LDPP at the FinNLP-2022 ERAI Task: Determinantal Point Processes and Variational Auto-encoders for Identifying High-Quality Opinions from a pool of Social Media Posts

Paul Trust
University College Cork
Cork, Ireland

Rosane Minghim
University College Cork
Cork, Ireland

Abstract

Social media and online forums have made it easier for people to share their views and opinions on various topics in society. In this paper, we focus on posts discussing investment related topics. When it comes to investment , people can now easily share their opinions about online traded items and also provide rationales to support their arguments on social media. However, there are millions of posts to read with potential of having some posts from amateur investors or completely unrelated posts. Identifying the most important posts that could lead to higher maximal potential profit (MPP) and lower maximal loss for investment is not a trivial task. In this paper, propose to use determinantal point processes and variational autoencoders to identify high quality posts from the given rationales. Experimental results suggest that our method mines quality posts compared to random selection and also latent variable modeling improves improves the quality of selected posts.

1 Introduction

The internet revolution and the social media era has made it easy for the public to create and share information including their opinions about certain aspects in society like politics (Chambers et al., 2015), economy (Pekar and Binner, 2017), finance (Chen et al., 2021b) and investment (Wang et al., 2020). When it comes to investment, it is now so simple for people to share their opinions about online traded items in online platforms, stock investment websites in real time. These numerous public comments are of great value in reflecting market conditions and making trading decisions.

The open nature of most of these online platforms means that anyone can share any information whether they are experts on the topic being discussed or not. This presents a serious challenge in identifying high quality opinions especially for critical purposes like investment from such a large crowd of mined results. When people are giving their investments opinions, they provide supporting augments which we define as rationales supporting their reasoning. In the paper, we use the rationales behind the view points by sorting out the opinions that would to higher maximal potential profit (MPP) and lower maximal loss (ML) (Chen et al., 2021a).

The majority of the previous studies have lied on the idea of large numbers using average results obtained from popular tasks for example opinion mining and sentiment analysis. The most recent and competitive approach by (Chen et al., 2021a) identifies posts of high quality by making an assumption that these posts will have similar characteristics as those written by experts.

In this work, we present an approach based on idea that high quality opinions are those that are less redundant but at the same time highly valuable. Unlike the previous approaches, we do not use any documents written by experts since they may not be available but rather only base on contextualized representations from the provided rationales using sentence transformers (Reimers and Gurevych, 2019). We select begin by identifying groups of similar opinions by performing joint dimensionality reduction and deep clustering on the embedding space of the opinions using variational autoencoders (Märtens and Yau, 2020). Determinantal point process (Kulesza and Taskar, 2010) are then used to select a set of representative opinions from the groups identified by variational autoencoders while maintaining high diversity among them.

2 Related Work

Social media and online data have exponentially grown in the last few years and many domains are trying it to leverage to their advantage. Some previous works have focused on utilizing user-generated data from social media (Ghosh Chowdhury et al.,

Proceedings of the Fourth Workshop on Financial Technology and Natural Language Processing (FinNLP), pages 136 - 140
December 8, 2022 ©2022 Association for Computational Linguistics

2019; Rouhizadeh et al., 2018), online forums (Wang et al., 2010), and e-commerce platforms (Backus et al., 2020). Most existing works aims to find clusters, topics, classes or categories from social media data (Jiang et al., 2019; Preoţiuc-Pietro et al., 2019). These approaches usually take into account the law of large numbers and simply average the results extracted from tasks such as opinion mining and sentiment analysis.

Few of the previous works have focused on evaluating opinion quality. Zhongyu and Yang used feature-based methods using textual information in the comments and social interaction related features (Wei et al., 2016). Ying and Duboue provided an annotated pilot dataset and used a vanilla neural network with semantic information to classification (Ying and Duboue, 2019). The most recent work is then one by (Chen et al., 2021a) that leverages high-accuracy models trained on documents written by experts and the crowd to mine high quality opinions from the crowd. In contract, to the existing work, we take a purely unsupervised approach using determinantal point processes and variational autoencoders without assuming access to documents written by experts.

3 Methodology

This section describes our proposed methodology to identify the most import important opinions from a pool of opinions provided by amateur investors. Our methodology is based on the assumption that the most important articles should not be redundant but at the same time should contain the most valuable information that could lead to higher MPP and lower ML.

3.1 Text Representation

We obtain embeddings for all the input sentences using a pre-trained SBERT (Sentence Bidirectional Encoder Representations from Transformers) (Reimers and Gurevych, 2019). SBERT is a modification of the pre-trained BERT networks using Siamese and triplet networks, which make it able to derive semantically meaningful sentence embeddings. This model was trained using Stanford Natural Language Inference(SNLI) and Multi-Genre Natural Language Inference (MNLI) datasets. SNLI contained $570,000$ annotated sentence pairs and MNLI contained 430000 annotated sentence pairs.

3.2 Latent Variable Modeling

In this section, we use a variational autoencoder (a likelihood based deep generative model) for identifying the most interesting groups from a large collection of online posts. Deep generative models define a joint probability distribution over a set of random variables composed of multiple layers of hierarchies. Our methodology is based on an assumption that online posts belong to a certain unobserved latent space, and it is only sufficient to read through only representative posts from the same group.

More formally, let $x = \{x^{(i)}\}_{i=1}^{N}$ be a dataset consisting of N Independent and identically distributed ($i.i.d$) samples of a variable x in a potentially high-dimensional space. We make an assumption that data is generated by some random process involving an unobserved continuous random variable z in a much lower dimensional space. Suppose that z has a normal prior distribution $z \sim \mathcal{N}(0,1)$ and that $f^{\theta}(z)$ is a family of deterministic functions given by deep neural networks. The process of latent variable modeling involves of two steps: (1) Latent variables z are generated from some prior distribution $p(z)$. (2) Observed variables x are generated from some conditional distribution $p(x|z)$.

The goal here is to learn the model distribution $p(x)$ to fit parameters of the true data distribution as well as possible. This is achieved by minimizing the Kullback-Leibler (KL) divergence between the two distribution equivalent to maximum likelihood objective. Our focus is on latent variable models, which define the marginal log-likelihood via a latent variable z

$$\log p(x) = \log \int p(x|z)p(z)dz \tag{1}$$

Assuming that conditional likelihood is described by the Gaussian likelihood just like the prior distribution; $x_i|z_i, \theta \sim \mathcal{N}(f^{\theta}(z_i), \sigma^2)$.

Estimating $\log p_{\theta}(x)$ involves an intractable integral, VAE instead optimizes maximizing a variational lower bound $\mathcal{L}_{VAE}(x)$ on the log-likelihood $\log p(x) \geq \mathcal{L}_{VAE}(x)$ where:

$$\mathcal{L}_{VAE} = E_{q(z|x)}[\log p(x|z)] - KL(q(z|x)||p(z)) \tag{2}$$

In a standard VAE, the variational approximation $q(z|x)$ is known as the encoder and the latent variable z is known as the decoder. Since our interest lies in identifying the groups of similar articles from where we can sample from, we use a decoder that has a mixture prior during the decoding process as proposed by Martens and Yau (2020).

The mixture prior is introduces a set of basis functions $f_{basis}^{(k)}$ parameterized by neural networks. The decoder in the standard VAE is replaced with a basis decoder network $f_{basis} : \mathcal{R}^{\mathcal{Q}} \to \mathcal{R}^{\mathcal{P}}$ with output mapped to the data via a categorical random variable, that is for every data dimension $m \in \{1, .., M\}$

$$f_{decoder}^{(m)}(z) = \sum_{k=1}^{K} w^{m,k}.f_{basis}^{(k)}(z) \tag{3}$$

where $\{w^{j,1}, ...w^{j,K}\} \sim Categorical(\lambda_1, ..., \lambda_K)$ (Märtens and Yau, 2020).

We identify the high quality opinions by selecting from the categorical distributions $Categorical(\lambda_1, ..., \lambda_K)$ using determinantal point process (DPP).

3.3 Determinantal Point Process

Let $\mathcal{S} = \{1, .., n\}$ denote a finite ground set containing n items corresponding to all sentences from one of the K groups identified by the variational autoencoder. Our goal is to find subsets of sentences $s \subseteq \mathcal{S}$ from all the K group that are most likely to lead to higher maximal potential profit (MPP) and lower maximal loss (ML).

A point process $\mathcal{P}$ on a discrete set $\mathcal{S}$ is a probability measure on 2^S (the set of all possible subsets of $\mathcal{S}$). $\mathcal{P}$ is called a determinantal point process if there exists a positive semi-definite matrix L indexed by elements of $\mathcal{S}$ such that if $S \sim \mathcal{P}$, we have

$$\begin{aligned} \mathcal{P}(Y; L) &= \frac{det(L_s)}{det(L+I)} \\ \sum_{s \subseteq \mathcal{S}} det(L_s) &= det(L+I) \end{aligned} \tag{4}$$

where $det(.)$ is the determinant of a matrix; I is the identity matrix; $L \in \mathcal{R}^{n \times n}$ is a positive semi-definite matrix known as $L-$ensemble. L_{ij} is a measure of the correlation between sentences i and j, L_s is a sub matrix of L containing only entries indexed by elements of $s \subseteq \mathcal{S}$.

We decompose the kernel matrix $L-$ensemble matrix assuming L is Gram matrix adopted from (Kulesza and Taskar, 2010): $L_{ij} = q_i.Z_{ij}.q_j$ where $q_i \in \mathcal{R}^+$ is a positive real number indicating the quality of a sentence and Z_{ij} is a measure of similarity between sentences i and j. Let $s = \{i, j\}$ be a summary containing only two sentences i and j, its probability $\mathcal{P}(Y; L)$ can be computed as:

$$\begin{aligned} \mathcal{P}(Y = \{i,j\}; L) &\propto det(L_Y) \\ &= \begin{vmatrix} q_i Z_{ii} q_i & q_i Z_{ij} q_j \\ q_j Z_{ji} q_i & q_j Z_{jj} q_j \end{vmatrix} \\ &= q_i^2.q_j^2.(1 - Z_{ij}^2) \end{aligned} \tag{5}$$

If two sentences i and j are similar to each other, denoted by Z_{ij}, then any subset containing both sentences will have low probability of inclusion. The selected subset S achieving the highest probability thus should contain a set of high-quality sentences while maintaining high diversity among the selected sentences via pairwise repulsion.

4 Experimental Results

4.1 Evaluation

The quality of the top retrieved opinions are evaluated using maximal potential profit (MPP) and lower maximal loss (ML). To calculate MPP and ML, we follow the opinion of the post on day t when entering the market at opening price on day $t+1$. The maximum possible profit and the maximum loss are traced during the backtesting period to find the unrealized return of the trading based on the opinions of amateur investors. For bullish opinions posted on day t, MPP and ML are calculated as shown in Equation 6 (Chen et al., 2021a):

$$\begin{aligned} MPP_{bullish} &= (max(H_{(t+1,T)}) - O_{t+1})/O_{t+1} \\ ML_{bullish} &= (min(L_{(t+1,T)}) - O_{t+1})/O_{t+1} \end{aligned} \tag{6}$$

where O_t represents the opening price of the day t, $H_{t,T}$ denotes a list of the highest price of the day t to day T, $L_{t,T}$ denotes a list of the lowest prices of day t to day T, and T is the last day of the back testing period.

For bearish opinions posted on day t, the MPP and the ML are calculated as follows in Equation 7 (Chen et al., 2021a):

$$\begin{aligned} MPP_{bearish} &= (O_{t+1} - min(L_{(t+1,T)}))/O_{t+1} \\ ML_{bearish} &= (O_{t+1} - max(H_{(t+1,T)}))/O_{t+1} \end{aligned} \tag{7}$$

4.2 Data

The dataset used for experiments in this paper was provided by the organizers of the shared task on

Model	Average MPP	Average ML
Random	11.94%	-17.28%
DPP	11.34%	-6.77%
DPP-VAE	**14.81%**	**-5.85%**

Table 1: Experimental results showing average Maximal potential profit (MPP) and Maximal Loss (ML) of the the top 10% of the posts identified by proposed method (DPP-VAE) and the comparison methods

Evaluating the Rationales of Amateur Investors (ERAI) organized at FinNLP: The Fourth Workshop on Financial Technology and Natural Language Processing at EMNLP 2022. The dataset consists of of social media posts from 2019/05/13 to 2019/06/13 consisting of 210 texts of investors' opinions written in text (Chen et al., 2021a).

4.3 Experimental Setup

We used sentence transformers library (Reimers and Gurevych, 2019) to obtain sentence representations of opinions. BasicVAE(Märtens and Yau, 2020) was used for latent variable modeling and implementing translation invariant variational autoencoder. Determinantal Point Processes (DPP) were implemented using submodlib library (Kaushal et al., 2022).

4.4 Discussion

In this section, we discuss the results obtained with our model (DPP-VAE) and other comparison methods in terms of average maximal potential profit (MPP) and lower maximal loss (ML). Table 1 summarizes the average results of the top 10% of the posts on the ERAI dataset. For interpretation purposes, the higher the MPP and the lower the average absolute ML, the better the model performance.

Our results demonstrate that selecting the most important and diverse opinions from the a pool of investor opinions can lead to a lower average ML (-6.77% against -17.28%). Contrary to our expectations, a naive random selection for our case can sometimes be better than a careful selection in terms of average MPP (11.94% versus 11.34%). The difference in average lower maximal loss (ML) between random selection and DPP could be possibly be attributed to the fact that DPP select the most diverse opinions. This can be seen as a more risk-averse strategy of investment which would lead to lower maximal loss but not necessary higher profits.

Our proposed methodology (DPP-VAE) which combines latent variable modeling and DPP registers a significant performance gain in terms of average MPP over naive random selection (14.81% against 11.94%) and average ML (-5.85% versus -17.28%). These performance differences re-enforces that careful selection rather random selection of opinions or articles to read is very key to achieve optimal results.

Our experimental results as demonstrated in Table 1 also demonstrate the importance of latent variable modeling in reducing redundancy over selected opinions from the crowd. The performance difference can be attributed to the fact that the marginal benefit of reading two important articles from the same groups (assumed to be similar) is much less than reading two important articles from different groups.

However, the proposed method (DPP-VAE) is still out-performed by the top method proposed in (Chen et al., 2021a) in terms of average ML (-2.46% versus -5.77%) and average MPP (17.61% versus 14.81%). The difference in performance can be attributed that methods leverages high accuracy models trained on documents written by experts to mine top of opinions. Much as their method is unsupervised but stylistic and semantic features learned from expert documents contributes significantly to their performance. Our method assumes no access to any documents written by experts which in most cases may not be available and thus takes a completely unsupervised approach.

5 Conclusion

In this paper, we propose an approach that identifies the most diverse and important opinions a large pool of opinions as high quality opinions. The proposed approach approach (DPP-VAE) combines variational auto-encoders and determinantal point process (DPP) to mine top quality opinions. The rationale behind our methods is that looking at diverse and well-represented opinions from a large crowd is more likely to lead to average maximal potential profit (MPP) and lower maximal loss (ML). Experimental results reveal that our proposed method improves over baseline determinantal point process (DPP) and also achieves significant performance gains over random selection.

Results further reveal that the gain obtained from our method (DPP-VAE) on average ML is much more than that obtained on average MPP. We attribute this to the fact that reading diverse opinions from different investors may be seen as a more risk averse strategy. As future work, it is important to experiment how guiding a determinantal point process selection with a few opinions written by experts would boost its performance and also extending the methodology beyond the ERAI dataset.

References

Matthew Backus, Thomas Blake, Jett Pettus, and Steven Tadelis. 2020. Communication and bargaining breakdown: An empirical analysis. Technical report, National Bureau of Economic Research.

Nathanael Chambers, Victor Bowen, Ethan Genco, Xisen Tian, Eric Young, Ganesh Harihara, and Eugene Yang. 2015. Identifying political sentiment between nation states with social media. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 65–75, Lisbon, Portugal. Association for Computational Linguistics.

Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2021a. Evaluating the rationales of amateur investors. In *Proceedings of the Web Conference 2021*, WWW '21, page 3987–3998, New York, NY, USA. Association for Computing Machinery.

Chung-Chi Chen, Hen-Hsen Huang, Hiroya Takamura, and Hsin-Hsi Chen, editors. 2021b. *Proceedings of the Third Workshop on Financial Technology and Natural Language Processing*. -, Online.

Arijit Ghosh Chowdhury, Ramit Sawhney, Rajiv Ratn Shah, and Debanjan Mahata. 2019. #YouToo? detection of personal recollections of sexual harassment on social media. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2527–2537, Florence, Italy. Association for Computational Linguistics.

Jyun-Yu Jiang, Xue Sun, Wei Wang, and Sean Young. 2019. Enhancing air quality prediction with social media and natural language processing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2627–2632, Florence, Italy. Association for Computational Linguistics.

Vishal Kaushal, Ganesh Ramakrishnan, and Rishabh Iyer. 2022. Submodlib: A submodular optimization library. *arXiv preprint arXiv:2202.10680*.

Alex Kulesza and Ben Taskar. 2010. Structured determinantal point processes. *Advances in neural information processing systems*, 23.

Kaspar Märtens and Christopher Yau. 2020. Basisvae: Translation-invariant feature-level clustering with variational autoencoders. In *International Conference on Artificial Intelligence and Statistics*, pages 2928–2937. PMLR.

Viktor Pekar and Jane Binner. 2017. Forecasting consumer spending from purchase intentions expressed on social media. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 92–101, Copenhagen, Denmark. Association for Computational Linguistics.

Daniel Preoţiuc-Pietro, Mihaela Gaman, and Nikolaos Aletras. 2019. Automatically identifying complaints in social media. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5008–5019, Florence, Italy. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Masoud Rouhizadeh, Kokil Jaidka, Laura Smith, H. Andrew Schwartz, Anneke Buffone, and Lyle Ungar. 2018. Identifying locus of control in social media language. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1146–1152, Brussels, Belgium. Association for Computational Linguistics.

Heyuan Wang, Tengjiao Wang, and Yi Li. 2020. Incorporating expert-based investment opinion signals in stock prediction: A deep learning framework. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 971–978.

Jia Wang, Qing Li, Yuanzhu Peter Chen, and Zhangxi Lin. 2010. Recommendation in Internet forums and blogs. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 257–265, Uppsala, Sweden. Association for Computational Linguistics.

Zhongyu Wei, Yang Liu, and Yi Li. 2016. Is this post persuasive? ranking argumentative comments in online forum. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 195–200, Berlin, Germany. Association for Computational Linguistics.

Annie Ying and Pablo Duboue. 2019. Rationale classification for educational trading platforms. In *Proceedings of the First Workshop on Financial Technology and Natural Language Processing*, pages 14–20, Macao, China.