Matrix factorization for co-training algorithm to classify human rights abuses

Ragini Gokhale
School of Computer Science and Electronic Engineering
University of Essex
Essex, United Kingdom
ragokh@essex.ac.uk

Maria Fasli
Institute for Analytics and Data Science
School of Computer Science and Electronic Engineering
University of Essex
Essex, United Kingdom
mfasli@essex.ac.uk

***Abstract*—In the human rights domain, there is need to filter, efficiently classify and prioritize the types of violation endured by victims in order to provide the necessary rehabilitation and support. However, the domain is dominated by unstructured data either from victims' accounts, doctors'/professionals' reports or available on line. Manual classification still prevails in this domain which is extremely time consuming and slow. This is a problem for non-government operated charities. To this end we have explored the application of the co-training algorithm in order to improve the performance of a semi-supervised learning algorithm by incorporating large amounts of unlabeled data into the training data set. However, it remains challenging to apply co-training on the data without two independent and self sufficient views.This paper puts forth a method of randomly dividing the available features to apply matrix factorization so as to discover latent features underlying the interactions between different kinds of entities present in a single view dataset. These labeled views balance the biased information in the dataset, but still satisfy the co-training assumptions. Alongside, the views are constrained such that pairs of labeled views create weak classifiers which in turn increase the prediction accuracy when combined. In the majority of cases, any classification tries to connect a single class to each sample or object. However, in the human rights domain, a victim can be subjected to more than one type of violation or abuse. This is multi-label classification where a sample can be assigned to more than one class. This paper aims to address all these aspects by bringing together a semi supervised classification model that relies on the effectiveness of matrix collaborative filtering in order to classify stories narrated by victims into one or more types of human rights abuses. Experimental results demonstrate the efficiency of this approach when applied on real-world stories from different victims.**

***Index Terms*—semi-supervised learning; co-training; multi-label classification; matrix factorization; human rights violations;**

I. Introduction

Although the massive growth of the Internet has provided access to unprecedented levels of data and information, typically such information is not well organised and this has resulted into the demand for structuring and systematizing data and information and how these are presented. Information retrieval, artificial intelligence, machine learning and user profiling techniques aim to make searching for documents efficient and more accurate in domains ranging from the public sector to general search on the web. However the application of such techniques is still lagging behind in the domain of human rights. Technology could potentially help in gathering new and different kinds of information to document human rights violations, especially in areas that are not safe and even inaccessible. Big Data analytics can use this data to analyze key trends and provide early warnings for critical issues before they occur, aiding the prevention and rapid response to humanitarian disasters.

Amongst the key areas in the human rights domain, classifying violations can be seen as a potential area where principles of artificial intelligence and machine learning can be deployed in order to assist recovery, progress and overall rehabilitation of victims and survivors from war-ridden or hate crime affected areas. For any learning tasks, while it is very expensive to obtain a sufficient amount of labeled data, unlabeled can be acquired comparatively easier and they are typically in abundance. The co-training algorithm [1], which is a semi-supervised learning approach, attempts to blend the insufficient labeled data and a large amount of unlabeled data to achieve better learning performance by training two classifiers from two conditionally independent and separate sets of labeled data. But it is difficult to segregate the survivor stories into such independent views, thus limiting the scope of the co-training model. According to [2], he proposes a greedy algorithm that the independence assumption on the views is discounted for, then co-training would still be applied under a

weaker independence assumption as effective as its conditional independence counterpart. His method is to maximize agreement on unlabeled data, which produces good results in a co-training algorithm for classification. Therefore an attempt is made in this paper, to randomly split the single-view data set into two views based on the labels and make the assumption that each view formed is self-sufficient for correct classification. The labeled and unlabeled data is computed into lower rank factorizations using a Singular Value Decomposition (SVD) which characterizes both stories and labels by vectors of factors inferred from label ranking patterns. Most of the classification problems investigated by machine learning algorithms are single-label classification problems. However, since a victim can be subjected to more than one violation, this paper extends the scope of the co-training model to multi-label classification. Experimental results on a real text dataset which is a collection of survivor stories collected by scraping the web, show that this approach attains better performance in contrast to some existing classification methods.

The rest of the paper is organized as follows: Section II, describes the research and work done using similar techniques of data mining and machine learning in various domains. Section III covers the technical details and the method used for the approach. The overview of the datasets followed by experimental findings is covered in Section IV. Section V discusses future work and the conclusions.

II. Related work

A. Co-training

Acknowledging the growing concern of having insufficient labeled data set to label and organize extensive unstructured data collected over the Internet in several domains, Blum and Mitchell [1] attempt to increase the amount of the labeled dataset using large amounts of the available unlabeled data with their proposition of the co-training algorithm. The co-training algorithm works by generating two classifiers trained on the input labeled data, which are then used to tag new unlabeled data. From this newly labeled data, the most confident predictions are added to the set of labeled data. In natural language learning, co-training has been applied to statistical parsing [3], reference resolution [4], part of speech tagging [5], and others, and has generally been found to bring improvement in cases where no additional unlabeled data are used.

The efficiency of the co-training algorithm per [1] relied on the conditional independence of the views that the training data could be split into. However in recent times, [2] demonstrated that overlooking the conditional independence of the views still maintained the effectiveness of the co-training algorithm. This was achieved by deploying a greedy algorithm that works under a weaker independence assumption to produce good results in labeling the data. This works as an added advantage for the model this paper is trying to put forth as it is extremely challenging to split a survivor stories into two conditionally independent views. In the sections that follow, this paper shows how to achieve new labels to tag unlabeled stories by relaxing the criteria and dividing the training dataset based on the number of discovered features.

B. Multi-label Classification

A simple classification problem is to categorize the given corpus into the available labels. Therefore classification can be based on the number of labels that can be assigned to any given sample. In some cases with mutually exclusive labels the classifier tends to classify the sample into just one label. This is single-label classification. When such classification is applied to victim stories (the focus of this paper), it leads to a single type of violation. This may result in incorrect medical care or reducing the chance of proper rehabilitation of the victim. Therefore it is important to look at victims' stories and categorize their narration into more than one violations which has led us to focusing on multi-label classification where each sample can be simultaneously associated with more than one class. As seen from recent studies, text classification is the primary scope for multi-label classification techniques [6], [7], [8], [9]. However, it is also finding significant applications in areas like bioinformatics [10], [11], [12],medical diagnosis [13], and scene classification [14], [15].

Using this approach benefits the subject of this paper because it focuses on specific supervised algorithms like SVC (kernel=linear), Decision Trees, K-Nearest Neighbours, which results in a better performance than its algorithm-independent counterparts like Instance based and Label decomposition.

C. Matrix factorization

Recommendations based on matrix factorization (MF) as proposed by Karen et al. [16] have achieved acceptable results for prediction ratings in cases where there is are partial ranked data. This is achieved by implementing low-rank matrix factorization [17], [18], [19] for the latent factor model that tries to find hidden patterns between the user matrix and the item matrix to

predict the user's ratings on previously unrated items. This method of recommendation has been used across many domains where users do not search reviews directly but are suggested products that would best match some definition of preference [20]. Current recommendation systems, such as the ones used by Netflix [21] or Amazon [22], rely on ratings to make recommendations. Another well known study area for recommendation using matrix factorization is analyzing different types of beers in [23]. Patterns can emerge which suggest the reasons why some people like the smell and taste of one beer and totally avoid others.

The ability of the method to be used on any domain benefits our approach in this paper because recommendations from the recommender system can be used to emphasize the different violations the unlabeled victim stories depict, by finding similar stories in the training set. This would help in labeling the unlabeled story according to the highest ranked labeled similar story and thus adding the newly identified labels to the labeled data and align with the principle of the co-training algorithm.

D. Human Rights

With the progress of technology and wider availability of mobile phones and other devices, it has become relatively easier to document, report, and monitor human rights violations, and subsequently analyze patterns and trends. Human rights research is venturing into utilizing the full potential of Big Data and Artificial Intelligence to try and address the challenges and difficulties faced in this domain due to the sensitivity of the subject. Although efforts are continuous, work in this area requires considerable resources with respect to time, financial investment, and expertise. Researchers have progressed with finding solutions for some of these issues [24], [25], [26], but there is an inherent need to strengthen the existing regulatory frameworks and have a human rights approach to the risks and limitations of using Big Data and this paper aims to support work in this area.

III. METHOD

A. Matrix factorization

With competition between internet-based organisations such as for instance Amazon, Netflix and Youtube intensifying, maintaining and increasing user satisfaction so that users can remain loyal is of paramount importance; typically exemplified through relating their preferences to the items they view, rank, shop and like. Organisations commonly use recommender systems that can analyze the trends or patterns exhibited by the user and associate the next best items for them. Such systems are based on two distinct methods:

- Content-based Filtering [27], which creates a profile for each user or product to characterize them;
- Collaborative Filtering [28], which relies on past user activities without any profile creation.

In this paper, the collaborative filtering method is exploited to find hidden and less explicit relationships between a few victim survivor stories that were labeled by experts in the human rights domain. The other reason to focus on this approach is because this method, being domain independent, can identify latent features underlying between two different kinds of stories and determine if they are similar or not. Collaborative filtering has two additional models, nearest neighbourhood model and the latent factor model, that can provide recommendations. Although the neighbourhood model assesses the user interaction with the preferred items based on items within the same neighborhood, this paper is making use of the second model where emphasis is given to users and items equally. This is because there is no direct relation that can be established between a story and the label based on the intensity, duration and other such factors that determine the level of abuse faced by the victim. Matrix factorization derives from the latent factor model and this paper takes advantage of this approach to compute additional labels using the small labeled dataset to classify the violations.

The idea behind this model is that in its basic form, both items and users (in this case, stories and labels) are distinguished by feature vectors based on trends and patterns, and that there must be some relationship in which the user (story) rates a particular item (label). In the current labeled dataset, the experts have rated some of the stories against each label but left the rest empty. The task now is to rate the remainder of the stories to predict ratings for new stories in order to classify them from the unlabeled set. Placing the stories and their labels in a two-dimensional matrix would help create a vector that describes the marked and unmarked stories that reveal the missing data. Applying this to the sparsely labeled set of victim stories and their respective rankings, the initial matrix for stories with their labels would look as shown in figure 1.

In figure 1, there are 14 unique story lines, each having 7 labels ranked from 1-10 (1-lowest to 10 -highest). Those which are marked as 0 are the ones which are not rated by the experts.

h

	story	Sensory Deprivation	Stress Techniques	temperature manipulation	Coercion	Intimidation	Sexual violence	Forced Exertion
0	Hooding with sandbags/cement bags	6	7	0	0	0	1	2
1	Blackened goggles	0	7	3	0	0	1	0
2	Forced silence/ duct tape on mouth	0	6	2	4	7	0	0
3	Prolonged sitting in required position	7	0	0	0	3	1	0
4	Prolonged standing/wall standing	7	5	0	4	3	1	6
5	Sexual violence to genitals	3	0	0	0	6	7	2
6	Molestation	3	4	1	0	0	7	2
7	Penetration using instruments	0	0	1	5	0	7	0
8	Pressed on hot surfaces	6	0	7	0	0	0	0
9	Stamping on victim	0	6	0	5	0	0	7
10	Dragging victim along ground	0	7	2	4	5	0	3
11	Simulated drowning	0	0	0	0	3	0	7
12	Verbal abuse	0	0	2	0	7	1	4
13	Detention in unbearably hot locations	6	3	7	0	0	0	0

Fig. 1. Matrix of user stories labeled by human rights experts.

B. Ranking

Matrix factorization models map both users and items to a joint latent factor space of dimensionality f, such that user-item (story-label) interactions are modelled as inner products in that space. Let $\mathbf{R}$ of size $|U| \times |D|$ be the matrix that contains all the ratings that the users (story) have assigned to the items. The task is to discover K latent features by finding out two matrices $\mathbf{P}(a|U| \times Kmatrix)$ and $\mathbf{Q}(a|D| \times Kmatrix)$ such that their product approximates $\mathbf{R}$:

$$\mathbf{R} \approx \mathbf{P} \times \mathbf{Q}^T = \hat{\mathbf{R}} \tag{1}$$

In this way, each row of $\mathbf{P}$ would represent the strength of the associations between a user (story) and the features. Similarly, each row of $\mathbf{Q}$ would represent the strength of the associations between an item (label) and the features. To get the prediction of a rating of an item (label) d_j by u_i, the dot product of the two vectors corresponding to u_i and d_j is calculated:

$$\hat{r}_{ij} = p_i^T q_j = \sum_{k=1}^{k} p_{ik} q_{kj} \tag{2}$$

However, in the above case, the chances of over-fitting the features is high. In order to reduce over-fitting and penalize complexities (if any), λ is added in as constant to control the regularization as follows:

$$e_{ij}^2 = (r_{ij} - \sum_{k=1}^{K} p_{ik} q_{kj})^2 + \frac{\lambda}{2} \sum_{k=1}^{K} (||P||^2 + ||Q||^2) \tag{3}$$

This constant, controls the magnitudes of the feature vectors of U and D to obtain a good approximation. This is applied to the sparsely labeled set of victim stories and their respective rankings and is re-executed to generate $\mathbf{P}$ and $\mathbf{Q}$. Once $\mathbf{P}$ and $\mathbf{Q}$ are derived from $\mathbf{R}$, their dot product is taken to produce the predicted ratings for the missing values.

Figure 2 describes how each story (by its id) is categorized by ranks into 7 different types of violations. There are some gaps seen in the figure, this means that these labels are either not ranked or have a very low rating. Hence, the task of predicting the missing ratings can be considered as filling in the blanks so that the values would be consistent with the existing ratings in the matrix to generate a robust training set for the classifiers to be trained on.

Figure 3 is the output of the dot product $\mathbf{P}$ and $\mathbf{Q}$, where$\mathbf{P}$ and $\mathbf{Q}$ are derived from the initial matrix shown in Figure 1. As seen in Figure 3, the new ranks are approximately close to those ranked by the human rights experts. Looking at this graph or the matrix it generates, similarities between stories can be ascertained. For example it is reasonable to say:

- story ids 4 and 5 which have the content as "Prolonged sitting in required position" and "Prolonged standing/wall standing" respectively with the same labels "Forced Exertion,Sensory Deprivation,Stress Techniques,temperature manipulation"
- story ids 6 and 7 are similar having content as "Molestation" and "Penetration using instruments" respectively with label as "Coercion, Intimidation, Sexual violence, Stress Techniques"

This supports the hypotheses of the paper which is that a large labeled dataset can be created using minimal input from experts and matrix factorization.

Updating the labeled set with new ranks, a selected few unlabeled samples (as described in the modified algorithm in section Co-training extension below) from the unlabeled data set are added to the training set to predict their rating in order to find recommended stories and the labels for them. As seen in figure 4, story ids 15 onwards are the newly added stories which are now ranked with respect to the earlier ranked stories. Additionally it can be seen that the newly added stories with id 15 and 16 are similar and thus labeled as "Coercion, Sexual Violence, temperature manipulation", where as story ids 19 and 20 are labeled as "Stress Techniques, Coercion, Forced Exertion'. Labels that have a ranking 3 and above out of

storyid	story	Sensory Deprivation	Stress Techniques	temperature manipulation	Coercion	Intimidation	Sexual violence	Forced Exertion
1	Hooding with sandbags/cement bags	6	7				1	2
2	Blackened goggles		7	3			1	
3	Forced silence/ duct tape on mouth		6	2	4	7		
4	Prolonged sitting in required position	7				3	1	
5	Prolonged standing/wall standing	7	5		4	3	1	6
6	Sexual violence to genitals	3				6	7	2
7	Molestation	3	4	1			7	2
8	Penetration using instruments			1	5		7	
9	Pressed on hot surfaces	6		7				
10	Stamping on victim		6		5			7
11	Dragging victim along ground		7	2	4	5		3
12	Simulated drowning					3		7
13	Verbal abuse			2		7	1	4
14	Detention in unbearably hot locations	6	3	7				

Fig. 2. Initial ranking of all the labels for each story id by the human right experts.

storyid	story	Sensory Deprivation	Stress Techniques	temperature manipulation	Coercion	Intimidation	Sexual violence	Forced Exertion
1	Hooding with sandbags/cement bags	5.95	6.69	8	3.76	0.77	1.19	2.4
2	Blackened goggles	9.16	6.62	3.32	5.38	8.26	1.19	7.37
3	Forced silence/ duct tape on mouth	5.84	5.89	1.88	4.3	6.91	4.39	4.44
4	Prolonged sitting in required position	7	6	4.92	4.2	2.97	1.01	4.66
5	Prolonged standing/wall standing	7.03	5.55	4.44	4.07	3.48	0.42	4.94
6	Sexual violence to genitals	3.03	5.18	0.51	3.35	5.94	7.03	2.02
7	Molestation	2.78	4.84	0.56	3.1	5.36	6.55	1.81
8	Penetration using instruments	5.42	6.71	1.16	4.69	8.32	7.09	4.06
9	Pressed on hot surfaces	6.09	7.39	6.87	4.32	0.2	3.36	2.77
10	Stamping on victim	9.32	6.08	5.32	4.85	4.61	0.65	7
11	Dragging victim along ground	4.69	6.44	2.44	4.15	5.24	6.33	2.92
12	Simulated drowning	9.9	6.67	6.87	5.06	2.99	0.25	6.99
13	Verbal abuse	5.83	4.24	1.46	3.54	6.37	1.32	4.89
14	Detention in unbearably hot locations	5.88	3.19	6.91	2.18	0.58	0.12	3.39

Fig. 3. New ranks for the stories after matrix factorization.

storyid	story	Sensory Deprivation	Stress Techniques	temperature manipulation	Coercion	Intimidation	Sexual violence	Forced Exertion
1	Hooding with sandbags/cement bags	3.2147273226	2.046006569	5.5909608046	5.9242696968	0.982876927	7.032166697	0.4174157072
2	Blackened goggles	3.9442946182	3.9843324972	5.1234939492	7.1067089121	0.996651268	6.9881835064	2.9785401435
3	Forced silence/ duct tape on mouth	4.4315010247	3.0116214957	6.6229118586	5.3906535757	4.903167234	6.1074687399	1.798351025
4	Prolonged sitting in required position	4.1230720701	6.144849837	2.9588177475	7.0038962152	1.029695498	4.8486013121	6.4944344626
5	Prolonged standing/wall standing	4.062694586	5.9568889995	3.0486611288	6.9690587001	0.954610467	4.9563341857	6.2167206022
6	Sexual violence to genitals	3.9336414641	1.9951026597	6.0864629048	2.9633373227	6.909618794	3.8537430608	1.0830950093
7	Molestation	3.9755364151	1.9765476467	6.2052313594	3.0219093764	6.959483939	3.9667833608	1.0206085795
8	Penetration using instruments	4.8172638615	2.5101945139	7.7146796952	4.7115552171	7.03771962	5.9795870945	1.0811290252
9	Pressed on hot surfaces	3.9046414596	6.1884069039	2.1873373887	6.0151519615	1.671735751	3.4744483708	6.9571413002
10	Stamping on victim	4.9587779801	6.9753055137	3.9762012417	8.2406481201	1.593884081	6.0132739146	7.1912062245
11	Dragging victim along ground	3.6834220406	3.2013498232	5.3136396356	6.4279223252	1.349532151	6.7783024007	1.9994847017
12	Simulated drowning	4.454539252	6.9702152151	2.9919284357	8.1256904394	0.293018068	5.5985612881	7.3759197659
13	Verbal abuse	4.5943064796	3.8582009841	6.9450603629	8.5409190044	1.042337086	9.2272914119	2.1016451906
14	Detention in unbearably hot locations	2.932418950	6.0065504874	0.5367366549	5.9551869684	0.997328234	3.0410517305	6.9786170995
15	Reprieve - The stories of three men executed by firing squad	2.2196904255	2.4541144773	2.5545977562	3.7223981086	0.860018441	3.3863205633	2.1139064976
16	Freedom from Torture - The true legacy of torture	2.2836297698	2.0814854755	3.058666158	3.584158564	1.329559973	3.6167726748	1.5294354538
17	Teresa Celia Meschiati	1.6702526517	0.7962468346	2.8960243096	2.0711632805	1.893346544	2.681508761	0.1101188087
18	testimonies of torture survivors	0.9252844495	0.7934193293	1.1644716144	1.0289504738	1.102283727	0.9959810888	0.6766465247
19	Tesfaye's Story	2.3425400171	3.2437163936	1.7720502541	3.3632600957	1.454982893	2.2424225213	3.4865378069
20	Saad's Story	2.7891462552	3.212030705	2.7988767054	3.8295944064	2.144390338	3.0914828881	3.1441968783

Fig. 4. New labels obtained after applying matrix factorization to the unlabeled data.

7 are considered good enough to be added to the initial labeled set . This process is repeated a finite number of times or until all the unlabeled data are labeled.

The advantage of using matrix factorization is that results can be inferred through using minimal input from experts (which is extremely time consuming as well as expensive). This assists in getting a larger labeled sample set to train the classifier.

C. *Recommending*

After building the neighborhood group of relevant rating of the missing labels in the labeled dataset, recommendations are generated for similar stories from the unlabeled set. Collaborative Filtering generates recommendations as either prediction, which is one numerical value, or recommendation, which is a list of top N labels. Since this paper proposes to use the latent factor model that recommends based on prior behaviour, similarities between the stories have to be determined.

Applying the cosine similarity along with (SVD) [18],

the similarities can be identified between the stories to predict the labels. The benefit of using both cosine similarity and SVD is to avoid negative correlation due to lack of data and handle the issues of scalability and sparsity.

$$cos(\boldsymbol{x}, \boldsymbol{y}) = \frac{\boldsymbol{x} \cdot \boldsymbol{y}}{||\boldsymbol{x}|| \cdot ||\boldsymbol{y}||} \tag{4}$$

SVD

$$X = U \times S \times V^T \tag{5}$$

Matrix X can be factorized to U, S and V. The U matrix represents the feature vectors corresponding to the stories in the hidden feature space and the V matrix represents the feature vectors corresponding to the labels in the hidden feature space. This calculates the vector which establishes the relationship between the stories and theirs labels. With this newly identified relationship between the stories and labels, the recommender lists top n stories similar to the story it needs recommendations for.

Figure 5 displays the top 5 recommendations for 4 randomly selected stories from the unlabeled dataset. As seen from the figure, each unlabeled story has 4 recommended stories and labels based on similar features that have been discovered using matrix factorization. Each of the recommended stories are ranked based on the average of individual ranks of each label. The figure thus display the 4 stories in descending order of their ranks. Out of the 4 recommendations, the story with the highest rank will be selected as the label for the unlabeled story.

D. Co-training extension

Blum and Mitchell [1] proposed a conditional independence assumption to combine labeled and unlabeled data to classify the content into relevant classes. Their algorithm is shown in Algorithm 1.

This paper extends the above algorithm by incorporating matrix factorization to first rank the missing labels. The newly ranked labeled set is used to predict ranks for some selected unlabeled examples to train the two classifiers to predict the classes for the entire unlabeled dataset. The trained classifiers are then executed multiple times over the entire unlabeled dataset to eventually predict their labels. The average prediction over several iterations is considered as the final prediction. This is because the training set is being randomly divided into two in order to have 2 views for the classifiers to run on and doing so might result in the training data, of either of the classifiers, not having enough data to be trained on which results into producing bad results.

Algorithm 1: Co-training Algorithm by Blum-Mitchell

Given:
a set L of labeled training examples
a set U of unlabeled examples
Create a pool U' of examples by choosing u examples at random from U
for $i \leftarrow 0$ to k by 1 do
  Use L to train a classifier h1on only the x1 portion of x
  Use L to train a classifier h2 on only the x2 portion of x
  Allow h1 to label p positive and n negative examples from U'
  Allow h2 to label p positive and n negative examples from U'
  Add these self-labeled examples to L
  Randomly choose 2p + 2n examples from U to replenish U'

Blum and Mitchell's algorithm works only for binary classification. This paper further broadens the scope of co-training to classify the unlabeled data into multiple labels.

The proposed algorithm to deal with multiple labels is shown in Algorithm 2.

IV. Experimental Setup

A. Dataset

A labeled dataset was created by experts in the human rights domain, identifying 10 different types of torture each describing the content for it. For example hooding with sandbags/cement bags, blackened goggles, plastic blindfold, sight deprivation by other means are types of sensory deprivation. In this, hooding with sandbags is the content and sensory deprivation is the label.

The unlabeled set is a 2.8 MB collection of 238 testimonies and survivor stories scraped from the Internet. This data was preprocessed in order to have meaningful content for classification. The normalization of this data includes removal of all the HTML/css/link tags, removal of stop words and punctuation and all text being converted into lower case.

B. Performance Evaluation

C. Classifiers

For thorough comparison, the proposed co-training classification model is evaluated against state-of-the-art

Unlabeled story	Recommemded story	Recommended label
Survivor Stories Building Freedom Brick by Brick	Hooding with sandbags/cement bags	Sensory Deprivation;Intimidation
Survivor Stories Building Freedom Brick by Brick	Forced silence/ duct tape on mouth	Sensory Deprivation;Intimidation
Survivor Stories Building Freedom Brick by Brick	Prolonged sitting in required position	Stress Techniques;Forced Exertion
Survivor Stories Building Freedom Brick by Brick	Sexual violence to genitals	Sexual Violence;Coercion;Intimidation
Survivor Stories Building Freedom Brick by Brick	Detention in unbearably hot locations	Sensory Deprivation;temperature manipulation
Youssef's Story	Hooding with sandbags/cement bags	Sensory Deprivation;Intimidation
Youssef's Story	Forced silence/ duct tape on mouth	Sensory Deprivation;Intimidation
Youssef's Story	Prolonged sitting in required position	Stress Techniques;Forced Exertion
Youssef's Story	Sexual violence to genitals	Sexual Violence;Coercion;Intimidation
Youssef's Story	Pressed on hot surfaces	Stress Techniques;temperature manipulation;Coercion
Peter's Story	Hooding with sandbags/cement bags	Sensory Deprivation;Intimidation
Peter's Story	Forced silence/ duct tape on mouth	Sensory Deprivation;Intimidation
Peter's Story	Prolonged sitting in required position	Stress Techniques;Forced Exertion
Peter's Story	Sexual violence to genitals	Sexual Violence;Coercion;Intimidation
Peter's Story	Detention in unbearably hot locations	Sensory Deprivation;temperature manipulation
Yonas' Story	Hooding with sandbags/cement bags	Sensory Deprivation;Intimidation
Yonas' Story	Prolonged sitting in required position	Stress Techniques;Forced Exertion
Yonas' Story	Sexual violence to genitals	Sexual Violence;Coercion;Intimidation
Yonas' Story	Detention in unbearably hot locations	Sensory Deprivation;temperature manipulation
Yonas' Story	Dr. Candeloro, Husband of Marta Garcia de Candeloro.	Sensory Deprivation,Stress Techniques,temperature manipulation

Fig. 5. Top 4 recommendations with labels for 5 randomly selected unlabeled stories

supervised classifiers, as well as some variants of co-training. The baselines considered are:

- **SVC** a widely used supervised text classifier. The linear kernel and onevs-rest scheme with TF-IDF weighting.
- **Decision Trees** - A popular classifier where trees are constructed where each node corresponds to a group of instances from the dataset. They can be adapted by taking multiple labels into consideration in decision functions.
- **MlKNN** - [11] propose a new multi-label learning algorithm based on K-Nearest Neighbours (KNN). This model uses a lazy-learning approach.
- **Logistic Regression** - In the multi-label case, Logistic Regression uses the one-vs-rest (OvR) scheme and uses the cross-entropy loss when the option is set to 'multinomial'.
- **GaussianNB** - A multi class text classifier which is used with the onevs-rest scheme provides apt multi label classification.
- **Random Forest** - A random forest is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting.
- **Co-training algorithm for SVC (kernel=linear), Logistic Regression, Decision Trees, Random Forest and NaiveBayes** - Using the co-training model to classify human right abuses/violations [29].

D. Metrics

For multi-label classification, to evaluate the performance, it is necessary to evaluate the ranking prediction of the most relevant documents for each category in order to quantify the quality of the predicted values. Thus all classifiers are evaluated for their macro-averaged F1. Another state-of-the-art performance evaluation criterion for multi-label classification is the Hamming Loss which measures the number of times a pair (instance, label) is misclassified.

$$F1 = \frac{2 \cdot precision \cdot recall}{precision + recall} \tag{6}$$

For recommender systems, the proposed algorithm is evaluated against the accuracy and coverage. For accuracy, the Root Mean Square (RMSE) is a popular state-of-the-art metric which is used for evaluating predicted ratings with actual ratings.

$$RMSE = \sqrt{\frac{1}{n}\Sigma_{i=1}^{n}\left(\frac{d_i - f_i}{\sigma_i}\right)^2} \tag{7}$$

The average precision score is also used to evaluate the effectiveness of the recommended values. It is the average of the maximum precisions at different recall values [30], [31].

$$AP = \sum_{n} (R_n - R_{n-1})\, P_n \tag{8}$$

where P_n and R_n are the precision and recall at the n-th threshold.

E. Results

Figure 6 provides a comparison overview of the various state of art multi-label classifiers, the co-training algorithm with label ranking and the proposed collaborative co-training algorithm. Based on the results, the paper

Algorithm 2: Proposed Co-training Algorithm

Given:
a set L of labeled training examples with only some samples ranked
Use MF rank the training examples
Create a pool U' by choosing u at random from U
Divide training set X into x1 and x2 randomly
for $i \leftarrow 0$ **to** k **by** 1 **do**
- **Use MF rank U' samples**
- **Use L to train h1 on the x1 portion of x**
- **Use L to train h2 on the x2 portion of x**
- **Use R to recommend top n recommendations for U'**
- **Compare recommendations with prediction for U' by h1 for x1 ||Select the highest ranked prediction as positive samples**
- **Compare recommendations with prediction for U' by h1 for x1 portion**
- **Select the lowest ranked prediction as negative samples**
- **Compare recommendations with prediction for U' by h2 for x2**
- **Select the highest ranked prediction**
- **Compare recommendations with prediction for U' by h2 for x2**
- **Select the lowest ranked prediction as negative samples**
- **Label p positive and n negative examples from U'**
- **Label p positive and n negative examples from U'**
- **Add these self-labeled examples to L**
- **Randomly choose 2p + 2n examples from U to replenish U'**

shows that overall the coverage error rate displayed by the improved co-training algorithm is much lower than its relevant counterparts. This indicates that the predicted labels (one or more) for a story match completely the truth value for it. The truth value for the stories was manually recorded by reading through each story one by one.

It is also seen from the results, that the macro F1 measure evaluates the ability of the algorithms to correctly identify the relevance of each label. RandomForest classifier combined with the proposed co-training algorithm provides the best results amongst all the classifiers used for the experiment. The result determines that the improved co-training RandomForest classifier has the highest chance to correctly identify the relevance of each label.

In a multi-label scenario such as this, Hamming Loss is an approximate prediction that can be efficiently computed from label-wise information and hence determines the accuracy of the classifier. As seen in Figure 6, when a GaussianNB classifier is teamed up with the proposed collaborative co-training classifier, it performs significantly better than the rest. This means that the percentage of true labels predicted matches exactly the manually recorded ground truth. This also aligns with the fact that Hamming Loss is a binary relevance method, which only trains a learner for each label without taking into account dependencies. The proposed collaborative co-training model relies on using the OneVsRest Classifier, whose approach is to fit one classifier per class.

ROC curves are another good measure for determining the accuracy of any classifier. As seen in figure 6, KNN with the collaborative co-training model classifier performs marginally better as compared to the others. This indicates that the ability of the new recommender model along with co-training is higher than standard state-of-the-art classifiers. This also can be seen in figure 7, that displays the accuracy of the newly proposed collaborative co-training algorithm paired with currently efficiently working classifiers. Thus the GaussianNB classifier predicts up to 70% correct labels for stories followed by the DecisionTree classifier.

As mentioned in section III-C, cosine similarity is used to build the recommender as it determines the similarity between two vectors by measuring the angle between them. Lower values of RMSE help confirm that the vectors are in the neighbourhood of each other and the aim of the recommender is to minimise the RMSE while predicting the labels for the stories. An RMSE of 0.5 implies that on average, the recommender is approximately 0.5 off with each prediction. However an RMSE over 0.6 is considered to be a good recommended prediction. Thus the classifier that performs the best recommending over 74% correct predictions is the GaussianNB classifier.

When comparing the labels predicted for the stories with the ground truth, figure 9 shows the variations of the labels for the stories. As seen from the figure, the predictions made by the GaussianNB classifier are relatively closer to the truth value as compared to the other classifiers.

Metrics	Average precision Score	Coverage Error	Hamming Loss	F1 macro averaging	ROC
Decision Tree	33.2186	7	0.5210084034	0.1743816837	0.4907631129
Decision Tree with co-training	20	7	0.5357142857	0.1746530617	0.52417144
Decision Tree with collaborative cotraining	32.8835	5.6764705882	0.5453501401	0.3602407427	0.5872473233
GaussianNB	33.0937	6.6078431373	0.6848739496	0.3588041085	0.4821626928
GaussianNB with co-training	22.6272	7	0.4285714286	0.2932653061	0.4285714286
GaussianNB with collaborative cotraining	33.7344	5.6862745098	0.7008403361	0.2624585094	0.5019943329
KNN	33.1666	6.4607843137	0.3767507003	0.1709825063	0.4886688667
KNN with co-training	20.0256	7	0.4142857143	0.22	0.5428571429
KNN with collaborative cotraining	33.4175	6.2549019608	0.3991596639	0.2902325899	0.5994271601
Logistic regression	33.5038	6.6568627451	0.3921568627	0.1635151639	0.5031892511
Logistic regression with co-training	22.3077	7	0.4285714286	0.1632653061	0.4885714286
Logistic regression with collaborative cotraining	33.5549	6.1117647059	0.4243697479	0.2574040281	0.5042276872
Random Forest	33.3333	7	0.5182072829	0.1766843747	0.5
Random Forest with co-training	18.8462	7	0.4	0.1176470588	0.5142857143
Random Forest with collaborative cotraining	33.2733	5.7352941176	0.531372549	0.3937449168	0.5956621861
SVC	33.4348	6.6568627451	0.393557423	0.1650969168	0.4998759381
SVC with co-training	23.0769	7	0.4282156863	0.2142857143	0.5242857143
SVC with collaborative cotraining	33.3333	5.862745098	0.4291596639	0.279327904	0.5

Fig. 6. Comparison of various state of art classification models with the proposed co-training classifier.

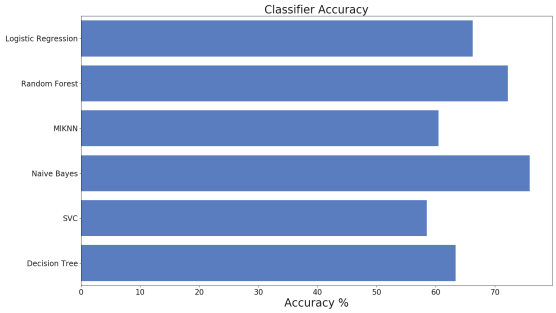

Fig. 7. Classifier accuracy for SVC (kernel=linear), LogisticRegression, MlKNN, RandomForest, DecisionTree, GaussianNB paired up with the proposed co-training classifier.

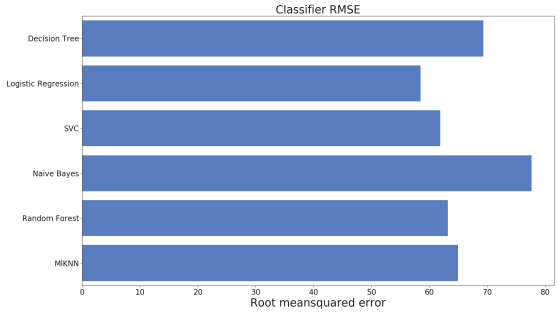

Fig. 8. Classifier RMSE for SVC(kernel=linear), Regression, MlKNN, RandomForest, DecisionTree, GaussianNB paired with the proposed co-training classifier

V. Conclusions

This paper extends the efficiency of the co-training algorithm by introducing matrix factorization to identify underlying features between stories and labels along with having a recommender system to predict top *n* similar stories for a story. Although both these algorithms have been providing efficient results independently, the experimental results in this paper, demonstrate the novelty of combining the advantages of the two algorithms - similarity and continuous learning by showing an improvement in classifying the stories and creating new labels without additional human support. This approach opens up new opportunities for retrieving similar patterns and information thus assisting in labeling the large unlabeled and unstructured data that are being collected and deposited on the Internet. This could also help reduce the input required and time from experts in a domain to manually label each story, thus help alleviate the problem of wrongly categorizing victim stories which has consequences for the rehabilitation and recovery of a victim. The proposed algorithm is currently evaluated in the human rights domain, however since collaborative filtering is domain independent, we believe that this approach has wider applicability and will provide the means to further improve the state-of-the-art recommendation techniques in other domains like gene recognition, cancer diagnostics, neuro-science, and finance.

Acknowledgement The authors would like to acknowledge the support of the Business and Local Government Data Research Centre (grant number ES/L011859/1) funded by the Economic and Social Research Council (ESRC) for undertaking this work.

References

[1] A. Blum and T. Mitchell, "Combining labeled and unlabeled data with co-training," in *Proceedings of the eleventh annual*

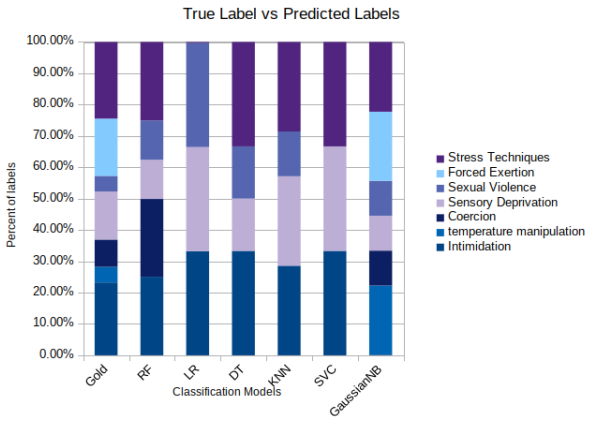

Fig. 9. True values(Ground truth) comparison with predicted values

conference on Computational learning theory. ACM, 1998, pp. 92–100.

[2] S. Abney, "Bootstrapping," in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ser. ACL '02. Stroudsburg, PA, USA: Association for Computational Linguistics, 2002, pp. 360–367. [Online]. Available: https://doi.org/10.3115/1073083.1073143

[3] A. Sarkar, "Applying co-training methods to statistical parsing," in *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies*. Association for Computational Linguistics, 2001, pp. 1–8.

[4] N. Aizenberg, Y. Koren, and O. Somekh, "Build your own music recommender by modeling internet radio streams," in *Proceedings of the 21st international conference on World Wide Web*. ACM, 2012, pp. 1–10.

[5] S. Clark, J. R. Curran, and M. Osborne, "Bootstrapping pos taggers using unlabelled data," in *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*. Association for Computational Linguistics, 2003, pp. 49–55.

[6] T. Gonçalves and P. Quaresma, "A preliminary approach to the multilabel classification problem of portuguese juridical documents," in *Portuguese Conference on Artificial Intelligence*. Springer, 2003, pp. 435–444.

[7] T. Joachims, "Text categorization with support vector machines: Learning with many relevant features," in *European conference on machine learning*. Springer, 1998, pp. 137–142.

[8] X. Luo and A. N. Zincir-Heywood, "Evaluation of two systems on multi-class multi-label document classification," in *International Symposium on Methodologies for Intelligent Systems*. Springer, 2005, pp. 161–169.

[9] R. McDonald, K. Crammer, and F. Pereira, "Flexible text segmentation with structured multilabel classification," in *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2005, pp. 987–994.

[10] A. Clare and R. D. King, "Knowledge discovery in multi-label phenotype data," in *European Conference on Principles of Data Mining and Knowledge Discovery*. Springer, 2001, pp. 42–53.

[11] M.-L. Zhang and Z.-H. Zhou, "A k-nearest neighbor based algorithm for multi-label classification," in *Granular Computing, 2005 IEEE International Conference on*, vol. 2. IEEE, 2005, pp. 718–721.

[12] A. Elisseeff and J. Weston, "A kernel method for multi-labelled classification," in *Advances in neural information processing systems*, 2002, pp. 681–687.

[13] A. Karalic and V. Pirnat, "Significance level based multiple tree classification," *Informatica*, vol. 15, no. 5, p. 12, 1991.

[14] M. Boutell, X. Shen, J. Luo, and C. Brown, "Multi-label semantic scene classification," Citeseer, Tech. Rep., 2003.

[15] X. Shen, M. Boutell, J. Luo, and C. Brown, "Multilabel machine learning and its application to semantic scene classification," in *Storage and Retrieval Methods and Applications for Multimedia 2004*, vol. 5307. International Society for Optics and Photonics, 2003, pp. 188–200.

[16] Y. Koren, R. Bell, and C. Volinsky, "Matrix factorization techniques for recommender systems," *Computer*, vol. 42, no. 8, pp. 30–37, Aug 2009.

[17] J. McAuley, J. Leskovec, and D. Jurafsky, "Learning attitudes and attributes from multi-aspect reviews," in *Data Mining (ICDM), 2012 IEEE 12th International Conference on*. IEEE, 2012, pp. 1020–1025.

[18] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl, "Incremental singular value decomposition algorithms for highly scalable recommender systems," in *Fifth International Conference on Computer and Information Science*. Citeseer, 2002, pp. 27–28.

[19] K. Yu, S. Zhu, J. Lafferty, and Y. Gong, "Fast nonparametric matrix factorization for large-scale collaborative filtering," in *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2009, pp. 211–218.

[20] J. L. Herlocker, J. A. Konstan, L. G. Terveen, and J. T. Riedl, "Evaluating collaborative filtering recommender systems," vol. 22, no. 1, pp. 5–53, 2004.

[21] Y. Zhou, D. Wilkinson, R. Schreiber, and R. Pan, "Large-scale parallel collaborative filtering for the netflix prize," in *International Conference on Algorithmic Applications in Management*. Springer, 2008, pp. 337–348.

[22] J. Mangalindan, "AmazonâĂŹs recommendation secret," *CNN Money http://tech. fortune. cnn. com/2012/07/30/amazon-5*, 2012.

[23] J. McAuley, J. Leskovec, and D. Jurafsky, "Learning attitudes and attributes from multi-aspect reviews," in *Data Mining (ICDM), 2012 IEEE 12th International Conference on*. IEEE, 2012, pp. 1020–1025.

[24] T. Brudholm, "Hate crimes and human rights violations," *Journal of Applied Philosophy*, vol. 32, no. 1, pp. 82–97, 2015.

[25] C. G. Weeramantry, *Human rights and scientific and technological development*. United Nations University Press, 1990.

[26] M. Ptaszynski, P. Dybala, T. Matsuba, F. Masui, R. Rzepka, and K. Araki, "Machine learning and affect analysis against cyber-bullying," *the 36th AISB*, pp. 7–16, 2010.

[27] R. Van Meteren and M. Van Someren, "Using content-based filtering for recommendation," in *Proceedings of the Machine Learning in the New Information Age: MLnet/ECML2000 Workshop*, 2000, pp. 47–56.

[28] P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, and J. Riedl, "Grouplens: an open architecture for collaborative filtering of netnews," in *Proceedings of the 1994 ACM conference on Computer supported cooperative work*. ACM, 1994, pp. 175–186.

[29] R. Gokhale and M. Fasli, "Deploying a co-training algorithm to classify human-rights abuses," in *2017 International Conference on the Frontiers and Advances in Data Science (FADS)*, Oct 2017, pp. 108–113.

[30] J. Davis and M. Goadrich, "The relationship between precision-recall and roc curves," in *Proceedings of the 23rd international conference on Machine learning*. ACM, 2006, pp. 233–240.

[31] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International journal of computer vision*, vol. 88, no. 2, pp. 303–338, 2010.