Detecting Polarization in Ratings: An Automated Pipeline and a Preliminary Quantification on Several Benchmark Data Sets

Mahsa Badami*, Olfa Nasraoui†, Welong Sun‡, Patrick Shafto §
Computer Engineering and Computer Science Department
University of Louisville, Louisville, Kentucky, USA
* *Email: Mahsa.Badami@Louisville.edu,* † *Olfa.Nasraoui@Louisville.edu,* ‡ *Wenlong.Sun@Louisville.edu*
§ *Department of Mathematics and Computer Science*
Rutgers University - Newark, Newark, NJ, USA
Email: Patrick.Shafto@Gmail.com

***Abstract*—Personalized recommender systems are becoming increasingly relevant and important in the study of polarization and bias, given their widespread use in filtering information spaces. Polarization is a social phenomenon, with serious consequences, in real-life, particularly on social media. Thus it is important to understand how machine learning algorithms, especially recommender systems, behave in polarized environments.**

In this paper, we study polarization within the context of the users' interactions with a space of items and how this affects recommender systems. We first formalize the concept of polarization based on item ratings and then relate it to the item reviews to investigate any potential correlation. We then propose a domain independent data science pipeline to automatically detect polarization using the ratings rather than the typical properties used to detect polarization, such as item's content or social network topology.

We perform an extensive comparison of polarization measures on several benchmark data sets and show that our polarization detection framework can detect different degrees of polarization and outperforms existing measures in capturing an intuitive notion of polarization. Our work is an essential step toward quantifying and detecting polarization in ongoing ratings and in benchmark data sets, and to this end, we use our developed polarization detection pipeline to compute the polarization prevalence of several benchmark data sets. It is our hope that this work will contribute to supporting future research in the emerging topic of designing and studying the behavior of recommender systems in polarized environments.

***Keywords*-Recommender System; Polarization; Controversy; Classifier; Big data; Sentiment Analysis; NLP**

I. Introduction

The growing popularity of online services and social networks and the trend to integrate Recommender Systems (RS) within most e-commerce applications and social media platforms to help filter data to the users, has led to a dynamic interplay between the information that users can discover and the algorithms that filter such information. This has given rise to several side effects, such as algorithmic biases [1], filter bubbles [2] and polarization [3]. For instance, polarization occurs when information happens to be related to controversial issues where a population is divided in groups with opposite opinions, fewer individuals with moderate opinions, and there is no clear majority in "ardent supporters" and "motivated adversaries" [3], [4], [5].

Polarization around controversial issues has arguably affected recommender systems (and vice-versa) due to the ill-fated consequences of this phenomenon [6], [7]. An effective and efficient recommender system needs to be able to apply the most suitable recommendation method even in the presence of a set of polarized items. Trust based recommender systems tries to offer a solution to this problem by defining a trust network for each user [8]. These types of recommender systems leverage the user's trust network's opinion on an item to finally decide whether to recommend an item. However, when such issues emerge on social media, we often observe the creation of 'echo chambers', where there is greater interaction between like-minded people who reinforce each other's opinion [6]. These individuals do not get exposed to the views of the opposing side, and this in turn results in polarization [1]. Allowing users to discover different viewpoints could allow them to develop unique tastes and diverse perspectives that are not limited by filter bubbles [9]. To this end, the preliminary work, presented in this paper, is motivated by the fundamental, yet challenging task of detecting polarized/controversial items across diverse platforms, rather than studying the evolution of polarization.

A. Contributions

Contrary to other models where polarization is based on either a social graph or item content, we propose a model that works with ratings in any domain and on a large scale. Ratings are more intuitive to work with since they directly capture the distribution of user opinions. In addition, an item itself is not polarized unless there are users with opposing opinion on the item, so the content may not be very reliable for polarization detection. In the absence of polarization, the distribution of opinions is either J-shaped or bell shaped. However, as polarization emerges, the resulting distribution shifts to a U-shaped distribution with two peaks

978-1-5386-2715-0/17/$31.00 ©2017 IEEE

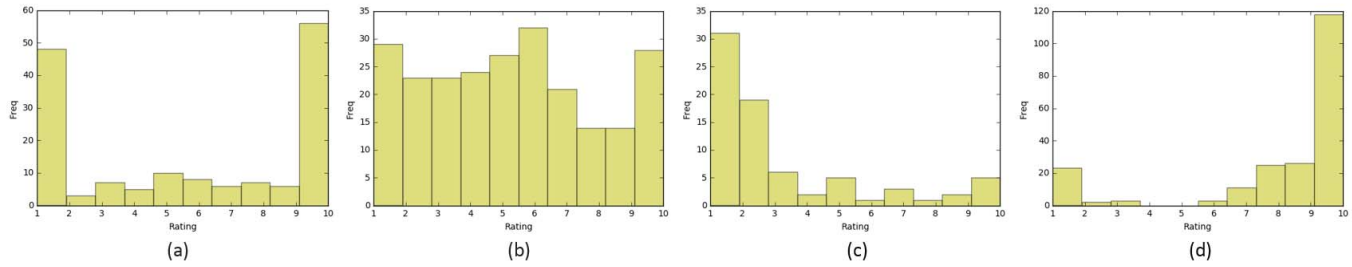


Figure 1: Different rating distributions for movies from from IMDb.com.

emerging around the two dominant and confronted opinions at the extreme sides of the rating scale [3], [10]. Different examples of such distributions are shown in figure 1.

We develop a new approach to quantify polarization based on three stages: (i) building an items' ratings histogram from user-item rating data; (ii) extracting a set of features from the histograms; (iii) training a polarization classifier based on a sample of annotated cases; and (iv) measuring the item-level polarization score. To verify our approach, we apply a multi-pooling text classifier, which combines both labeled comments/reviews and a word lexicon to understand the item's polarization and its possible relation with the item's comments.

We offer a systematic lightweight pipeline with simple yet effective features based on an item's ratings that can be used in any domain. Due to the simplicity, generality, and speed of this framework, it can be used with any recommender system, in diverse platforms such as news-reading and public-debate scenarios. We use human intuition to capture polarization, in other words "what" causes a human to be undecided about the overall opinions on this item by looking at the ratings distribution.

Polarization is an important phenomenon, with serious consequences in real-life, particularly on social media. Thus it is important to understand how machine learning algorithms, especially recommender systems, behave in a polarized environment; and to this end it is important, as a prerequisite step, to quantify polarization in existing and new data sets. Our contribution is an essential step toward quantifying and detecting polarization in ongoing ratings on generic platforms and in benchmark data sets, on a large scale. Hence it can support future research in the emerging topic of designing and studying the behavior of recommender systems in polarized environments.

The rest of the paper is organized as follows. Section II reviews related work. Section III presents our polarization detection approach. Section IV presents preliminary results, and finally we conclude in Section VI.

II. Related Work

Research on polarization in recommender systems is rapidly emerging as an important inter-disciplinary topic [11], [1], [6], [12], with some efforts to decrease online polarization, especially in recommender systems [6], [2], [13]. Although various models have been proposed, and from different perspectives, there is not yet a general agreement on how to define polarization. Polarization has been investigated from a network perspective mainly using the social network structure, content and sentiment of discussions, in order to compute a polarization score [13]. However, this type of information is expensive to extract or not always available. Hence, another line of work has studied polarization based on the ratings provided by users on items.

A simple but naive way to detect a polarized item is to inspect the standard deviation on the item's ratings. However, this is unable to distinguish between a flat and U-shaped rating distribution which represents diversity and polarization, respectively. To overcome this limitation, [8] considered the standard deviation of adjacent scores with respect to the total number of received ratings, while [14] presented a polarization measure based on the geometric mean of likes and dislikes' distributions to investigate the existence of a local and a global regime. Following a similar intuition, [3] used a general formula as a function of difference in the two opposite population sizes and their distance. They applied their methodology to a Twitter conversation about the late Venezuelan president, Hugo Chavez, to study the effect of polarization.

In addition, [10] studied the temporal evolution of reviews. To measure polarization, they used standard deviation, average and kurtosis of a given item's ratings. Kurtosis is the fourth central moment of the ratings distribution and captures bi-modality. The study presented a model to show that users adopt more extreme opinions when they disagree with the average opinion and this is one reason behind polarization that increases over time.

Although the aforementioned recent efforts tried to detect polarization, they lack an algorithmic approach that works in a domain-independent manner, and they used metrics that are either too simple to fully capture an item's polarization or too expensive and complicated to be used as a component of other algorithms, such as recommender systems. In addition, since existing efforts compute a score, this score must be binarized based on a threshold to finally decide if an item

is polarized or not. Finding this threshold is a challenging task as it is domain specific.

To overcome these limitations, we developed a data-driven machine learning pipeline to learn an optimal polarization classifier using engineered features that are based on rating distributions. When these features are used to build and train a classifier, the result is a Polarization Detection Classifier (PDC) that works in a domain- and language-independent manner to detect and quantify polarization in a variety of domains. The similarity between the above-mentioned approaches and our work is limited to the intuition of detecting a U-shaped ratings distribution. However, our methodology is the first ratings based polarization detection classifier which is predominantly different than those found in other studies. In order to handle the case where no individual ratings are available (e.g. IMDb), we resort to the preliminary extra step of mapping reviews to ratings to build a ratings' histogram. In addition, we evaluate our polarization pipeline on diverse real world data, and demonstrate that it outperforms existing polarization scores.

A. Sentiment Classification

Movie reviews contain emotional expressions by users about movies. Understanding the sentiment of movie reviews is critical to get first-hand information about the movies. Many prior efforts in this field followed a knowledge-based approach, which simply used the sentiment-polarity labeled words defined beforehand in lexicons and classified the text according to its linguistic patterns. An alternative approach is based on training machine learning models, on documents labeled with positive/negative sentiment, to learn text classifiers that distinguish the sentiments in domain-specific documents [15]. This solved the problem of adaptation for different and changing domains, however manual annotation requires too much human labor. Prem et al. proposed an approach that combined lexical knowledge and a multinomial Naive Bayesian classifier model [16] to build successful sentiment classifiers for complex practical problems [17].

III. Proposed Pipeline

We propose a methodology to estimate the polarization of an item using a combination of features computed based on the item's rating distribution. Our methodology consists of two main steps: first, we construct a feature set from the histograms; and then, we train a binary classifier model to estimate the polarization score for each item. Figure 2 shows the proposed polarization detection pipeline along with the sentiment classifier component.

A. Problem Definition

Definition 1 - Polarization: *Given an environment* $G = (U, I, R)$*, user* $u \in \mathbb{R}^{1\times n}$ *had rated item* $i \in \mathbb{R}^{m\times 1}$ *with rating* $r_{ui} \in \mathbb{R}^{m\times n}$ *on a scale of* x *to* y*. Item* i *'s polarization score* ϕ_i *is defined as the spread of its ratings* r_i*. We say the item is polarized if* $\phi_i \geq \delta$*.*

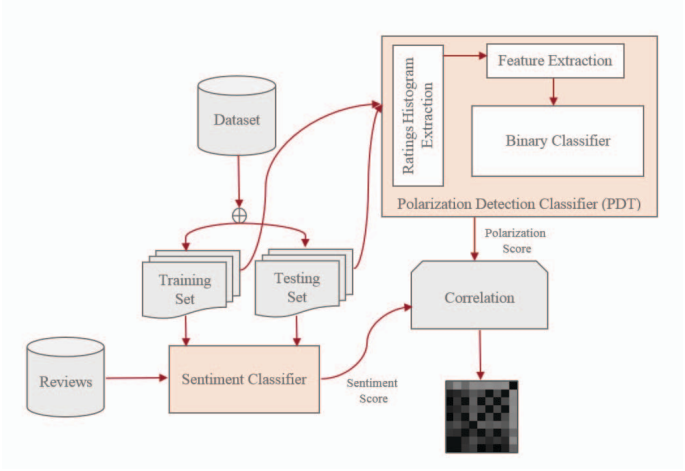


Figure 2: Proposed Polarization Detection Pipeline

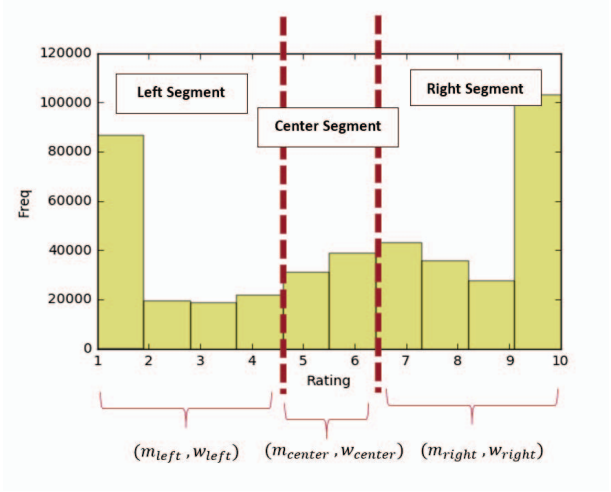


Figure 3: Proposed feature set 1, based on an item's rating histogram

B. Polarization Detection Classifier

1) Feature Extraction: It is important to distinguish between "polarization" and "diversity", see figure 1a and 1b, respectively. Although both are social phenomena and can be sometimes related, they are not the same. A polarized item is only diverse in two opposite directions and this is sometimes considered a negative phenomenon in social studies. On the other hand, diversity is generally considered desirable as a social good since it represents different points of views or different varieties of choices [18]. By focusing on the representation of the polarization phenomena within a distribution of user ratings, we introduce a set of features which capture polarization from the perspective of how the users react to the items, and not from the similarity or difference between the items' descriptions.

In general, there are three types of rating distributions, namely J-shaped, U-shaped and flat-shaped distributions, as

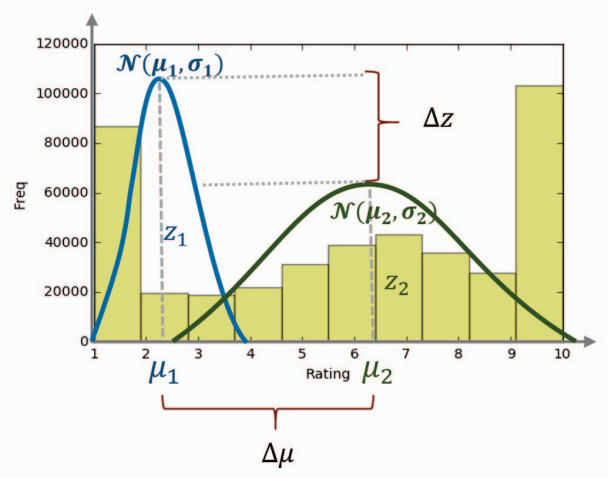


Figure 4: Proposed feature set 2, based on an item's rating histogram

shown in figure 1. As we can see, in a U-shaped distribution, the ratings are divided between the two extreme ratings. For this reason, we divide the distribution into *left*, *center* and *right* segments to identify these conceptual distinctions in the histograms of rating distributions. The segmentation can vary depending on the application and scale. For each segment, we compute the average rating and the count of ratings which is represented in Equation 1; we call it *Feature Set One*. The intuition behind taking into account the popularity of each segment is that polarization for an item with only a few ratings is probably not as significant as polarization for an item with many ratings. Our proposed features clearly capture the conceptual distinctions between polarized and unpolarized items based on the user ratings only, and hence are applicable to any domain with any rating scale. Figure 3 illustrates the extracted features using a movie's rating histogram.

$$F_1 = \{(w_{left}, m_{left})\} \cup \{(w_{center}, m_{center})\} \cup \{(w_{right}, m_{right})\}$$

$$\begin{cases} m_{left} = \sum_{j=1}^{4}(h_j) & w_{left} = \frac{\sum_{j=1}^{4}(h_j \times r_j)}{m_{left}} \\ m_{center} = \sum_{j=5}^{6}(h_j) & w_{center} = \frac{\sum_{j=5}^{6}(h_j \times r_j)}{m_{center}} \\ m_{right} = \sum_{j=7}^{10}(h_j) & w_{right} = \frac{\sum_{j=7}^{10}(h_j \times r_j)}{m_{right}} \end{cases}$$

$$\textit{where rating, } r_j \textit{ is on scale from 1 to 10 and}$$
$$h_j = \textit{number of items with ratings } r_j, \forall j \in [1, 10] \quad (1)$$

We extract another set of features, *Feature Set Two*, inspired by the literature [10], [4], [14], [3]. We start by estimating the best number of Gaussian distributions that can be fitted to the item's rating histogram. The next feature is based on an assumption that an extremely polarized item has two peaks in its rating histogram. As shown in figure 4, we first fit two Gaussian distributions, $\mathcal{N}(\mu_1, \sigma_1), \mathcal{N}(\mu_2, \sigma_2)$ and then compute the following features:

$$\begin{gathered} \Delta\mu = |\mu_1 - \mu_2| \\ \Delta z = |z_1, z_2|, \\ \text{where } z_1 \text{ is the peak of } \mathcal{N}(\mu_1, \sigma_1) \\ \text{and } z_2 \text{ is the peak of } \mathcal{N}(\mu_1, \sigma_1) \end{gathered} \quad (2)$$

In addition, we calculate the similarity between the two fitted Gaussian distributions using the Chi-square measure [19].

C. Classification Model

After extracting the feature set for each item, we train a binary classifier, which needs ground truth labels. To do so, we asked multiple experts to manually categorize each item into the polarized or non-polarized class. To aid in annotation, a web application was created. Each item's ratings histogram was shown to at least 3 annotators, where they were given the definition of polarization and were instructed to identify the polarized items. In case of a tie, we use a majority voting scheme to decide the final category of the item. We rely on human intuition to detect polarization. A human is able to tell if an item is polarized or not by simply looking at its rating histogram, often without needing to know the item's detailed information. Hence, for the sake of removing any bias, we hid all of the item's information, except for the rating histogram, from the annotators.

We trained different binary classifiers and finally chose the Random Forest Classifier [20] due to its high predictive accuracy and strength in handling imbalanced datasets. After the classification, the predicted probability of belonging to the 'polarized' class is considered as the polarization score for an item.

D. Item Review Sentiment Classification

In order to check the possible relationship between item polarization and reviews, we performed sentiment analysis on the reviews from each item, e.g. movie. The label for each review was based on the rating given by the user to the movie. We discretized the ratings into a binary scale, i.e. 1 if the rating is higher than 5, and 0 otherwise. From our observation, building the review sentiment classifier on the whole review data led to under-fitting, since our collected movies contain various genres, requiring distinct lexicons. On the other hand, building a sentiment classifier for each movie led to over-fitting, since it does not lead to good generalization for unseen movies. Therefore, we built a different sentiment classifier for each movie genre. When a movie has multiple genres, we decide its final sentiment label by voting based on the predicted labels from all

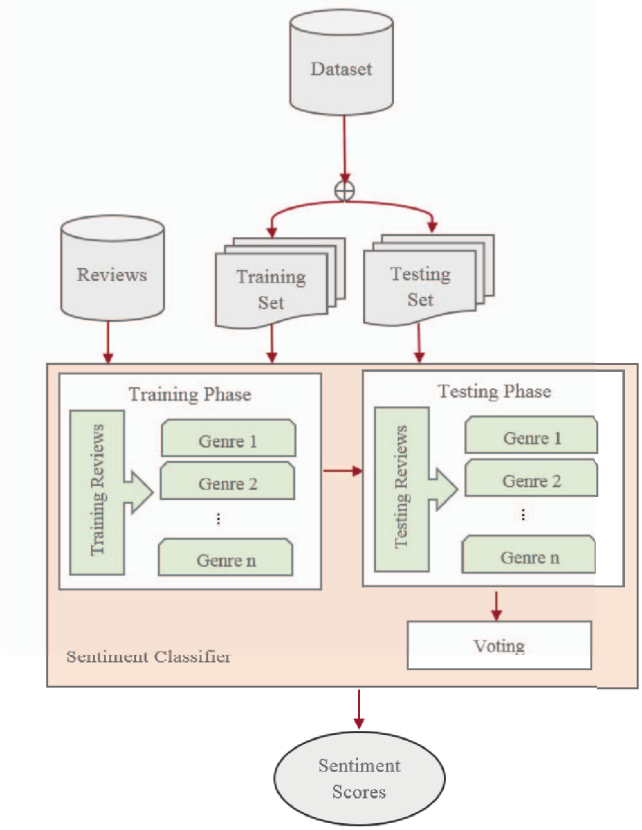


Figure 5: Proposed Polarization Detection Pipeline

genre classifiers. Figure 5 shows the steps of the developed sentiment classification approach.

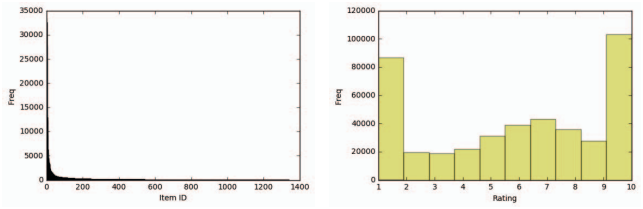
We used a multi-pooling review sentiment classifier, which combines both lexical background knowledge and labeled reviews, due to its demonstrated efficiency and adaptability to diverse contexts [16].

To understand the correlation between polarization and arousal and valence, we adopted the approach in [14], [21]. We used a lexicon of affective norms for valence (V) and arousal (A) of about 14,000 English words [22]. We averaged valence (V) and arousal (A) scores across all reviews for each movie, and assigned the score to that movie.

IV. Experiments

A. Datasets

We evaluate our pipeline on different domains ranging from books to various available movie datasets. The available datasets have a major limitation in that several have none or few polarized items. For example [23], [24] introduced polarity datasets that contain 2,000 and 50,000 movies, respectively. However, these datasets consider only reviews and ratings associated with these reviews. This is not enough for identifying polarization since not all users provide reviews; hence, by not considering those users, we lose informative ratings. Moreover, [24] hand-picked only 30 reviews, with an even number of positive and negative reviews, instead of considering all the reviews for each movie. This data set is therefore an artifact subset of the entire data.



(a) Item Frequency (b) Rating Frequency

Figure 6: Frequency of movies in the dataset and total ratings histogram

Due to the above limitations, we constructed a balanced collection of movie ratings from IMDb, by crawling polarized movies based on their histograms. After pruning and annotating the dataset, we ended up with 612 polarized movies, each with at least 50 ratings; then, we crawled all of the reviews for these movies. We also crawled an almost equal number of randomly selected non-polarized movies from different genres and years. Similar to other datasets, the movie popularity distribution follows a log-normal distribution. In the interest of providing a benchmark for future work in this area, we will release this dataset to the public [1].

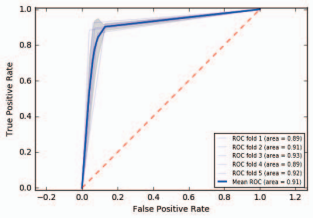
The proposed IMDb dataset contains 1,340 movies and 427,074 ratings. The data was collected the last week of March 2017. Each movie has a rating on a scale from 1 to 10. Figure 6a and 6b show the movies' frequencies and rating frequencies, respectively.

We first build a rating histogram of each item in the datasets. Figure 10 shows a snapshot of a movie histogram along with 2 Gaussian distributions of our constructed dataset. Then, we label each item as polarized or non-polarized using our annotation methodology as shown in figure 7 (more description in Section III-C). Next, we extract the proposed feature set for the items. We report 5-fold cross validation results to make our results comparable with others in the literature. We then train our proposed Polarization Detection Classifier using the Random Forests algorithm.

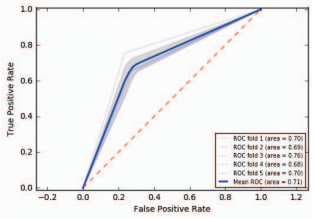
We compare our Polarization Detection Classifier (PDC) to several recent models [4], [10], [14], [3]. All of these methods use metrics that measure the polarization score for each item based on various criteria. In order to compare them with our classifier, we learned a threshold for each metric based on a validation set, and then, using this threshold, we decided whether an item was polarized or not. In addition, we tuned any additional parameters -e.g.

[1] Available here.

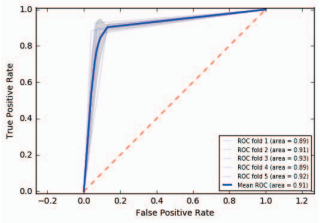
Figure 7: Annotation Platform



(a) ROC feature Set 1



(b) ROC feature Set 2



(c) ROC for both of the feature Sets

Figure 8: ROC comparison for the proposed feature sets

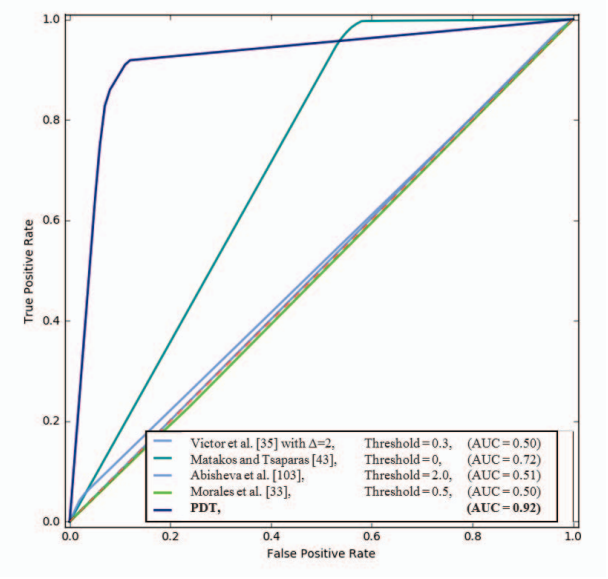


Figure 9: ROC comparison

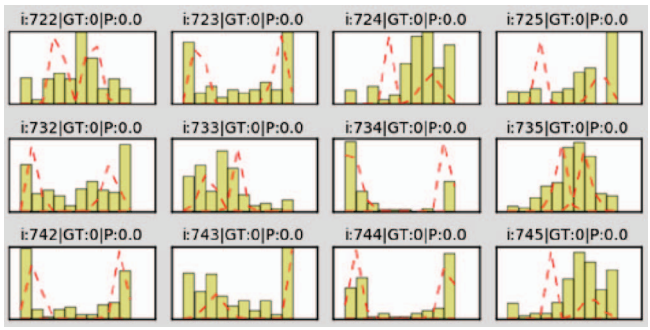


Figure 10: Histograms of 12 randomly selected movies from the crawled IMDb dataset

for [8] to fully capture polarization. Figure 9 shows the Area Under Curve (AUC) of the Receiver Operating Characteristic (ROC) curve[25] for each method. The proposed polarization classifier achieves the highest AUC = 0.92.

We compared these two sets of features to decide which one to choose for further experiments. Figure 8a shows the results of the classifier using feature set 1, 2, and both of them, respectively. As we can see, considering all features (10 features) does not improve the accuracy significantly. Considering only feature set 2 decreases the accuracy by about 20%. These results verify that feature set 1 is simple, fast, and accurate for detecting polarized items, based only on how they have been rated by users.

Figure 11 shows the Precision, Recall, and F-score [20] for the polarized and non-polarized class. PDC outperforms the other approaches significantly. To compare the total performance achieved by PDC, we also list the AUC for ROC curves in Table I which shows that PDC significantly outperforms the other methods in terms of AUC.

Table I shows the time taken by PDC compared to the other methods. PDC is significantly faster, even though we excluded the time needed for finding the best threshold for other polarization scores, and our approach has a training phase.

Our new approach is a valuable asset in the different domains where polarization matters, such as recommender systems. To illustrate this, we automatically quantify the polarization of several benchmark data sets for recommender systems, using our developed polarization detection pipeline to process these large data sets without the need to label them manually or to retrain the classifier.

1) Polarization versus Sentiment: For the review sentiment classifier, we built 13 review sentiment classifiers, one for each genre. We noticed that each genre has a different number of instances for training; therefore, the accuracy varied depending on the genre. The accuracy is generally around 0.75 ± 0.15 with Area Under Curve (AUC) 0.7 ± 0.1 for all genres. However, some genres resulted in

Table I: A comparison of AUC and time complexity for different methods

	PDC	Victor et al.[8]	Morales et al.[3]	Matakos and Tsaparas [10]	Abisheva et al.[14]
AUC	**0.92**	0.5	0.72	0.5	0.51
Time(sec.)	**1.87**	51.21	3.92	63.93	5.27

Table II: Detecting Polarized Items of Different Domains

	Book-Crossing	Amazon books	Epinions	IMDb-Polar	Movie Lens	Netflix	IMDb-PDC
Num. of items	1,340	5,912	103	3,581	19,344	10,524	1,340
Num. of ratings	49,317	25,230	1,561	25,000	138,493	19,847,947	427,074
Num. of polarized items	2	12	9	3,581	4	5	612

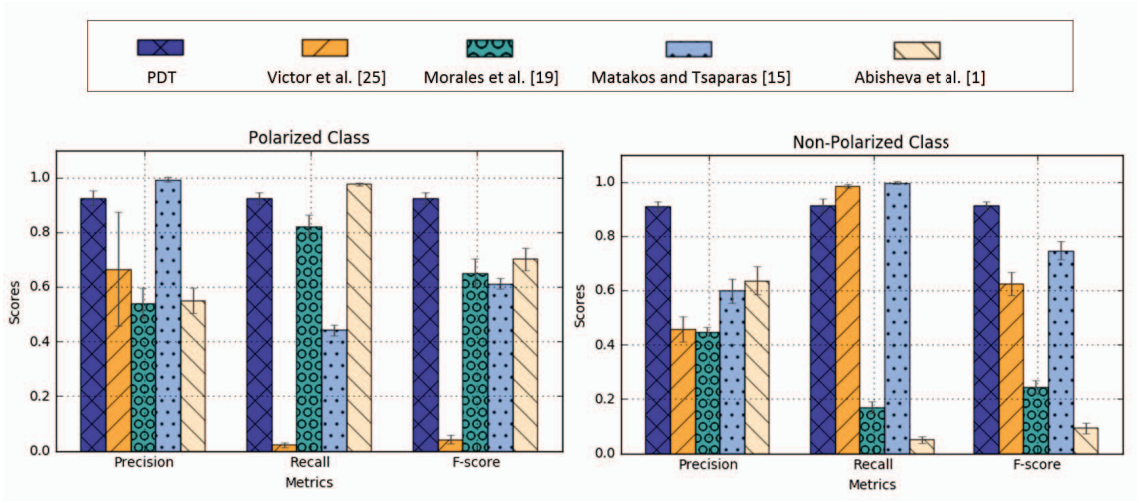


Figure 11: Performance of PDC compared to existing polarization metrics on polarized and non polarized data. PDC achieves a competitive tradeoff of precision and recall regardless of polarization

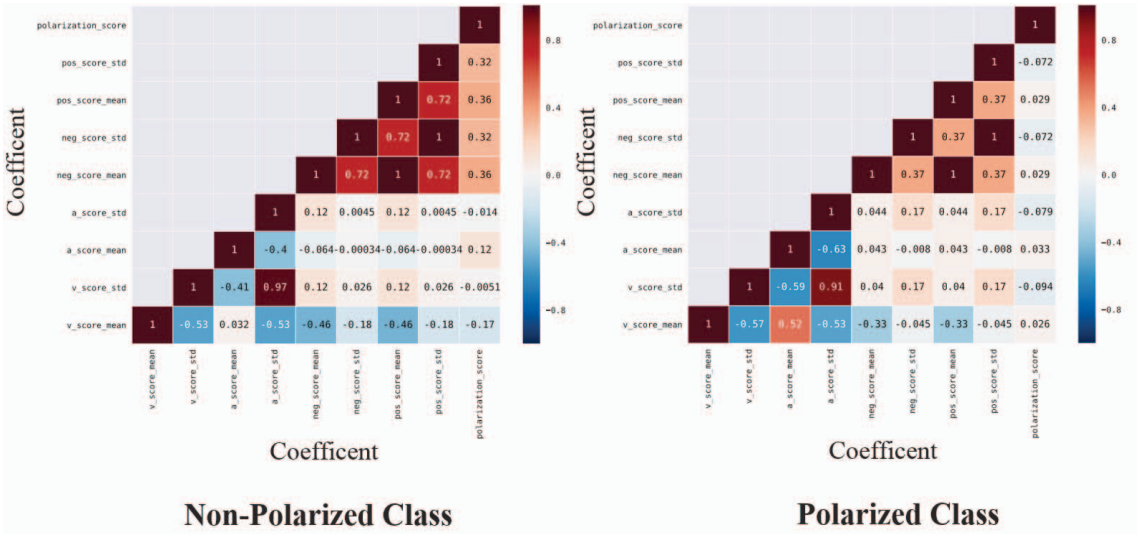


Figure 12: Heatmaps showing Pearson correlation coefficients between text reviews, polarization and emotion (v=Valence, a=Arousal) in polarized and non-polarized items.

low accuracy, e.g. 'Horror' and 'Thriller'. One possible reason is that lexical knowledge gives high negative weight to words that are actually positive in horror/thriller movies. For example, "This movie is very scary!" expresses positive sentiment in horror movies, but the lexical knowledge gives a high negative sentiment to this review. As mentioned before, Valence and Arousal is highly related to the emotions of people as well. For example, 'This is a good movie' has lower arousal score than the sentence 'This is a very good movie!', meanwhile they have similar valence scores because they both express a positive feeling.

To understand the relationship between the ratings-based polarization score and review sentiments, we compute the Pearson correlation [26] between the polarization score and sentiment score of each movie. To do so, we first needed to aggregate review sentiments for each movie. We considered both average and standard deviation of review sentiments. The reason for considering standard deviation is that if a movie's sentiment has a larger standard deviation, this means that the reviews contain different sentiment scores. In other words, it shows that there is some disagreement between reviews that may indicate a higher polarization. However, the sentiments' disagreement does not always mean that extreme opinion (polarization) is present. It may simply show diverse sentiments regarding that movie. We investigate these assumptions in Figure 12, which shows different correlations for polarized and non-polarized movies. The correlation between the polarization score and review sentiments is close to zero, showing that reviews by themselves are not enough to automatically infer polarization status; and that furthermore, due to being domain-dependent, reviews are unable to fully infer polarization in generic domains. However, in agreement with the theory in [14], arousal sentiment is associated with higher levels of the polarization score, due to the hidden strong emotions in the reviews . In addition, there is a stronger correlation between the sentiments themselves and the polarization score in non-polarized movies. This is due to the fact that the non-polarized movies have many reviews (at least 100 for each movie) compared to the polarized movies. Finally, we can conclude based on our experiments and analysis, that although text-based item reviews are able to slightly point to the polarization, they are not sufficient to fully capture polarization.

2) Other Benchmark Data Sets: In this section, we apply our PDC pipeline on various recommender system benchmark data sets, including the Netflix Prize [27], IMDb-Polar[24], Book-Crossing [28], MovieLens (20M) [29] and Amazon books (for several random users)[30]. Table II shows the percentage of polarized items in each dataset after removing items with less than 50 ratings. As we can see, there are only a few polarized items for most datasets. This confirms the fact that the available benchmark datasets do not always fully capture the realistic distribution of polarized user-item ratings. Hence, they may not be suitable to study polarization's interplay with recommender systems. Polarization often emerges during a discussion about a controversial topic when there are two groups of individuals with extreme opinions. A typical example is the e-commerce site Opinions.com, in which users can evaluate other users by including them in their 'Web of Trust' [8]. The available benchmark on Opinions.com has been used for trust-based recommender systems mainly for recommending items that the trust network disagrees with. However, our large scale quantification of this benchmark data set shows that most of the items are diverse ratings (flat ratings histogram) rather than polarized ratings (U-shaped ratings histogram), as shown in table II.

V. Conclusions

In real-life, polarization is an important phenomenon, with serious consequences, particularly on social media. Thus it is important to understand how machine learning algorithms, especially recommender systems, behave in a polarized environment; and to this end it is important to quantify polarization in existing and new data sets. We presented a domain independent data science pipeline to automatically detect polarization using the ratings. Our polarization detection framework was shown to detect different degrees of polarization and to outperform existing measures in capturing an intuitive notion of polarization. Our work is an essential step toward quantifying and detecting polarization in ongoing ratings and in benchmark data sets, and to this end, we used our developed polarization detection pipeline to compute the polarization prevalence of several benchmark data sets. It is our hope that this work will contribute to supporting future research in the emerging topic of designing and studying the behavior of recommender systems in polarized environments.

Acknowledgment

This work was partially supported by the National Science Foundation through grant NSF-154998.

References

[1] P. Dandekar, A. Goel, and D. T. Lee, "Biased assimilation, homophily, and the dynamics of polarization," *Proceedings of the National Academy of Sciences*, vol. 110, no. 15, pp. 5791–5796, 2013.

[2] Q. V. Liao and W.-T. Fu, "Can you hear me now?: mitigating the echo chamber effect by source position indicators," in *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*. ACM, 2014, pp. 184–196.

[3] A. Morales, J. Borondo, J. C. Losada, and R. M. Benito, "Measuring political polarization: Twitter shows the two sides of venezuela," *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 25, no. 3, p. 033114, 2015.

[4] P. Victor, C. Cornelis, M. De Cock, and A. Teredesai, "Trust-and distrust-based recommendations for controversial reviews," in *Web Science Conference (WebSci'09: Society On-Line)*, no. 161, 2009.

[5] C. R. Sunstein, "The law of group polarization," *Journal of political philosophy*, vol. 10, no. 2, pp. 175–195, 2002.

[6] K. Garimella, G. D. F. Morales, A. Gionis, and M. Mathioudakis, "Balancing opposing views to reduce controversy," *arXiv preprint arXiv:1611.00172*, 2016.

[7] D. J. Isenberg, "Group polarization: A critical review and meta-analysis." *Journal of personality and social psychology*, vol. 50, no. 6, p. 1141, 1986.

[8] P. Victor, C. Cornelis, M. De Cock, and A. Teredesai, "A comparative analysis of trust-enhanced recommenders for controversial items." in *ICWSM*, 2009.

[9] B. P. Knijnenburg, S. Sivakumar, and D. Wilkinson, "Recommender systems for self-actualization," in *Proceedings of the 10th ACM Conference on Recommender Systems*. ACM, 2016, pp. 11–14.

[10] A. Matakos and P. Tsaparas, "Temporal mechanisms of polarization in online reviews," in *Advances in Social Networks Analysis and Mining (ASONAM), 2016 IEEE/ACM International Conference on*. IEEE, 2016, pp. 529–532.

[11] S. A. Munson and P. Resnick, "Presenting diverse political opinions: how and how much," in *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM, 2010, pp. 1457–1466.

[12] Y. Mejova, A. X. Zhang, N. Diakopoulos, and C. Castillo, "Controversy and sentiment in online news," *arXiv preprint arXiv:1409.8152*, 2014.

[13] K. Garimella, G. De Francisci Morales, A. Gionis, and M. Mathioudakis, "Quantifying controversy in social media," in *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*. ACM, 2016, pp. 33–42.

[14] A. Abisheva, D. Garcia, and F. Schweitzer, "When the filter bubble bursts: collective evaluation dynamics in online communities," in *Proceedings of the 8th ACM Conference on Web Science*. ACM, 2016, pp. 307–308.

[15] O. Nasraoui and B. L. B. Masand, "Advances in web mining and web usage analysis," in *8th international workshop on the Web, WebKDD*. Springer, 2006.

[16] P. Melville, W. Gryc, and R. D. Lawrence, "Sentiment analysis of blogs by combining lexical knowledge with text classification," in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2009, pp. 1275–1284.

[17] P. Melville, V. Chenthamarakshan, R. D. Lawrence, J. Powell, M. Mugisha, S. Sapra, R. Anandan, and S. Assefa, "Amplifying the voice of youth in africa via text analytics," in *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2013, pp. 1204–1212.

[18] A. Said, B. J. Jain, and S. Albayrak, "Analyzing weighting schemes in collaborative filtering: cold start, post cold start and power users," in *Proceedings of the 27th Annual ACM Symposium on Applied Computing*. ACM, 2012, pp. 2035–2040.

[19] Y. Ma, X. Gu, and Y. Wang, "Histogram similarity measure using variable bin size distance," *Computer Vision and Image Understanding*, vol. 114, no. 8, pp. 981–989, 2010.

[20] T. K. Ho, "Random decision forests," in *Document Analysis and Recognition, 1995., Proceedings of the Third International Conference on*, vol. 1. IEEE, 1995, pp. 278–282.

[21] J. A. Russell and L. F. Barrett, "Core affect, prototypical emotional episodes, and other things called emotion: dissecting the elephant." *Journal of personality and social psychology*, vol. 76, no. 5, p. 805, 1999.

[22] A. B. Warriner, V. Kuperman, and M. Brysbaert, "Norms of valence, arousal, and dominance for 13,915 english lemmas," *Behavior research methods*, vol. 45, no. 4, pp. 1191–1207, 2013.

[23] B. Pang and L. Lee, "A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts," in *Proceedings of the 42nd annual meeting on Association for Computational Linguistics*. Association for Computational Linguistics, 2004, p. 271.

[24] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts, "Learning word vectors for sentiment analysis," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*. Association for Computational Linguistics, 2011, pp. 142–150.

[25] D. M. Powers, "Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation," 2011.

[26] S. T. Garren, "Maximum likelihood estimation of the correlation coefficient in a bivariate normal model with missing data," *Statistics & probability letters*, vol. 38, no. 3, pp. 281–288, 1998.

[27] Y. Koren, R. Bell, and C. Volinsky, "Matrix factorization techniques for recommender systems," *Computer*, vol. 42, no. 8, 2009.

[28] C.-N. Ziegler, S. M. McNee, J. A. Konstan, and G. Lausen, "Improving recommendation lists through topic diversification," in *Proceedings of the 14th international conference on World Wide Web*. ACM, 2005, pp. 22–32.

[29] F. M. Harper and J. A. Konstan, "The movielens datasets: History and context," *ACM Transactions on Interactive Intelligent Systems (TiiS)*, vol. 5, no. 4, p. 19, 2016.

[30] J. Leskovec and A. Krevl, "SNAP Datasets: Stanford large network dataset collection," http://snap.stanford.edu/data, Jun. 2014.