run both Apriori and Apriori Inverse Algorithm on the dummy data. We get the following results;

In Apriori Algorithm. Showing Frequent languages spoken, from the data of languages spoken in Slovakia and the support threshold. The **number of candidates** Apriori counted were **100**, where as the **number of Frequent Itemsets** Apriori showed were **36**. The **memory and time used** by Apriori Inverse is **8.43mbs** and **16ms** respectively. And stopped at itemset size 5.

In Apriori Inverse Algorithm. Showing Rare languages spoken, from the data of languages spoken in Slovakia and the support threshold. The **number of candidates** Apriori Inverse counted were **80**, where as the **number of Rare Itemsets** Apriori Inverse showed were **23**. The **memory and time used** by Apriori Inverse is **8.42mbs** and **7ms** respectively. And stopped at itemset size 4 because there are no candidates.

"Table. I" shows the patterns of Apriori and Apriori Inverse. The first half of the table shows the intersecting patterns/common patterns of Apriori and Apriori Inverse, and other half is non-common patterns.

Comparing the two algorithms; Apriori and Apriori Inverse, we deduce that if we want Frequent items, then there is no perfect fit other than Apriori. Where as, it is not a good fit for rare item because Apriori includes perfect+imperfect patterns. Where as Apriori Inverse extracts perfectly rare patterns by saving memory and computational time.

TABLE I
APRIORI ALGORITHM AND APRIORI INVERSE INTERSECTION ON DUMMY DATA

| ***Pattern Apriori*** | *Support* | ***Pattern Apriori Inverse*** | *Support* |
|---|---|---|---|
| *anglicky, francuzsky* | *1* | *anglicky, francuzsky* | *1* |
| *anglicky, madarsky* | *1* | *anglicky, madarsky* | *1* |
| *anglicky, rusky* | *1* | *anglicky, rusky* | *1* |
| *anglicky, slovensky* | *3* | *anglicky, slovensky* | *3* |
| *anglicky, spanielsky* | *1* | *anglicky, spanielsky* | *1* |
| *anglicky, taliansky* | *1* | *anglicky, taliansky* | *1* |
| *anglicky, vietnamsky* | *1* | *anglicky, vietnamsky* | *1* |
| *francuzsky, rusky* | *1* | *francuzsky, rusky* | *1* |
| *slovensky, vietnamsky* | *1* | *slovensky, vietnamsky* | *1* |
| *anglicky,francuzski,ruski* | *1* | *anglicky,francuzski,ruski* | *1* |
| *anglicky,slovensky,vietnamsky* | *1* | *anglicky,slovensky,vitnamsky* | *1* |
| *nemecky, slovensky, vietnamsky* | *1* | | |
| *anglcky,nmicky,slovnsky,vtnamsky* | *1* | | |
| *anglicky, nemecky* | *8* | | |
| *nemecky, rusky* | *1* | | |
| *nemecky, taliansky* | *2* | | |
| *nemecky, vietnamsky* | *1* | | |
| *anglicky, nemecky, vietnamsky* | *1* | | |
| *anglicky, nemecky, slovensky* | *2* | | |
| *anglicky, nemecky, spanielsky* | *1* | | |
| *anglicky, nemecky, taliansky* | *1* | | |
| *nemecky, slovensky* | *3* | | |
| *nemecky, spanielsky* | *1* | | |

The efficiency of Apriori Inverse is measured in terms of the Perfect Item Count, which is much lesser and accurate in Apriori inverse, time used is 2times lesser than Apriori and memory usage is also comparatively lesser than Apriori.

## II. Body

### 1. Exploratory Data Analysis

Exploratory Data Analysis **-** Analyse and Summarize the data sets. Exploratory Data Analysis aids in giving a better understanding of dataset, its variables and their relationships. [1].

Following are the usage of EDA: [2]

- Get answers by manipulating resources of data.
- Discover patterns with the help of desired answers.
- Identify relationships between the variables.
- Spot main and basic errors.
- Detect outliers.
- Identify anomalies.
- Hypothesis Testing.
- Testing the assumptions we made, if they are true or not.

The processed information is now to be displayed, for that there are 4 types of EDA representations:

[3] **1. Uni-variate Non-graphical -** It consist of single variable that assists in data description and pattern detection.

**2. Multi-variate Non-graphical -** It consist of multiple variables and shows relationships between two or more variables.

**3. Uni-variate Graphical -** This representation shows graphs using single variable. Such as histogram, stem-and-leaf-graph, box plot.

**4. Multi-variate Graphical -**This representation uses multiple variables to form understandable graphs elaborating relationships between multiple sets of data.

### 2. Frequent Pattern Mining

Frequent Pattern Mining **-** Data analyst comes across many uncertain, inappropriate and imprecise data. However, mining frequent patterns from that data is a popular data science task. [4]

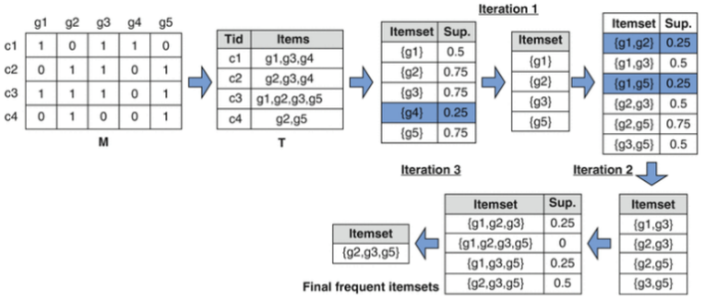
**1. Apriori Algorithm**

This algorithm supports the concept of monotonicity. That is, if item sets I appears S threshold times then it belongs to frequent pairs. If not, then it is not a frequent pair. Below are the list of concepts for understanding Apriori algorithm:

**1. Items and Transaction -** The number of elements in the data set are called items and the set of subsets of items are called transactions. Each transaction identifies items' subsets.

**2. Frequent Items -** A support threshold is set, and if the frequency of the set of items equals or exceeds the support threshold, it is called frequent item.

**3. Support -** Number of transactions in which the item sets are there. [5]



|    | g1 | g2 | g3 | g4 | g5 |
|----|----|----|----|----|----|
| c1 | 1 | 0 | 1 | 1 | 0 |
| c2 | 0 | 1 | 1 | 0 | 1 |
| c3 | 1 | 1 | 1 | 0 | 1 |
| c4 | 0 | 1 | 0 | 0 | 1 |

M

| Tid | Items |
|-----|-------|
| c1 | g1,g3,g4 |
| c2 | g2,g3,g4 |
| c3 | g1,g2,g3,g5 |
| c4 | g2,g5 |

T

Iteration 1

| Itemset | Sup. |
|---------|------|
| {g1} | 0.5 |
| {g2} | 0.75 |
| {g3} | 0.75 |
| {g4} | 0.25 |
| {g5} | 0.75 |

| Itemset |
|---------|
| {g1} |
| {g2} |
| {g3} |
| {g5} |

| Itemset | Sup. |
|---------|------|
| {g1,g2} | 0.25 |
| {g1,g3} | 0.5 |
| {g1,g5} | 0.25 |
| {g2,g3} | 0.5 |
| {g2,g5} | 0.75 |
| {g3,g5} | 0.5 |

Iteration 2

| Itemset |
|---------|
| {g1,g3} |
| {g2,g3} |
| {g2,g5} |
| {g3,g5} |

Iteration 3

| Itemset | Sup. |
|---------|------|
| {g1,g2,g3} | 0.25 |
| {g1,g2,g3,g5} | 0 |
| {g1,g3,g5} | 0.25 |
| {g2,g3,g5} | 0.5 |

| Itemset |
|---------|
| {g2,g3,g5} |

Final frequent itemsets

Fig. 2. Frequent Pattern Mining algorithm [5]

**2. PCY Algorithm**

PCY algorithm is the extension and a smarter version of Apriori algorithm. In Apriori algorithm memory slots in he main memory were idle as they could be used and two steps could be done simultaneously. In pass 1, most the memory was idle so PCY made use of that idle memory to keep count of the buckets in which pairs of items are hashed. So item count and bucket count are done simultaneously in one pass. In pass 2 it gives a condition that the candidate pairs must satisfy. PCY aids in faster computation, better memory usage and optimal solution due to hash functions. [6]

PCY algorithm is the hash-based improvement of Apriori. The standard model of association rule hconsists of ”items” and ”baskets” where the basket contains purchased items. We want to extract frequent items. [17]

**3. FP Growth Algorithm**

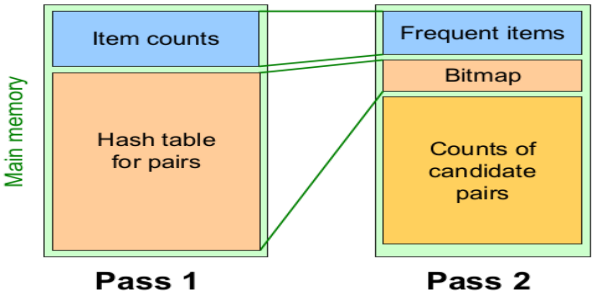
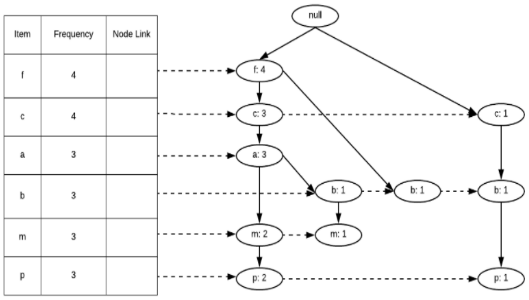


Fig. 3. Memory utilization in PCY algorithm [6]

Frequent item mining is used to discover co-occurring items. FP Growth is currently one of the fastest approach of mining frequent items. FP Growth Algorithm supports ”Divide and Conquer” approach. this is a smarter version of PCY algorithm. It aids in even faster computation, better memory usage and optimal solutions. It represent database in a tree-form called frequent pattern tree. Unlike Apriori this algorithm needs only two scans because data is stored in compact version and paring is not done, so memory and computation time is saved. [7]



| Item | Frequency | Node Link |
|------|-----------|-----------|
| f | 4 | |
| c | 4 | |
| a | 3 | |
| b | 3 | |
| m | 3 | |
| p | 3 | |

Fig. 4. FP Growth algorithm [7]

### 3. Rare Pattern Mining

Rare Pattern Mining **-** In this algorithm we define irregular and not-common rules with low support and high confidence. [14]

**1. Apriori Inverse Algorithm**

In the light of the algorithm ”Apriori Algorithm” that extracts frequent patterns, we propose another algorithm to extract rare patterns. In this algorithm the minimum support is set very low which produces a large number of trivial frequent item sets. With the help of which we discover rare patterns by ignoring all candidate item sets above maximum support threshold. Rare patterns are the ones that appear in the data below the user defined threshold, which are said to be of no interest. However, we have seen that using that rare group of people, we can gain more insights and more knowledge. [15]

## III. Literature Review

### A. Social Media Marketing and how it started

Back in circa 2004, social media led to social media marketing. It was a new step to a totally new domain which extended its roots as much as it has become a basic necessity now for businesses, promotions, and daily routine tasks like buying and selling. Mainly businesses use it to target the set of audience and bring traffic to their content, which benefits the audience by being treated as per their interests, likes and dislikes. [8]

Automation was not the only aspect that turned the marketing upside down, but back in 2004 a Harvard student made a website in his room ”The Facebook”. He didn’t intended to make it public and only emails with ”.edu” extension could join. Later it expanded and was made available worldwide, without any age, gender, work limitations. Then along with Facebook other social networks emerged like twitter, Instagram, Linked-In etc..

So automation along with social media sites and connections raise the concept of Social Media Marketing. [9]

### B. Big Data and how it started

The term Big Data was coined in circa 2005 which is referred as large amount of data, nearly impossible to manage by BI tools. After the advent of social sites, the data kept on increasing at an immense rate. So there emerged a need to manage that data. This was the time when ”Hadoop” was created by Yahoo! built on Google MpaReduce for streaming and continues data. [9]
In the past decade there has been an enormous increase in Big Data startups. All the data from devices, applications is coming in the form of stream which need to be organised using Big Data analytic to get insights, relationships, patterns and deeper understanding of our audience interests and performance.
**Let The Big Data Era Begin!**

### C. Big Data and Social Media Marketing

One of the most apparent and personal way in which Big Data combined with Social Media affects our lives is personalized advertisement, media, music, shows, movies and entertainment that you consume daily. Basically data plays such an important role in social media directly related to our lives that it has become impossible to point out things that does not involve the both of them. [10]
McKinsey says ”Brands that have used customer analytics have seen a mega increase of **126% profit** improvement over their competitors. [10]
Below are a few ways Big Data is used to improve Social Media Marketing:

- Cloohawk - A Social Media Marketing tool, collects your data, apply marketing strategies, understand your data, collects insights and then prompts you to follow your interests and build a relationship with like-minded group of people. [11]
- Social Mention - A Social Media Marketing tool, studies tone of the customer, likes, dislikes, emotions, shares, hashtags and then Sentiment Analysis is done, dividing like-minded people further into specific groups. [11]
- Shortstack - A Social Media Marketing tool, targets audience based on groups of gender, interest, profession, likes and shopping habits to set up personalised campaigns. [11]
- Social Blade - When you aim to collect data about competitors Social Blade is used. It gets your competitor brand’s frequency of publishing, messages, customer’s sentiments. This data is used to compare your growth with competitors and collect weaknesses of customers to gain traffic. [11]

Big Data is using data to compete with their competitors across industry. relying solely on Big Data, won’t make you win the competitor race. Big Data combined with Social Media Marketing plays a tough competition. [12]

## IV. Experiments

### A. Dataset Collection

Data collection is the process of gathering information of interest, in an established systematic fashion that enables one to answer stated research questions and evaluate outcomes. In our research we have used Pokec Social Network Dataset [13]. Pokec is a social networking platform popular in Slovakia with 1.6 million users. It is publicly available under Stanford Network Analysis Project, which they have obtained through web crawling in May 2012. The dataset has information about user profiles and user relationships among each other. We have only focused on user profile data. [12]

### B. Preliminary Dataset Exploration

User profile dataset in its raw form contains 1632803 records and 59 features. Upon inspection of different columns individually it is found that features in “Table. II” have no missing values.

TABLE II
Features without Null Values

| *public* | *user-id* | *completion-percentage* |
|---|---|---|

Meanwhile features in “Table. III” contain garbage values.

TABLE III
Features with Garbage Values

| | | | |
|---|---|---|---|
| *fun* | *politics* | *music* | *cars* |
| *relationships* | *movies* | *education* | *sports* |
| *travelling* | *health* | *computers internet* | *life-style* |
| *art-culture* | *hobbies-interest* | *science-technologies* | *companies-brands* |

Numeric feature like age is also analyzed producing the distribution in “Fig. 5”

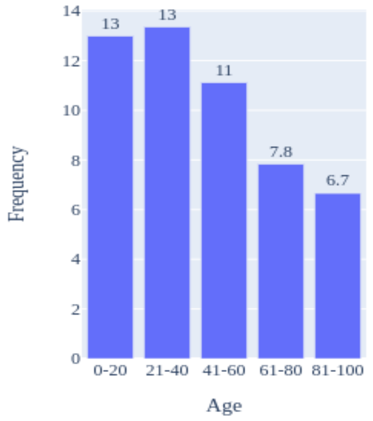


Fig. 5. Frequency Distribution of Age Feature

Age contains outliers in form of users having age 0 and age above 100. Apart from that random values from different features are extracted to get the intuition of the feature quality.

### C. Data Cleaning

Data cleaning plays a significant role in the overall effectiveness of results in any data analysis task. Clean data can ensure that the outcome of the algorithm is correct and optimal, on the basis of which we can make right decisions.

We have performed several data cleansing activities to bring our data in desired form. First we performed dimensionality reduction and removed those features which are not interesting for the analysis we are going to perform. We have also removed all the features which contain garbage values.

For dealing with outliers in age feature we have entirely removed users with age greater than equal to 100 and replaced age for users having age equal to 0 with median age.

Features having unstructured text data are stripped off of punctuation, extra spaces, symbols and emoticons. Completed-Level-Of-Education feature is categorised into four distinct values. Region feature is dissolved into region and cities as two separate features rather than one.

Last-login feature having date datatype is converted into categorical variable. Users who have logged in before 2012 are categorised as inactive while other are labelled as active users.

The feature completion-percentage having numerical data, represents the number of features having non null values with respect to all the features for each record. We have used this feature to decide which rows can be dropped without losing too much information. Records having completion-percentage less than 60 are dropped from the dataset.

### D. Exploratory data analysis

Elementary insights from the data are obtained by performing exploratory data analysis on combination of several features.

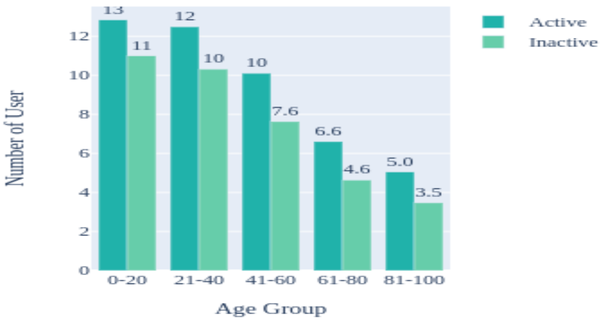


Fig. 6. Distribution of Active and Inactive Users in Different Age Groups

“Fig. 6” shows the distribution of active and inactive users in different age groups. Every age group has greater number of active users as compared to inactive users. Age group 81-100 has the least difference between active and inactive users.

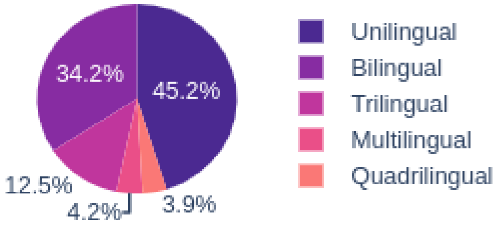


Fig. 7. Distribution of Multi-Lingualism Among Pokec Users

“Fig. 7” shows how the users of Pokec are distributed according to the number of languages they speak. Majority of users are unilingual that means they speak only one language. Then comes users who speak two languages. Least number of users speak upto four languages.

“Fig. 8” displays the top 5 most popular languages spoken by users.

These insights help us in identification of diverse user groups which act like distinct targets of social media marketing.

### E. Frequent Pattern Mining

In Frequent Pattern Mining, we used three algorithms;

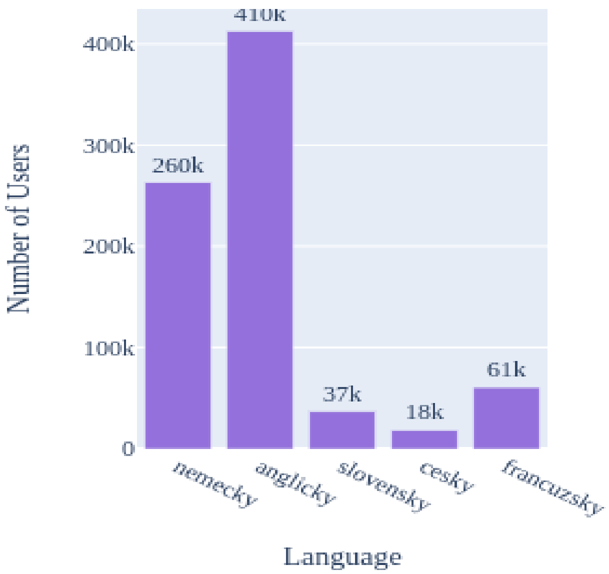


Fig. 8. Top 5 Spoken Languages Among Pokec Users

- A-Priori Algorithm.
- FP Growth.
- PCY.

### 1. A-Priori Algorithm.

For A-Priori Algorithm, we first apply Association Rule and extract co-relation. the metric used in Association Rule is ”Lift”. The algorithm of association rule returns the lift value itself. The higher the value of lift, the stronger the co-relation. the lower the value, the weaker the co-relation. The support threshold used in this algorithm is 0.01.
Below are the visualizations using A-Priori Algorithm:
In the above visualization, y-axis is the increasing lift value, x-axis is the different education levels and the bars represent different age groups in the bins of 20s.
We conclude that in the age group 0-20 majority level of education is ”Basic”, which is justified keeping in view the age group. And it serves the maximum lift-value so it has the strongest co-relation.
On the other hand we conclude that in the age group 20-40, majority are university going, and a few are serving apprentice; meaning running their own small business for income. This claim is also justifies seeing the lift value, as university going group serves larger lift value and stronger co-relation.

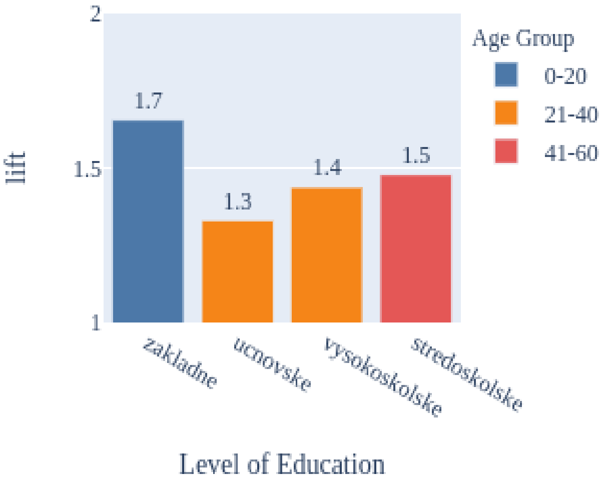


Fig. 9. Correlation Analysis Between Education and Age - Using A-Priori Algorithm

The interesting thing we come across after visualization is that the high-school age group is 41-60. And it bears a high value of lift and a stronger co-relation. So we get to know that the maximum level of education in that era was only high-school. So the maximum education one could get is the high-school. And the value of co-relation justifies this claim.

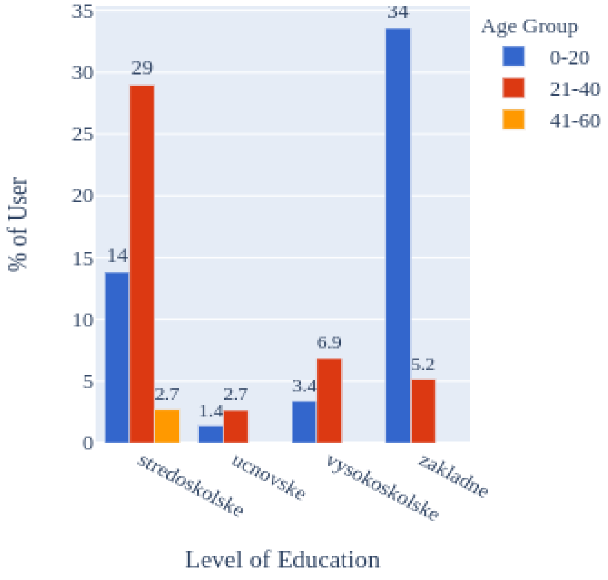


Fig. 10. Pattern of Education among Different Age Groups - Using A-Priori Algorithm

In the above visualization, y-axis represent the percentage of users, x-axis is the different education levels and the bars represent different age groups in the bins of 20s.

In this visualization, we conclude that in the are ”Zaklandne” 33% users belong to ”basic” level of education. and 5% belongs to high-school. Similarly in the are ”vysokoskolske” 6.8% users belong to high-school and 4% belongs to basic level of education.

In the above visualization, the focused groups are region-wise education level. We can use these in our social media marketing for audience targeting and organizing personalized campaigns.

### 2. FP Growth Algorithm.

In FP Growth Algorithm, we extracted frequent patterns by using support threshold 0.01 and we got ”spacial patterns” with the help of these spacial patterns we got to know the region-wise social media users of Slovakia. Below are the visualizations using FP Growth Algorithm:

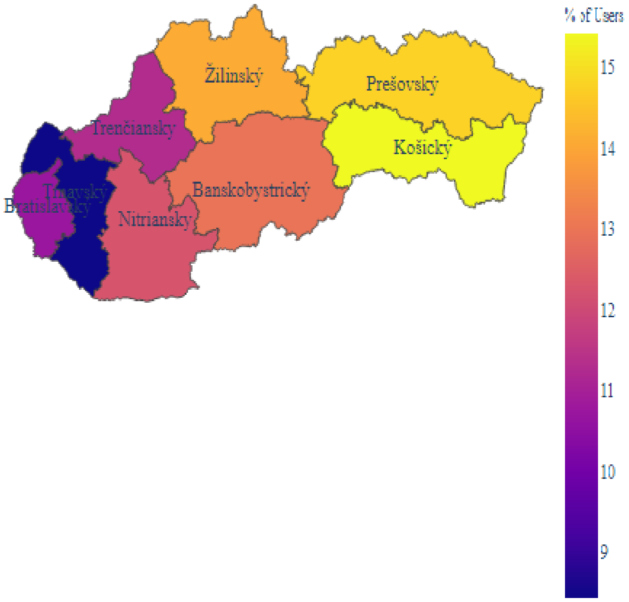


Fig. 11. Region wise Pattern of Users - Using FP GROWTH Algorithm

In “Fig. 11” we can tell about concentration of users according to area. We don’t see a symmetry that larger the area, greater the number of users. But we can see that there can be more number of users in a small area, compared to the number of users in a large area. Number of users are not directly proportional.

“Fig. 12” shows the most frequent age group in different regions. We get the most frequent age group using support threshold value. The region having highest support bears the most interesting pattern. This visualization says that the 0-20 are group is the highest in most of the areas and 21-40 age group is concentrated in only one specific area.

In “Fig. 13”, y-axis represent the percentage of users, x-axis are the regions and the colors of bars represent different age groups in the bins of 20s.

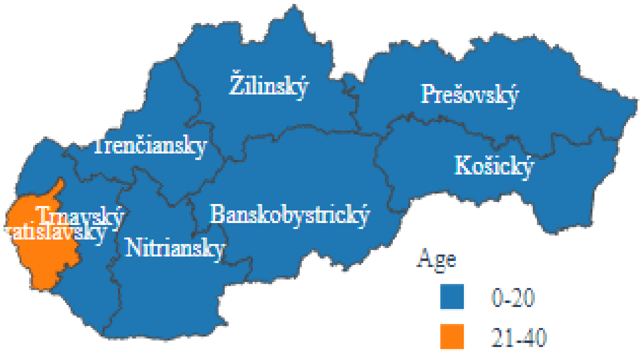


Fig. 12. Regional most Frequent age group Pattern- Using FP Growth Algorithm

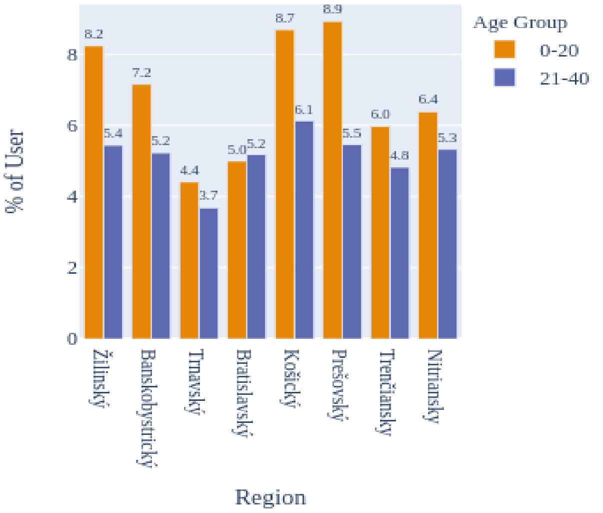


Fig. 13. Region wise Frequent Pattern of Different Age Groups- Using FP Growth Algorithm

This visualization says about the age groups present in different regions and frequent age groups are determined by number of users in percentage.In the region ”Zilinsky” 8.2% out of 10% are in the age group 0-20and 5.4% out of 10% are in the age group 21-40.

In the above visualizations, the data constraints and focused groups are region-wise Frequent Pattern of Different Age Groups. We can use these in our social media marketing for audience targeting and organizing personalized campaigns.

### 3. PCY Algorithm.

In PCY we extract constraint based patterns using support threshold 0.03. Below are the visualizations using PCY Algorithm:

“Fig. 14” explains the most popular education level in particular regions. In the city Nitriansky the most popular

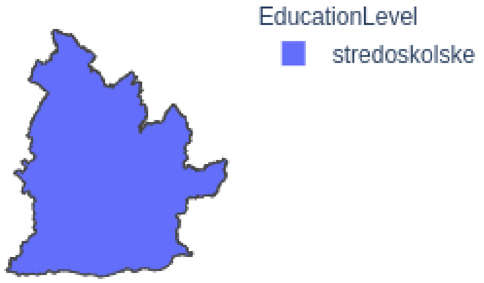


Fig. 14. Frequent Education Pattern in Nitriansky Region- Using PCY Algorithm

education level is ”zakladne” translation: ”Basic” level of education. We got this using threshold 0.03.

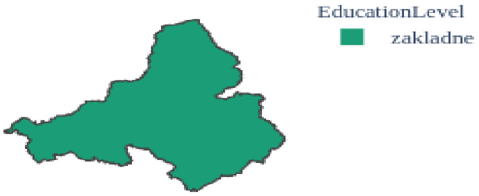


Fig. 15. Frequent Education Pattern in Trenčiansky Region- Using PCY Algorithm

“Fig. 15” explains the most popular education level in particular regions. In the city Trenčiansky the most popular education level is ”stredoskolske” translation: ”high-school” level of education. We got this using threshold 0.03.

In the above visualizations, the data constraints and focused groups are region-wise Frequent Pattern of Education Levels. We can use these in our social media marketing for audience targeting and organizing personalized campaigns.

### *F. Rare Pattern Mining*

In Rare Pattern Mining, we used one algorithm;

- A-Priori Inverse Algorithm.

Apriori Inverse is used to extract irregular rules by ignoring all candidate itemsets above maximum support threshold. Rare patterns are the ones that appear in the data below the user defined threshold, which are said to be of no interest. However, we have seen that using that rare group of people, we can gain more insights and more knowledge by gaining interests of people with different opinions that were being overlooked by only using frequent patterns. [14]

#### 4. A-Priori Inverse Algorithm.

In the light of the algorithm ”Apriori Algorithm” that extracts frequent patterns, we propose another algorithm to extract rare patterns. In this algorithm the minimum support is set very low which produces a large number of trivial frequent item sets. And then Apriori Inverse is used to extract irregular rules by ignoring all candidate itemsets above maximum support threshold. [15]

The methodologies used in rare pattern mining gives a brief explanation of different methodologies for mining rare patterns followed by experimental analysis. [16] A comparison between number of attempts based on rare and frequent pattern mining techniques is illustrated in “Table. IV”

Sporadic Rules are the ones that fall in the range of user-defined maximum and minimum support. Further there are two types of sporadic rules;

- Perfectly Sporadic Rules **-** They have no subset above maximum support.They are the ones that rarely occur.
- Imperfectly Sporadic Rules **-** They have subset above maximum support.

Apriori Inverse finds all Perfectly Sporadic Rules since we have totally inverted the concept of Apriori by excluding all frequent items and extracting all rare items using maximum support and minimum confidence. [14]

“Table. IV” shows only the rare languages spoken, from the data of languages spoken in Slovakia and the support threshold.

### *G. Data Processing with Apache Kafka*

Upto this point we have performed analysis to gain insights on categorical and numerical features. The dataset contains numerous features that have unstructured textual information that represents users’ opinion on respective topics. If utilized we can gain rich insights from them. But processing textual information is a compute intensive task. For that purpose we have used Apache Kafka which is a framework popularly applied in processing streaming data.

#### 5. Streaming Data using Apache Kafka.

We have focused on a single textual feature that is I-like-watching-movies that describes the preferred setting in which the user like to watch movie. We have divided that data into chunks that are supplied as stream of data to our

TABLE IV
APRIORI INVERSE ALGORITHM ON FULL DATA

| Patterns | Support |
|---|---|
| cesky, slovensky | 219 |
| francuzsky, nemecky | 148 |
| francuzsky, rusky | 328 |
| francuzsky, slovensky | 127 |
| francuzsky, spanielsky | 293 |
| francuzsky, taliansky | 309 |
| japonsky, rusky | 122 |
| japonsky, taliansky | 203 |
| madarsky, slovensky | 261 |
| nemecky, polsky | 622 |
| nemecky, rusky | 3050 |
| nemecky, slovensky | 1578 |
| nemecky, spanielsky | 1491 |
| nemecky, svoj | 172 |
| nemecky, taliansky | 3295 |
| polsky, rusky | 194 |
| polsky, spanielsky | 102 |
| rusky, slovensky | 106 |
| rusky, spanielsky | 312 |
| rusky, taliansky | 490 |
| spanielsky, taliansky | 268 |
| nemecky, rusky, spanielsky | 202 |
| nemecky, rusky, taliansky | 119 |
| nemecky, spanielsky, taliansky | 140 |

algorithm in Kafka. This way an intensive task is broken down manageable process.

The first algorithm we have applied is Edit Distance, used to find the difference between words or sentences on the basis of edit distance between them.

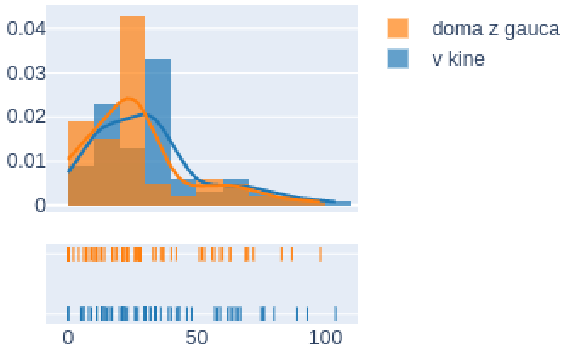


Fig. 16. Frequency Distribution of Edit Distances of User Opinions

"Fig. 16" shows contrast of two different frequency distributions. Each distribution represents different ranges of edit distance in which other user opinions lie with respect to specific user opinion about movies setting preferences. the spread of frequencies of edit distances. The range in orange shows that most opinions are 20-29 edit distance away from opinion At Home and the range in blue shows that most opinions are 30-39 edit distance away from opinion In Cinema.

Calculating edit distance of each user opinion against every other is computationally expensive. In order to improve our implementation, we can First calculate Jaccard Similarity among the users to create groups of similar users. After that we can calculate the edit distance for similar users only.

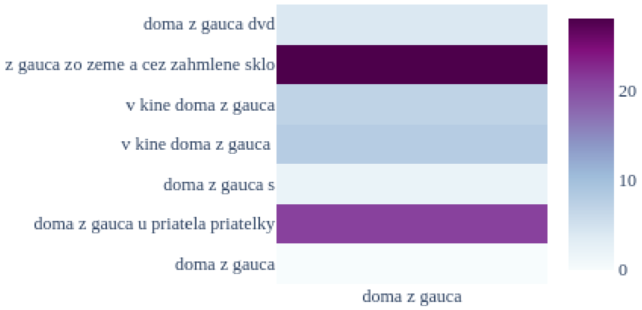


Fig. 17. HeatMap of Edit Distances of User Opinions

"Fig. 17" shows the plot of edit distances of all other opinions of similar users against opinion doma z guaca. The most distant similar user opinion can be seen with the darkest color.

## V. CONCLUSION

Social media platforms are one of the largest sources of data. This data is rich in information that can be used in variety of ways, but processing and analysing this data is an intensive task. Big Data Analysis methodologies are applied to handle such data. We have used social media data and applied Big Data Analysis methodologies on it to find insights. These insights have helped us in dividing users into groups based on similar interests, regions, education, spoken languages and user activity. After grouping users, we can approach each group individually and entertain them according to their interests. This helps boost businesses by showing people what they want to see and by organizing personalized campaigns to targeted audience.The insights we have collected are on an high level. In future they can be narrowed down to granular level.

## REFERENCES

[1] J. H. Friedman and J. W. Tukey, "A Projection Pursuit Algorithm for Exploratory Data Analysis," in IEEE Transactions on Computers, vol. C-23, no. 9, pp. 881-890, Sept. 1974, doi: 10.1109/T-C.1974.224051.

[2] T. Blascheck, M. John, K. Kurzhals, S. Koch and T. Ertl, ”VA2: A Visual Analytics Approach for Evaluating Visual Analytics Applications,” in IEEE Transactions on Visualization and Computer Graphics, vol. 22, no. 1, pp. 61-70, 31 Jan. 2016, doi: 10.1109/TVCG.2015.2467871.

[3] V. A. Moraes Carvalho, N. Spolaôr, E. A. Cherman and M. C. Monard, ”A framework for multi-label exploratory data analysis: ML-EDA,” 2014 XL Latin American Computing Conference (CLEI), 2014, pp. 1-12, doi: 10.1109/CLEI.2014.6965166.

[4] Aguilar-Ruiz J., Rodríguez -Baena D., Alves R. (2013) Frequent Pattern Mining. In: Dubitzky W., Wolkenhauer O., Cho KH., Yokota H. (eds) Encyclopedia of Systems Biology. Springer, New York, NY. https://doi.org/10.1007/978-1-4419-9863-7-1114/

[5] J. Liu, ”The Analysis of Water and Health Database by One Improved A-Priori Algorithm,” 2009 International Conference on Environmental Science and Information Application Technology, 2009, pp. 101-103, doi: 10.1109/ESIAT.2009.64.

[6] I. Sandler and A. Thomo, ”Large-Scale Mining of Co-occurrences: Challenges and Solutions,” 2012 Seventh International Conference on P2P, Parallel, Grid, Cloud and Internet Computing, 2012, pp. 66-73, doi: 10.1109/3PGCIC.2012.38.

[7] M. Chen, X. Gao and H. Li, ”An efficient parallel FP-Growth algorithm,” 2009 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery, 2009, pp. 283-286, doi: 10.1109/CYBERC.2009.5342148.

[8] Vinerean2013TheEO,The Effects of Social Media Marketing on Online Consumer Behavior,Simona Vinerean and I. Cetină and L. Dumitrescu and M. Țichindelean,International Journal of Biometrics, 2013,8,66

[9] B. J. W. Sayyed and R. Gupta, ”Social Media Impact: Generation Z and Millenial on the Cathedra of Social Media,” 2020 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), 2020, pp. 595-600, doi: 10.1109/ICRITO48877.2020.9197995.

[10] M. Gastaldi, ”Integration of mobile, big data, sensors, and social media: Impact on daily life and business,” 2014 IST-Africa Conference Proceedings, 2014, pp. 1-10, doi: 10.1109/ISTAFRICA.2014.6880670.

[11] K. Kuang, M. Jiang, P. Cui, H. Luo and S. Yang, ”Effective Promotional Strategies Selection in Social Media: A Data-Driven Approach,” in IEEE Transactions on Big Data, vol. 4, no. 4, pp. 487-501, 1 Dec. 2018, doi: 10.1109/TBDATA.2017.2734102.

[12] C. A. Steed, M. Drouhard, J. Beaver, J. Pyle and P. L. Bogen, ”Matisse: A visual analytics system for exploring emotion trends in social media text streams,” 2015 IEEE International Conference on Big Data (Big Data), 2015, pp. 807-814, doi: 10.1109/BigData.2015.7363826.

[13] Jure Leskovec and Andrej Krevl, miscsnapnets, SNAP Datasets: Stanford Large Network Dataset Collection, http://snap.stanford.edu/data,jun,2014

[14] Koh Y.S., Rountree N. (2005) Finding Sporadic Rules Using Apriori-Inverse. In: Ho T.B., Cheung D., Liu H. (eds) Advances in Knowledge Discovery and Data Mining. PAKDD 2005. Lecture Notes in Computer Science, vol 3518. Springer, Berlin, Heidelberg. https://doi.org/10.1007/1143091913

[15] Millham R., Agbehadji I.E., Yang H. (2021) Pattern Mining Algorithms. In: Fong S., Millham R. (eds) Bio-inspired Algorithms for Data Streaming and Visualization, Big Data Management, and Fog Computing. Springer Tracts in Nature-Inspired Computing. Springer, Singapore. https://doi.org/10.1007/978-981-15-6695-04

[16] Borah, A., Nath, B. Rare pattern mining: challenges and future perspectives. Complex Intell. Syst. 5, 1–23 (2019). https://doi.org/10.1007/s40747-018-0085-9

[17] Ullman J.D. (2000) A Survey of Association-Rule Mining. In: Arikawa S., Morishita S. (eds) Discovery Science. DS 2000. Lecture Notes in Computer Science, vol 1967. Springer, Berlin, Heidelberg. https://doi.org/10.1007/3-540-44418-11