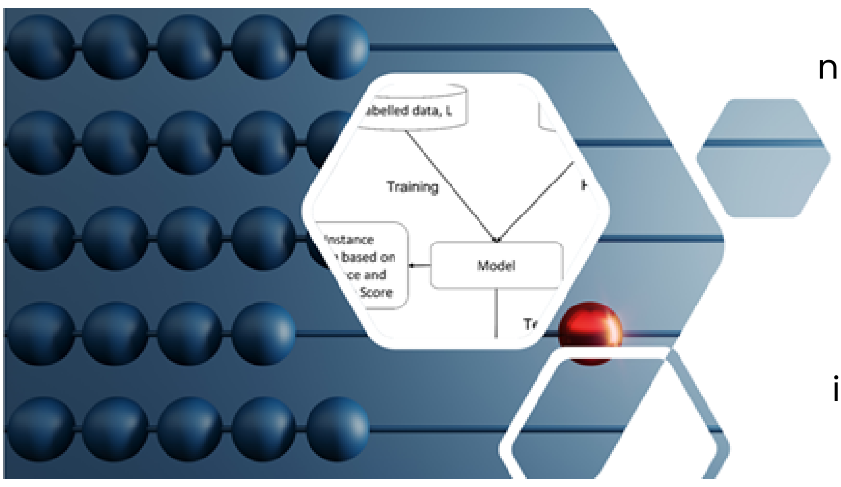
INCORPORATING INFORMATIVE SCORE FOR INSTANCE SELECTION IN SEMI-SUPERVISED SENTIMENT CLASSIFICATION



Training deep models requires large numbers of annotated texts with sentiment-related labels. Such requirement poses a challenge in scenario when early-stage sentiment analysis is required but data sources available are small as they are acquired in incremental manner. Early-stage sentiment classes prediction is important to find out the society well-being from service, product or event feedbacks.

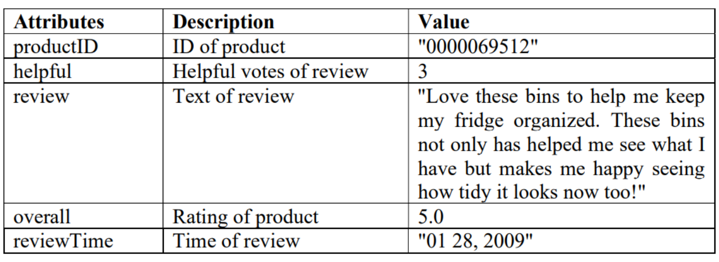
Objective

(1) Enable good sentiment classification with a limited amount of annotated data by investigating hybrid of semi-supervised learning and deep learning techniques.

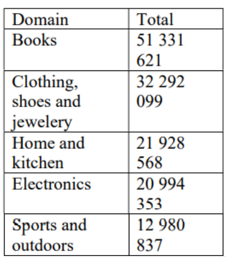
(2) incorporate sentiment-based quality criteria to improve the selection of labelled/unlabeled data for training deep learning model.

(3) maintain the stability of classification model by determining optimal ratio of labelled/unlabelled and optimal parameters of the criteria for the classifier

Method Overview



Attributes	Description	Value
productID	ID of product	"0000069512"
helpful	Helpful votes of review	3
review	Text of review	"Love these bins to help me keep my fridge organized. These bins not only has helped me see what I have but makes me happy seeing how tidy it looks now too!"
overall	Rating of product	5.0
reviewTime	Time of review	"01 28, 2009"



Domain	Total
Books	51 331 621
Clothing, shoes and jewelery	32 292 099
Home and kitchen	21 928 568
Electronics	20 994 353
Sports and outdoors	12 980 837

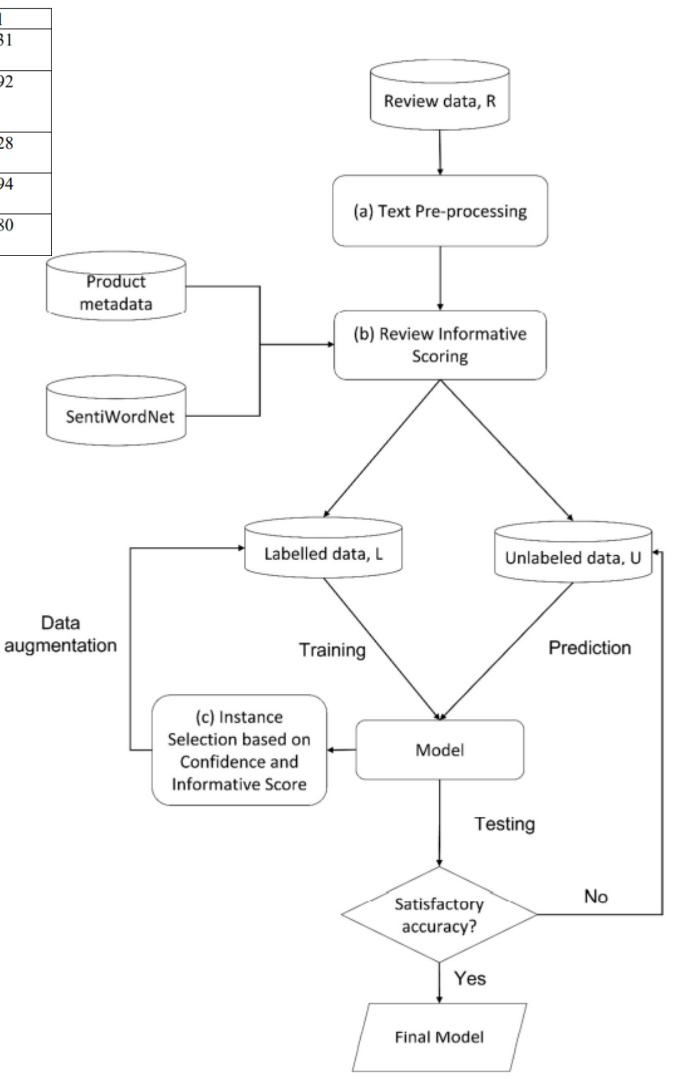
Sample Review Data & Size

$$S(I) = w_1 S(C) + w_2 S(P), where\ w_1 + w_2 = 1$$

Domain Relevancy Equation

$$x_1 confidence + x_2 S(I), where\ x_1 + x_2 = 1$$

Instance Selection Equation



Source: https://www.mstsolutions.com/technical/sentiment-analysis/

The framework of Semi-supervised Learning with Domain Relevancy Factor

Key Findings

1 87.26% SV

78.35% SSV

82.51% SSV+DR

CNN Model on Book Domain

With only **40 percent** of the labeled data used in training, the performance of semi-supervised models (SSV+DR) is **comparable** to the supervised models (SV) that use fully labeled data.

2 40:60 86.78%

CNN Model

40 percent labeled data with 60 percent unlabeled is the best split .

3 0.5cf:0.5df 89.52%

RNN Model

0.5 confidence factor + 0.5 domain factor gives the best performance.

Conclusion

Automation of the data annotation with a small amount of labeled data Reduction of dependency on supervised models. Enabling harnessing of deep learning model with the labeled data.

Dr. Gan Keng Hoon, Dr. Tan Tien Ping & Prof. Dr. Rosni Abdullah
School of Computer Sciences

FRGS/1/2018/ICT02/USM/02/9