Draft version September 17, 2022
Typeset using LATEX default style in AASTeX631

Market-basket analysis (Project 2)*

Elaheh Esfandi [1]

[1] *Master Student of Milan University, data science in economic (DSE), Matriculation number: 963762, elaheh.esfandi@studenti.unimi.it.*

ABSTRACT

Conflict in Ukraine escalated on 24 February 2022 when the Russian military entered the country. by using "Apriori Algorithml" method, we tried to do the Market-basket analysis with Ukraine Conflict Twitter data set and comparing the time changes of "basket items" and it is frequency according to the changes that may happen in social minds during the war by using text and hashtags of available twits data set in kaggle.

results of this study has been proved changes in the way of using words to express the feeling by social during the war." stop the war" and "we are all Ukrainian" has the most probability to come to gather along all Twits hashtag. while the "Ukraine Russia War" and "Ukrainian Arm" has the least probability to be used with each other.

Keywords: Market-basket analysis, Ukraine Conflict, frequent item set.

1. INTRODUCTION

In October 2021, Russia began moving troops and military equipment near its border with Ukraine, reigniting concerns over a potential invasion. Conflict in Ukraine escalated on 24 February 2022 when the Russian military entered the country. Heavy fighting is currently taking place throughout the country and that is resulting in civilian casualties and significant damage to infrastructure and buildings, especially in "Mariupol" and "Chernihiv" cities.

* "Algorithms for massive data" project report, 2022-2023, Released on Jun, 12th, 2022

This specific situation got so many attentions from people around the world. people mostly used their social media accounts to show their anger and their support against this conflict. As we know, in this millennium the power of social media like Instagram, Facebook or Twitter is not deniable. We can conclude that the study of the people‘s behavior around the world who are using theses media and with their behavior they may effect on the political act, which is somehow vital to be considered. For the aim of analyzing people‘s behavior on the specific issue of "Ukraine Conflict" by using the tweeter data, we propose" the Related concepts" in market- basket analysis. For this matter, items are words, and baskets are documents (e.g., Web pages, blogs, Twit). A basket/document contains those items/words that are present in the document. If we look for sets of words that appear together in many documents, the sets will be dominated by the most common words.

The novelty of this study lies in the following endeavors. firstly, we are showing the method(algorithm) to implement a system finding frequent item set (market-basket analysis). Secondly, by using this method which we will introduce in 3rd section, we will try to do the Market-basket analysis with the Ukraine Conflict Twitter data set. thirdly, another aim of this study is to compare the time changes of "basket items" and it is frequency according to the social minds during the war

The rest of the paper is organized as follows. The next section deals with the Research data and data preparation. In Section 3, we will go through the algorithm and Experimental results and the most important key finding. Finally, Section 4 provides Concluding remarks of the using algorithm, general framework and the Results and finding of the model.

2. RESEARCH DATA

We will use the Ukraine Conflict Twitter data set [1], which is published on Kaggle, with attribution required; we will go through "the overview of the chosen data set" and "how data have been organized for experiments". This data set is only a fraction of a fraction of the Twits in the real-world.

This data set contains at least 51 the text column of the CSV files which belongs to the date of twits monitoring the current ongoing Ukraine-Russia conflict. As it said in the Kaggle, "Data sets will be updated every day between 1 am to 3am UTC". So, the model outcome of the update data set is different from day to day. with the python coding, I tried to link the Kaggle database to the googlecolab [2], so I could run the codes whenever I want without the need of the changing in the repository and observe any different. after entering the data-set from the Kaggle and unzip them, we faced with the mass of the data. so only two data-set (first of April and September 12th) from above have been chosen to analysis.

according table(1), two important features of this data is the "text" and the "hashtag". for the aim of this study, we could use two different ways to find the data of the baskets.

1. case (1): using each of all twits text and try to implement the frequency analysis, find the frequent word and use it in our "market basket analysis".

2. case(2): using the hashtag of each text in our twits and do the "market basket analysis" on the hashtags.

base on each of these situation, next step for preparing the data-set for our target analysis is to do the "tokenize", removing all the numbers, emoji, website address, punctuation and in the case of using the hashtags (case 2), we have to remove two words of "text" and "indices" too. at the end, for the mater of using the "Apriori Algorithm" from market basket model, we had to transform the data set into a list data set.

3. EXPERIMENTAL RESULTS

In this section, we will go through the applied pre processing techniques, the metrics were used for evaluating performances, and the experimental methodology like algorithms and their implementations. then experimental results as plots and/or tables will be present.

3.1. *The Market-Basket Model*

With massive amounts of data continuously being collected and stored, many industries are becoming interested in mining such patterns from their databases. The discovery of interesting correlation relationships among huge amounts of business transaction records can help in many business decision-making processes such as catalog design, cross-marketing, and customer shopping behavior analysisHan et al. (2012).

[1] https://www.kaggle.com/datasets/bwandowando/ukraine-russian-crisis-twitter-dataset-1-2-m-rows/

[2] https://colab.research.google.com/drive/1km1nQeawSU_0tlkQQQTuYqoaIy3lzsKj?usp=sharing/this link is the link of the goolecolab of this paper.

Table 1. data description.

heads	0	1
username	Barrie360	silpantipolo
acctdesc	What Barrie's talking about. From local news t...	This is the official Twitter account of San Is...
location	Barrie, Ontario	Diocese of Antipolo
following	678.0	9.0
followers	16818.0	14.0
totaltweets	106996.0	398.0
usercreatedts	2011-06-03 03:17:10.000000	2021-07-13 09:03:02.000000
tweetid	1509319562654720000.0	1509319562658865152.0
tweetcreatedts	2022-03-31 00:00:00.000000	2022-03-31 00:00:00.000000
retweetcount	0.0	0.0
text	Humanitarian drive in Midhurst sends medicine,...	This morning, LetUsPray for RussiaUkraine th...
hashtags	['text': 'Ukraine', 'indices': [63, 71]]	['text': 'LetUsPray', 'indices': [14, 24], ...
language	en	en
coordinates	NaN	NaN
favorite$_{c}ount$	0.0	0.0
extractedts	2022-03-31 00:09:25.807372	2022-03-31 01:12:01.286851

Source: the outcome of this paper

Market basket analysis is a process that looks for relationships of objects that "go together" within the business context. practically, market basket analysis goes beyond the supermarket scenario from which its name is derived. Market basket analysis is the analysis of any collection of items to identify affinities that can be exploited in some manner David (2013).

The most frequent approach are to apply the Apriori algorithm and The eclat algorithm. Instead of the breadth-first approach that the "Apriori" algorithm uses to identify frequent item sets, "eclat" uses a depth-first approach. It looks at each item, identifies the transaction IDs for the transactions in which that item appears, and makes a list of those IDs. It then looks for intersections among those lists for the various items and calculates support based on the intersections.The eclat algorithm can be faster, but it also can be memory-intensive as it constructs and uses the lists at these intermediary steps. according to all these reason, for achieving the goal of this study, the "Apyori "package in python were used.

3.2. *Apriori Algorithml*

The Apriori algorithm is a seminal algorithm for mining frequent item sets for Boolean association rules. It explores the level-wise mining Apriori property that all nonempty subsets of a frequent item set must also be frequent. At the k‘th iteration (for k¿1), it forms frequent k-item set candidates based on the frequent (k 1)-item sets, and scans the database once to find the complete set of frequent k-item sets, LBhandari et al. (2015).

The first step in market basket analysis is to infer association rules, which express which products are frequently purchased together, and in our cases is words. according to Bhandari et al. (2015),Association rule mining consists of first finding frequent item sets (sets of items, such as A and B, satisfying a minimum support threshold, or percentage of the task-relevant tuples), from which strong association rules in the form of are generated. These rules also satisfy a minimum confidence threshold (a pre specified probability of satisfying B under the condition that A is satisfied). Associations can be further analyzed to uncover correlation rules, which convey statistical correlations between item sets A and B.

In the algorithm function, there is some elements, that will be describe each as fallows:

• Support: Support is the basic probability of an event to occur. If we have an event to buy product A, Support (A) is the number of transactions which includes A divided by total number of transactions.

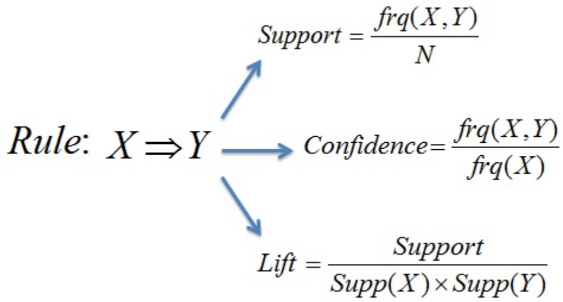

Figure 1. Measures to evaluate association rules. Source:Li (2017)

Table 2. result of apriori algorithm for 22 of February, case(1).

Left Hand Side	Right Hand Side	Support	Confidence	Lift
Kalashnikov	surreal	0.005272	0.977191	179.745521
Kalashnikov	bear	0.005272	0.977191	162.585550
bear	surreal	0.005272	0.877157	161.345070
Kalashnikov	sounds	0.005272	0.977191	150.406339
OpKremlin	broadcast	0.004751	0.894792	149.667771
sounds	surreal	0.005272	0.811449	149.258783
FckPutin	OpKremlin	0.004827	0.767253	144.510457
Kalashnikov	prepare	0.005272	0.977191	141.351877
prepare	surreal	0.005272	0.762600	140.273404
bear	sounds	0.005273	0.877387	135.044732

Source: the outcome of this paper

• Confidence: The confidence of an event is the conditional probability of the occurrence; the chances of A happening given B has already happened.

• Lift: This is the ratio of confidence to expected confidence. The probability of all the items in a rule occurring together (otherwise known as the support) divided by the product of the probabilities of the items on the left and right side occurring as if there was no association between them. The lift value tells us how much better a rule is at predicting something than randomly guessing. The higher the lift, the stronger the association.

figure(1)is shown Clearly the statistic behind theses factors.

we will take all steps of the "Apriori" algorithm. at the beginning , we will Define the minimum support and confidence for the association rule. then Take all the subsets in the transactions with higher support than the minimum support. the next step is to Take all the rules of these subsets with higher confidence than minimum confidence. then we will Sort the association rules in the decreasing order of lift. at the end, we will Visualize the rules along with confidence and support.

3.3. *data analysis result*

"Apriori Algorithml" were used in 2 cases(case(1):using text and case (2):using hashtags). as we mentioned before, another aim of this study is comparing the time changes of "basket items" and it is frequency according to the social minds during the war. for achieving this goal, 22 of February and 18 of august were chosen.

table(2) and table(3) are showing the probability of 2 chosen words(right and left hand side) according to the case(1), in all twits of February 22 th and 18 of august. the result has been ordered by the most "ratio of confidence" to the least one. according to table(2), support of 5.2 percent of for Rule means that 5.2 percent of all the transactions

Table 3. result of priory algorithm for the august 18th, case(1).

Left Hand Side	Right Hand Side	Support	Confidence	Lift
ArsonColeUkraine	NEWsampRTs	0.005219	1.000000	191.591667
Embargo	Infanticide	0.005480	0.984375	178.202879
AndrzejDuda	Infanticide	0.005524	0.976923	176.853846
Infanticide	Murder	0.005524	1.000000	176.853846
AndrzejDuda	Embargo	0.005480	0.969231	174.090505
Embargo	Murder	0.005480	0.984375	174.090505
AndrzejDuda	Murder	0.005524	0.976923	172.772604
Infanticide	oleksiireznikov	0.005524	1.000000	171.574627
Embargo	oleksiireznikov	0.005480	0.984375	168.893773
Anti	Infanticide	0.005524	0.927007	167.817518

Source: the outcome of this paper

Table 4. result of priory algorithm for the February 22th, case(2).

Left Hand Side	Right Hand Side	Support	Confidence	Lift
Kiev	Kyiv	0.005597	0.297179	3.998680
WWIII	RussiaUkraineConflict	0.011040	0.323328	5.824241
UkraineInvasion	UkraineRussia	0.009067	0.325729	5.900204
UkraineUnderAttack	UkraineRussia	0.010772	0.246037	4.456670
UkraineWar	UkraineRussia	0.007631	0.296250	5.366221
UkraineWar	UkraineUnderAttack	0.008427	0.327153	7.472372
russia	ukraine	0.010695	0.587674	18.020824

Source: the outcome of this paper

under analysis show that ”Kalashnikov” and ”surreal” are coming together. A confidence of 97 percent means that 97 percent of the twitter user who wrote a ”Kalashnikov” also wrote the ”surreal”. while according to table(3), support of 5 percent of for Rule means that 5.48 percent of all the transactions under analysis show that ”Embargo” and ”Infanticide” are coming together. A confidence of 99 percent means that 99 percent of the twitter user who wrote a ”Embargo” also wrote the ”Infanticide”. by comparing the outcome of the first observed day(table(2)) and the last observed day(table (3)), base on the text mining of relative twits, we can understand two word or phrase of ”Embargo” and ”Infanticide” have the most possibility to come to gather according to that day. according to the Oxford dictionary ”Infanticide” means a person who kills an infant, especially their own child and ”Embargo” is an official ban on trade or other commercial activity with a particular country. but why most of the people try to use this two word together in that specific day of Russian and Ukraine conflict? after this observation, i went through the historical event of that exact day. i came across the news and the picture of Andrzej Duda the Poland president and the interview about embargo, murder and Infanticide.

on the other hand, we can see in table(4), two chosen word phrases of ”WWIII” and ”Russia Ukraine Conflict” has the most probability to come to gather along all Twits hashtag. while the ”Kiev” and ”Kyiv” has the least probability to come to gather.

according to table(4), support of 11 percent of for Rule means that 11 percent of all the transactions under analysis show that ”WWIII”and ”Russia Ukraine Conflict” are coming together. A confidence of 95 percent means that 32 percent of the twitter user who wrote a ”WWIII” also wrote the ”Russia Ukraine Conflict”.

Table 5. result of priory algorithm for the august 18th, case(2).

Left Hand Side	Right Hand Side	Support	Confidence	Lift
Fascism	Infanticide	0.005651	0.920290	162.855072
Infanticide	MLRS	0.005651	1.000000	161.683453
BB	Movie	0.006363	1.000000	157.160839
BB	Vote	0.006363	1.000000	154.993103
BB	saudi	0.006363	1.000000	154.993103
Movie	Vote	0.006363	1.000000	154.993103
Movie	saudi	0.006363	1.000000	154.993103
Vote	saudi	0.006363	0.986207	152.855268
BB	twitter	0.006363	1.000000	148.834437
Infanticide	StandUpForUkraine	0.005651	1.000000	148.834437

Source: the outcome of this paper

while according to table(5), support of 6.3 percent of for Rule means that 6.3 percent of all the transactions under analysis show that ”BB” and ”Vote” or ”BB” and ”saudi” are coming together. A confidence of 100 percent means that 100 percent of the twitter user who wrote a ”BB” also wrote the ”saudi” or ”vote”.

by comparing the outcome of the first observed day(table(4)) and the last observed day(table (5)), starting with the ”WWIII”and ”Russia Ukraine Conflict” hashtags together, it seems that people try to convince the world this is going to be mater for all of us to ”Ukraine Invasion” by using two by two of these hashtags.then after so many month the most coming together hashtags became ”Vote”, ”BB” and ”saudi”. which shows the concern of the news of Saudi Arabia about the exporting oil and the vote of the NATO for the Ukraine.

4. CONCLUDING REMARKS: COMMENTS AND DISCUSSION ON THE EXPERIMENTAL RESULTS

the aim of this study was to find the relation between most frequent words in tweets about conflict between Ukraine and Russia. the data were gathered from Kaggle website and we decided to do it in the way of market basket analysis.

The most frequent approach in market basket analysis is to apply the Apriori algorithm, which starts out by generating the frequent item sets for your data with a minimum number of items k, which you can set. It decides which item sets are frequent by requiring them to meet a minimum level of support (explained above). Then, those frequent item sets are partitioned (divided) and re-combined repetitively and the support calculated for each combination, until no more item sets can be created.

the results in both cases of text analysing and hashtags analysing show the relevant change in people opinion during the conflict. in the beginning, social believe the conflict is about to start so heavily, by using the two hashtags of ”Kalashnikov” and ”prepare”. while due to the using of ”vote” and ”Saudi” hashtags to gather, the population shows their worries about the consequences of this conflict on other nations. the frequency of using two sentences of ”Infanticide” and ”Murder” shows the fear of the public about this conflict by being so long.

the outcome of text mining for 22 of February shows that ”Kalashnikov”, ”prepare” ”sound” and ”surreal” are the most used words that appear to gather two by two. ”The Russia-Ukraine war could ultimately serve as a demonstration of how authoritarian regimes can sow the seeds of their own downfall.” was one of the most famous news of those days. which brings the idea of ”Putin’s war in Ukraine shows the limits of authoritarianism”Benedikter (march 2022).

APPENDIX

A. DECLARATION BY AUTHOR

“I declare that this material, which I 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 my 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.“

B. ALL THE AVAILABLE LINKS FOR CODES BY AUTHOR

GitHub:https: https://github.com/elahehesfandi/Market-basket-analysis/
Google-colab :https://colab.research.google.com/drive/1LbHBq-ibJYw3co0RJ0XTNvwxL3pk9df2?usp=sharing

REFERENCES

Benedikter, R. march 2022. https://blogs.lse.ac.uk/europpblog/2022/03/18/putins-war-in-ukraine-shows-the-limits-of-authoritarianism/

Bhandari, A., Gupta, A., & Das, D. 2015, Procedia Computer Science, 46, 644, doi: https://doi.org/10.1016/j.procs.2015.02.115

David, L. 2013, in Business Intelligence, Second Edition, 271

Han, J., Kamber, M., & Pei, J. 2012, in Data Mining (Third Edition), third edition edn., The Morgan Kaufmann Series in Data Management Systems (Boston: Morgan Kaufmann), 243–278, doi: https://doi.org/10.1016/B978-0-12-381479-1.00006-X

Li, S. 2017, Towards Data Science, on Medium, first Edition, 1