Market-Basket Analysis for Ukraine Conflict

Fatma Er, 967585

December 09, 2022

Abstract

On February 24, 2022, Russia invaded Ukraine, what is now known as the Russo-Ukrainian War, and this garnered international attention and sparked public conversations on social media. Twitter is one of the most popular online social networks today and one of the main sources of communication and dissemination in the online world. It has been used in the past to analyze crises in the world of politics. Data collected during such a crisis may reflect broad public and sentiment. Additionally, data can also be used to examine different campaigns used by various organizations to build public opinion. In this study, we aimed to develop a scientific and consistent model the market basket analysis method was used to analyze the distribution of tweets during the event and the public's view of the situation in order to understand the Russia-Ukraine Conflict. Apriori Algorithm and FP Growth, two of the most commonly used market basket analyses , are used.The results obtained from the models were compared and tested.

Key words: Market-Basket Analysis, Ukraine Conflict,Russo-Ukrainian War, Twitter ,Apriori Algorithm and FP Growth.

1 Introduction

The Russia-Ukraine conflict escalated in February 2022 after Russia recognized Ukraine's two separatist regions - the Donetsk People's Republic and the Luhansk People's Republic (Hernandez). Following this, the Senate of the Russian Federation authorized the use of military force in these areas on February 22, 2022. On February 24, the Russian government began invading Ukraine, which it called a special military operation.

On February 24, 2022, Russia invaded Ukraine. In the days that followed, reports kept flooding in from layman to news anchors of a conflict quickly escalating into war. Russia faced immediate backlash and condemnation from the world at large. While the war continues to contribute to an ongoing humanitarian and refugee crisis in Ukraine, a second battlefield has emerged in the online space, both in the use of social media to garner support for both sides of the conflict and also in the context of information warfare(Chen and Ferrara, 2022).

Association algorithms are widely used in retail analysis of transactions, recommendation engines, and online click stream analysis across web pages, etc. One of the popular applications of this technique is called market basket analysis, which finds co-occurrences of one retail item with another item within the same retail purchase transaction (Kotu and Deshpande, 2019).

Market Basket Analysis (MBA) is an accidental transaction pattern that purchasing some products will affect the purchasing of other products. MBA is used to predict what products that customer interested in. MBA has three parameters which are support, confidence, and lift. Support is a proportion of event B because of event A. Confidence is a probability event B happened because of event A dependently. Lift is a probability of event B happened because of event A independently (Halim et al., 2019).

There are different stages that are attached with the market basket analyses. For this study, firstly pre-processing of the text preparing (Tokenization, Normaliszation, Lemmatization, Clean Stopwords), after this process using Apriori ,FP Growth,SETM Algorithm and AIS algorithms for undertandin support,confidence,lift and compare results. In this project, we prefereed Apriori and FP Growth algoritms. The input to our models are related Ukraine Conflict daily twitts. We compared the performance of multiple algorithms with different features and explored why several models achieved better performance.

The paper is organized as follows: Section II describes the related work and background acknowledge about Market Basket Basket Analysis and Section III introduces the dataset and features we are using in this project. The methods, models and text processing section IV, followed by presenting the results and discussion details about each model and draw some perspectives about the future work section V.

2 RELATED WORKS

In this section, studies published in the literature on market basket analysis are reviewed.

(Mustakim et al., 2018) proposed a model on Market Basket Basket Analysis with FP-Growth algorithm is proposed to determine the layout and planning of goods availability.The application of FP-Growth algorithm proved to be useful in generating many and informative association rules to find out the consumer spending pattern at Berkah Mart in Pekanbaru.they used 8,307 items of goods sold there in December 2017, where the number of transactions that come out in a day average of 400 transactions, in a single transaction there is 1 item and maximum there are 20 items.The experimental results show that the application of FP-Growth and Apriori algorithm for the analysis of consumer spending pattern at Berkah Mart can increase the overall income or profit, but it is recommended to use FP Growth algorithm which maximally have the process speed in rule form and have superior support and confidence value a priori algorithm.

(Patil and Khot, 2022) proposed a discussion on Association rules and use of Apriori principle used for Market basket Analysis. Data mining provides the way to use précised information from the large dataset. Association rules find the relationship between items by analyzing the data and provide the accurate solution to the retailer to make better business decisions.

(Patil and Khot, 2022) proposed a model using a market basket analysis method to see the relationship (rules) between a set of selling attributes.The aim of their study is to determine the pattern of relationships in the transactions that take place. The data used is the transaction data of outdoor goods. Analysis The Apriori algorithm used and the frequent pattern growth (FP-growth) algorithm and Association Rules. The results of their study consist of 10 rules in the Apriori algorithm and 4 rules in the FP-Growth algorithm. The relationship pattern or association rule created is included in the article "if a consumer buys a portable stove, it is possible that portable gas will also be purchased" at the strength level of the rules with a minimum support of 0.296 and confidence 0.774 at Apriori and 0.296 and 0.750 at FP-Growth.

3 DATASET AND FEATURES

3.1 Ukraine Conflict Twitter Dataset

In this project, the Ukraine Conflict Twitter Dataset, which contains 276 text files, was used. 1-day (10 March-2022) data was used for easy processing of the algorithm result.The file contains 460075 line tweet.

Data set has 17 features, we used four of them "userid","text","language", "retweetcount".

The dataset,consists has no null line and different languages.

As we can see from Figure 2, English tweets count 303.769. We will continue our analyses with English tweets. For working with tweets need also cleaning data from duplicate . Deleted duplicate tweets and our data count reached 92.842 . After this process our aim to know most important topic for people and for reason we deleted less tweets that has retweets then 5 (['retweetcount'] >5). All these process we continued to analysing with 15.727 row tweets. For the spark code I used a different structure for Filtering data which lies in the above 50% quantile range of retweet count, should be

	userid	text	language	retweetcount
4557550	382303683	This #Russia-dropped bomb would flatten a buil...	en	6572
4557551	1433449200470032385	#Zelensky:\n\nThere was little Bandera there? ...	en	25
4557552	3689308343	@SecBlinken If the horrific attacks on #Ukrain...	en	1
4557553	567289542	Thousands of residents from the northeastern #...	en	0
4557554	1276373615466799104	In 2 hours, we are #streaming 3 #indiegames fr...	en	1
...	...	...	...	...
5017620	46023832	'Fading' \n\n"War is only a cowardly escape fr...	en	9
5017621	27004493	#UKRAINE WHAT IS USA doing to take in refugee ...	en	0
5017622	2612775241	🔴 Due to russian aggression against #Ukraine, ...	en	75
5017623	867067128	ULTIMA HORA: #Rusia acaba de atacar el Institu...	es	596
5017624	885276544746352640	#StopRussia #StopRussianAggression #StopWar #S...	und	1

460075 rows × 4 columns

Figure 1: Data/Tweets Structure

```
en      303769
fr       26510
de       23488
it       23256
es       20751
und      19080
pl        5799
tr        5417
th        4368
ja        4288
Name: language, dtype: int64time:
```

Figure 2: Language Distribution (Top)

unique as per user id and text. For sample(fraction = 0.010,seed = 3000)) changed the Fraction Argument to take different sample sizes.

3.2 Data Processing Cleaning Tweets

In this part we need to prepare out tweets for analysing. We need to using NLP (natural language processing) steps.I used two different type of for library for processing cleaning the tweets. First pandas libraries and second spark NLP. Aim is learn the different system and understand which one is better or which case and algorithms more effective.

They are the steps for cleaning my tweets.

Step A : Converting html entities

Step B : Removing "@user" from all the tweets

Step C : Changing all the tweets into lowercase

Step D : Apostrophe Lookup

Step E : Short Word Lookup

Step F : Emoticon Lookup

Step H : Replacing Special Characters with space

Step I : Replacing Numbers (integers) with space

Step J : Removing words whom length is 1

Generally these steps are called Tokenization , Normaliszation, Lemmatization, Clean Stopwords.

The cleanin stopword process , we created Custom Stopword list. Because allready know this tweets related Ukraine and Russian .

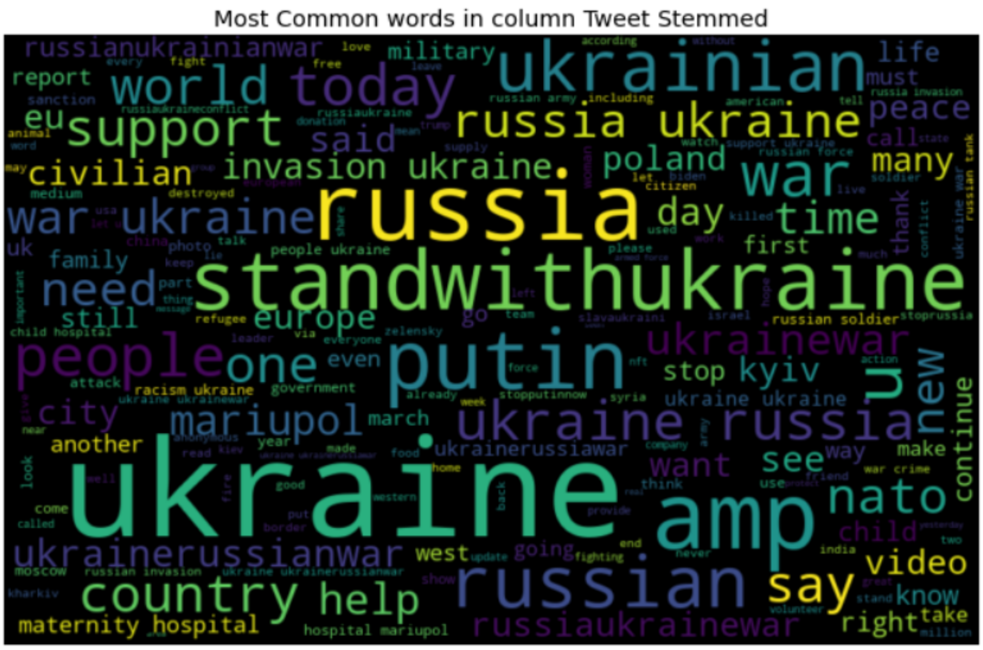

Figure 3: Before Custom Stopword

Figure 3 shows the Word clouds without cleaning our custom stopword. Also we decided our custom stopword with this word cloud.

custom$_{s}top_{w}ord_{l}ist = ['ukraine','russia','ukrainian','russian','chinese','u']$

Figure 4 shows the Word clouds with cleaning our custom stopword.

Figure 5 and 6 show the text before starting market basket analyses.

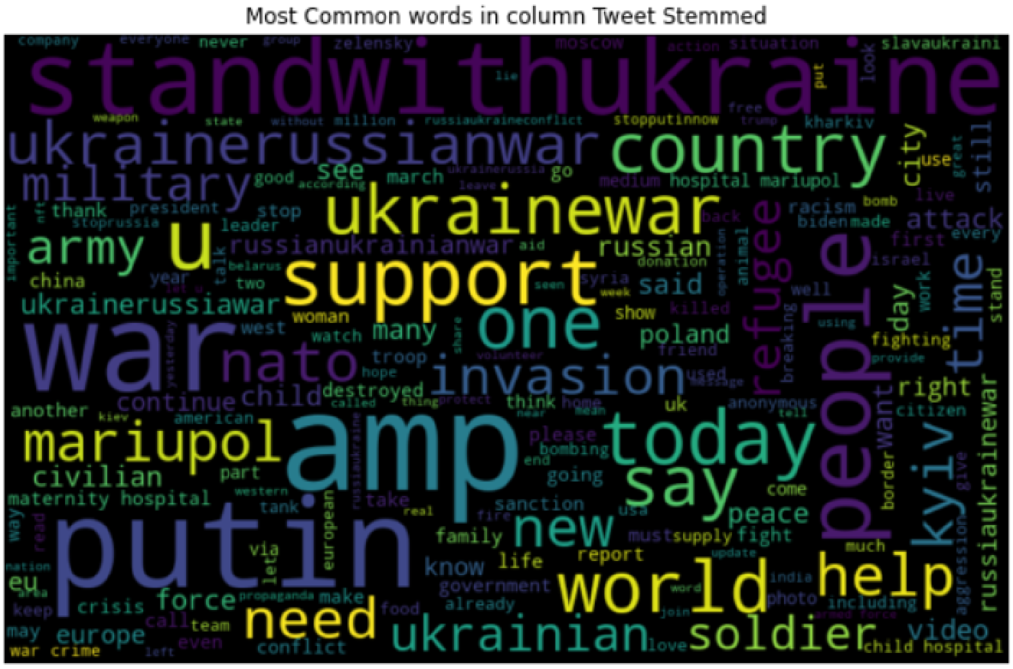

Figure 4: After Custom Stopword

	text	clean_tweet	tweet_token	tweet_token_filtered	tweet_stemmed	tweet_lemmatized
4557585	i'm still trying to work out how i decided the...	am still trying to work out how decided the ou...	[am, still, trying, to, work, out, how, decide...	[still, trying, work, decided, outcome, world,...	still tri work decid outcom world championship...	still trying work decided outcome world champi...
4557594	#russianinvasion: "we don't know if we'll be a...	russianinvasion we do not know if we will be a...	[russianinvasion, we, do, not, know, if, we, w...	[russianinvasion, know, alive, tomorrow, even,...	russianinvas know aliv tomorrow even evacu tak...	russianinvasion know alive tomorrow even evacu...
4557615	pakistani PK student asma shafique thanking in...	pakistani student asma shafique thanking india...	[pakistani, student, asma, shafique, thanking,...	[pakistani, student, asma, shafique, thanking,...	pakistani student asma shafiqu thank india res...	pakistani student asma shafique thanking india...

Figure 5: Cleaned with pandas

```
+--------------------+--------------------+--------------------+----------+
|                text|      finished_ngrams|              unique|word_count|
+--------------------+--------------------+--------------------+----------+
|@BorisJohnson Unf...|[borisjohnson, un...|[borisjohnson, un...|         9|
|Raise your hand i...|[raise, hand, wan...|[raise, hand, wan...|        15|
|Brutal beating of...|[brutal, beat, pr...|[brutal, beat, pr...|        13|
|@DmytroKuleba @ua...|[dmytrokuleba, ua...|[dmytrokuleba, ua...|        25|
|@nytimes Lynsey i...|[nytimes, lynsey,...|[nytimes, lynsey,...|        19|
+--------------------+--------------------+--------------------+----------+
only showing top 5 rows
```

Figure 6: Cleaned with Spark NLP

4 METHODS

This section describes Market Basket Analysis.

4.1 Market Basket Analysis and Algorithms

Frequent itemset mining leads to the discovery of relationships and correlations between items in large transactional or relational datasets. With the huge amount of data constantly collected and stored, many industries are starting to become interested in extracting such patterns from their databases. Disclosure of “Correlation Relationships” across multiple transaction records can be helpful in many decision-making processes, such as catalog design, cross-marketing, and customer exchange analysis(Kadlaskar).

Market Basket Analysis (MBA) is considered to be a popular data mining tool to improve many business decisions(Jirapatsil and Phumchusri, 2022).

Types of Market Basket Analysis in Data Mining. There are three types of market basket analysis in data mining:

Descriptive Market Basket Analysis This approach of the most popular Market Basket Analysis in Data Mining draws its conclusions from previous data. The analysis does not make any predictions, but instead uses statistical approaches to rate the relationship between items. It uses the unsupervised data mining model.

Predictive Market Basket Analysis In Data Mining this Market Basket analysis uses supervised data mining models such as classification and regression. Its primary purpose is to imitate the market to find out what’s causing things to happen. Essentially, it takes into account products purchased in a particular order to calculate cross-selling.

Differential Market Basket Analysis This type of market basket analysis in data mining helps in the analysis of competitors. To uncover fascinating patterns in consumer behavior, stores compare purchase histories between seasons, periods, days of the week, and other variables(Ganiyu).

There are Multiple Techniques and Algorithms Used in Market Basket Analysis.

1)Apriori Algorithm

2)AIS Algorithm

3)SETM Algorithm

4)FP Growth Algorithm

1. Apriori Algorithm

Market Basket Analysis can be implemented by using the Apriori algorithm . This algorithm determines the frequent data or item set from the transaction database. It determines the frequent item sets from the database using candidate item set generation(Rao and Kiran, 2021).

The Apriori Algorithm was proposed by Agrawal and Srikant in 1994. Apriori is designed to work in databases containing transactions (for example, collections of items purchased by customers, or details of a website visit or IP addresses) (Jirapatsil and Phumchusri, 2022).

The Apriori algorithm is a level-based, breadth-first algorithm that counts operations containing prior knowledge of frequent item set properties. Apriori uses an iterative approach known as level-level search, in which n itemsets are used to discover (n+1) itemsets. The Apriori property is used here to increase the efficiency of frequent itemsets by level. The apriori property insists that all non-empty subsets of a frequent itemset must also be frequent. This is due to the anti-monotonic nature of the support measure. Support for a set of items never exceeds Support for its subsets. A two-stage joining and pruning process is done iteratively (Annie and Kumar, 2012).

Apriori algorithm for Frequent Itemset Mining

Cd_n : Candidate itemset of size n
L_n: frequent itemset of size n
L_1 = {frequent items};
For (n=1; L_n != ϕ; n++)
Do begin
Cd_{n+1} = candidates generated from L_n;
For each transaction T in database do
Increment the count of all candidates in Cd_{n+1} that are contained in T
L_{n+1}= candidates in Cd_{n+1} with min_support
End
Return $\cup_n L_n$

Figure 7: Apriori algorithm for Frequent Itemset Mining

Limitations of the Apriori Algorithm

Although the Apriori algorithm is clear and simple, it has some weaknesses. The main limitation is very frequent item sets, low minimum support, or costly wasting of time to hold large item sets and large number of candidate sets. For example, if there are 104 of the frequent 1 item set, it should generate more than 107 candidates with length 2 to be sequentially tested and accumulated. Also, detecting frequent patterns of size 100 (eg) v1, v2. . . v100 would require generating 2100 candidate item-sets, which is costly and wastes candidate generation time. Therefore, it will check many candidate item sets and also iteratively scan the database to find candidate item sets. Apriori will be very low and inefficient when memory capacity is limited to a large number of processes(Al-Maolegi and Arkok, 2014).

2) AIS Algorithm

The AIS algorithm is the first algorithm developed and published by Agrawal, Imielinski and Swami in 1993 to generate all common product sets in the database. The algorithm is focused on increasing the functionality of databases to make decision support queries. It carries the restriction of ordering the product names in the database from A to Z(Dunham et al., 2001).

The AIS algorithm creates multiple passes over the entire database or transactional data. Scans all processes on each pass. On the first pass, it counts the support of individual items and then determines which ones are frequent in the database. Huge sets of items in each pass are scaled up to create candidate itemsets. After each scan of a transaction, it identifies common itemsets between those itemsets of the previous pass, and then items from that transaction(Kadlaskar).

Advantage: It is convenient to use the AIS algorithm to find out if there is a relationship between the items.

Disadvantage: The main disadvantage of the AIS algorithm is that it generates too many candidate sets after it turns out to be small.

3)SETM Algorithm

SETMs were proposed by Houtsmal in 1995 for aggregation of frequent clusters(Dunham et al., 2001).

This Algorithm is quite similar to the AIS algorithm. The SETM algorithm creates batch migrations on the database. On the first pass, it counts the support of individual items and then determines which ones are frequent in the database. Then it also creates candidate itemsets by extending the large itemsets of the previous pass. In addition, the SETM algorithm recalls TIDs (transaction IDs) of transactions created with candidate itemsets(Kadlaskar).

Advantage: When generating candidate itemsets, the SETM algorithm arranges the candidate itemsets sequentially along with their TID(transaction ID).

Disadvantage: There is a relationship to the Tid for each item set, so it requires more space to store a large number of TIDs.

SETM algoritması da AIS algoritmasında olduğu gibi bir çok kez tarama yapar. Bu iki algoritmada gereksiz aday oluşturduğu için çok tercih edilen algoritmalar değildir(Kumbhare and Chobe, 2014).

4)FP Growth Algorithm

FP Growth is known as Frequent Pattern Growth Algorithm. The FP growth algorithm is a concept that represents data in the form of an FP tree or Frequent Pattern. So FP Growth is a method of Mining Frequent Item Sets. This algorithm is an improvement over the Apriori Algorithm. No candidate generation is required to create a frequent pattern. This frequent pattern tree structure preserves the relationship between sets of items(Kadlaskar).

The FP-growth algorithm is used to overcome the two disadvantages of the Apriori algorithm. FP growth requires the creation of the FP tree. This requires two passes. FPgrowth uses the divide and conquer strategy. Requires the database to be scanned twice. First, it calculates a list of frequently used items in descending order (F-List) and sorted by frequency during the initial database scan. In the second scan, the database is compressed into an FP-tree [6]. This algorithm performs recursive mining on the FP-tree. There is a problem with finding frequent sets of items that are iteratively converted into searching and tree building. Favorite itemsets are created with only two passes from the database and no lead generation process. The pattern generation process has two sub-processes: generating the FP-tree and generating the pattern from the FP-tree(Kumbhare and Chobe, 2014).

4.2 Understanding Association Rule Terminology

Association rule mining uses special terminology to refer to the items on either side of the rule(Clarke). Association analysis measures the strength of co-occurrence between one item and another. The objective of this class of data science algorithms is not to predict an occurrence of an item, like classification or regression algorithms do, but to find usable patterns in the co-occurrences of the items. Association rules learning is a branch of unsupervised learning processes that discover hidden patterns in data, in the form of easily recognizable rules(Kotu and Deshpande, 2019).

The model outcome of an association analysis can be represented as a set of rules, like the one below:

{Item A} - {Item B}

This rule indicates that based on the history of all the transactions, when Item A is found in a transaction or a basket, there is a strong propensity of the occurrence of Item B within the same transaction. Item A is the antecedent or premise of the rule and Item B is the consequent or conclusion of the rule. The antecedent and consequent of the rule can contain more than one item,like Item A and Item C. To mine these kinds of rules from the data, previous customer purchase transactions would need to be analyzed. In a retail business, there would be millions of transactions made in a day with thousands of stock keeping units, which are unique for an item(Kotu and Deshpande, 2019).

Support

The support metric indicates how frequently the itemset occurs within the dataset(Clarke).The support of an item is simply the relative frequency of occurrence of an itemset in the transaction set(Kotu and Deshpande, 2019).

Confidence

The confidence value tells you how often the rule proves to be true(Clarke).The confidence of a rule measures the likelihood of the occurrence of the consequent of the rule out of all the transactions that contain the antecedent of the rule.Confidence provides the reliability measure of the rule(Kotu and Deshpande, 2019). Confidence of the rule (X-Y) is calculated by

Confidence (X-Y) = Support (X U Y) / Support (X)

Lift

Lift is essentially the support (or the probability of all the items in a rule occurring together) divided by the product of the probabilities of the items on either side appearing if there was no association. The higher the lift, the stronger the association between the antecedent and the consequent(Clarke). Though confidence of the rule is widely used, the frequency of occurrence of a rule consequent (conclusion) is largely ignored. In some transaction itemsets, this can provide spurious scrupulous rule sets because of the presence of infrequent items in the rule consequent. To solve this, the support of a consequent can be put in the denominator of a confidence calculation. This measure is called the lift of the rule(Kotu and Deshpande, 2019).

Life (X-Y)= Support (X U Y) / Support (X) * Support (Y)

Rule Generation

1. Finding all frequent itemsets. For an association analysis of n items it is possible to find $2^n - 1$ itemsets excluding the null itemset. As the number of items increase, there is an exponential increase in the number of itemsets. Hence it is critical to set a minimal support threshold to discard less frequently occurring itemsets in the transaction universe.

2. Extracting rules from frequent itemsets. For the dataset with n items it is possible to find $3^n + 2^n + 1 + 1$rules This step extracts all the rules with a confidence higher than a minimum confidence threshold.

Hyperparameters

minSupport - The minimum support of an item to be considered in a frequent itemset.

minConfidence - The minimum confidence for generating an association rule from an itemset.

This two-step process generates hundreds of rules even for a small dataset with dozens of items. Hence, it is important to set a reasonable support and confidence threshold to filter out less frequent and less relevant rules in the search space. The generated rules can also be evaluated with support, confidence, lift, and conviction measures. In terms of computational requirements, finding all the frequent itemsets above a support threshold is more expensive than extracting the rules. Fortunately, there are some algorithmic approaches to efficiently find the frequent itemsets. The Apriori and Frequent Pattern (FP)-Growth algorithms are two of the most popular association analysis algorithms (Kotu and Deshpande, 2019).

4.3 Models, Market Basket Analyses

As mentioned earlier, we perform the market basket analysis with Ukraine Conflict Twitter Dataset, which contains 1-day tweets(March-2022).

In the scope of this study,two methods that are widely used in the literature were used for the market basket analysis. These methods are; Apriori Algorithm and FP Growth Algorithm. These algorithms are coded with the Python and pandas, numpy, nltk,spark-nlp, pyspark, FPGrowth, apriori libraries were used for analysis.Also reporting and interpretation of the consequent, confidence, lift, support were made. Processed the text data with Tokenization , Normaliszation, Lemmatization, Clean Stopwords.

4.4 Market Basket Analysis with Apriori Algorithm

The results obtained from the Apriori Algorithm are given in figure 8 and 9.

	support	itemsets	length
82	0.020220	(war, standwithukraine)	2
81	0.030076	(war, putin)	2
80	0.022445	(hospital, maternity)	2
79	0.027850	(hospital, mariupol)	2
52	0.022827	(see)	1
...	...	...	...
26	0.036371	(like)	1
25	0.023463	(life)	1
24	0.052394	(kyiv)	1
23	0.022255	(know)	1
41	0.021237	(poland)	1

Figure 8: Sort Values Min Support=0.02

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
3	(maternity)	(hospital)	0.023844	0.043301	0.022445	0.941333	21.739133	0.021413	16.307364
0	(hospital)	(mariupol)	0.043301	0.055446	0.027850	0.643172	11.599958	0.025449	2.647083
2	(hospital)	(maternity)	0.043301	0.023844	0.022445	0.518355	21.739133	0.021413	2.026713
1	(mariupol)	(hospital)	0.055446	0.043301	0.027850	0.502294	11.599958	0.025449	1.922215
5	(putin)	(war)	0.136771	0.155974	0.030076	0.219898	1.409838	0.008743	1.081943
4	(war)	(putin)	0.155974	0.136771	0.030076	0.192825	1.409838	0.008743	1.069445
7	(standwithukraine)	(war)	0.133655	0.155974	0.020220	0.151284	0.969935	-0.000627	0.994475
6	(war)	(standwithukraine)	0.155974	0.133655	0.020220	0.129637	0.969935	-0.000627	0.995383

Figure 9: Association Rules Min Threshold=0.02

For Apiori algorithm, we used a minimum support of 0.02 since lower values resulted in irrelevant items while the higher values returned a few items. Figure 8 tell us that there are frequent itemsets of different lengths. Figure 9 tell us that for 10th of March tweets mostly related (maternity,hospital), (Putin,war), (Standwithukraine, war).

4.5 Market Basket Analysis with FP Growth Algorithm

The results obtained from the FP Growth Algorithm are given in figure 10 and 11.

```
+--------------------+----+
|               items|freq|
+--------------------+----+
|          [sanction]|  19|
|             [fight]|  19|
|           [refugee]|  25|
|              [cant]|  22|
|               [see]|  32|
|              [even]|  23|
|        [stoprussia]|  27|
|  [russiaukrainewar]|  57|
|            [please]|  35|
|              [look]|  21|
|             [money]|  18|
|              [last]|  21|
|             [every]|  24|
|              [need]|  29|
|              [take]|  40|
|  [standwithukraine]| 139|
|[standwithukraine...|  23|
|[standwithukraine...|  21|
|          [mariupol]|  50|
|            [attack]|  26|
+--------------------+----+
only showing top 20 rows
```

Figure 10: Sort Values Min Support=0.02

antecedent	consequent	confidence	lift	support
[putin]	[standwithukraine]	0.11864406779661017	0.7673454456773564	0.02335928809788654
[putin]	[people]	0.10734463276836158	1.1626846368524948	0.021134593993325918
[putin]	[country]	0.1016949152542373	2.4709115895556573	0.020022246941045607
[putin]	[nato]	0.11864406779661017	1.3852080123266566	0.02335928809788654
[putin]	[war]	0.1751412429378531	1.1166806907881557	0.034482758620689655
[putin]	[stop]	0.11864406779661017	1.9046610169491527	0.02335928809788654
[putin]	[ukrainewar]	0.11864406779661017	0.978541439900482	0.02335928809788654
[country]	[putin]	0.4864864864864865	2.4709115895556573	0.020022246941045607
[standwithukraine]	[war]	0.16546762589928057	1.0550028062656258	0.025583982202447165
[standwithukraine]	[putin]	0.1510791366906475	0.7673454456773564	0.02335928809788654
[stop]	[war]	0.35714285714285715	2.2771023302938196	0.02224694104560623
[stop]	[putin]	0.375	1.9046610169491525	0.02335928809788654
[ukrainewar]	[ukrainerussianwar]	0.1834862385321101	3.054706082229018	0.02224694104560623
[ukrainewar]	[war]	0.1834862385321101	1.1698874357472835	0.02224694104560623
[ukrainewar]	[putin]	0.1926605504587156	0.978541439900482	0.02335928809788654
[nato]	[putin]	0.2727272727272727	1.3852080123266564	0.02335928809788654
[putins]	[war]	0.5142857142857142	3.2790273556231	0.020022246941045607
[people]	[war]	0.24096385542168675	1.5363581987524566	0.02224694104560623
[people]	[putin]	0.2289156626506024	1.1626846368524946	0.021134593993325918
[war]	[standwithukraine]	0.16312056737588654	1.0550028062656258	0.025583982202447165

Figure 11: Association Rules Min Threshold=0.02

The hyperparameters used in FPGrowth are minimum support (0,02), minimum confidence(0,02).We calculated the Market Basket Measures like Support, Lift, and Confidence. For 10th of March with 0,02 value of algoritm, The most published tweet words are Standwithukraine, war, nato , people .

5 CONCLUSION

In this research paper, market basket analysis with Ukraine Conflict Twitter Dataset, which contains 1-day tweets(March-2022). Apriori Algorithm and FP Growth Algorithm are used for analyse.

Most important part of our data analysis is text cleaning and pre-processing.This step involved running text through many pre-processing steps including tokenization, lemmatization, normalization and stopword cleanser. This is crucial as this step removes all unwanted noises and only keep the most relevant data.

The overall results of these models have different performance.Since FP-Growth doesn't require creating candidate sets explicitly, it is magnitudes faster than the alternative Apriori algorithm. We used PySpark implementation of FP-growth and mlxtend implementation of A-priori algortihm. Although we can't really compare them because the dataset sizes in the two cases are different, the resulting frequent item tables are similar.

For future work on this, better text preprocessing techniques can be used to further remove irrelevant information from the dataset. Also, a more sophisticated and high-spec machine can be used to perform the analysis on a larger scale to gain better insights.

6 Declaration

I declare that this material, which we now submit for assessment, is entirely my own work and has not been taken from the work of others, save and to the extent that such work has been cited and acknowledged within the text of our work. I understand that plagiarism, collusion, and copying are grave and serious offences in the university and accept the penalties that would be imposed should I engage in plagiarism, collusion or copying. This assignment, or any part of it, has not been previously submitted by me or any other person for assessment on this or any other course of study.

References

M. Al-Maolegi and B. Arkok. An improved apiori algorithm for association rules. *International Journal on Natural Language Computing (IJNLC)*, 3(1):1–9, 2014.

L. C. Annie and A. Kumar. Market basket analysis for a supermarket based on frequent itemset mining. *IJCSI International Journal of Computer Science Issues,*, 9(5):257–264, 2012.

E. Chen and E. Ferrara. Tweets in time of conflict: A public dataset tracking the twitter discourse on the war between ukraine and russia. *, Information Sciences Institute*, pages 1–5, 2022.

M. Clarke. How to use the apriori algorithm for market basket analysis. URL `https://practicaldatascience.co.uk/data-science/how-to-use-the-apriori-algorithm-for-market-basket-analysis`.

M. Dunham, Y. Xiao, L. Gruenwald, and Z. Hossain. A survey of association rules. *Researchgate*, pages 1–65, 2001.

I. S. Ganiyu. Market basket analysis in data mining simplified 101. URL `https://hevodata.com/learn/market-basket-analysis-in-data-mining/`.

S. Halim, T. Octavia, and C. Alianto. Designing facility layout of an amusement arcade using market basket analysis. *ScienceDirect*, (161):623–629, 2019.

J. Hernandez. Npr: Why luhansk and donetsk are key to understanding the latest escalation in ukraine. URL `https://www.npr.org/2022/02/22/1082345068/why-luhansk-and-donetsk-are-key-to-understanding-the-latest-escalation-in-ukrain`.

P. Jirapatsil and N. Phumchusri. Market basket analysis for fresh products location improvement: A case study of e-commerce business warehouse. *MSIE 2022, April 28–30, 2022, Chiang Mai, Thailand*, pages 1–28, 2022.

A. Kadlaskar. A comprehensive guide on market basket analysis. URL `https://www.analyticsvidhya.com/blog/2021/10/a-comprehensive-guide-on-market-basket-analysis/`.

V. Kotu and B. Deshpande. Association analysis. *Science Direct,*, pages 1–22, 2019.

T. Kumbhare and P. S. Chobe. An overview of association rule mining algorithms. *International Journal of Computer Science and Information Technologies,*, 5(1):927–930, 2014.

Mustakim, D. Herianda, and A. Ilham. Market basket analysis using apriori and fp-growth for analysis consumer expenditure patterns at berkah mart in pekanbaru riau. *Journal of Physics*, pages 1–10, 2018.

B. Patil and L. Khot. A study on market basket analysis using apriori algorithm. *Gogte Institute of Technology Belagavi*, pages 1–5, 2022.

A. Rao and J. Kiran. Market basket analysis for fresh products location improvement: A case study of e-commerce business warehouse. *Int J Syst Assur Eng Manag*, pages 1–6, 2021.