The Thirty-Second AAAI Conference
on Artificial Intelligence (AAAI-18)

ROAR: Robust Label Ranking for Social Emotion Mining

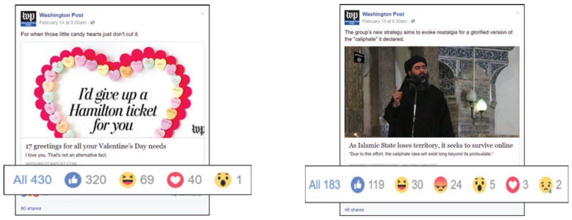
Jason (Jiasheng) Zhang, Dongwon Lee
College of Information Sciences and Technology
The Pennsylvania State University, USA
{jpz5181,dlee}@ist.psu.edu

Abstract

Understanding and predicting latent emotions of users toward online contents, known as *social emotion mining*, has became increasingly important to both social platforms and businesses alike. Despite recent developments, however, very little attention has been made to the issues of nuance, subjectivity, and bias of social emotions. In this paper, we fill this gap by formulating social emotion mining as a *robust label ranking* problem, and propose: (1) a robust measure, named as G-mean-rank (GMR), which sets a formal criterion consistent with practical intuition; and (2) a simple yet effective label ranking model, named as ROAR, that is more robust toward unbalanced datasets (which are common). Through comprehensive empirical validation using 4 real datasets and 16 benchmark semi-synthetic label ranking datasets, and a case study, we demonstrate the superiorities of our proposals over 2 popular label ranking measures and 6 competing label ranking algorithms. The datasets and implementations used in the empirical validation are available for access[1].

Introduction

It has become increasingly important for businesses to better understand their users and leverage the learned knowledge to their advantage. One popular method for such a goal, so-called *social emotion mining*, is to mine users' digital footprints to unearth users' "emotions" toward particular products or services on social platforms. Users' latent emotions can be indirectly peeked via various channels–e.g., low star rating given to an Amazon review, angry comments left to a YouTube video, upvote to a news story in Reddit, or retwitting a friend's post. In particular, we note one recently-introduced function to social platforms where users may select one emoticon, out of many choices, to more precisely express their emotions. Facebook introduced this function in 2016, while Chinese news portal, Sina, supports a similar function. Two Facebook posts are shown in Fig. 1 as examples. Then, a natural question is whether one can predict the emotions expressed as emoticons in such a setting.


Washington Post
February 14 at 5:00am
For when those little candy hearts just don't cut it.
I'd give up a Hamilton ticket for you
17 greetings for all your Valentine's Day needs
I love you. That's not an alternative fact.
WASHINGTONPOST.COM
90 shares

Washington Post
February 14 at 6:30am
The group's new strategy aims to evoke nostalgia for a glorified version of the "caliphate" it declared.
As Islamic State loses territory, it seeks to survive online
"Due to this effort, the caliphate idea will exist long beyond its proto-state."
48 shares

(a) (b)

Figure 1: Two Washington Post Facebook posts with different emoticon reactions @ www.facebook.com/washingtonpost/

Most existing research on social emotion mining focuses on extracting informative features to infer emotions from data (Jia, Chen, and Yu 2009; Lin and Chen 2008; Lin, Yang, and Chen 2008; Tang and Chen 2011; Bai et al. 2012; Lei et al. 2014; Zhu et al. 2014; Zhang et al. 2014). On the one hand, as taken by most present works, predicting one dominant emoticon as a *classification* problem may fail to catch the nuance of human emotions. For example, two posts in Fig. 1 share the same top-2 dominating emoticons, *like* and *haha*, rendering such a classification approach be less useful. On the other hand, the subjectivity makes predicting the exact composition of emotions as a *regression* problem to be less useful too. For instance, in Fig. 1(a), reporting the emotion of users as $69/430$ $haha$, $40/430$ $love$ and $1/430$ wow conveys little extra information than simply saying that users feel more $haha$ than $love$ and few wow. Therefore, to reflect the nuance and subjectivity of human emotions, we propose to formulate social emotion mining as a **label ranking** problem, where the emotions of users toward a given post are represented by a ranking among a set of emotion labels. In this way, for nuance, the number of all possible emotions becomes $d!$, as opposed to d in a classification framework, where d is the size of emotional label set. For subjectivity, only relative rather than absolute strength of different emotional labels is mined.

The label ranking problem asks *if one can learn a model to annotate an instance with a ranking over a finite set of predefined labels*. Label ranking can be seen as a specific type of the preference learning problem (Hüllermeier et al. 2008) in AI. However, in the case of social emotion mining, some labels may be preferred, causing a skewed distribution of chosen labels. For example, ordinary Facebook users (i.e., posters) tend to post more happy stories and their friends (i.e., readers) are more willing to give positive feedback

[1] http://pike.psu.edu/download/aaai18/

such as $like$ or $haha$. Therefore, the ranking distribution is highly biased toward those rankings with positive labels ranked higher than negative ones. However, posts with dominating negative labels are usually more informative. None of existing label ranking methods has considered this "imbalance" issue.

Although there have been methods to address the imbalance issue in classification, as will be illustrated in next section, imbalance in label ranking is still anything but trivial due to its large and nontrivial target space (i.e., $d!$ correlated possible rankings). To the best of our knowledge, we are the first to point out and give a formal definition of imbalance in label ranking and the first to formulate social emotion mining as a "robust" label ranking problem. Toward this challenge, we make two contributions: (1) we first show the inadequacy of popular performance measures in label ranking to handle the imbalanced data, propose a novel robust performance measure, named as G-mean-rank (GMR), and experimentally demonstrate the superiority of GMR over existing measures; and (2) we propose a novel robust label ranking model, ROAR, for imbalanced data without any re-sampling or costs as hyper-parameters, and show that ROAR outperforms 6 competing models, in real-life Facebook emoticon prediction task and achieves competitive performance in semi-synthetic benchmark label ranking data sets.

Related Works

There are three classes of label ranking methods. First, label-wise methods (Har-Peled, Roth, and Zimak 2002; Dekel, Manning, and Singer 2003; Cheng, Henzgen, and Hüllermeier 2013) treat label ranking as the regression problem for the relevant score of each label or position of ranking. Second, pair-wise methods (Hüllermeier et al. 2008; Cheng et al. 2010; 2012; Destercke 2013; Grbovic, Djuric, and Vucetic 2013) decompose label ranking problem to binary classification problem for each pair of labels and then aggregating pairwise results into rankings. Third, list-wise methods employ different ranking distance measures to directly predict rankings without decomposing, such as Mallows model (Mallows 1957) based methods (Cheng, Hühn, and Hüllermeier 2009; Zhou et al. 2014), Plackett-Luce model based method (Cheng, Hüllermeier, and Dembczynski 2010) and weighted distance model (Shieh 1998) based methods (Lee and Philip 2012; 2010). Our proposed solution, ROAR, belongs to the third class.

Previous label ranking works (Har-Peled, Roth, and Zimak 2002; Dekel, Manning, and Singer 2003; Hüllermeier et al. 2008; Cheng, Hühn, and Hüllermeier 2009; Cheng et al. 2010; 2012; Cheng, Henzgen, and Hüllermeier 2013; Busa-Fekete, Hüllermeier, and Szörényi 2014) typically evaluate preformance using ranking distance measure such as *Kendall tau correlation* (Kendall 1948) or Spearman's rank correlation (Spearman 1904). On the other hand, social emotion mining works (Lin and Chen 2008; Lin, Yang, and Chen 2008; Lei et al. 2014; Zhang et al. 2014) typically measure performance using metrics from information retrieval community, such as $ACC@k$ and $nDCG@k$ (Järvelin and Kekäläinen 2002), emphasizing the intuition that higher ranked positions are more informative. Similarly, some rank modeling works in statistics (Lee and Philip 2012; 2010; Shieh 1998) weight the distance between ranks to model such bias. Note that the bias there is rewarding heterogeneity of different ranking positions, rather than bias in ranking distribution considered in this work. Hence, in imbalanced data, those performance measures are not good enough.

Imbalanced data problem has been previously investigated under the classification framework (He and Garcia 2009). Popular methods include random sampling (Batista, Prati, and Monard 2004; Japkowicz and Stephen 2002) and cost-sensitive methods (Chawla, Japkowicz, and Kotcz 2004; Weiss 2004; Maloof 2003). Both methods try to first obtain balanced data from original imbalanced data so that the problem is reduced to the balanced classification. However, these methods involve tricky hyper-parameter tuning, especially in multi-class classification (Sun, Kamel, and Wang 2006), which will become even more severe in label ranking framework. Besides, there is nontrivial correlation among rankings rather than independent labels in classification. Hence it is hard to determine a sampling parameter or a cost for each ranking. In contrast, the robust label ranking method proposed in this work is free of hyper-parameters related to data imbalance.

Preliminaries

Social Emotion Mining

Here we formulate social emotion mining as the label ranking problem. Given a post x in social media, with $x \in \mathcal{X}$ as feature vector, and a set of emotional labels $\mathcal{Y} = \{y_1, y_2, ..., y_d\}$, called emoticons, the goal is to associate the post with an aggregated emotion of crowd $\phi(\mathbf{x})$ it triggers, represented by the emoticons. As argued, we choose $\phi(\mathbf{x})$ to be a ranking over the emoticon set, $\phi(\mathbf{x}) = (\phi_1(\mathbf{x}), \phi_2(\mathbf{x}), ..., \phi_d(\mathbf{x}))$, where $\phi_i(\mathbf{x}) \in \mathcal{Y}$ and $\phi_i(\mathbf{x}) \neq \phi_j(\mathbf{x})$, $\forall i \neq j$. $\phi_i = y_l$ indicates that label y_l ranks on position i. For consistent annotation, a ranking position vector is defined as $\pi(\mathbf{x}) = (\pi_{y_1}(\mathbf{x}), \pi_{y_2}(\mathbf{x}), ..., \pi_{y_d}(\mathbf{x}))$, where $y_i \in \mathcal{Y}$ and $\pi_{y_i} \in \{1, 2, ..., d\}$, which means that label y_i ranks on position $\pi_{y_i}(\mathbf{x})$. With a ranking, the represented emotion consists of more of emoticons ranking higher and less of those lower. Therefore, the social emotion mining is formulated as a label ranking problem.

Problem 1 (Label Ranking) *Find a mapping $f\colon \mathcal{X} \to \Omega_d$, where Ω_d is the set of all possible rankings over a label set of size d, such that given an instance with feature vector $\mathbf{x}$, predict ranking $\hat{\phi}(\mathbf{x}) = f(\mathbf{x})$.*

Imbalance in Label Ranking

In social emotion mining context, imbalance in data refers to the characteristics of data where documents with some emotional reactions are rarer than those with others. In the context of label ranking, it means that instances with some rankings are rarer than those with others. As for a formal definition of this intuition, a naive choice is treating different rankings as different classes and the problem reduces to a classification problem. However, classification framework ignores the fact that different rankings are not independent or equal-interval with each other. Instead, therefore, here imbalance is defined based on pairwise comparisons.

Given any pair of labels $\{y_i, y_j\}$, $y_i, y_j \in \mathcal{Y}$, and an instance ν, a pairwise comparison function is defined as:

$$I_\nu(y_i, y_j) = \begin{cases} 1, & \text{if } \pi_{y_i} < \pi_{y_j} \text{ for instance } \nu \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

Then, for each label pair y_i, y_j, imbalance in data distribution $\mathcal{D} = \{(x, \phi(x))\} \subset \mathcal{X} \times \Omega_d$ can be seen as the difference between $\sum_{\nu \in \mathcal{D}} I_\nu(y_i, y_j)$ and $\sum_{\nu \in \mathcal{D}} I_\nu(y_j, y_i)$. Since a ranking consists of pairwise comparisons of all pairs, a single-value imbalance measure for label ranking, $IMBA$-$rank$ (or $IMBA$ without ambiguity), of $\mathcal{D}$ is defined as:

$$IMBA(\mathcal{D}) = \frac{1}{2} \sum_{i,j=1, i \neq j}^{d} \left| log\left(\frac{\sum_{\nu \in \mathcal{D}} I_\nu(y_i, y_j) + 1}{\sum_{\nu \in \mathcal{D}} I_\nu(y_j, y_i) + 1}\right) \right| . \tag{2}$$

When data is perfectly balanced, $IMBA$-$rank$ should be 0. The more imbalanced the data is, the larger $IMBA$-$rank$ gets.

Robust Performance Measure

We first show that commonly used performance measures in label ranking are no longer adequate in imbalance case, and then introduce a robust one.

Previous Measures for Label Ranking

One of the most popular performance measures in label ranking community is *Kendall's tau correlation* (Kendall 1948). The correlation tau for two rankings $\{\pi, \hat{\pi}\}$ is formally defined as:

$$tau = \frac{C(\pi, \hat{\pi}) - D(\pi, \hat{\pi})}{C(\pi, \hat{\pi}) + D(\pi, \hat{\pi})}, \tag{3}$$

where $D(\pi, \hat{\pi}) = |\{(i,j)|i < j, \pi_{y_i} > \pi_{y_j} \wedge \hat{\pi}_{y_i} < \hat{\pi}_{y_j}\}|$ and $C(\pi, \hat{\pi}) = |\{(i,j)|i < j, \pi_{y_i} > \pi_{y_j} \wedge \hat{\pi}_{y_i} > \hat{\pi}_{y_j}\}|$ denote the number of discordant and consistent-ordered pairs of labels between two rankings, respectively. To emphasize the importance of higher positions in ranking, previous works on social emotion mining usually use $ACC@k$ as performance measure. The $ACC@k$ of an instance is defined as:

$$ACC@k(\phi, \hat{\phi}) = I(\phi_i = \hat{\phi}_i | \forall i \in \{1, 2, ..., k\}), \tag{4}$$

where $I()$ is the indicator function.

Concerning one pair of labels and two candidate ranking positions, the imbalanced label ranking problem reduces to imbalanced classification problem. Both tau and $ACC@k$ consider only true fractions without distinguishing true positives and true negatives, which has been well known to be inadequate in imbalanced classification (He and Garcia 2009). This is similarly true for other label ranking performance measures, such as Spearman's rank correlation and $nDCG@k$.

For a better illustration, consider a toy data set as an example. Here the label set $\mathcal{Y} = \{y_i | i \in \{1, 2, 3, 4\}\}$ with $d = 4$. The dataset contains 100 instances where 90 of them are associated with rank $\phi^{\mathbf{9}} = (y_1, y_2, y_3, y_4)$ while the rest are associated with $\phi^{\mathbf{1}} = (y_1, y_4, y_2, y_3)$. Then, a trivial label ranking model predicts all instances to be with rank $\hat{\phi} = \phi^{\mathbf{9}}$ for this toy data set. Then, the performance of the trivial model is $tau \approx 93\%$ and $ACC@k = 90\%$ if $k = 2$, which is relatively high compared with a perfect model with $tau = 100\%$ and $ACC@k = 100\%$. Hence, both measures help little in recognizing such trivial solution in imbalanced data or giving sufficient attention to minority rankings.

Robust Measure for Label Ranking: GMR

As shown above, it is critical to distinguish negative and positive classes for performance measure in imbalanced classification problem. Similarly in label ranking problem, we first decompose it into pairwise comparison classification problem. For each ordered pair of labels (y_i, y_j), the class of instance ν is defined as Positive if $I_\nu(y_i, y_j) = 1$ or Negative if $I_\nu(y_j, y_i) = 1$. Since the Negative class of ordered pair (y_i, y_j) is the same as Positive class of (y_j, y_i), only positive class for each ordered pair is considered. Hence only recall can be defined. Similar to classification, the recall for ordered pair (y_i, y_j) in data set $\mathcal{D}$ is defined as

$$recall_{\mathcal{D}}(y_i, y_j | f) = \frac{\sum_{\nu \in \mathcal{D}} I_\nu(y_i, y_j) I_{\hat{\nu}}(y_i, y_j) + 1}{\sum_{\nu \in \mathcal{D}} I_\nu(y_i, y_j) + 2}, \tag{5}$$

where $I_{\hat{\nu}}(y_i, y_j)$ is the pairwise comparison function for predicted ranking for instance ν, and the extra $+1$ term in numerator and $+2$ term in denominator are smooth terms. To combine recalls of different pairs, inspired by G-Mean (Sun, Kamel, and Wang 2006) for imbalanced multi-class classification, geometric mean is used and *G-mean-rank* (GMR) for data set $\mathcal{D}$ is defined as:

$$GMR_{\mathcal{D}}(f) = \sqrt[P]{\prod_{i \neq j}^{d} recall_{\mathcal{D}}(y_i, y_j | f)}\,, \tag{6}$$

where P is the number of ordered pairs involved. For each pair of labels, GMR is the same as G-$mean$ for two-class classification. Hence according to the definition of imbalance in label ranking, GMR is insensitive to imbalanced label ranking data.

Using the aforementioned toy dataset again, the performance of the trivial algorithm in terms of GMR is $GMR \approx 53\%$, which is not high compared to 97% for a perfect one. Therefore, GMR rightfully gives sufficient penalty to a trivial solution which is not supposed to perform well, a behavior that we intended.

Is GMR Superior to Previous Measures?

To show the robustness of GMR compared to two popular measures–i.e., tau and $ACC@k$, we apply the idea of toy example above to real datasets. We use four Facebook post datasets with emoticon set size of $d = 6$, whose detail can be found in Empirical Validation section later.

We extend the idea of the trivial model in the toy example by designing a naive model, denoted as **NAIVE**, that assigns the most common ranking in a training set to all instances in a test set regardless of their feature values. As the datasets

Table 1: Comparison of GMR with $ACC@3$ and tau using real datasets

		Datasets			
Measures	Methods	ROU	NYT	WSJ	WaPo
$IMBA$		3.03	1.94	2.96	2.14
ACC@3	RPC	0.850	0.185	0.191	0.182
	KNN-PL	0.770	0.125	0.169	0.144
	NAIVE	0.837	0.0532	0.171	0.0976
	Gain (%)	-3	191	5	67
tau	RPC	0.933	0.497	0.595	0.541
	KNN-PL	0.929	0.495	0.584	0.544
	NAIVE	0.932	0.399	0.567	0.503
	Gain (%)	-0.1	24	4	8
GMR	RPC	0.241	0.429	0.345	0.358
	KNN-PL	0.408	0.458	0.513	0.461
	NAIVE	0.152	0.0796	0.0881	0.0770
	Gain (%)	**113**	**457**	**387**	**432**

we use are rankings converted from number of votes for different emoticons, the most common ranking (i.e., the output) in NAIVE is set as the ranking of emoticons according to the number of accumulated votes in the training set. Therefore, NAIVE is a "dumb" solution and is not supposed to work well.

Next, we choose 2 state-of-the-art models, **RPC** and **KNN-PL**, as examples of good models, whose detail will be explained in Robust Label Ranking Model section. The idea is that a robust performance measure should be able to clearly distinguish good models (e.g., RPC and KNN-PL) from bad ones (e.g., NAIVE) even when a dataset is severely unbalanced.

The result is shown in Table 1. For $ACC@k$, k is set as 3 to mimic the behavior of Facebook, where only top-3 emoticons of posts are shown by default. The row, Gain (%), in Table 1 shows the *average* improvement of two good models over NAIVE in terms of three different measures. In all four datasets, the improvement in terms of GMR is always far larger than $ACC@3$ and tau, which illustrates the robustness of GMR. Table 1 also shows the $IMBA$-$rank$ of each dataset as $IMBA$. Note that the improvement in terms of $ACC@3$ and tau decreases as $IMBA$-$rank$ increases. For the most imbalanced dataset, ROU, the improvement in terms of both $ACC@3$ and tau is even negative, which indicates that GMR is capable of capturing the fact that two state-of-the-art models far outperform a naive poorly-designed model.

Now we are ready to formally define robust label ranking problem.

Problem 2 (Robust Label Ranking) *Find a mapping* f: $\mathcal{X} \rightarrow \Omega_d$*, for data distribution* $\mathcal{D}$*, with large* $IMBA$-$rank(\mathcal{D})$*, such that* $GMR_{\mathcal{D}}(f) \geq GMR_{\mathcal{D}}(f'), \forall f' : \mathcal{X} \rightarrow \Omega_d$.

Robust Label Ranking Model

Competing models

In this work, to our best knowledge, we consider all existing state-of-the-art label ranking models as follows.

- Ranking by Pairwise Comparison (RPC) (Hüllermeier et al. 2008): It predicts pairwise order for each pair of labels using logistic regression and then combines them into ranking output with Borda count (de Borda 1781).
- Label-Wise Decomposition (LWD) (Cheng, Henzgen, and Hüllermeier 2013): It predicts position probability distribution of each label and then combines them to minimize expected Spearman's footrule (Diaconis and Graham 1977).
- Soft Multi-Prototype (SMP) (Grbovic, Djuric, and Vucetic 2013): It fits label ranking data with multiple prototypes both in feature and ranking space, and combines prototypes into ranking prediction given feature values.
- K-Nearest-Neighbor with Plackett-Luce model (KNN-PL) (Cheng, Hüllermeier, and Dembczynski 2010): It predicts ranking by aggregating rankings of instances whose feature values are nearest to given feature value. The aggregation is based on Plackett-Luce model.
- K-Nearest-Neighbor with Mallows model (KNN-M) (Cheng, Hühn, and Hüllermeier 2009): It is the same with KNN-PL except the aggregation is based on Mallows model (Mallows 1957).
- Log-Linear model (LogLinear) (Dekel, Manning, and Singer 2003): It learns utility functions for each label via pairwise comparison and sorts labels by utility function values into ranking. Here the utility function is adopted from (Hüllermeier et al. 2008), in which case LogLinear is equivalent to the Constraint Classification algorithm (Har-Peled, Roth, and Zimak 2002).
- Label Ranking Tree (LRT) (Cheng, Hühn, and Hüllermeier 2009): It is a decision tree method whose induction is based on Mallows model (Mallows 1957).

Robust Label Ranking Model: ROAR

Now, we propose a robust label ranking model, named as **ROAR** (RObust lAbel Ranking), which is a simple, efficient, and effective tree based model. The performance measure GMR is difficult to be directly optimized, as it is not an average over some performance measure for each instance. Hence an alternative learning objective function, an induction criterion in decision tree, is proposed. This supports the model searching for finest structure in feature and target space without overfitting, which makes it robust against imbalanced data.

Learning. To learn a decision tree, a general algorithm begins with all instances in the root node. Then, it partitions the training data recursively, by one-dimension splits according to the comparison between thresholds and a feature value. The decision tree in this work is a binary tree.

The threshold and the feature for each split are selected by exhaustive search so that the sizes of the neighborhoods in the target space, estimated by training data in the resultant child nodes, become the smallest. The size of a neighborhood is estimated by the impurity of the set of rankings in a node. One intuition about the impurity of a set of rankings is the impurity of labels on each ranking position. Because labels are independent of each other, for a given position, we choose the popular Gini index, and the Gini index of a tree node T for position i is defined as:

$$Gini_i(T) = \sum_{y \in \mathcal{Y}} \frac{(n_i(T) - n_{iy}(T))n_{iy}(T)}{n_i(T)^2}, \qquad (7)$$

where $n_{iy}(T) = \sum_{\nu \in T} I(\phi_i(\mathbf{x}_\nu) = y)$ is the number of instances with label y ranking on position i and $n_i(T) = \sum_{y \in \mathcal{Y}} n_{iy}(T)$ denotes the number of instances with any label ranking on position i. Then the impurity for rankings of the node T can be measured by weighted sum of Gini index for each position, that is,

$$Gini(T) = \sum_{i=1}^{d} \frac{n_i(T)Gini_i(T)}{|T|}, \qquad (8)$$

which is called point-wise Gini index. For parent node T and potential child nodes T^- and T^+, the split criterion is defined as

$$criterion = |T|^{-1}(|T^+|Gini(T^+) + |T^-|Gini(T^-)). \qquad (9)$$

The stopping criterion is straightforward. The partitioning stops when no further partitioning is possible, that is, when there is no partitioning whose $criterion$ is smaller than $Gini(T)$ for current node T.

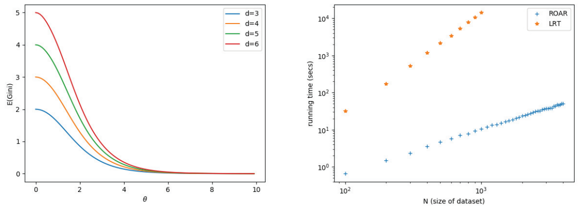
Prediction. Here for ROAR, we use a position-wise ranking aggregation method. From highest to lowest ranking position, given a position, it assigns the label that has not been assigned and appears most frequently at that position, to each position. When there is no such label for a position, it resorts to label distributions of other positions, from highest to lowest and does the same.

Consistency with Ranking Theory. This point-wise Gini index is consistent with our intuition about the purity of a set of rankings. To show that, we have to assume a measurement of the size of a neighborhood $\Omega(T)$ around a point in Ω_d, noting that the center ranking π_0 is unknown. Mallows model (Mallows 1957) is a popular assumption of probability model of rankings, using the annotation from (Cheng, Hühn, and Hüllermeier 2009), defined as

$$P(\pi|\theta, \pi^0) = \frac{exp(-\theta D(\pi, \pi^0))}{\psi(\theta)}, \qquad (10)$$

where $\psi(\theta) = \sum_{\pi \in \Omega_d} exp(-\theta D(\pi, \pi^0))$ is a normalization constant, θ the spread parameter, π^0 the center ranking and $D(\cdot, \cdot)$ the distance between two rankings, which is the number of discordant pairs between two rankings. Assuming that the rankings in T are independently generated according to Mallows model, the spreading parameter θ measures the size of $\Omega(T)$. Under the independence assumption, the expectation of point-wise Gini index for node T is

$$\begin{aligned} E(Gini(T)) &= E(\sum_{i}^{d} \frac{n_i(T)Gini_i(T)}{|T|}) \\ &= \sum_{i}^{d} E(Gini_i(T)), \end{aligned} \qquad (11)$$



(a) E(Gini) versus θ

(b) Running time

Figure 2: Left figure illustrates the consistency between Gini index and impurity measure in Mallows model; right figure shows the efficiency of ROAR compared with LRT.

$$E(Gini_i) = \frac{(n_i - 1)}{n_i}(1 - \sum_{y \in \mathcal{Y}} P(\phi_i = y)^2). \qquad (12)$$

According to Mallows model, without loss of generality, we assume the center ranking $\phi_i^0 = y_i$, $\forall i \in \{1, 2, ..., d\}$. Then ranking ϕ with $\phi_i = y_j$, is with probability $P(\phi) = \frac{\mathcal{O}(exp(-|i-j|\theta))}{\psi(\theta)}$. Therefore, for large enough θ, $P(\phi_i = y_j) = \mathcal{O}(exp(-|i-j|\theta))$. Hence, when θ is larger, which is when the spread of Mallows model is smaller, then the probabilities $P(\phi_i = y_j)$ over $y_j \in \mathcal{Y}$ are more skewed toward smaller $|i - j|$. Therefore according to eq. 12, $E(Gini_i)$ is smaller, so is $E(Gini(T))$, as illustrated in Fig. 2a with different sizes of label set in the limitation of $n_i \to \infty$. Therefore, $Gini$ is a good estimator of the impurity of rankings in a node.

Time Complexity. In ROAR, the amortized running time for each potential partition is $\theta(d^2)$, constant in terms of $|T|$. Hence the running time for each induction of a tree node is $\Theta(m|T|(log|T| + d^2))$. In contrast, LRT takes $\Omega(|T|)$ steps for each potential partition, so that the running time for each induction of a tree node is $\Omega(m|T|^2 d^2)$. The running time of two methods applied to data of different sizes are shown in Fig. 2b. As LRT becomes prohibitively slow as data gets large, it is not considered in following empirical validation.

Empirical Validation

We attempt to validate if: (1) our proposed G-mean-rank is superior to two popular label ranking measures in imbalance datasets, which has been done in Robust Performance Measure section; and (2) our proposed ROAR outperforms 6 competing label ranking models. The datasets and implementations used in the empirical validation are available for access[2].

Datasets and Set-Up

In this work, we use emoticon clicks data of Facebook posts. For each post, there are six emoticon labels, {like, love, haha, wow, sad, angry}. Each user (i.e., reader) can select one of the six labels for each post. For evaluating NAIVE in Robust Performance Measure section, we use the number of votes for labels per post as the input. To obtain rank-

[2] http://pike.psu.edu/download/aaai18/

Table 2: Summary of four datasets

	ROU	NYT	WSJ	WaPo
# posts	$17,394$	$4,684$	$7,464$	$6,117$
$IMBA$	3.03	1.94	2.96	2.14
$\# like$	$834K$	$7.99M$	$2.44M$	$3.81M$
$\# love$	$14K$	$578K$	$105K$	$222K$
$\# haha$	$3,281$	$434K$	$130K$	$248K$
$\# wow$	$2,610$	$328K$	$84K$	$179K$
$\# sad$	$2,430$	$786K$	$70K$	$332K$
$\# angry$	678	$1,07M$	$93K$	$549K$

Table 3: Pairwise comparison matrix of WaPo

	$like$	$love$	$haha$	wow	sad	$angry$
$like$	–	$6,117$	$6,093$	$6,115$	$5,982$	$5,906$
$love$	0	–	$2,994$	$2,654$	$2,978$	$3,116$
$haha$	23	$2,003$	–	$1,872$	$2,415$	$2,295$
wow	2	$2,623$	$3,036$	–	$3,093$	$2,968$
sad	130	$2,203$	$2,104$	$1,717$	–	$1,880$
$angry$	209	$2,014$	$1,979$	$1,821$	$2,040$	–

Table 4: Summary of results on Facebook posts datasets (* means significance level of 0.1, and ** 0.01)

		Datasets			
	Methods	ROU	NYT	WSJ	WaPo
tau	RPC	0.933	0.497	0.595	0.541
	LWR	0.937	0.499	0.603	**0.562**
	SMP	0.933	0.495	0.601	0.547
	KNN-PL	0.929	0.495	0.584	0.544
	KNN-M	0.928	0.504	0.595	0.550
	LogLinear	0.935	0.488	0.593	0.537
	ROAR	**0.954****	**0.634****	**0.612***	0.554
GMR	RPC	0.241	0.429	0.345	0.358
	LWR	0.295	0.433	0.289	0.390
	SMP	0.246	0.351	0.257	0.247
	KNN-PL	**0.408***	0.458	0.513	0.461
	KNN-M	0.387	0.455	0.468	0.435
	LogLinear	0.209	0.287	0.253	0.203
	ROAR	0.343	**0.680****	**0.534***	**0.478***

ing, for each post, the labels are sorted according to their number of votes. If the number is zero for some labels, they are considered ranked at an extra tail position attached to the normal ranking without preference to each other and the ranking positions without labels are treated as missing. There are four data sets: (1) public posts from random ordinary users, denoted as ROU (Random Ordinary Users); (2) New York Times (NYT)[3] posts; (3) the Wall Street Journal (WSJ)[4] posts; and (4) the Washington Post (WaPo)[5] posts. We have crawled all four sets of posts in 2016 after Facebook introduced six emoticons.

As our focus is on the evaluation of our two proposals for the robust label ranking problem (instead of finding effective features), we avoid sophisticated features (e.g., user related or network structure based), and instead use fundamental textual features, extracted via AlchemyLanguage API (by IBM Watson Lab). For posts in ROU, only posts with text are included, and the document emotion of the text given by AlchemyLanguage is used as features. For posts in other three sets, if there is a link to external original full news, the document emotion of the full news is used as feature, and otherwise, only the text in posts is used. The returned document emotion from AlchemyLanguage consists of $[0, 1]$ scores, for five emotion dimensions, "anger", "joy", "fear", "sadness" and "disgust". The scores measure the amplitude of each emotion conveyed by the text. Then the four data sets are with the same feature and target format, that is, $\mathcal{Y} = \{like, love, haha, wow, sad, angry\}$ with $d = 6$ and $m = 5$ dimensional feature space.

The details of four data sets are shown in Table 2. Comparing ROU and the other three sets, the $IMBA$-$rank$ of ROU is much higher. This is due to the fact that readers of the posts from ordinary users are usually their friends, who tend to give positive feedback, $\{like, love, haha\}$ rather than negative one, $\{sad, angry\}$ All our datasets are significantly imbalanced in that $like$ or other positive labels are more frequent. This is partially due to the interface limitation such that users have to hover their mouse over the *Like* button to be able to select other emoticons. To illustrate imbalance in label ranking more clearly, we show the pairwise comparison matrix of WaPo, as Table 3, where the number in each entry (y_i, y_j) counts the number of posts (support) with y_i being higher ranked than y_j. For instance, there are only 209 posts where $angry$ is ranked higher than $like$ compared with $5,906$ in contrast.

We also use 16 semi-synthetic data sets obtained by converting benchmark multi-class classification using Naive Bayes and regression data using feature-to-label technique from the UCI and Statlog repositories into label ranking (Cheng, Hühn, and Hüllermeier 2009). These data sets are widely used as benchmark in label ranking works.

Results

Competing Models

First we test models on Facebook posts data sets, which are imbalanced. All results are obtained with 5-fold cross validation. We compare ROAR with 6 existing state-of-the-art label ranking models, RPC[6], LWD, SMP[7], KNN-PL[8], KNN-M and LogLinear.

[3] www.facebook.com/nytimes/

[4] www.facebook.com/wsj/

[5] www.facebook.com/washingtonpost/

[6] There is only one minor modification. In case of missing labels, missing label pairs indicate that they are not preferred to each other, as abstention means. Therefore, missing label pairs are assigned a small weight α, and counted as one preference relation for each order. The weight is picked empirically, $\alpha = 1/64$ used in this work, and results do not appear sensitive to the weight for a range of α values.

[7] There is a hyperparameter k (Grbovic, Djuric, and Vucetic 2013), which is set to default value 100 for Facebook data, and slightly smaller than number of all possible rankings for each semi-synthetic data.

[8] Two KNN based model takes default $K = 20$.

Table 5: Summary of results on semi-synthetic data sets

	tau							**GMR**						
Data	RPC	LWR	SMP	KNN-PL	KNN-M	LogLinear	**ROAR**	RPC	LWR	SMP	KNN-PL	KNN-M	LogLinear	**ROAR**
glass	0.877	**0.884**	0.476	0.809	0.807	0.812	0.851	0.785	**0.795**	0.471	0.582	0.625	0.625	0.772
authorship	0.919	0.912	0.636	**0.931**	0.930	0.497	0.873	**0.912**	0.910	0.571	0.903	0.903	0.808	0.870
pendigits	0.928	0.930	0.488	0.934	0.933	0.539	**0.936**	0.945	0.948	0.523	0.949	0.951	0.828	**0.954**
elevators	0.725	**0.789****	0.753	0.724	0.726	0.556	0.715	0.732	0.769	0.618	0.751	0.739	**0.793****	0.768
segment	0.927	0.950	0.123	0.940	0.940	0.733	**0.956***	0.948	0.963	0.220	0.955	0.958	0.885	**0.969****
wine	0.914	0.910	0.944	0.936	**0.948**	**0.948**	0.892	0.901	0.898	0.908	0.909	0.917	**0.917**	0.887
vowel	0.623	0.750	0.503	0.746	0.749	0.558	**0.796****	0.768	0.844	0.517	0.812	0.834	0.766	**0.870****
cpu	0.445	0.462	-0.010	0.496	**0.497**	0.358	0.370	0.534	0.672	0.035	0.672	**0.693****	0.675	0.656
vehicle	0.844	**0.860**	0.817	0.843	0.835	0.756	0.833	0.906	**0.908**	0.870	0.895	0.896	0.839	0.894
housing	0.667	0.685	0.469	0.647	0.661	0.606	**0.775****	0.806	0.817	0.632	0.798	0.806	0.784	**0.866****
iris	0.884	**0.982***	0.760	0.956	0.960	0.804	0.929	0.890	**0.927****	0.823	0.918	0.919	0.847	0.903
stock	0.750	0.850	0.697	**0.900**	0.899	0.643	0.888	0.859	0.910	0.820	**0.936**	0.936	0.806	0.932
calhousing	0.243	0.243	0.256	0.325	0.334	0.190	**0.338**	0.576	0.583	0.558	0.626	0.640	0.589	**0.663****
wisconsin	**0.626****	0.510	0.056	0.477	0.486	0.563	0.311	**0.777****	0.721	0.289	0.697	0.702	0.747	0.628
bodyfat	**0.292**	0.282	0.139	0.227	0.231	0.272	0.074	**0.623**	0.618	0.509	0.593	0.591	0.617	0.525
fried	**1.000**	0.990	0.470	0.902	0.905	0.994	0.879	**1.000**	0.995	0.697	0.951	0.952	0.997	0.939
# of common wins in both measures in ascending order: SMP=0, KNN-PL=1, KNN-M=1, LogLinear=1, RPC=3, LWR=3, and **ROAR=5**														

For evaluation, as we have shown the superiority of GMR in imbalanced data, here GMR is used. For consistency with previous label ranking works, results in terms of tau are also included in Tabel. 4.

Table 4 shows that ROAR achieves significantly better performance in all four data sets except ROU in terms of GMR. In ROU dataset, ROAR loses only to two KNN based methods. As pointed out previously, that posts emotion extracted from posts in ROU may not be meaningful enough, hence lack of structural correlation between feature and target favors instance-based learning method such as KNN. Hence the experiment shows that ROAR outperforms other models in real-world imbalanced label ranking data.

Next, ROAR and other label ranking models are applied to benchmark semi-synthetic data sets, evaluated by tau and GMR. As shown in Table. 5, ROAR achieves competitive results against other models and wins in the most data sets in terms of both tau and GMR.

Case study

What does it mean that an model performs better in terms of GMR in imbalanced label ranking data sets? Here we use WaPo data set result to answer it. Because imbalance measure of label ranking data $IMBA{-}rank$ is defined based on imbalance between two orders of each pair of labels (eq. 2), here we want to know whether an model can recall those minority orders in imbalanced pairs. In WaPo (Table. 3), we choose $(haha, like)$, $(sad, like)$ and $(angry, like)$ three minority pair orders, where (y_i, y_j) means y_i ranks higher than y_j. The number of posts with each of these orders is 23, 130 and 209, respectively, compared with that of those opposite, 6 093, 5 982 and 5 906. The $recall$ of each pair order is shown in Fig. 3. It is obvious that ROAR is superior than any other models. Actually any models except ROAR do not recall any posts with those minority pair orders, which is why they get same $recall$. Hence ROAR works better in recalling minority pair orders, as it achieves highest GMR in WaPo (Table. 4). However, this advantage is not well appreciated by tau as shown in Table. 4.

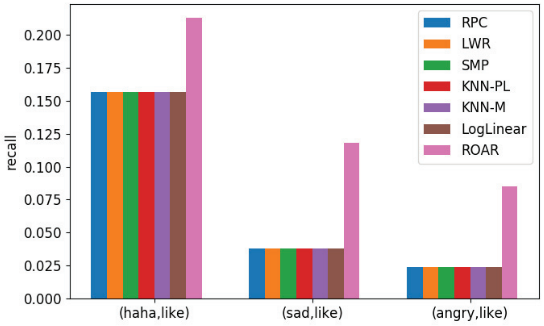


Figure 3: $recall$ of minority pair orders. Actually any models except ROAR do not recall any posts with those minority pair orders, which is why they get same $recall$ (not vanishing due to the smooth term in $recall$ definition).

Conclusion

In this work, we formally define robust label ranking problem for social emotion mining. To overcome the challenges, we first propose a robust measure, GMR, as the criterion for the problem. Both synthetic and experimental analysis show the superiority of GMR over popular measures such as *Kendall's tau correlation* and $ACC@k$. Then, we also propose a robust model, ROAR, and empirically validate its superiority over 6 competing label ranking models in Facebook posts data sets and benchmark semi-synthetic data sets.

Acknowledgement

The authors would like to thank the anonymous referees for their valuable comments and helpful suggestions. This work was supported in part by NSF CNS-1422215, NSF IUSE-1525601, NSF CNS-1742702 and Samsung GRO 2015 awards.

References

Bai, S.; Ning, Y.; Yuan, S.; and Zhu, T. 2012. Predicting readers emotion on chinese web news articles. In *ICPCA-*

SWS, 16–27. Springer.

Batista, G. E.; Prati, R. C.; and Monard, M. C. 2004. A study of the behavior of several methods for balancing machine learning training data. *ACM Sigkdd Explorations Newsletter* 6(1):20–29.

Busa-Fekete, R.; Hüllermeier, E.; and Szörényi, B. 2014. Preference-based rank elicitation using statistical models: The case of mallows. In *ICML*, 1071–1079.

Chawla, N. V.; Japkowicz, N.; and Kotcz, A. 2004. Editorial: special issue on learning from imbalanced data sets. *ACM Sigkdd Explorations Newsletter* 6(1):1–6.

Cheng, W.; Rademaker, M.; De Baets, B.; and Hüllermeier, E. 2010. Predicting partial orders: ranking with abstention. In *ECML-PKDD*, 215–230. Springer.

Cheng, W.; Hüllermeier, E.; Waegeman, W.; and Welker, V. 2012. Label ranking with partial abstention based on thresholded probabilistic models. In *NIPS*, 2501–2509.

Cheng, W.; Henzgen, S.; and Hüllermeier, E. 2013. Labelwise versus pairwise decomposition in label ranking. In *LWA*, 129–136.

Cheng, W.; Hühn, J.; and Hüllermeier, E. 2009. Decision tree and instance-based learning for label ranking. In *ICML*, 161–168. ACM.

Cheng, W.; Hüllermeier, E.; and Dembczynski, K. J. 2010. Label ranking methods based on the plackett-luce model. In *ICML*, 215–222.

de Borda, J. C. 1781. Memoire sur les élections au scrutin. *Histoire de l'Academie Royale des Sciences*.

Dekel, O.; Manning, C. D.; and Singer, Y. 2003. Log-linear models for label ranking. In *NIPS*.

Destercke, S. 2013. A pairwise label ranking method with imprecise scores and partial predictions. In *ECML-PKDD*, 112–127. Springer.

Diaconis, P., and Graham, R. L. 1977. Spearman's footrule as a measure of disarray. *Journal of the Royal Statistical Society. Series B (Methodological)* 262–268.

Grbovic, M.; Djuric, N.; and Vucetic, S. 2013. Multi-prototype label ranking with novel pairwise-to-total-rank aggregation. In *IJCAI*.

Har-Peled, S.; Roth, D.; and Zimak, D. 2002. Constraint classification for multiclass classification and ranking. In *NIPS*.

He, H., and Garcia, E. A. 2009. Learning from imbalanced data. *TKDE* 21(9):1263–1284.

Hüllermeier, E.; Fürnkranz, J.; Cheng, W.; and Brinker, K. 2008. Label ranking by learning pairwise preferences. *Artificial Intelligence* 172(16):1897–1916.

Japkowicz, N., and Stephen, S. 2002. The class imbalance problem: A systematic study. *Intelligent data analysis* 6(5):429–449.

Järvelin, K., and Kekäläinen, J. 2002. Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems (TOIS)* 20(4):422–446.

Jia, Y.; Chen, Z.; and Yu, S. 2009. Reader emotion classification of news headlines. In *NLP-KE 2009.*, 1–6. IEEE.

Kendall, M. G. 1948. *Rank correlation methods.* Charles Griffin & Co. Ltd., London.

Lee, P. H., and Philip, L. 2010. Distance-based tree models for ranking data. *Computational Statistics & Data Analysis* 54(6):1672–1682.

Lee, P. H., and Philip, L. 2012. Mixtures of weighted distance-based models for ranking data with applications in political studies. *Computational Statistics & Data Analysis* 56(8):2486–2500.

Lei, J.; Rao, Y.; Li, Q.; Quan, X.; and Wenyin, L. 2014. Towards building a social emotion detection system for online news. *Future Generation Computer Systems* 37:438–448.

Lin, K. H.-Y., and Chen, H.-H. 2008. Ranking reader emotions using pairwise loss minimization and emotional distribution regression. In *EMNLP*, 136–144. Association for Computational Linguistics.

Lin, K. H.-Y.; Yang, C.; and Chen, H.-H. 2008. Emotion classification of online news articles from the reader's perspective. In *WI-IAT'08.*, volume 1, 220–226. IEEE.

Mallows, C. L. 1957. Non-null ranking models. i. *Biometrika* 44(1/2):114–130.

Maloof, M. A. 2003. Learning when data sets are imbalanced and when costs are unequal and unknown. In *ICML-2003 workshop on learning from imbalanced data sets II*, volume 2, 2–1.

Shieh, G. S. 1998. A weighted kendall's tau statistic. *Statistics & probability letters* 39(1):17–24.

Spearman, C. 1904. The proof and measurement of association between two things. *The American journal of psychology* 15(1):72–101.

Sun, Y.; Kamel, M. S.; and Wang, Y. 2006. Boosting for learning multiple classes with imbalanced class distribution. In *ICDM*, 592–602. IEEE.

Tang, Y.-j., and Chen, H.-H. 2011. Emotion modeling from writer/reader perspectives using a microblog dataset. In *Proceedings of IJCNLP Workshop on sentiment analysis where AI meets psychology*, 11–19.

Weiss, G. M. 2004. Mining with rarity: a unifying framework. *ACM Sigkdd Explorations Newsletter* 6(1):7–19.

Zhang, Y.; Zhang, N.; Si, L.; Lu, Y.; Wang, Q.; and Yuan, X. 2014. Cross-domain and cross-category emotion tagging for comments of online news. In *SIGIR*, 627–636. ACM.

Zhou, Y.; Liu, Y.; Gao, X.-Z.; and Qiu, G. 2014. A label ranking method based on gaussian mixture model. *Knowledge-Based Systems* 72:108–113.

Zhu, C.; Zhu, H.; Ge, Y.; Chen, E.; and Liu, Q. 2014. Tracking the evolution of social emotions: A time-aware topic modeling perspective. In *ICDM*, 697–706. IEEE.