Faculty of Science
SRINAKHARINWIROT UNIVERSITY

COM13

Predictive text on social media to detect depression

Krittanut Sangsri, Ritaporn Thammawat and Werayuth Charoenruengkit
Bachelor of Science program in Computer Science

Abstract
Nowadays the people use social media for communication and convey feelings that can't speak openly in real life. This project uses the data from social media to analyze several factors associated with depression such as time, words that show signs of depression, sentiment word and emoji. This project uses pre-train models to create 300 - dimensional vectors from text and language indication to develop models to identify the sign of depression. This research retrieves Thai sentences from Twitter 21,905 sentences to create models. Several exploratory data analysis are presented to depict the characteristics of language of the dataset. The models are evaluated by 30% from the dataset. The best result of a depression detection is at 94% accuracy.

Background
The statistical data of the World Health Organization in 2015 indicate that patients with depression are about 4.4% from the global population. which is the main cause is chemicals in the brain or another cause for example, disappointment and loss. Nowadays, people can access the internet more and use social media for communication and convey feelings that can't speak openly in real life.

Methodology

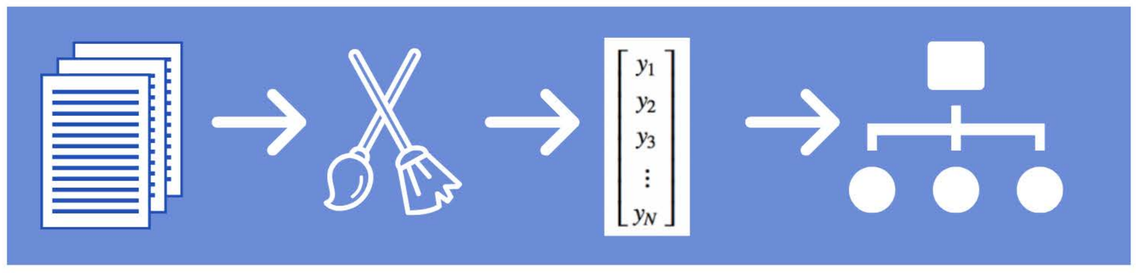
1 Thai language : scarped Twitter dataset (21,905 rows)
- Depress 10,232 rows
- Non-depress 11,673 rows

English Data : Public dataset (6,622 rows)
- Depress 3,967rows
- Non-depress 2,655rows

2 Exploratory data analysis such as word cloud and check type of data for looking for signs of depression that relative with related research. Clean data by remove emoji, hashtag, link, special characters and mention .

3 Word embedding by using FastText and Word2vec.Then split the dataset into Train 70% test 30%.

4 Using data for learning with three classification algorithms, which is Support vector machine, Logistic regression and Random forest.



Knowledge
Word Embedding is an open-source, free and lightweight library that allows users to learn text representations and text classifiers.
- FastText
- Word2Vec

Classification technique

Logistic regression(LR) is a fundamental classification technique. It belongs to the group of linear classifiers and somewhat similar to polynomial and linear regression.

Support Vector Machine(SVM) is a linear model for classification and regression problems. The algorithm creates a line or a hyperplane which separates the data into classes.

Random Forest(RF) is a supervised learning algorithm. The forest it builds, is an ensemble of decision trees, usually trained with the bagging method. The general idea of the bagging method is a combination of learning models increases the overall result.

Experiment result

Figure1 : Word Cloud monogram Non-depression

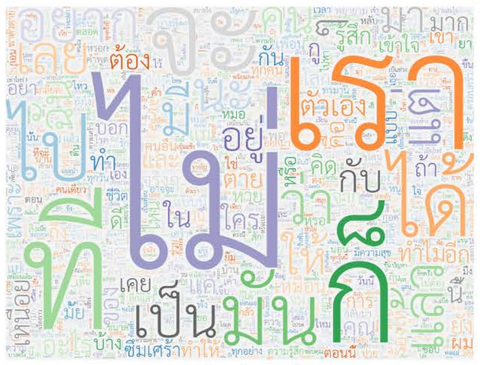


Figure2: Word Cloud monogram Depression

English Language

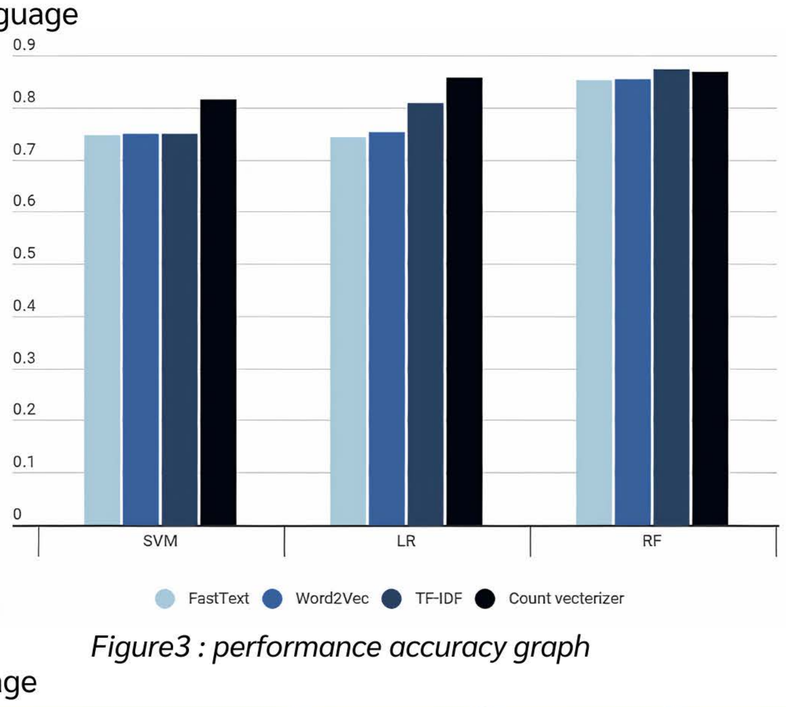


Figure3 : performance accuracy graph

Thai Language

Algorithm	Accuracy	Precision	Recall
Support Vector Machine	0.939	0.956	0.928
Logistic Regression	0.915	0.932	0.907
Random Forest	0.902	0.910	0.905

Figure4 : performance accuracy table

Conclusion
Summary of performance from research and develop the prediction message on social media for detect sentences indicate signs of depression. Using various algorithms for learning and evaluate the prediction. The best result in English dataset is add features such as positive word, negative word and words that show signs of depression etc. and use TF-IDF as a word embedding then use random forest as classification algorithm that get accuracy is 87.3%.

In Thai dataset have used FastText as word embedding and learning with various algorithms which is the best result of accuracy is 94% by using svm as classification algorithm.

Reference
[1] T. Mikolov, G. Corrado, K. Chen, and J. Dean, "Vector Space," pp. 1–12.'
[2] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching Word Vectors with Subword Information," 1996.
[3] R. N. Ramirez-Esparza, C. K. Chung, E. Kacewicz, and J. W. Pennebaker, "The psychology of word use in depression forums in english and in spanish: Testing two text analytic approaches," ICWSM 2008 - Proc. 2nd Int. Conf.
[4] จันทิมา อังคพณิชกิจ. (2563). ภาษาการสื่อสารและโรคซึมเศร้า การสำรวจเบื้องต้นเพื่อเข้าใจโรคซึมเศร้าและผู้มีภาวะซึมเศร้าในสังคมไทย.กรุงเทพฯ:อมรินทร์ พริ้นติ้ง.