Learning Multi-Label Classification from Data Annotated with Unique Labels

Gargi Roy, Lipika Dey

roy.gargi@tcs.com, lipika.dey@tcs.com

TCS Research, India

My pre-Ph.D work

A. Problem Definition

Classification is a popular analytical technique to mine business insights from consumer generated texts like support emails, customer complaints etc.

1. Challenges

- Most manual annotation is single labeled and often noisy - text containing multiple issues usually gets single label based on the most important issue or the one occurring first while ignoring the others
- Noisy, skewed text with overlapping classes
- Need for explainable techniques for traceability of a decision

2. Example: *"Dear Sir I m buy a new (mobile phone name) on (date). on the box (service provider (SP) name) free data offer and i used already a (SP name) GSM sim (sim number) and i use this sim in the (phone name) but data offer 500mb/month for 6 month not activate on my (SP name) no. and i call (SP name) customer care they don't answer my problem. and i go to nearest (SP name) store they not listen my problem properly. so i m kindly request you plz solved my problem soon."*- Annotation: **Internet**, Actual issues present: **Sim, Internet, Customer Care**

B. Solution Methodologies

1. Text pre-processing: Noise cleaning, stop word removal, stemming

2. Supervised Term weighting: (i) Word's *class-discriminating power* using Inverse Gravity Moment and (ii). Word's *class-representative power (CRP)*.

$$w^d(t_k) = (CRP)\cdot\left(1+\lambda\cdot\left(\frac{f_{k1}}{\sum_{r=1}^{p} f_{kr}\cdot r}\right)\right), CRP = \left(\frac{t_k^d}{\#terms\ in\ d}\right), log(t_k^d+1), \sqrt{t_k^d} \quad (1)$$

3. Proposed Classification Method: First, a class membership distribution $\mathbf{X}$ is generated then further analysis finds significant classes in $\mathbf{X}$ with confidence

ALGORITHM 1: ComputeClassMembershipDistribution(D, T, d)

Input : D, T, d
Output: $\{o_1, o_2, ...o_p\}$ for d

Compute $\mathbf{Y_{p\times n}}$ from D, $\widehat{y}_{ij} \leftarrow v_{ij}/\sum_{j=1}^{n} v_{ij}$ where $v_{ij} \leftarrow \Sigma_{document\ d' \in D\ has\ label\ i}(w(t_j^{d'}))$;
if *if imbalanced data* **then**
  $\widehat{y}_{ij} = \widehat{y}_{ij}/max(\widehat{y}_{ij=1\ to\ n})$
Compute $\mathbf{Z_{p\times n}}$ from D where $\widehat{z}_{ij} = \widehat{y}_{ij}/\sum_{i=1}^{p}\widehat{y}_{ij}$;
Calculate term weight vector $\mathbf{N_{1\times m}}$ from d;
for d compute membership value for each class in matrix $\mathbf{O_{p\times 1}}$ where $\mathbf{O}_{p\times 1} = \mathbf{Z}_{p\times m}\mathbf{N^T}_{m\times 1}$;
Normalize class membership values, $o_i(d)^{updated} = o_i(d)/\sum_{i=1}^{p} o_i(d)$

- $x_\mu, \bar{x}, \sigma^2, \gamma, \kappa$: maximum value, mean, variance, skewness and kurtosis of $\mathbf{X}$
- $\dot{x}_i = \frac{x_i - \bar{x}}{x_i}\cdot 100$
- $\psi(x_i) = \frac{\dot{x}_i}{100}\cdot p\cdot\sigma^2\cdot|\gamma+\kappa|$
- $\mathbf{X}_i^\eta = \{y_i\} such\ that\ y_i \in \mathbf{X}\ and\ y_i \in (x_i, \frac{x_i\cdot(100-\eta)}{100})$

First four categorization is done for $|\sum_{i=1}^{p}\psi(x_i)| > (0+\rho), x_i \in \mathbf{X}$.

1. *Single Label with Very High confidence (SLVH)*
$$((\dot{x}_\mu > \alpha)\wedge(\psi(\dot{x}_\mu) > 0))\wedge((|\mathbf{X}_\mu^\eta| = 0)\vee((|\mathbf{X}_\mu^\eta| > 1)\wedge((\nexists x_i \in \mathbf{X}_\mu^\eta)\wedge(\dot{x}_i > \alpha)))) \quad (2)$$
2. *Multi-Label with High confidence (MLH)*
$$((\dot{x}_\mu > \alpha)\wedge(\psi(\dot{x}_\mu) > 0)\wedge(|\mathbf{X}_\mu^\eta| > 1))\wedge(\exists x_i \in \mathbf{X}_\mu^\eta\wedge(\dot{x}_i > \alpha)\wedge(\psi(\dot{x}_i) > 0)) \quad (3)$$
3. *Single Label with Medium confidence (SLM)*
$$((\alpha \geq \dot{x}_\mu > \beta)\wedge(\psi(\dot{x}_\mu) > 0))\wedge((|\mathbf{X}_\mu^\eta| = 0)\vee((|\mathbf{X}_\mu^\eta| > 1)\wedge(\nexists x_i \in \mathbf{X}_\mu^\eta\wedge(\alpha \geq \dot{x}_i > \beta)))) \quad (4)$$
4. *Multi-Label with Medium confidence (MLM)*
$$((\alpha \geq \dot{x}_\mu > \beta)\wedge(\psi(\dot{x}_\mu) > 0)\wedge(|\mathbf{X}_\mu^\eta| > 1))\wedge(\exists x_i \in \mathbf{X}_\mu^\eta\wedge(\alpha \geq \dot{x}_i > \beta)\wedge(\psi(\dot{x}_i) > 0)) \quad (5)$$
5. *Reject Classification for LOW Confidence (RCLC)*
$$(\beta \geq \dot{x}_\mu)\wedge(|\sum_{i=1}^{p}\psi(x_i)| \approx 0, \forall x_i \in \mathbf{X}) \quad (6)$$

After the final label set determination, the confidence score is computed and normalized.

$$s = (Avg(\psi(x_i), \forall x_i \in output\ label\ set) * (|\sum_{i=1}^{p}\psi(x_i)|, \forall x_i \in \mathbf{X}) \quad (7)$$

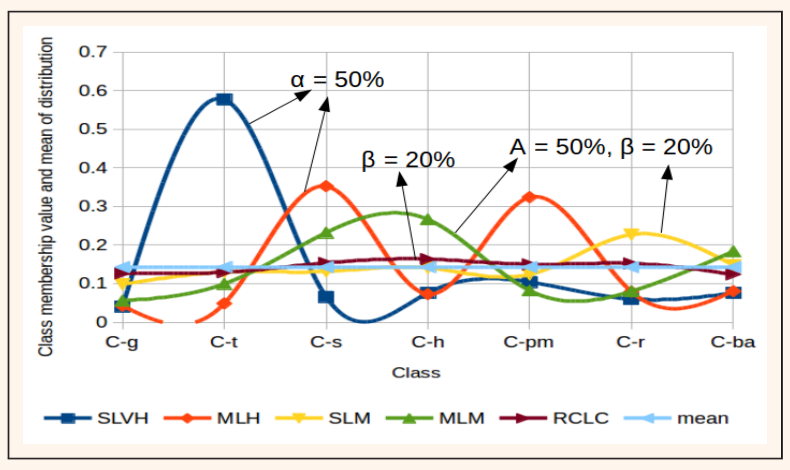
C. Results and conclusion

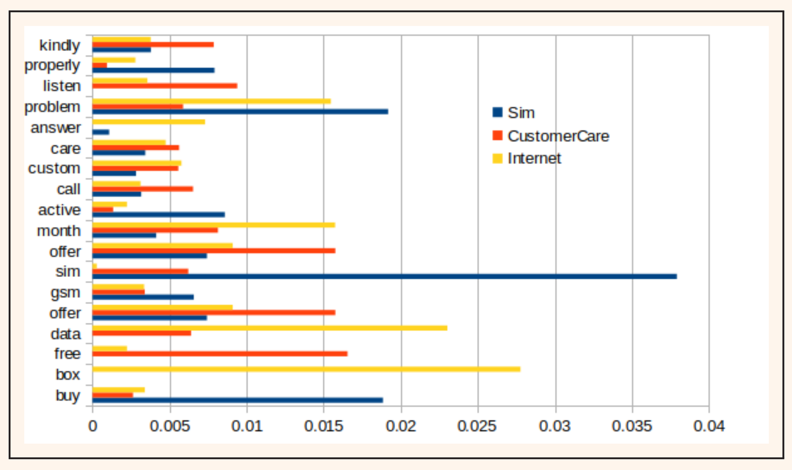
1. Evaluation scheme:

- No available baseline for unstructured data
- Standard performance measures computed using first label in ten fold cross validation setup
- Performance measures updated using 2nd, 3rd label
- Comparison of classification performance using first label with state-of-the-art
- Manual inspection and label similarity analysis for multi-label output
- Method is extended for structured data results compared with existing baseline

Classifier	Macro F1	Accuracy
Naive Bayes	83	
Rocchio	78.6	
K-NN	81.2	
SVM	78.19	
L Square	83.05	
SVM (CS&T)	82.4	
LR (CS&T)	81.5	
RSV-NN	83	
GE1-MNB	63	
MaxEnt	79	
LSTM		82
LM-LSTM		84.7
SA-LSTM		84.4
SC-LSTM-P		82.98
CNN2		80.19
Our	**84.7**	**84.87**

Table 1: Performance on 20 Newsgroup data.

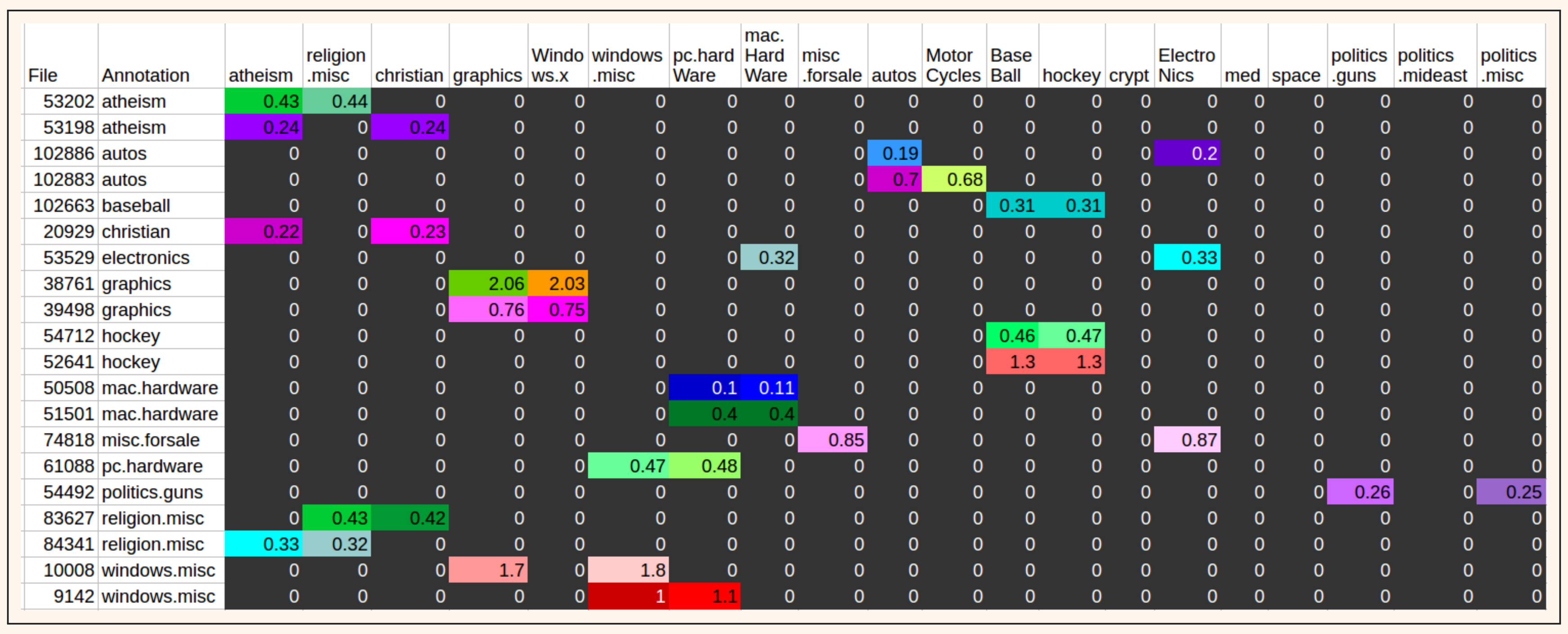




Left Figure 1: Sample distributions from several confidence catagories;
Right Figure 2: Prediction (Sim, Internet, CustomerCare) interpretation of the example

Dataset	Performance measures	Term weighting scheme NTF	LTF-IGM	NTF-IGM	RTF-IGM
20 News (Full)	Macro F1	82.99	84.1	84.1	84.2 (84.7)
	Accuracy	83.53	84.49	84.49	**84.56 (84.87)**
ScienceNews	Macro F1	95.59	96.82	96.67	96.79
	Accuracy	95.6	**96.82**	96.67	96.8
DisjointNews	Macro F1	97.35	98.29	98.32	98.28
	Accuracy	97.35	98.3	**98.32**	98.28
CompScNews	Macro F1	83.46	85.73	86.27 (86.3)	85.68
	Accuracy	83.58	85.8	**86.32 (86.35)**	85.74
HR (internal)	Macro F1	80.37	85.15	84.47 (85.27)	85.08
	Accuracy	82.21	84.86	**84.89 (85.61)**	84.56
Telecom	Macro F1	60.35	61.02	63.1 (64.1)	60.4
	Accuracy	69.6	64	**70.4 (71.26)**	69
IMDB	Macro F1	86.06	87.63	87.89	87.62
	Accuracy	86.07	87.64	**87.9**	87.62
RT	Macro F1	75.34	79.03	78.85	79.05
	Accuracy	75.35	79.04	78.87	**79.06**

Table 2: Prediction performances using first label for text datasets. Model training for imbalanced data is used for HR (internal) data.



File	Annotation	atheism	religion .misc	christian	graphics	Windo ws.x	windows .misc	pc.hard Ware	mac. Hard Ware	misc .forsale	autos	Motor Cycles	Base Ball	hockey	crypt	Electro Nics	med	space	politics .guns	politics .mideast	politics .misc
53202	atheism	0.43	0.44	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
53198	atheism	0.24	0	0.24	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
102886	autos	0	0	0	0	0	0	0	0	0	0.19	0	0	0	0	0.2	0	0	0	0	0
102883	autos	0	0	0	0	0	0	0	0	0	0.7	0.68	0	0	0	0	0	0	0	0	0
102663	baseball	0	0	0	0	0	0	0	0	0	0	0	0.31	0.31	0	0	0	0	0	0	0
20929	christian	0.22	0	0.23	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
53529	electronics	0	0	0	0	0	0	0	0.32	0	0	0	0	0	0	0.33	0	0	0	0	0
38761	graphics	0	0	0	2.06	2.03	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
39498	graphics	0	0	0	0.76	0.75	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
54712	hockey	0	0	0	0	0	0	0	0	0	0	0	0.46	0.47	0	0	0	0	0	0	0
52641	hockey	0	0	0	0	0	0	0	0	0	0	0	1.3	1.3	0	0	0	0	0	0	0
50508	mac.hardware	0	0	0	0	0	0	0.1	0.11	0	0	0	0	0	0	0	0	0	0	0	0
51501	mac.hardware	0	0	0	0	0	0	0.4	0.4	0	0	0	0	0	0	0	0	0	0	0	0
74818	misc.forsale	0	0	0	0	0	0	0	0	0.85	0	0	0	0	0	0.87	0	0	0	0	0
61088	pc.hardware	0	0	0	0	0	0.47	0.48	0	0	0	0	0	0	0	0	0	0	0	0	0
54492	politics.guns	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.26	0	0.25
83627	religion.misc	0	0.43	0.42	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
84341	religion.misc	0.33	0.32	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10008	windows.misc	0	0	0	1.7	0	1.8	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9142	windows.misc	0	0	0	0	0	1	1.1	0	0	0	0	0	0	0	0	0	0	0	0	0

Figure 3: Few instances with multi-label output from 20 Newsgroups dataset.

2. Comparitive study for structured data:

- Method extended for structured data - data scaling, data standardization, binarization of the categorical features, missing value handling, mtual information for feature selection
- Result compared on UCI datasets with fuzzy rule induction technique of KNIME
- Similarities are seen in results