23RD IEEE/ACM International Symposium on
Cluster, Cloud and Internet Computing
Bangalore, India | May 1- 4, 2023

Spark based Parallel Frequent Pattern Rules for Social Media Data Analytics

Shubhangi Chaturvedi
PhD Research Scholar
IIITDM Jabalpur

Sri Khetwat Saritha
Assistant Professor
NIT Bhopal

Animesh Chaturvedi
Assistant Professor
IIIT Dharwad

Introduction

- Social media websites like Facebook, Twitter, Instagram etc. produces data tremendously.
- Analysis of data can produce hidden patterns.
- Mining patterns helps in
  - Elections
  - Personal and corporate benefits
  - Business decision
  - Marketing
  - Customer service
  - Reputation management
  - Sales

Association Rule Mining

- Items purchased on per transaction basis or over a certain period are known as basket data [1][2].
- Method for discovering interesting relations between variables in large databases.
- Market Basket analysis Example:- {Bread, Butter} => {Jam}

  Promotional pricing or Product Placement

Support:- frequency (or percentage) of customer will buy both bread and butter.

$$Supp(X) = | \{t \in T; X \subseteq t\} | / | T |$$ Where, T is set of transactions

Confidence:- conditional probability if customer buy bread and butter, then jam will also be purchased.

$$Conf(X => Y) = Supp(X \cup Y) / Supp(X)$$

If support present then go for confidence.

[1] Rakesh Agrawal, Tomasz Imieliński, and Arun Swami. "Mining association rules between sets of items in large databases". ACM SIGMOD Record. Vol. 22. No. 2. ACM, (1993).

[2] Rakesh Agrawal, and Ramakrishnan Srikant. "Fast algorithms for mining association rules." Proceeding of the 20th Int. Conf. Very Large Data Bases, VLDB. Vol. 1215. 1994.

Frequent Patterns

- As described by Han, Jian and Yiwen [3]
- Let a transactional database $T = <t_1, t_2, t_3 \ldots . t_n>$
  - set of items are $I = \{i_1, i_2, i_3 \ldots . i_m\}$
  - where T_i $(i \in 1 \ldots . n)$ is a transaction which contains a set of items in I.
- If A is set of items,
  - then support of pattern A is the number of transactions containing A in T.
- If support is above certain minimum threshold ξ,
  - then pattern A is said to be frequent.
- Finding the complete set of frequent patterns

[3] Han Jiawei, Jian Pei, and Yiwen Yin. "Mining frequent patterns without candidate generation". ACM SIGMOD Record. Vol. 29. No. 2. ACM, (2000).

SPMF (Sequential Pattern Mining Framework)

- The construction of FP-Tree is done in a second scan.
- For every item,
  - support value is calculated,
  - those items whose support is greater than threshold a FP-Tree is constructed
  - FP-Tree helps to generate rules.
- For experimentations,
  - we used SPMF by Fournier-Viger et al. [4][5].

[4] P. Fournier-Viger, et al. "SPMF: a java open-source pattern mining library." J. Mach. Learn. Res. 15.1 (2014): 3389-3393.
[5] "Mining frequent itemsets using the FP-Growth algorithm", Accessed on Feb 2023 https://www.philippe-fournier-viger.com/spmf/FPGrowth.php

Resource Challenges in FP-Growth:

- According to Li et al. [6] FP-Growth suffers from following resource challenges:
  - **Storage:**
    - The FP-tree can be huge.
  - **Computation distribution:**
    - The steps of FP-Growth are not parallelized.
  - **Costly communication:**
    - FP-Trees synchronization have interdependency.
  - **Support value:**
    - Overflow of storage may occur because of generation of FP-Tree.

[6] Li, Haoyuan, et al. "PFP: Parallel FP-growth for query recommendation." *Proceedings of the 2008 ACM conference on Recommender systems*. ACM, 2008.

Parallel Frequent Pattern Growth

1) ***Sharding :*** Division of data into parts.
2) ***Parallel Counting :*** MapReduce to count the support values of all items that appear in database.
3) ***Grouping Items :*** All items |I| on F-List divided into Q groups.
   - Group list (G-list) and group-id (gid)
4) ***Parallel FP-Growth :***
   - ***Mapper-Generating group dependent transactions***
     - Mapper instance fed with a shard of DB, Reads the G-list, Outputs one or more Key- value pair
   - ***Reducer – FP-Growth on dependent shards***
     - Groups all corresponding group-dependent transactions.
     - Process group dependent shard one by one. Builds a local FP-tree.
5) ***Aggregating :*** Outputs corresponding top-K mostly supported patterns.

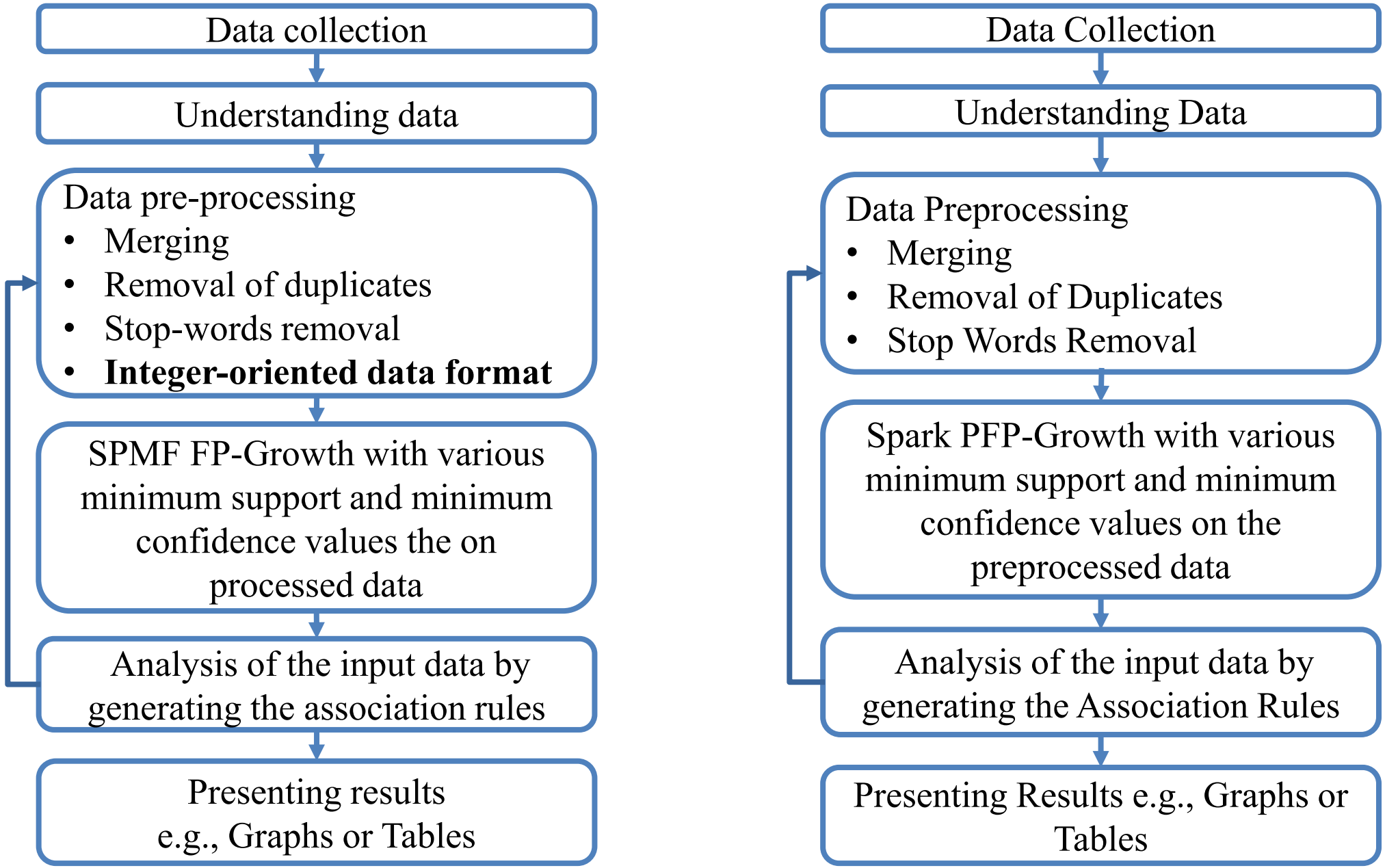
[6] Li, Haoyuan, et al. "PFP: Parallel FP-growth for query recommendation." *Proceedings of the 2008 ACM conference on Recommender systems*. ACM, 2008.

PFP: Parallel FP-Growth on Spark Framework

- In Spark ML-Library (MLLib), a parallel version of FP-growth called
  - PFP: Parallel FP-Growth
- PFP distributes the work of growing FP-trees based on the suffixes of transactions.
- More scalable than a single-machine implementation.
- PFP partitions computation,
  - where each machine executes an independent group of mining tasks

[6] Li, Haoyuan, et al. "PFP: Parallel FP-growth for query recommendation." *Proceedings of the 2008 ACM conference on Recommender systems*. ACM, 2008.

Flowchart of the FP-Growth and the PFP-Growth



Flow Chart of experiment performed the FP-Growth (LHS) and the PFP-Growth (RHS) mining on social media dataset.

Frequent Pattern Rules for Social Media Data Analytics

- Number of users on social media are increasing.
- An applied process to mine social media dataset to retrieve frequent patterns (as rules) in cost effective time.
  - The experiment is performed on three social media datasets.
    - Large Movie Reviews Dataset
    - Political Social Media Dataset
    - Disasters on Social Media Dataset
  - Experiments are also performed for both
    - Frequent Pattern (FP) Growth using SPMF,
    - Parallel FP (PFP) Growth using Spark.
- The parallel computation is achieved with the help of scalable Apache Spark environment.
- PFP-Growth does not require preprocessing on the dataset to generate rules.

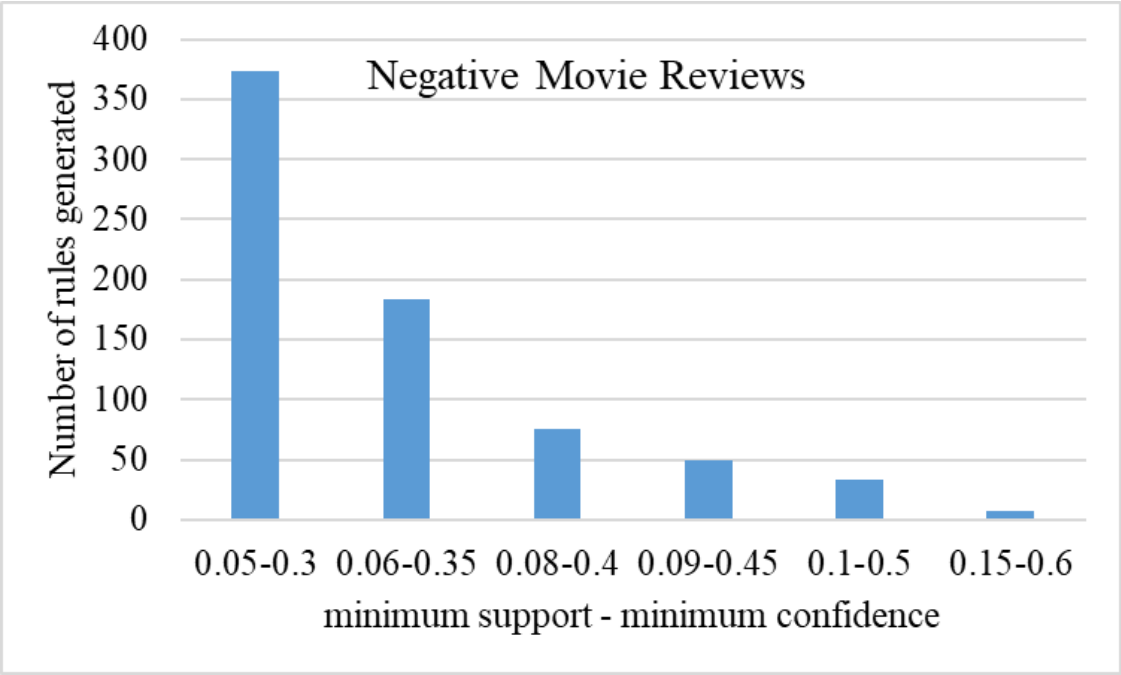
Social Media Datasets

- "Large Movie Review Dataset", Accessed on Feb 2023
  http://ai.stanford.edu/~amaas/data/sentiment/

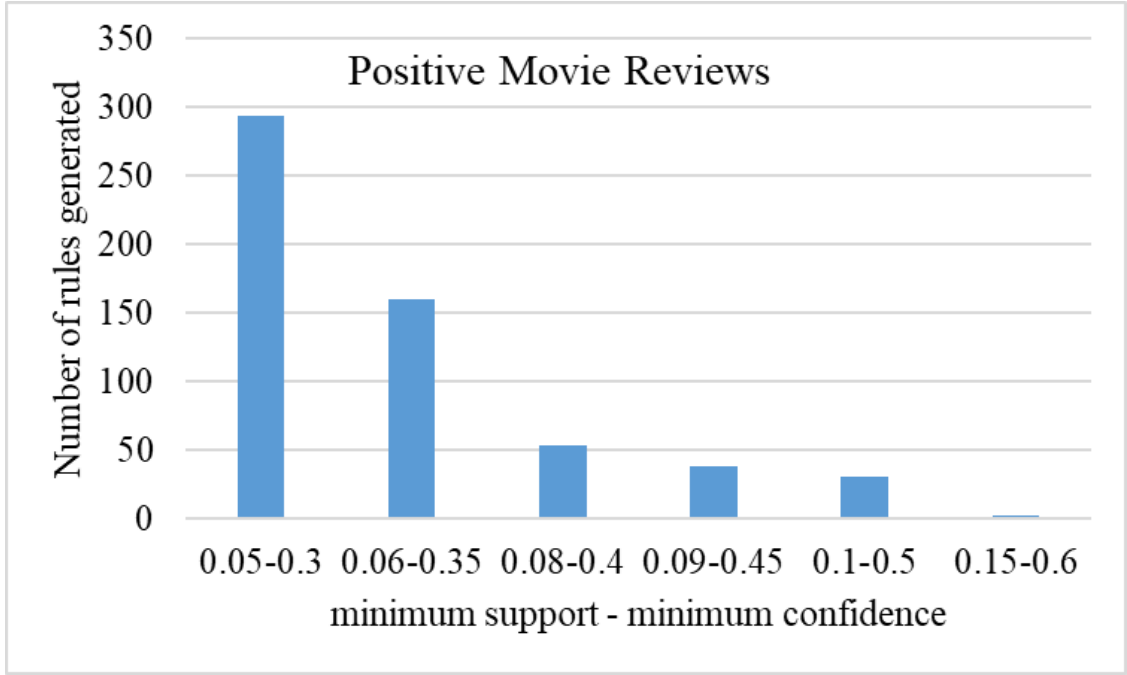
- "Classification of political social media", Accessed on Feb 2023
  https://mlbazaar.github.io/dataset/Political-media-DFE

- "Disasters on social media", Accessed on Feb 2023
  https://mlbazaar.github.io/dataset/socialmedia-disaster-tweets-DFE

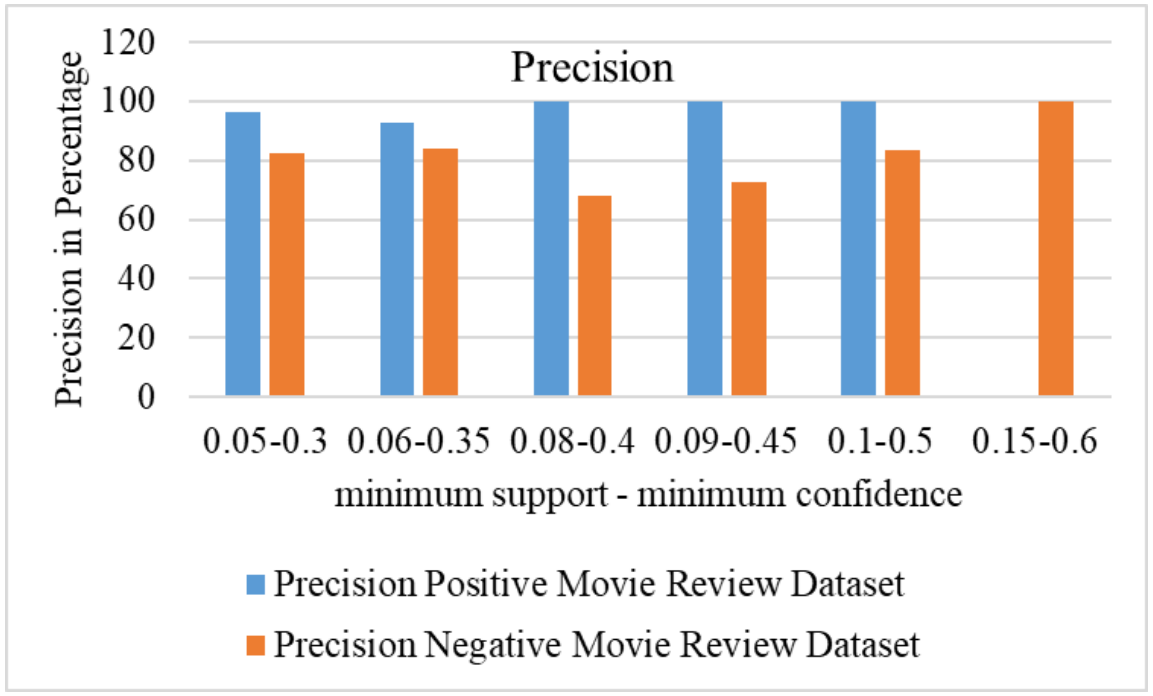
Experiments



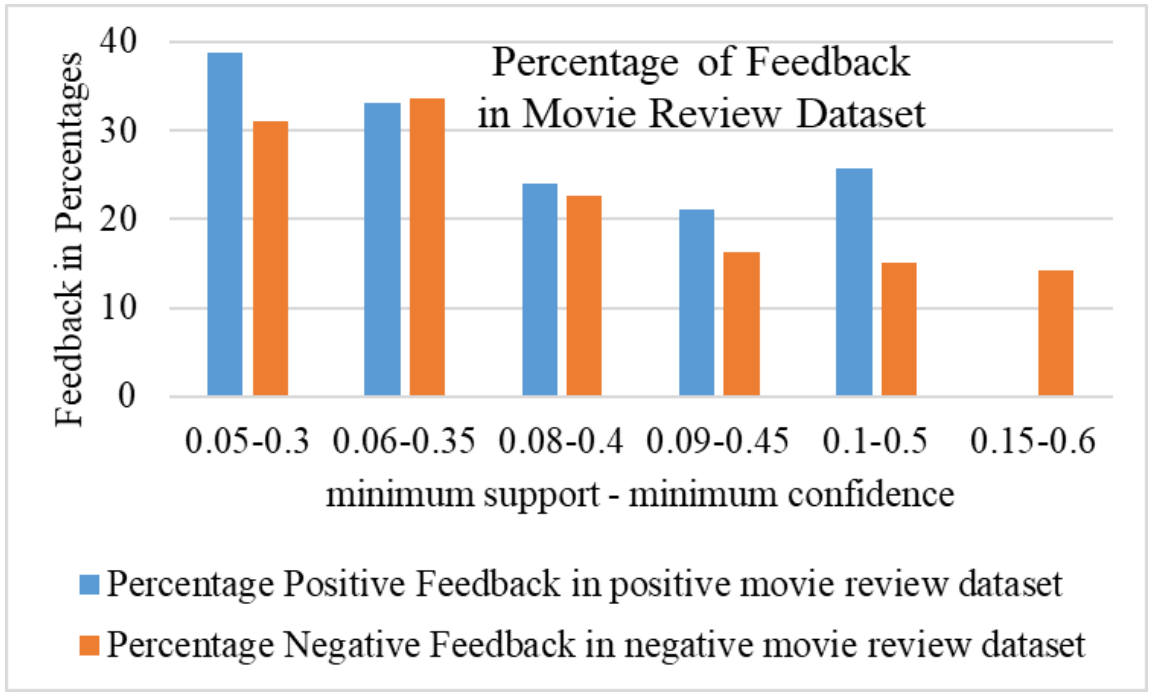
Showing number of rules generated using various minimum support-confidence values for PFP-Growth on Negative Movie Review Dataset.



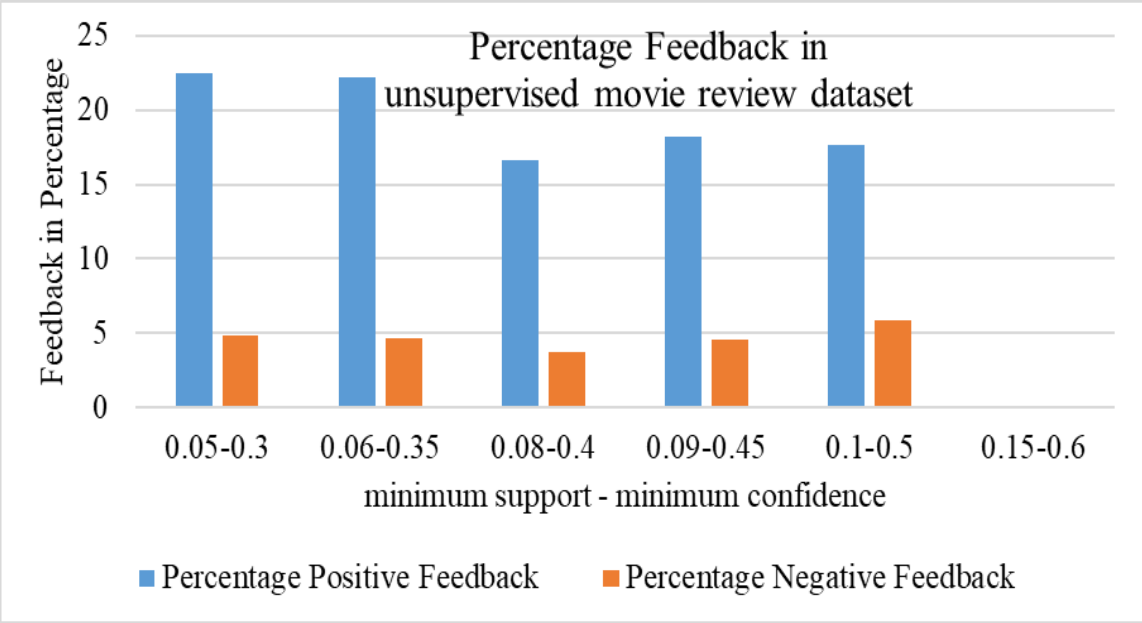
Showing number of rules generated using various minimum support-confidence values for PFP-Growth on Positive Movie Review Dataset



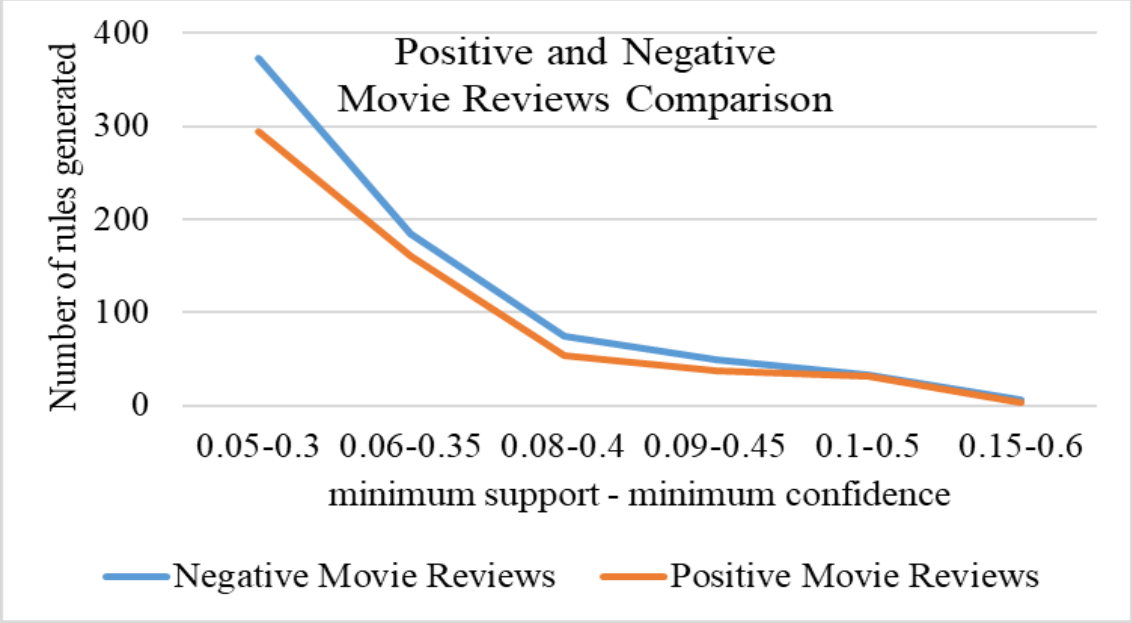
A bar-graph showing precision comparison in percentage for Positive and Negative Movie Review Dataset.



A bar-graph showing percentage of positive and negative feedback in positive and negative movie review dataset.



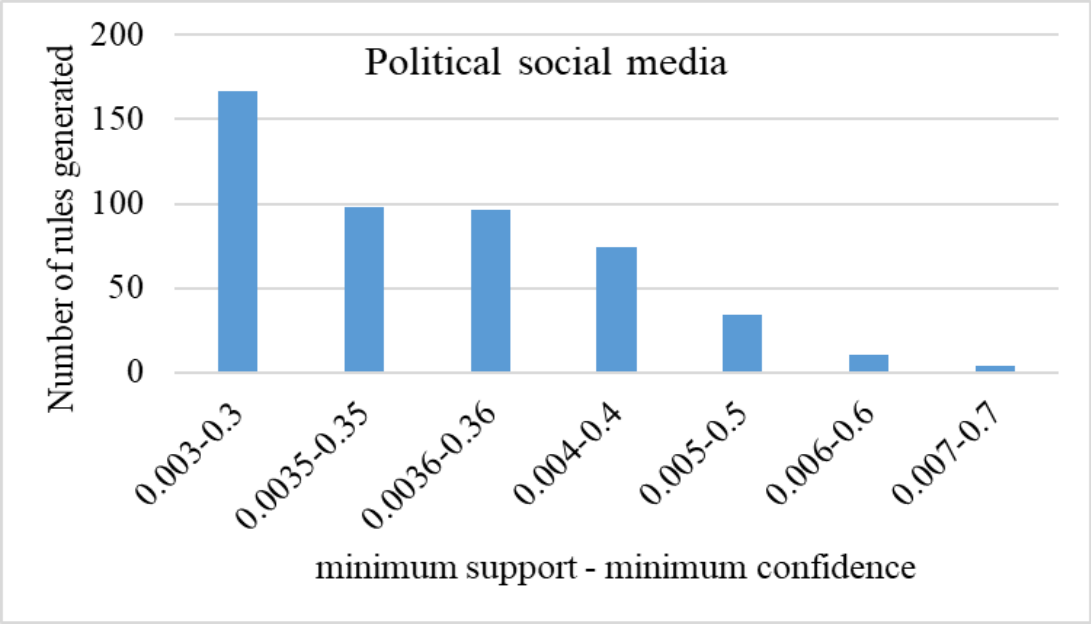
Bar-graph showing percentage feedback in unsupervised movie reviews.



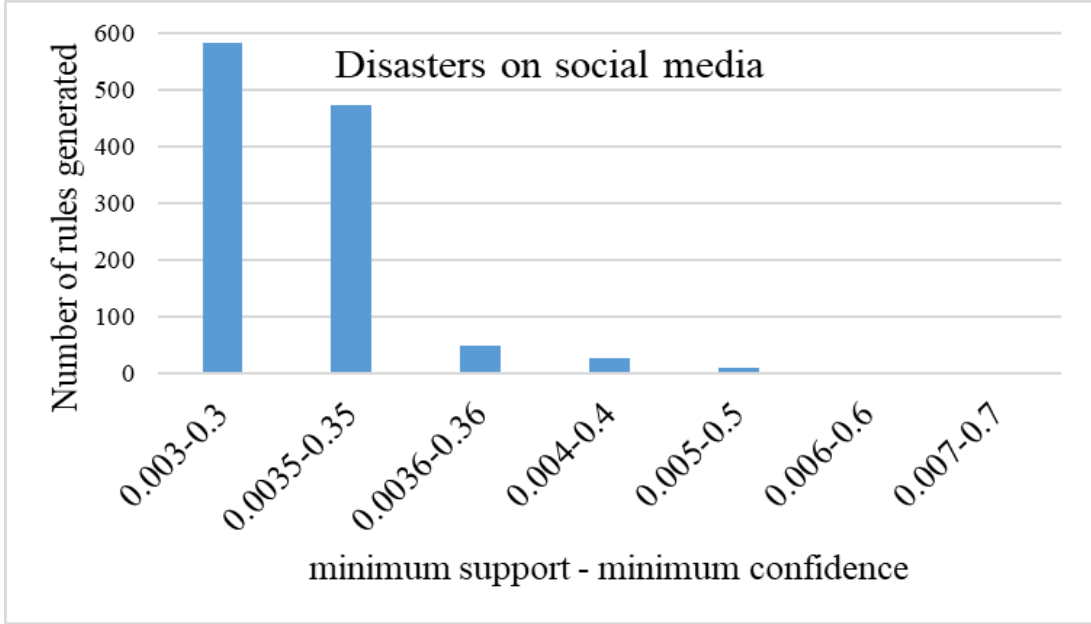
A series-graph showing the number of rules generated using various minimum support-confidence values for PFP-Growth execution for Positive and Negative Movie Review Dataset.

Analysis of Positive and Negative Movie Reviews

- [characters, bad, movie] together occurred 744 times
- [bad, film] occurred 2436 times
- Similarly, with minimum support = 0.05 and minimum confidence = 0.3 rules such as : [watching, movie] → [bad] with minimum confidence 0.4275 were found,
- This means
  - when "watching" and "movie" occurs together
  - then "bad" will also occur,
  - as they are associated with each other.
- An experiment was also performed on a positive movie review dataset where various rules were generated and words used in the rules are positive feelings of audiences about movies in positive reviews.



It shows the number of rules generated using various minSup-minConf values for PFP-Growth execution for Political Social Media Dataset.



Showing number of rules generated using various minSup-minConf values for PFP-Growth execution for Disasters Social Media Dataset.

Analysis of Political Social Media Dataset

- Data mining on political social media dataset can be helpful in
  - Making decisions related to applied policies or for future expectations of voters.
  - Analyzing drawbacks in implemented policies.
- With support = 0.003 [obama, congress, president], together occurred 18 times and [enforcement, law] occurred 20 times.
- For minimum support = 0.004 and minimum confidence = 0.4:

  [passed, act] → [house], 0.6285;

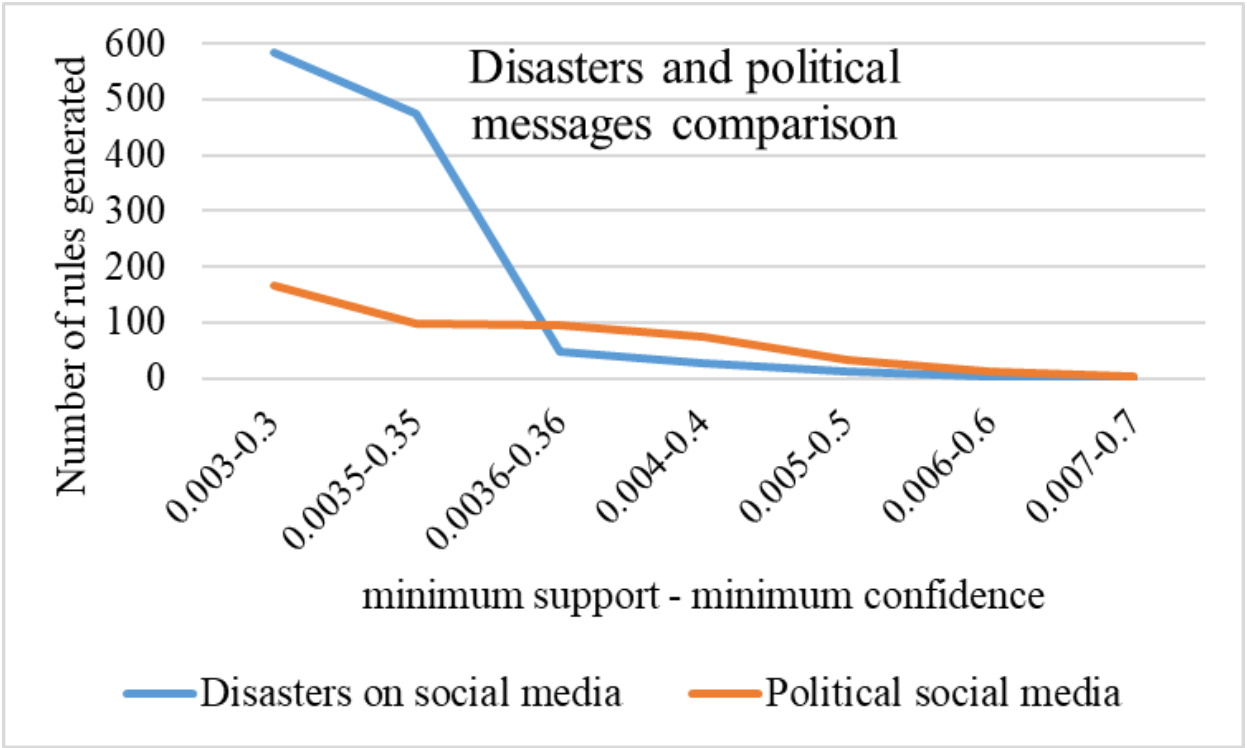
  [passed, bill] → [house], 0.5111 was generated.

Analysis of Disasters on Social Media Dataset

- During various disasters, social media has appeared as a platform to provide awareness.
- It provides information about the disaster, to get connected to the community, family, and friends.
- Social media sometimes leads to false or fake news too.
- With minimum support = 0.003 and minimum confidence = 0.3 rule such as:

  [homes, northern, california] → [wildfire] with confidence 0.973

- This means
  - when "homes", "northern" and "california" occurs together
  - then "wildfire" will also occur, as they are associated with each other.



A series-graph showing the number of rules generated using various minimum support-confidence values for PFP-Growth execution for Disasters and Political Social Media messages Dataset.

Time Comparison of PFP-Growth and FP-Growth

Datasets	Time required in FP-Growth			PFP-Growth
	Preprocess	mining	total	
Negative Movie Reviews	201 minutes 31 seconds	1311 milli seconds	202 min (approx.)	52 seconds
Positive Movie Reviews	177 minutes 49 seconds	1408 milli seconds	178 min (approx.)	48 seconds
Political Social Media Dataset	22 minutes 14 seconds	433 milli seconds	23 min (approx.)	18 seconds
Disaster Social Media Dataset	28 minutes 30 seconds	245 milli seconds	29 min (approx.)	11 seconds

Related Works

Authors [references]	Contributions
Chaturvedi and Sri Khetwat [7]	discussed parallel frequent pattern mining method; helpful in determining the interesting patterns in social media.
Chaturvedi et al. [8] [9] [10] [11]	performed network rule mining and pattern analysis using novel threshold: minimum stability (minStab).
Zhuang et al. [12]	proposed an approach on movie review mining in which each of the sentences of review, they found a feature of opinion pairs, and then found whether the opinion is positive or negative.
Braun et al. [13]	presented a survey of Spark based pattern mining on the big data generated IoT with fog computing.
Filgueira et al. [14]	presented a Spark based architecture SparkFlow for Genome data analytics.

Conclusion

- To mine three datasets of social media, we used the two approaches
  - Frequent Pattern Growth (FP-Growth) and
  - Parallel Frequent Pattern Growth (PFP-Growth).
- We observed that FP-Growth needs an appropriate input format for the execution.
- We observed some pre-processing time is required only once,
  - then it performs mining faster.
- Whereas, PFP-Growth does not require any pre-processing to generate interesting rules.
- The rules information is helpful in decision-making.

References

[1] Rakesh Agrawal, Tomasz Imieliński, and Arun Swami. "Mining association rules between sets of items in large databases". ACM SIGMOD Record. Vol. 22. No. 2. ACM, (1993).

[2] Rakesh Agrawal, and Ramakrishnan Srikant. "Fast algorithms for mining association rules." Proceeding of the 20th Int. Conf. Very Large Data Bases, VLDB. Vol. 1215. 1994.

[3] Han Jiawei, Jian Pei, and Yiwen Yin. "Mining frequent patterns without candidate generation". ACM SIGMOD Record. Vol. 29. No. 2. ACM, (2000).

[4] P. Fournier-Viger, et al. "SPMF: a java open-source pattern mining library." J. Mach. Learn. Res. 15.1 (2014): 3389-3393.

[5] "Mining frequent itemsets using the FP-Growth algorithm", Accessed on Feb 2023 https://www.philippe-fournier-viger.com/spmf/FPGrowth.php

[6] Li Haoyuan, et al. "PFP: Parallel FP-Growth for query recommendation". Proceedings of the 2008 ACM Conference on Recommender Systems. ACM, (2008).

[7] S. Chaturvedi, and S. K. Saritha. "Parallel Frequent Pattern Mining on Natural Language-Based Social Media Data." Emerging Technologies in Data Mining and Information Security: Proceedings of IEMIS 2018, Volume 3. Springer Singapore, 2019.

References

[8] A. Chaturvedi, A. Tiwari, and N. Spyratos. “minStab: Stable network evolution rule mining for system changeability analysis”. IEEE Trans. on Emerging Topics in Computational Intelligence 5.2 (2019): 274-283.

[9] A. Chaturvedi, A. Tiwari, and N. Spyratos. “System Network Analytics: Evolution and Stable Rules of a State Series.” arXiv preprint arXiv:2210.15965 (2022).

[10] A. Chaturvedi, and A. Tiwari. “System evolution analytics: Evolution and change pattern mining of inter-connected entities.” 2018 IEEE Int. Conf. on Systems, Man, and Cybernetics (SMC). (pp. 3877-3882) IEEE, 2018.

[11] A. Chaturvedi. “Call Graph Evolution Analytics over a Version Series of an Evolving Software System.” 37th IEEE/ACM Int. Conf. on Automated Software Engineering. 2022.

[12] L. Zhuang, F. J., and X.-Y. Zhu. “Movie review mining and summarization.” Proceedings of the 15th ACM Int. Conf. on Information and Knowledge Management. ACM, 2006.

[13] P. Braun, et al. "Pattern mining from big IoT data with fog computing: models, issues, and research perspectives." 2019 19th IEEE/ACM Int. Symp. on Cluster, Cloud and Grid Computing (CCGRID). IEEE, 2019.

[14] R. Filgueira, et al. "SparkFlow: Towards high-performance data analytics for spark-based genome analysis." 2022 22nd IEEE Int. Symp. on Cluster, Cloud and Internet Computing (CCGrid). IEEE, 2022.

ขอบคุณ
Thai

Grazie
Italian

תודה רבה
Hebrew

धन्यवादः
Sanskrit

ಧನ್ಯವಾದಗಳು
Kannada

Ευχαριστώ
Greek

Thank You
English

Gracias
Spanish

Спасибо
Russian

Obrigado
Portuguese

شكراً
Arabic

https://sites.google.com/site/animeshchaturvedi07

Merci
French

多謝
Traditional
Chinese

धन्यवाद
Hindi

Danke
German

多谢
Simplified
Chinese

நன்றி
Tamil
Tamil

ありがとうございました
Japanese

감사합니다
Korean