Evaluation of Conflicts among Reddit Posts using Clustering Techniques

Srinidhi Gopalakrishnan
Khoury College of Computer Sciences
Northeastern University
Boston Massachusetts USA
gopalakrishnan.s@husky.neu.edu

Nanditha Sundararajan
Khoury College of Computer Science
Northeastern University
Boston Massachusetts USA
sundararajan.n@husky.neu.edu

ABSTRACT

Reddit is an online discussion platform where posts related to a variety of subjects are created and shared among 330 million monthly users. While most of the posts comply with the 'Reddiquettes', approximately 10% of the posts generate conflicts in different groups or subreddits due to their negative sentiment. Clustering subreddits based on features of posts created in them and identifying magnitude of conflicts directed towards subreddits in the same or other clusters using a self-defined metric called 'Conflict Score', will help the Reddit Security and Post-Control Community in controlling network traffic and prevention of conflicts.

INTRODUCTION

Reddit is an American social news aggregation, web content rating and discussion website. Registered users submit content to the site as posts, images and links [1]. Posts consists of a Title and a Body. The posts can be upvoted or downvoted by other users based on their sentiment towards the post. The users can also choose to share it in other user-created groups called subreddits.

Subreddits pertain to different users' interests and are centralized around topics such as sports, politics, religion, etc. Users can join multiple subreddits and can post their views in them. According to the Reddit content policy [2], posts should follow certain guidelines to avoid causing disputes among users. For example, a content is prohibited if it 'incites or encourages violence' or 'threatens, harasses or bullies others'. Despite such regulations, users continue to post content which, when shared among Reddit communities, cause a chain of conflicts and written abuses. Identification of such interactions may help the Reddit community in facilitating peaceful interactions among all users.

To make this facilitation easier, we aim to group subreddits into clusters based on the features extracted from the titles of their posts. Further, we consider posts which are shared in other subreddits to assess the magnitude of conflict generation. Through the creation of clusters, we capture the interaction of closely related subreddits with those classified otherwise.

This interaction is defined by a self-defined metric, 'Conflict Score', which is an aggregation of number of posts leading to conflicts. From this, we create a weighted, directed graph with vertices as clusters generated and edges with weight as conflict score between the clusters.

We present the results to the Reddit Security and Post-Control Community in the form of the weighted, directed graph that explains which Reddit communities create the most offensive posts and trigger negative emotions. This enables the Post-Control communities to take measures such as issuing an official warning to users contained in that subreddit, keeping posts hidden from affected subreddit groups, or blocking them permanently from the Reddit community.

RELATED WORK

Earlier work on capturing community interaction and conflicts on the web involved manual tagging of subreddits and early detection of conflicts between communities using LSTM models [3]. Also, extracting inter-community conflicts in Reddit has been explored without grouping of subreddits and directly constructing a conflict graph based on upvote and downvote counts in subreddit posts [4]. However, manual tagging is time consuming and involves human intervention. Bias in tagging and hard classification of subreddits leads to inaccurate results. Also, there is a high chance that important features in defining posts are not considered during manual tagging.

We hope to close the gaps in the previous approaches by automating the grouping of subreddits into clusters and by deploying dimensionality reduction techniques that consider all important features in a lower dimension space. Future classification of subreddits can be based on the cluster profiles created in this paper.

BACKGROUND INFORMATION

In the context of our proposal, the definition of posts as a function of the words is crucial in our ability to cluster them.

We employ the Linguistic Inquiry and Word Count dictionary to analyze the text within the title of the posts and break it down into features such as percentage of words belonging to certain categories or topics (e.g., LIWC_Bio represents percentage of words related to biological processes or terms such as digest, health, pain).

Due to the presence of a large number of LIWC features (~100 features), it is important to use a dimensionality reduction technique that can help discover latent features and represent the data in lower dimensions. It can also be used to visualize features of the posts in two dimensions and understand similarities among subreddits in that space.

The representation of data after dimensionality reduction ensures that the clustering algorithm groups the subreddits into well-defined clusters. By well-defined, we mean that the points in the same cluster are close to each other and are far away from points in other clusters.

Finally, for better interpretation, we construct a directed, complete graph that explains the interactions among clusters. This is done by computing the 'Conflict Score' and using it as edge weights in the graph.

PROPOSED APPROACH

Given a list of LIWC features for a post generated in a source subreddit and shared in a target subreddit, we propose to group the subreddits along with other similar subreddits and give a measure of the magnitude of conflicts from one group of subreddits to other groups of subreddits.

Source subreddits: subreddit where post link originates

Target subreddits: subreddit where post link ends

Conflict score(x,y) = 1-mean(c) where c = -1, if post from x creates conflicts in y, 1 otherwise.

Given the data, we iterate through the following dimensionality reduction techniques:

- PCA: Statistical technique that uses orthogonal transformation to convert a set of possibly correlated variables to linearly uncorrelated components. It uses Eigen transformations to compute the loadings of each of the features on to the principal components after normalization of initial data [5].
- Factor Analysis: Technique to describe variability among observed, correlated variables in terms of lower number of unobserved variables called factors incorporating more domain specific assumptions [6].
- SVD: Procedure that decomposes data into matrices that can be reduced to explain data in lower dimensions.

  $$M_{[m\,x\,n]} = U_{[m\,x\,r]}\Sigma_{[r\,x\,r]}V^*{}_{[r\,x\,n]}$$

  Where M: Input data matrix, U: Left singular vectors, Σ : Singular values, V^*: Right singular vectors
- UMAP: Technique that nonlinearly embeds high dimensional data with fuzzy topological structure into low dimensional manifold. Though it is primarily used in visualization in two dimensional spaces, for our project, we are treating it as a dimensionality reduction technique [7].

Following dimensionality reduction, we iterate through the clustering techniques:

- K-Means++: Method to partition data into 'K' clusters in Euclidian space with initial cluster centroids placed far away from each other. This uses Expectation-Maximization to find the local minimum on the sum of squared distance between the data points and the cluster centroids [8].
- Spectral Clustering: Technique that employs graph theory and identifies communities of nodes in a graph based on connectivity among points. This works well even for data represented in non-Euclidian space [9].
- Gaussian Mixture Model: A soft clustering technique that fits Gaussian components on data points. This uses Expectation-Maximization to find the optimal Gaussian components given the data [10].
- Hierarchical Clustering: A bottom-up approach that groups clusters based on their similarity. Unweighted average linkage is considered as the similarity metric [11].

The clusters generated are evaluated using the below scores:

- Silhouette score: Indicates similarity between the objects of its own cluster (cohesion) compared to other clusters (separation) [12].

  $$s(i) = \begin{cases} 1 - a(i)/b(i), & \text{if } a(i) < b(i) \\ 0, & \text{if } a(i) = b(i) \\ b(i)/a(i) - 1, & \text{if } a(i) > b(i) \end{cases}$$

  where, a(i) is the mean distance between point 'i' and all other data points in the same cluster as it and b(i) is the smallest mean distance of 'i' to all points in any other cluster, of which 'i' is not a member.

  The Silhouette score varies between -1 to 1, the closer it is to 1, the better is the quality of cluster.
- Davies Bouldin score: This is a function of the ratio of intra-cluster scatter to the inter-cluster separation.

  $$DB = \frac{1}{N}\sum_{i=1}^{N} max_{j\neq i}\frac{S_i + S_j}{M_{i,j}}$$

  where, S_i is the measure of scatter within the cluster C_i, $M_{i,j}$ is a measure of separation between the clusters C_i and C_j and N is the number of clusters. Values of DB closer to 0 indicate better clusters [13].

After finalizing the clusters, we define Conflict score(x,y) between the clusters x and y = 1 - mean(c) where c = -1, if post from cluster x creates conflicts in cluster y, 1 otherwise.

Using the conflict scores between the clusters as edge weights, we construct a weighted, directed graph G with clusters as vertices.

Algorithmic Representation of Proposed Approach:

Step 1: Perform random sampling of data
Step 2: Normalize sampled data and check distribution and correlation of features
Step 3: Perform feature selection based on results from Step 2
Step 4: Perform dimensionality reduction using above mentioned techniques
Step 5: Cluster on the low dimensional data; calculate number of points in each cluster, Silhouette score, Davies Bouldin score
Step 6: Select clustering algorithm that gives most uniform and well-defined clusters
Step 7: Interpret cluster meaning through feature distribution
Step 8: Construct graph with vertices as cluster labels and edges as calculated conflict score

EXPERIMENTS

Experimental Setup:

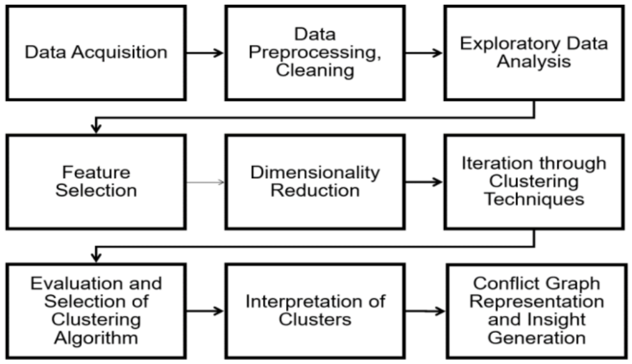

Data Acquisition
Data Preprocessing, Cleaning
Exploratory Data Analysis
Feature Selection
Dimensionality Reduction
Iteration through Clustering Techniques
Evaluation and Selection of Clustering Algorithm
Interpretation of Clusters
Conflict Graph Representation and Insight Generation

Fig. 1 : Diagrammatic representation of workflow

Data Acquisition:

The dataset for this project was obtained from the Stanford Network Analysis Project (SNAP) library [14] which was developed to perform research on large social and information networks. It contains the following fields:

Feature name	Feature Description
Source Subreddit	Subreddit where link originates
Target Subreddit	Subreddit where link ends
Timestamp	Time of post
Link_sentiment	Label indicating conflict between source and target subreddit. -1 if there is conflict,1 otherwise
LIWC features	Features representing fraction of words relating to family, money, technology, etc. from Linguistic Inquiry and Word Counts Dictionary (LIWC)

Fig. 2 : Description of Dataset

The number of data points in this dataset is 500,000.

Data Preprocessing and Cleaning:

The dataset was sampled randomly to accommodate non-linear algorithms such as Hierarchical Clustering. Also, not all Target subreddits were present among the list of Source subreddits. Hence, they were appended as Source subreddits to the dataset with features of their original Source subreddits. The LIWC features which were present in the dataframe as a single vector per row, were expanded to 86 features. The variables were standardized using Min-Max Scaling technique to bring them to identical scale. There were no null values in the data and all features were converted to their appropriate data types.

Exploratory Data Analysis

We observed the distribution of LIWC features on the entire dataset and compared it with the distributions after separating it into conflict-causing and normal posts. Below were the conclusions drawn:

- Word count of conflict-causing posts = 1.5 * Word count of normal posts.
- Higher negative sentiment in conflict-causing posts compared to normal posts – more swear words and words with negative connotation.
- Most LIWC variables have value 0 for an average of 90% of different datapoints. This is expected as we cannot have posts that talk about all topics. Some posts are skewed towards certain topics which increases the value of LIWC feature of that topic.

  This means that applying K-Means or GMM directly would not yield good results as all clusters would have similar profiles – all features tending to 0.

Feature Selection

Since the features used in clustering had to signify topics that were independent from each other, highly correlated features were removed and 14 features were retained. The retained features are:

Features	Meaning	Examples
LIWC_Swear	Swear words	damn, shit
LIWC_Social	Social behavior	mate, talk
LIWC_Neg_Emo	Negative sentiments	hurt, worried
LIWC_Pos_Emo	Positive sentiments	happy, love
LIWC_Percept	Perception	look, feeling
LIWC_Bio	Biological	eat, pain
LIWC_Relativ	Relativity	motion, area
LIWC_Work	Work related	job, majors
LIWC_Achieve	Achievements	win, success
LIWC_Leisure	Leisure activities	cook, movie
LIWC_Home	Home related	kitchen, landlord
LIWC_Money	Monetary	audit, cash
LIWC_Relig	Religion related	altar, church
LIWC_Death	Death related	bury, kill

Fig. 3 : Retained features where all features pertain to proportion of corresponding words

Dimensionality Reduction

PCA was applied on the dataset and the number of components was reduced from 14 to 11 as shown in Figure 4. On further inspection of loadings of features onto the principal components and clustering with the components, we observed that different clusters were defined by similar principal components as shown in Figure 5.

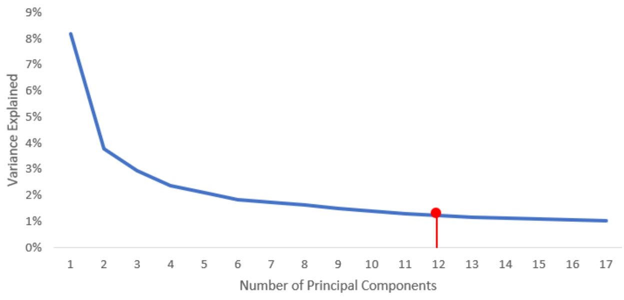
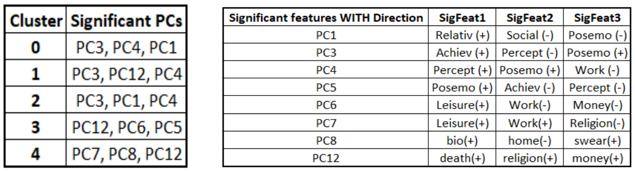


Fig.4 : Scree plot for PCA



Cluster	Significant PCs
0	PC3, PC4, PC1
1	PC3, PC12, PC4
2	PC3, PC1, PC4
3	PC12, PC6, PC5
4	PC7, PC8, PC12

Significant features WITH Direction	SigFeat1	SigFeat2	SigFeat3
PC1	Relativ (+)	Social (-)	Posemo (-)
PC3	Achiev (+)	Percept (-)	Posemo (+)
PC4	Percept (+)	Posemo (+)	Work (-)
PC5	Posemo (+)	Achiev (-)	Percept (-)
PC6	Leisure(+)	Work(-)	Money(-)
PC7	Leisure(+)	Work(+)	Religion(-)
PC8	bio(+)	home(-)	swear(+)
PC12	death(+)	religion(+)	money(+)

Fig.5 : K-means post PCA, Significant features of each Principal component

Similar results were obtained when Factor analysis was applied on the data. Thus, it was required to reduce the data to two dimensions using UMAP.

Iterations through Clustering techniques

Dim Redctn Algorithm	Clustering Algorithm	Number of Clusters	Silhouette Score	Davies Bouldin Score	Notes
PCA	Kmeans++	7	0.1912	1.6601	Low Silhouette,Hard to interpret
UMAP	Hierarchical	5	0.3088	0.9126	Unbalanced number of points in cluster
UMAP	Spectral	4	0.397	0.8571	Low Silhouette compared to Kmeans++
UMAP	GMM	5	0.4016	0.8241	Low Silhouette compared to Kmeans++
UMAP	Kmeans++	5	0.4172	0.8044	Good Silhouette,Easy to interpret reduced features

Fig.6 : Select iterations with evaluation metrics

For K-means++ and GMMs, the number of clusters was determined using the K-means Scree plot. For Spectral and Hierarchical clustering, the number of clusters was determined by evaluating the Silhouette score for each number of clusters and selecting the value which produced the highest Silhouette score.

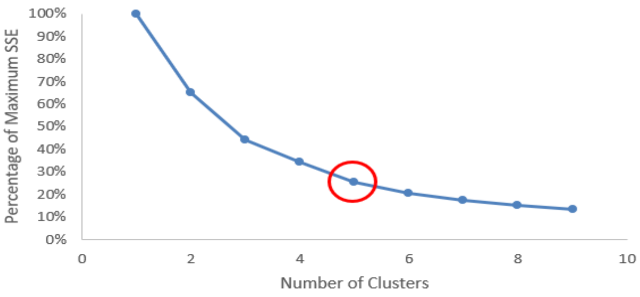
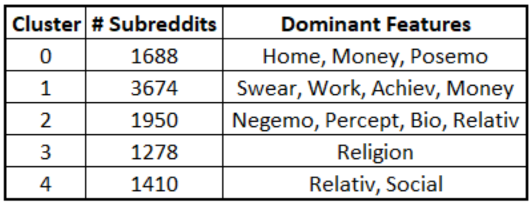


Fig.7 : K-means Scree plot (K=5)

We observe that UMAP followed by K-means++ gave well-defined, interpretable clusters based on LIWC feature distribution and a good Silhouette score of 0.4172.

Interpretation of Clusters



Cluster	# Subreddits	Dominant Features
0	1688	Home, Money, Posemo
1	3674	Swear, Work, Achiev, Money
2	1950	Negemo, Percept, Bio, Relativ
3	1278	Religion
4	1410	Relativ, Social

Fig.8: Final Cluster characteristics

For each cluster, dominant features were identified based on deviation of mean from complete data and used to define cluster labels. For example, in Cluster 0, the feature LIWC_Home has a mean value 55% greater than the entire data. Similarly, the dominant features were derived for all 5 clusters.

Insights and Conflict Graph Representation

Based on the results, we conclude that we cannot differentiate between offensive posts for which the Source subreddits are to be blamed and non-controversial posts which were misinterpreted as offensive by the Target subreddits. The conflict graph in Figure 9 simply shows whether the directed interactions between clusters created conflicts or not.

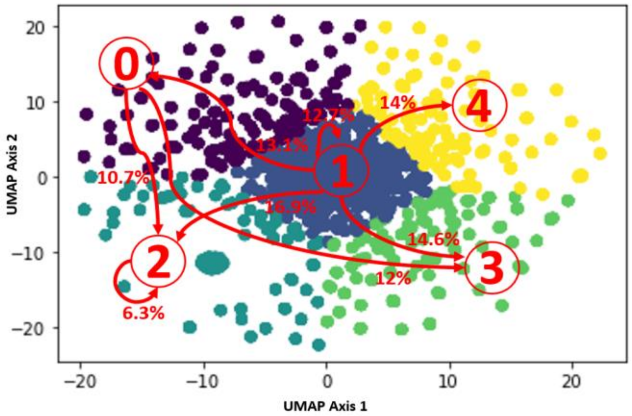


Fig.9 : Graphical Representation of Conflict Score

From the graph, it is inferred that the cluster which talked about Monetary and Materialistic subjects (Cluster 1), though in a positive sense, triggered conflicts or was misinterpreted by other clusters. Also, the clusters which talked about Monetary and Materialistic subjects and Work Politics were found to cause conflicts among themselves and other clusters (Clusters 0 and 1). Cluster 2 which has Negative Emotions as one of the dominant features, seemed to create conflicts irrespective of source of posts. Thus, the Post-Control Community can be recommended to monitor the activities of users belonging to this cluster.

Surprisingly, posts related to Religion or Social Causes (Clusters 3 and 4) were non-controversial among Reddit communities.

CONCLUSIONS AND FUTURE WORK

In this project, we overcame the shortcomings of previous Reddit conflict resolution techniques that involved manual interventions by proposing an approach of automating the grouping of similar subreddits using dimensionality reduction and clustering algorithms. We discovered latent communities that were important in identifying conflict magnitude and paths. We defined the scope of the problem statement and identified the appropriate data to achieve our goal. After data preprocessing and exploratory data analysis, we iterated through various dimensionality reduction techniques and clustering algorithms and found that UMAP followed by K-means++ gave a good Silhouette score and interpretable clusters.

With the interpretations, we calculated the conflict scores and provided a graphical representation of the subreddits and the magnitude of conflicts among them.

However, the absence of the actual text data proved to be detrimental in the acquisition of certain crucial features that would have further enhanced the results.

Thus, extracting more text-based features and potentially user demographic information might produce significant results. Also, newer variations of clustering algorithms such Affinity Propagation [15], Mean Shift [16] and dimensionality reduction algorithms like t-SNE [17] could prove to be effective in solving Reddit conflicts compared to existing methods.

REFERENCES

[1] Ohlheiser, Abby (2016). "Reddit will limit the reach of a pro-Trump board and crack down on its 'most toxic users'". Washington Post. ISSN 0190-8286. OCLC 2269358. Archived from the original on 2017-01-14.

[2] https://www.redditinc.com/policies/content-policy

[3] Kumar, Hamilton, Leskovec, Jurafsky (2018). "Community Interaction and Conflict on the Web", arXiv:1803.03697v1.

[4] Datta, Adar (2019). "Extracting Inter-Community Conflicts in Reddit". Proceedings of the Thirteenth International AAAI Conference on Web and Social Media (ICWSM 2019).

[5] Abdi. H. & Williams, L.J. (2010). "Principal component analysis". Wiley Interdisciplinary Reviews: Computational Statistics. 2 (4): 433–459. arXiv:1108.4372. doi:10.1002/wics.101.

[6] Timothy A. Brown (2006). "Confirmatory Factor Analysis for Applied Research Methodology in the social sciences". Guilford Press.

[7] McInnes, Healy, Melville (2018). "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction". arXiv:1802.03426.

[8] Arthur, D.; Vassilvitskii, S. (2007). "k-means++: the advantages of careful seeding" (PDF). Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms. Society for Industrial and Applied Mathematics Philadelphia, PA, USA. pp. 1027–1035.

[9] Ng, Andrew Y and Jordan, Michael I and Weiss, Yair (2002). "On spectral clustering: analysis and an algorithm" (PDF). Advances in Neural Information Processing Systems.

[10] Carl Edward Rasmussen (2000). "The Infinite Gaussian Mixture Model". In Advances in Neural Information Processing Systems 12 S.A. Solla, T.K. Leen and K.-R. Muller (eds.), pp. 554–560, MIT Press.

[11] Rokach, Lior, and Oded Maimon (2005). "Clustering methods." Data mining and knowledge discovery handbook. Springer US. 321-352.

[12] Peter J. Rousseeuw (1987). "Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis". Computational and Applied Mathematics. 20: 53–65. doi:10.1016/0377-0427(87)90125-7.

[13] Davies, David L.; Bouldin, Donald W. (1979). "A Cluster Separation Measure". IEEE Transactions on Pattern Analysis and Machine Intelligence. PAMI-1 (2): 224–227. doi:10.1109/TPAMI.1979.4766909.

[14] https://snap.stanford.edu/data/soc-RedditHyperlinks.html - Extracted dataset of directed connections between Subreddits.

[15] Wang, Zhang, Li, Zhang, Guo (2007). " Adaptive Affinity Propagation Clustering". Acta Automatica Sinica, 33(12):1242-1246.

[16] Yizong Cheng (1995). " Mean Shift, Mode Seeking, and Clustering". IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 17, NO. 8.

[17] van der Maaten, Hinton (2008). " Visualizing Data using t-SNE". Journal of Machine Learning Research 9 2579-2605.