Generating insights about financial asks from Reddit posts and user interactions

Sachin Thukral
TCS Research
Delhi, India
sachi.2@tcs.com

Suyash Sangwan
TCS Research
Delhi, India
suyashsangwan16@gmail.com

Vipul Chauhan
TCS Research
Delhi, India
vipul.chauhan5@gmail.com

Arnab Chatterjee
TCS Research
Delhi, India
arnab.chatterjee4@tcs.com

Lipika Dey
TCS Research
Delhi, India
lipika.dey@tcs.com

Abstract—As an increasingly large number of people turn to platforms like Reddit, YouTube, Twitter, Instagram, etc. for financial advice, generating insights about the content generated and interactions taking place within these platforms have become a key research question. This study proposes content and inter-action analysis techniques for a large repository created from social media content, where people's interactions are centered around financial information exchange. We propose methods for content analysis that can generate human-interpretable insights using topic-centered clustering and multi-document abstractive summarization. We share details of insights generated from our experiments with a large repository of data gathered from subreddit for personal finance. We have also explored the use of ChatGPT and Vicuna for generating responses to queries and compared them with human responses. The methods proposed in this work are generic and applicable to all large social media platforms.

Index Terms—social network analysis, behavior, finance, Reddit

I. INTRODUCTION

Several recent studies have shown that, unlike their prede-cessors, the millennials and the generations next are looking for financial advice on social media. Rather than friends or family, they prefer to rely on the advice of domain experts vetted by their peers. A recent article published by Forbes [1] states that nearly 80% of Americans in the age group of 18 to 41 get their financial advice from social media, among which Reddit and YouTube are the most trusted platforms. Independent surveys conducted among students and young professionals from Indonesia [2] and Uzbekistan [3] also high-lighted the strategic role played by social media in imparting financial literacy to this group, thereby influencing their finan-cial acts. Given the above context, there is increasing interest in gaining deeper insights into the communities themselves. While the survey-based studies presented responses from a limited number of users, text analysis techniques can be

applied to gather large-scale insights about the users, user-generated content, the nature of user participation, and the type of interactions manifesting within these communities. Business analysts and strategy planners take a keen interest in these insights as it provides them with a glimpse of the needs and psyche of their potential consumers, much beyond their customer base, and thereby provide them actionable intelligence to design new products, increase customer base as well as improve customer satisfaction.

In this paper, we have proposed methods for analyzing large social communities to derive insights about content and user interactions from them. We present a pipeline for completely automated analysis of textual content, along with the application of generative techniques to create human interpretable insights. We have also proposed a few novel measures for measuring interactions that can help in assessing the probable interest in and influence of generated content. The key contributions of the paper are summarized as follows:

1. A discovery-driven text analysis framework using topic modeling and topic-driven clustering is proposed for analyzing posts and comments from a large social media community. A novel text clustering algorithm has been proposed in Sec. II-A1. The proposed algorithm is capable of discovering cohesive and distinctive topical clusters from large repositories of text data.

2. We propose a few novel measures to capture user interaction patterns, which provide deeper contextual insights about how users interact with different kinds of content, in Sec. II-B.

3. We have created a large repository of publicly available data containing personal finance-related discussions, from a dedicated subreddit named *r/personalfinance* which has around 18 million members. The repository was analyzed using the proposed pipeline and a lot of interesting insights has been obtained. The results, presented in Sec. III reveal insights about financial issues of persistent interest, emerging issues, and also the nature of human engagement with content.

The content analysis and interaction assessment measures proposed in this paper are fairly generic and can be applied to analyze exchanges over any large social media platform.

II. A FRAMEWORK FOR ANALYZING LARGE SOCIAL PLATFORMS

The proposed social sensing framework consists of two key components whose functions and desirable qualities are defined as follows:

(a). Content sensing and insight generation module - this module is responsible for analyzing large repositories of unstructured text in an unsupervised manner. Given that content generation happens in an unrestricted fashion, *discoverability* is an important aspect of this component. Another desired feature of such a solution would be to ensure that all significant issues present in a repository have been identified.

(b). Interaction sensing module - this module measures user participation in terms of different activities permitted on a social media platform like posting, commenting, upvoting or downvoting, sharing, etc. These measures help in computing the amount of user interest in content based on the extent of interactions taking place around it, and thereby help in quantifying its utilization.

A. Content Sensing using Topical Clustering

We have used the Latent Dirichlet Allocation Machine Learning for Language Toolkit (LDA MALLET) for extracting topics [4] from the posts. The MALLET topic model package consists of a fast and scalable implementation of Gibbs sampling, thereby making it highly efficient for analyzing large repositories. We argue that fewer documents should end up with equal probability for all topics, and hence *skewness* of topic distribution for the posts is used to determine the optimal number of topics.

For each post P_i, let $T_1, T_2, \ldots, T_k$ be the topical distribution of all k topics given by LDA MALLET, and $\mu(P_i)$, $\sigma(P_i)$ and $\mathcal{M}(P_i)$ be their mean, standard deviation and median respectively. 'Skewness' is a measure of the asymmetry of the probability distribution of a real-valued random variable around its mean, and is defined as:

$$\text{Skewness}(P_i) = \frac{3(\mu(P_i) - \mathcal{M}(P_i))}{\sigma(P_i)}.$$

A positive skewness, in which the average topical distribution is greater than the median value of the distribution indicates that at least one of the topics is significantly present. A zero or negative skewness on the other hand indicates that none of the topics is significantly present in the post. For a given k, let W_k denote the count of posts with negative skewness measures. Starting with a low positive integer value for k, say 2, LDA is run repeatedly to extract k topics, $Skewness(P_i)$ is computed for all posts P_i, W_k that denotes the number of posts with negative skewness is computed and compared with earlier values of k. The value for k at which the minimum value of W_k is obtained, is accepted as the number of topics. This ensures that a minimum number of posts are left with equal assignment of topic probabilities, essentially indicating that their key topics cannot be determined correctly.

1) A hybrid text clustering algorithm: Topic modeling provides high-level grouping of content, but to obtain deeper insights for deriving actionable intelligence, we now propose a topical clustering algorithm that groups together semantically similar posts within the same topic. This algorithm is executed only for representative posts of each topic, which, for a topic T_j are chosen as those in which $P(T_j) > \tau$, where $\tau = 1/(k - \sqrt{k})$. The threshold was determined through experimentation and found to be capable of identifying representative posts.

The proposed clustering algorithm is incremental in nature. Starting with a random data point as the first cluster center, it adds on cluster centers judiciously, one at a time, ensuring that the clusters created around each center are cohesive within themselves and sufficiently distinct from each other.

Cluster cohesivity and distinctiveness are measured using the "Word Mover's Distance (WMD)", which is known to capture both semantic and syntactic similarities between text documents effectively. WMD is a hyperparameter-free measure that computes the dissimilarity between two text documents as the minimum distance that the embedded words of one document need to "travel" to reach the embedded words of another document. The travel cost of one word x_i to another word x_j is defined by $c(i,j)$, the Euclidean distance between the embeddings of x_i and x_j, and is computed as $c(i,j) = ||x_i - x_j||_2$. The word mover distance between two documents A and B is computed as a cumulative function of word frequencies and the minimum distance required for each word from the source document A to a word in the target document B. It is computed as $min_{T \geq 0}(\sum_{i,j=1}^{n} \kappa_{ij} * c(i,j))$, where κ_{ij} denotes how many times x_i in document A transforms into word x_j in document B.

We modify and extend the above formula to compute distances among multiple sets of documents, where each set eventually represents a cluster. Rather than using all words, we use a collection of keywords and phrases, identified using the YAKE tool [5] to represent a set of documents. We replace the frequency of terms with the term frequency-inverse document frequency measure ($TF-IDF$), which can effectively capture the significant strings for each set. Keywords are embedded using Word2Vec.

Let us assume that there are n sets of documents, each denoted by S_j, where $j = 1, \ldots, n$. Let R denote the union of all keywords extracted by YAKE. For each set S_j, compute $TF-IDF$ score for all elements $w \in R$ as follows:

$$\text{TF} - \text{IDF}(w,j) = tf(w,j) \times \log \frac{n}{df_w}$$

where $tf(w,j)$ = number of occurrences of w in S_j,
n = total number of sets,
df_w = number of sets that contain w.
Select the top m keywords of S_j sorted with respect to their $TF-IDF$ scores to represent the set. We have used $m = 10$ for our experiments.

For each pair of sets S_i and S_j the word mover distance

between them is computed as

$$WMD(S_i, S_j) = \sum_{w=1}^{m} TF - IDF(w, j) \times c_w(i, j),$$

where $c_w(i, j)$ = minimum traveling cost of any keyword w in S_i to a keyword in S_j.

The incremental clustering algorithm works as follows. Each post is converted into a single document vector using SBERT embeddings. Starting with the entire set of documents D, a random point is chosen to represent the first cluster center. The goal of the algorithm is to find subsequent cluster centers as far away as possible from all existing centers. The steps followed are presented and explained below:

- Let c_1 denote the randomly picked up first cluster center.
- Calculate the distance of all documents $d \in D$ from c_1 using the Euclidean distance between their SBERT embeddings.
- Select the furthest point from c_1 as the second cluster center c_2,

$$\mathbf{c_2} = \max_{d \in D} ||d - c_1||_2.$$

- Assign each document $d_i \in D$ to the cluster whose centre is closest to it, yielding two clusters C_1 and C_2. Recompute their cluster centers z_1 and z_2 as mean of all vectors assigned to the respective clusters.
- Compute $\chi = WMD(C_1, C_2)$ as defined earlier.
- At this point, posts are divided into two clusters and $k = 2$. Repeat the following steps till χ stops increasing.
- For each document $d \in D$, compute $\delta_{max}(d) = max_k(||d - z_k||_2)$.
- Choose document p for which $\delta_{max}(p) \geq \delta_{max}(d)$ for all $d \in D$, as the new cluster center. Increase k to $k+1$.
- Redistribute all documents $d_i \in D$ to the cluster whose center is closest to it, and recompute the cluster centers z_k as means of all vectors assigned to them.
- Compute $\chi = \frac{1}{k} \sum_{i,j=1,...,k} (WMD(C_i, C_j))$, as the average word mover distance among all k clusters. If new χ is greater than earlier χ, then we can say clusters have moved further apart, hence this is better than the previous state.

Cluster-level insights are generated using a summary of summaries approach. Posts are first summarized using OpenAI ChatCompletion model **gpt-3.5-turbo**. Cluster summaries are generated from these summaries using the same model. A few sample summaries are presented later along with other results in Sec. III.

B. Measuring Platform Engagement

For a financial platform, while post content revolves around financial issues, products, services, questions or express needs for advice etc., comments in response to the posts made by authors other than the post-creator, captures the latter's interest and contribution in the post content. Voting also indicates user interest, though in a passive way. The following measures capture user interest in a post :

Active Engagement around a post is measured by the number of comments received by a post from other users. These could be in the form of suggestions, opinions, answers, or asking more questions. For computing the active engaged users (authors) on a post, we have taken the number of unique users that have commented on a post. Active engagement for a topic is computed by the average number of comments the topic has gotten over all of its representative posts.

Passive Engagement around a post is measured by the total score received by the post and its comments. For Reddit, the total score is given by the difference between upvotes and downvotes. The exact number of each is not known. Since each user can vote only once, the score provides the minimum number of users who must have engaged with the post, which are our passive-engaged users (voters). Passive engagement for a topic is computed as the average score the topic has received over all its representative posts and comments.

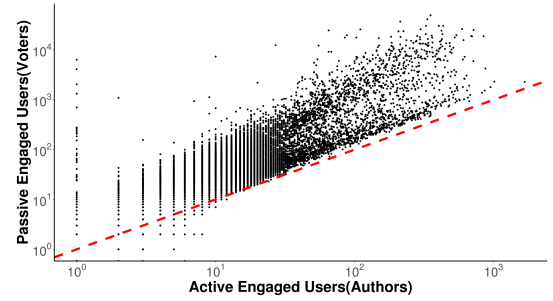**Total Engagement** is obtained by adding the above two quantities.

Fig. 1. Engagement of Active users vs Passive users

Fig. 1 shows a plot of active versus passive engagements computed from a financial subreddit dataset that contains $134, 521$ posts and $1, 521, 940$ comments made by $237, 718$ users. The plot shows that passive engagement for the majority of the posts is much higher than active engagement. The red dotted line indicates the equilibrium between active and passive engagements. Though the engagements have mostly garnered overall positive scores, a few dots below the red dotted line depicts posts that have garnered more downvotes than upvotes. Interestingly, their engagement levels are also very high. The vertical line of dots at the extreme left shows the number of votes received by posts, which may not have received a single comment. Manual inspection revealed that these are announcements or advertisements for financial services and products, thus indicating that people like to engage more with people than advertisers on these platforms.

III. Experiments, Results and Observations

We now present insights from a large dataset which contains data for a two-year period ranging from 1st July 2020 to 30th June 2022. The data was gathered from the subreddit *r/personalfinance*, using *Pushshift API* [6], and divided into two subsets, each of one-year duration. Table I presents a comprehensive account of users, posts, and comments for the

two time periods. It shows an increase in the amount of content as well as number of users creating them over the period. Fig. 2 presents the topics extracted for the two timelines, along

Fig. 3. Topic-wise comparison of (Left) active and (Right) passive engagement in both the timelines

Timeline	Posts	Comments	Users
July 2020 - June 2021	134,521	1,521,940	237,782
July 2021 - June 2022	131,045	2,041,536	252,572

with their respective strengths. The topic names have been manually assigned based on the representative topic words returned by LDA. A number of topics like *credit cards, stocks, annual tax, etc.* are present throughout. A new topic related to *health insurance* has emerged in the second period, clearly due to the prevalence of the pandemic. Topics like *savings* and *loans* disappeared as independent topics in the second period.
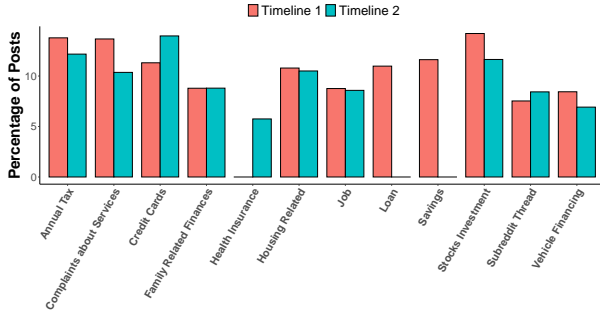

Fig. 2. Topic-wise percentage of Posts in two timelines

Fig. 3 presents the levels of active and passive engagement with the different topics over the two time periods. It may be noted that while for most of the topics, both the active and passive engagement levels have increased over the years, passive engagement shows around five times increase for topics related to *Family Related Finances*, and *Complaints about Banking Operations*. Clearly, these platforms are helping people learn from mutual experiences. The increase in passive engagement also shows that a lot of users, who may not have shared their own queries, are following the discussions, presumably to get enriched. Thus, the platform is also playing a role in spreading awareness about managing personal finance.

On clustering the representative posts for each topic for each timeline, each topic yielded five to ten clusters. Cluster summaries generated for all using ChatGPT were also compared with an independently generated manual summary for validation. It was found that the abstractive cluster summaries provide a succinct view of content. More details about a few key topic clusters obtained from the summaries are shared below to provide an idea about the nature of financial knowledge exchange occurring within the platform.
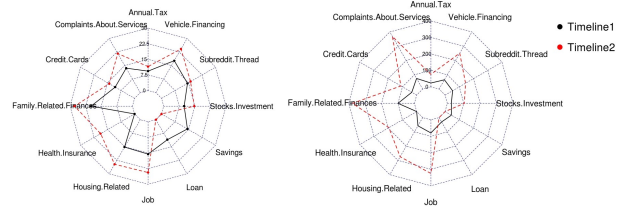
Credit Cards: User queries about credit card management clusters are around (i) strategies for improving credit scores and changing credit limits, (ii) canceling credit cards after missing payments (iii) vehicle refinancing with a credit card, and (iv) paying medical debt with a credit card.

Stocks and Investment: Clusters are around (i) investment queries about Vanguard, Fidelity, and 401k options, (ii) operational issues with brokerage accounts, (iii) investment advice for retirement accounts, and (iv) seeking expert advice on portfolios and fund diversification.

Family Related Finances: Posts in this topic cluster are around (i) financially troubled parents, (ii) users suffering due to past family debts, (iii) seeking financial advice for mental health issues in family, and (iv) issues related to inheritance of properties and wealth.

Complaints about Services: Discussions are centered around (i) issues with online refunds and purchases, (ii) identity theft, and (iii) challenges faced while closing bank accounts due to fraud. As people share their personal experiences, the discussions play an active role in creating awareness about fraud and scams.

Job: These discussions are around (i) concerns about promotion and raise, (ii) seeking advice on new job offers, (iii) exploring saving options after a new job, and (iv) issues with current employer and advice sought on leaving job.

Health Insurance: Discussion around health insurance schemes stood out as a unique topic during the second timeline. The clusters found are (i) seeking opinions on health plans, (ii) advice on recovering high medical bills charges, and (iii) sharing personal experiences about insurance plans.

Since the content in the platform is mostly around seeking and providing advice, we wanted to understand how a chatbot's response to these queries might compare with human responses. A small subset of the top 20 posts from each topic of Timeline 1 were fed to ChatGPT and Vicuna [7], and the responses were manually analyzed. The key observations about the differences in nature of human and chatbot responses are summarized in Table II.

In summary, human responses on social platforms are shaped by social psychology and individual judgment. These are not based on surface-level interpretation of queries. They seek clarification and show empathy. Multiple answers provide diverse perspectives. Shared experiences connect users. Passive engagement helps in acquiring financial knowledge from these platforms. In a possible futuristic scenario, if chatbots

Human Responses	Machine-Generated Responses
Large variation in length of responses	Responses are either very long or very short
Humans tend to ask questions for clarification before answering	Lacking in this ability
Ability to make decisions on user's behalf	List pros and cons and not take responsibility
May disagree with user's viewpoint	Mostly acquiescent to user's stance
More responses provide multiple viewpoints	One answer - single viewpoint
Shows empathy	Lacks Empathy
Share their own experiences	Only give to do instructions

are chosen as the only way to talk about financial needs, the positive aspects of passive learning and obtaining diverse viewpoints from social media conversation would be lost. This may lead to biased judgments and reduced financial awareness.

IV. RELATED WORK

Behavioral Finance is the study of investors' psychology while making financial decisions. This involves analysis of market-level data, like returns and trading volume, to study the trading behavior of investors. Social media captures the activity of individuals, interactions among them, or more precisely, the complex behavior of a society. There are studies on Twitter [8] that attempt to predict returns, volatility, trading volume, and survey sentiment indices – suggesting that mining social media may provide valuable actionable intelligence to investors. Another study suggests that differences in personalities drive decision-making in different ways [9]. Individuals interact with each other in many ways, but determining how they interact and uncovering the function of social patterns can be done using social network analysis. A recent review on social network analysis can be found in [10]. LDA has been used predominantly in finding the discussion topics in text data. In both of these [11], [12], topics were extracted using LDA from Tweets, and it was reported that LDA is able to find the most relevant and accurate topics. Another recent study focuses on finding the discussion topics in the finance domain using Reddit data [13]. It used RAKE to find the keywords and then distributed these keywords into topics. They have a very small data set to study for this domain as compared to ours. Recently, LLMs are also being used to study financial behavior through social media. One such study on Reddit has generated sentiment labels for the posts, using LLMs for understanding the market sentiment and being able to build a market sentiment model around it [14]. Another study uses LLMs for financial time series forecasting on the NASDAQ

stocks [15]. Another recent work in which users are able to develop a finance-specific LLM model named BloombergGPT, but they have said it is yet not up to the benchmark [16].

V. CONCLUSION

Users in social discussion platforms related to finance interact with each other expressing their opinions, concerns, or points of view, and also sharing their experiences. We have proposed a discovery-driven analytical framework deploying topic modeling and topic-driven clustering to derive insights about content, and also assess different types of human engagements. A large repository of public data was curated from the subreddit *r/personalfinance* and insights derived from it using the proposed framework are presented. Further, we also try to assess a futuristic scenario where users may directly engage with a chatbot for answering their financial queries, rather than engage in a platform. A concern that emerges is that while social discussions involve a multitude of people and their varying points of view, one-on-one conversations with chatbots will lack multiple perspectives. Conversations may become biased in unknown ways, and its effect will be manifested on the financial activities of users. A lot of users, who are not posing questions, but following the conversation will also be affected. This might affect the level of financial knowledge in a wider population.

One of our future tasks would be to analyze the responses in more detail and assess biases, if any. For example, we find that users often pose queries about how brands compare with each other to possibly obtain a more balanced view. Our intent is to assess whether the responses are controlled by the brands themselves. We would also like to conduct more studies to understand the demography better, for more actionable intelligence.

REFERENCES

[1] J. Egan, "Nearly 80% Of Young Adults Get Financial Advice From This Surprising Place," https://www.forbes.com/advisor/investing/financial-advisor/adults-financial-advice-social-media/, 2023.

[2] H. Yanto, N. Ismail, K. Kiswanto, N. M. Rahim, and N. Baroroh, "The roles of peers and social media in building financial literacy among the millennial generation: A case of indonesian economics and business students," *Cogent Social Sciences*, vol. 7, no. 1, p. 1947579, 2021.

[3] G. Isomidinova, J. S. K. Singh, and K. Singh, "Determinants of financial literacy: a quantitative study among young students in tashkent, uzbekistan," *Electronic Journal of Business & Management*, vol. 2, no. 1, pp. 61–75, 2017.

[4] R. Rehurek and P. Sojka, "Gensim–python framework for vector space modelling," *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic*, vol. 3, no. 2, p. 2, 2011.

[5] R. Campos, V. Mangaravite, A. Pasquali, A. Jorge, C. Nunes, and A. Jatowt, "YAKE! Keyword extraction from single documents using multiple local features," *Information Sciences*, vol. 509, pp. 257–289, 2020.

[6] Pushshift, "Calling API for dumping dataset," https://pushshift.io/api-parameters, 2021, accessed: 2021-01-01.

[7] W.-L. Chiang, Z. Li, Z. Lin, Y. Sheng, Z. Wu, H. Zhang, L. Zheng, S. Zhuang, Y. Zhuang, J. E. Gonzalez *et al.*, "Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality," *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2023.

[8] N. Oliveira, P. Cortez, and N. Areal, "The impact of microblogging data for stock market prediction: Using twitter to predict returns, volatility, trading volume and survey sentiment indices," *Expert Systems with applications*, vol. 73, pp. 125–144, 2017.

[9] M. A. Sattar, M. Toseef, and M. F. Sattar, "Behavioral finance biases in investment decision making," *International Journal of Accounting, Finance and Risk Management*, vol. 5, no. 2, p. 69, 2020.

[10] D. E. Froehlich and J. Brouwer, "Social network analysis as mixed analysis," in *The Routledge Reviewer's Guide to Mixed Methods Analysis*. Routledge, 2021, pp. 209–218.

[11] H. Yin, S. Yang, and J. Li, "Detecting topic and sentiment dynamics due to Covid-19 pandemic using social media," in *International Conference on Advanced Data Mining and Applications*. Springer, 2020, pp. 610–623.

[12] D. R. Prabhakar Kaila and D. A. Prasad, "Informational flow on twitter–corona virus outbreak–topic modelling approach," *International Journal of Advanced Research in Engineering and Technology (IJARET)*, vol. 11, no. 3, 2020.

[13] V. Karpenko, K. Mukhina, D. Rybakova, I. Busurkina, and D. Bulygin, "A study of personal finance practices. the case of online discussions on reddit." in *IMS*, 2021, pp. 206–211.

[14] X. Deng, V. Bashlovkina, F. Han, S. Baumgartner, and M. Bendersky, "What do llms know about financial markets? a case study on reddit market sentiment analysis," *arXiv preprint arXiv:2212.11311*, 2022.

[15] X. Yu, Z. Chen, Y. Ling, S. Dong, Z. Liu, and Y. Lu, "Temporal data meets llm–explainable financial time series forecasting," *arXiv preprint arXiv:2306.11025*, 2023.

[16] S. Wu, O. Irsoy, S. Lu, V. Dabravolski, M. Dredze, S. Gehrmann, P. Kambadur, D. Rosenberg, and G. Mann, "Bloomberggpt: A large language model for finance," *arXiv preprint arXiv:2303.17564*, 2023.