Handling Imbalanced Dataset in Multi-label Text Categorization using Bagging and Adaptive Boosting

Prepared by

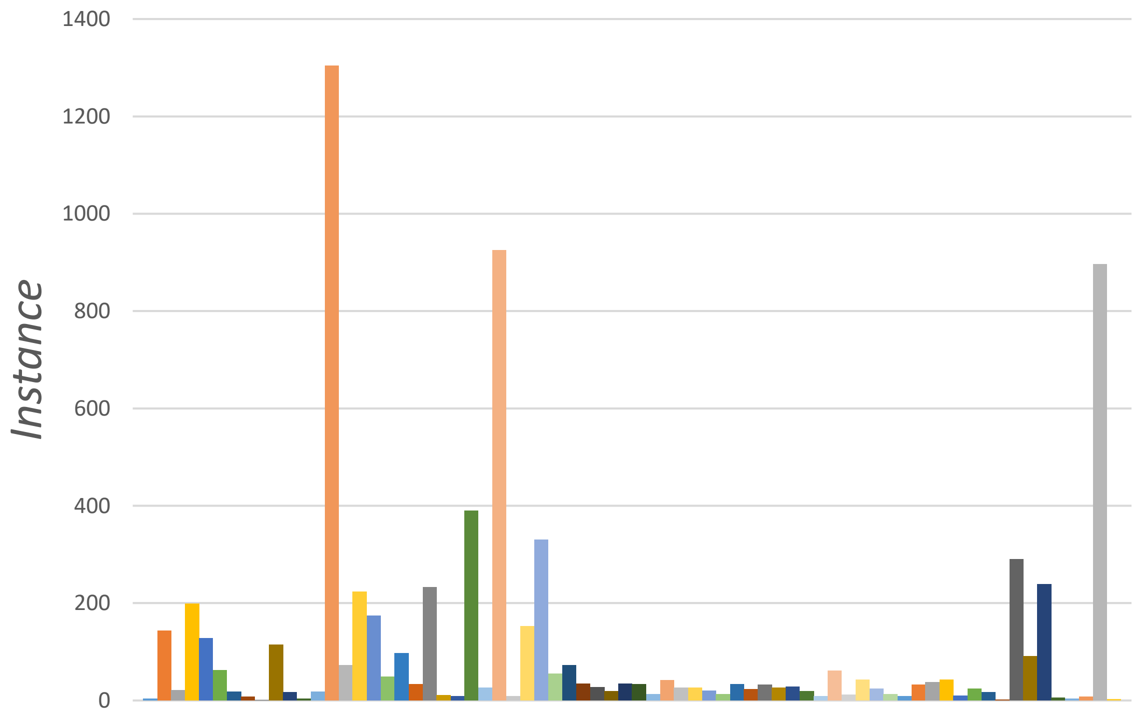
Genta Indra Winata

Masayu Leylia Khodra

Overview

Background → Related work → Objectives → Methods → Result → Conclusion

Background



LAYANAN ASPIRASI DAN PENGADUAN ONLINE RAKYAT

President's Delivery Unit for Development Monitoring and Oversight Indonesia

1. Imbalanced dataset distribution

classifier tends to be **weighted down** by the **majority of the data** and **ignore** the **minority**

Few research in **multi-label text categorization** which handles this issue

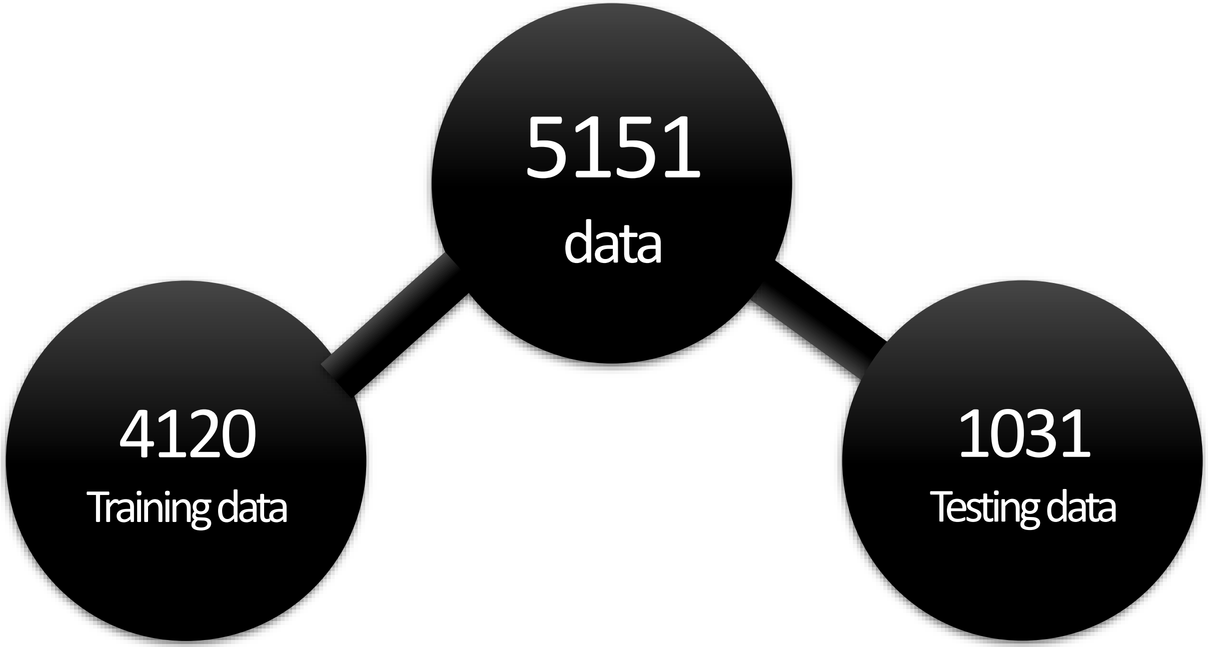
2. Lack of Density

70 labels

21% minority label

Many abbreviations and informal words found

Dataset



LAPOR!
LAYANAN ASPIRASI DAN PENGADUAN ONLINE RAKYAT

Each instance comprises features such as *id*, *complaint text*, complaint *topic* and *a set of label*.

Example

Text

"Di stasiun KA Kiaracondong (**jln** Kiaracondong antara kebaktian **s d** kantor Polisi kebon jayanti) ada **gepeng** anak usia sekolah **ngelem**, bicaranya kasar **tdk** **karuan**, suka ganggu penumpang wanita! **Trm ksh**!"

"In KA Kiaracondong station (Kiaracondong street between kebaktian and kebon jayanti police station), there were homeless students who did glue sniffing, spoke harshly and harassed women. Thanks"

Label Target

Dinas Sosial (Dinsos) Kota Bandung

Satuan Polisi Pamong Praja (Satpol PP) Kota Bandung

Related work

Previous research

Fauzan and M. L. Khodra, "Automatic Multilabel Categorization using Learning to Rank Framework for Complaint Text on Bandung Government," in *International Conference of Advanced Informatics: Concept, Theory and Application (ICAICTA)*, Bandung, 2014.

Use **LAPOR dataset (consists of complaints data)**

Best performer : Label PowerSet (LP) with SMO weak classifier

Problem : Did not handle imbalanced dataset.

Related work (2)

Previous research

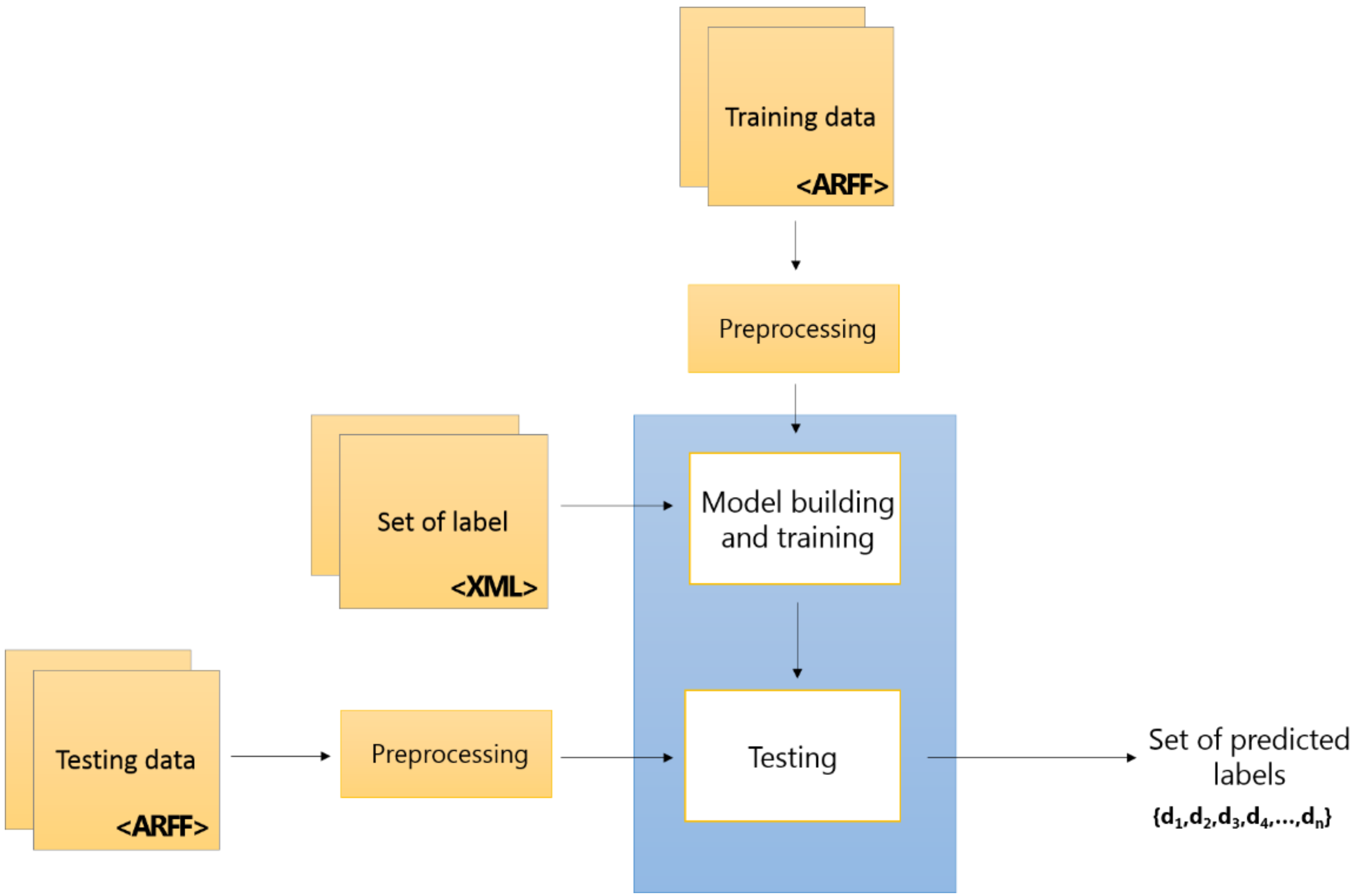
Syaripudin and Khodra states adaptive boosting is able to handle imbalanced dataset, especially for single-label categorization.[1]

[1] A. Syaripudin and M. L. Khodra, "A Comparison for Handling Imbalanced Datasets," in *International Conference of Advanced Informatics: Concept, Theory and Application (ICAICTA)*, Bandung, 2014.

Objectives

1. Compares between **handling imbalanced dataset techniques** with **baseline results**
2. Improves **multi-label text classification performance**

Architecture



Methods

1. Text Processing Techniques

2. Multi-label Text Categorization

Problem transformation techniques

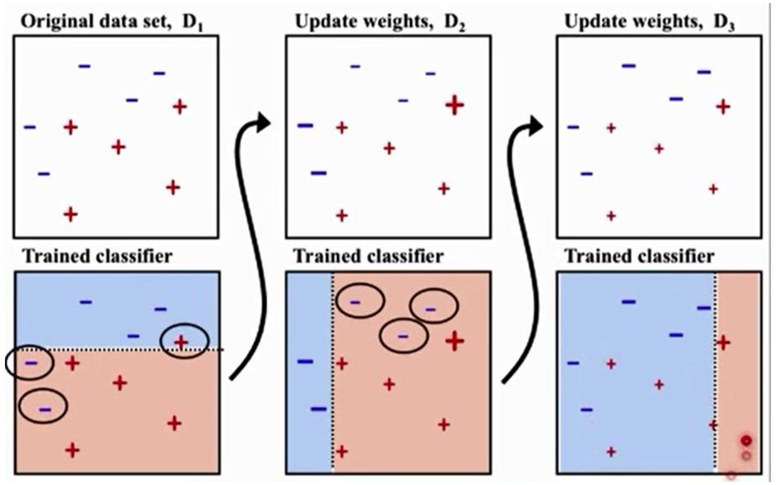
- Binary Relevance **(BR)**
- Label PowerSet **(LP)**

Methods (2)

3. Imbalanced Dataset Handling Algorithm

a. Adaptive Boosting (AdaBoost.MH)

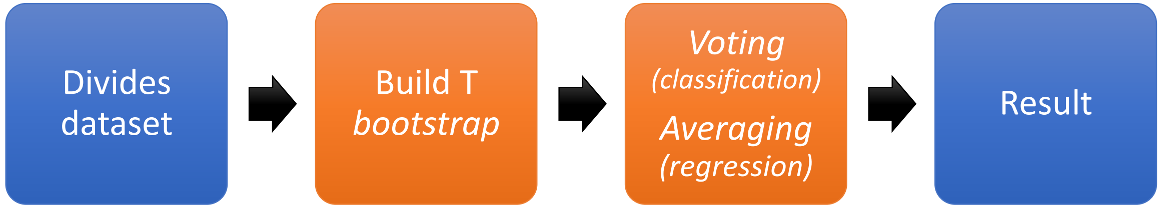
an algorithm to find highly accurate classification rule by combining many weak hypotheses[1]



Source: Ensembles: Boosting, Prof. Alexander Ihler

b. Bagging (Bagging.ML)

an algorithm for generating multiple bootstraps and use the aggregation average over all bootstraps to predict a class.[2]



[1] R. E. Schapire and Y. Singer, "BoosTexter: A Boosting-based System for Text Categorization," *Machine Learning Volume 39 Issue 2-3,* vol. 39, no. 2-3, pp. 135-168, 2000.
[2] L. Breiman, "Bagging Predictors," *Machine Learning,* vol. 24, no. 2, pp. 123-140, 1996.

Metric Evaluation

1. Hamming Loss
2. Subset Accuracy
3. Example-based Accuracy
4. Micro-averaged F-measure

$$hamming\_loss(h) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{Q}\left|h(\mathbf{x_i})\Delta\mathcal{Y}_i\right|$$

$$subset\_accuracy(h) = \frac{1}{N}\sum_{i=1}^{N} I(h(\mathbf{x_i}) = \mathcal{Y}_i)$$

$$accuracy(h) = \frac{1}{N}\sum_{i=1}^{N}\frac{\left|h(\mathbf{x_i})\cap\mathcal{Y}_i\right|}{\left|h(\mathbf{x_i})\bigcup\mathcal{Y}_i\right|}$$

$$micro\_F_1 = \frac{2 \times micro\_precision \times micro\_recall}{micro\_precision + micro\_recall}$$

$$micro\_precision = \frac{\sum_{j=1}^{Q} tp_j}{\sum_{j=1}^{Q} tp_j + \sum_{j=1}^{Q} fp_j} \quad micro\_recall = \frac{\sum_{j=1}^{Q} tp_j}{\sum_{j=1}^{Q} tp_j + \sum_{j=1}^{Q} fn_j}$$

Results

Hamming Loss

Weak classifier	Baseline		AdaBoost.MH	Bagging.ML (BR)	Bagging.ML (LP)
	BR	LP			
Decision Stump	**0.0152**	0.0247	0.0197	N/A	0.0250
J48	0.0133	0.0188	**0.0131**	0.0150	0.0170
Random Forest	**0.0132**	0.0179	0.0146	N/A	N/A
Naive Bayes	0.0420	**0.0159**	0.0327	N/A	N/A
SMO	**0.0144**	0.0148	0.0197	0.0150	0.0150

Results (2)

Subset Accuracy

Weak classifier	Baseline		AdaBoost.MH	Bagging.ML (BR)	Bagging.ML (LP)
	BR	LP			
Decision Stump	**0.3346**	0.2318	0.0039	N/A	0.2320
J48	0.4074	0.4016	**0.4277**	0.3750	0.3800
Random Forest	**0.4103**	0.3986	0.3337	N/A	N/A
Naive Bayes	0.2144	**0.4219**	0.0145	N/A	N/A
SMO	0.4200	**0.4588**	0.0039	0.4000	0.4490

Results (3)

Example-based Accuracy

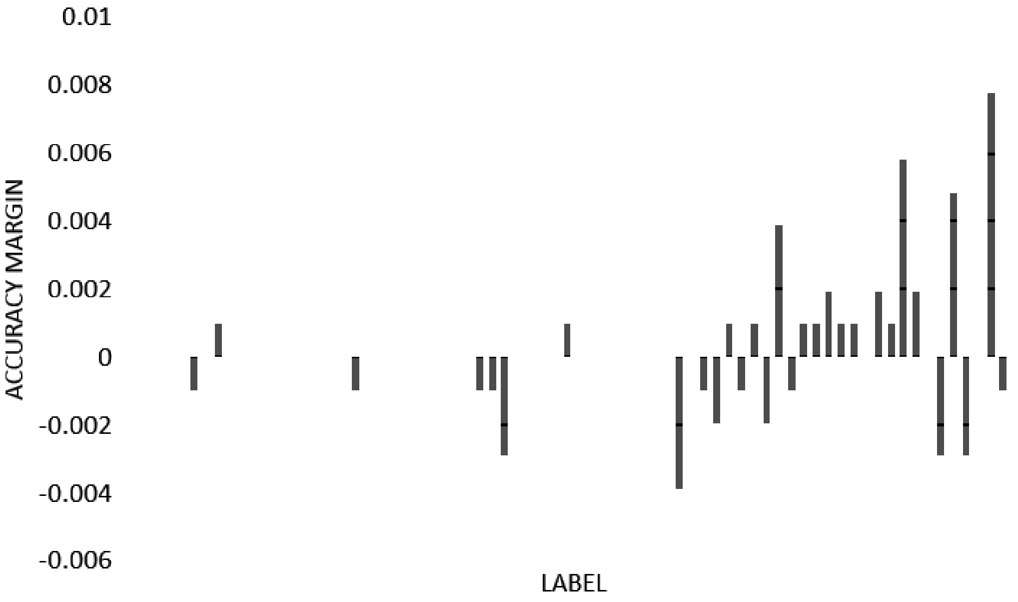
Weak classifier	Baseline BR	Baseline LP	AdaBoost.MH	Bagging.ML (BR)	Bagging.ML (LP)
Decision Stump	**0.4123**	0.2726	0.0039	N/A	0.2730
J48	0.5151	0.5034	0.5301	**0.5520**	0.5400
Random Forest	**0.5080**	0.4907	0.3847	N/A	N/A
Naive Bayes	0.4098	**0.5570**	0.0417	N/A	N/A
SMO	0.5556	0.5821	0.0039	0.5740	**0.5850**

Results (4)

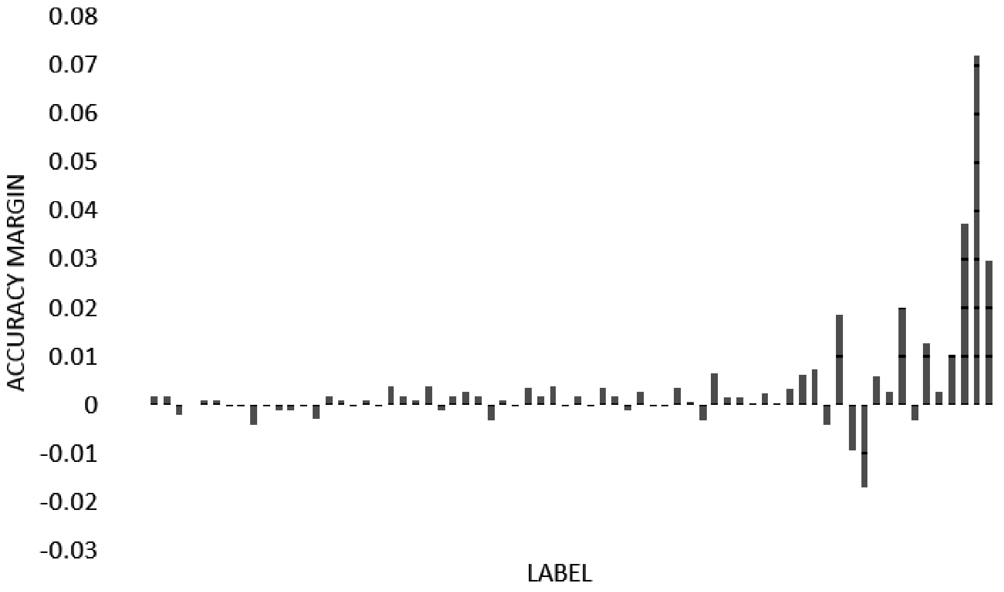
Micro-averaged F-measure

Weak classifier	Baseline		AdaBoost.MH	Bagging.ML (BR)	Bagging.ML (LP)
	BR	LP			
Decision Stump	**0.4919**	0.2716	0	N/A	0.2720
J48	0.5950	0.5055	0.6044	**0.6240**	0.5700
Random Forest	**0.5839**	0.4988	0.4704	N/A	N/A
Naive Bayes	0.3889	**0.5801**	0.1094	N/A	N/A
SMO	0.6095	0.5977	0	**0.6270**	0.6040

Results (5)



AdaBoost.MH-J48 with BR-J48



Bagging.ML-LP-SMO with LP-SMO

Results (6)

Bagging.ML-LP (SMO)	Subset accuracy 0.4490 Example-based accuracy 0.5850
AdaBoost.MH (J48)	Hamming loss 0.0131
Bagging.ML-BR (SMO)	Micro-averaged F-measure 0.6270

Conclusion

1. Handling imbalanced dataset improves **categorization performance** for particular **weak classifiers**. J48 for AdaBoost.MH and SMO for Bagging.ML.
2. AdaBoost.MH and Bagging.ML **increases majority label accuracy**. Adaptive Boosting only **increases one minority label** and bagging **boosts most of minority labels**.