* Description
* This report is written for the second project of the subject knowledge technology.
* The purpose of this project is to analysis the sentiment of the tweets using different data mining approaches and acquire knowledge from the results. After the feature engineering, four machine learning approaches are applied in this project including Naïve Bayes, Decision Tree, Random Forest, and KNN. Then, a comparison of the efficiency for each method will be present with the limitation of the project. Finally, the acquired knowledge from analysis will be shown.
* Background
  + Data set
    - The tweet collections used in this project is a sub-sample gathered from Twitter and own by a research group (Rosenthal, 2017). There are around 23000 data points in the train.txt which is the training data for the models, and around 5000 data points in the dev.txt which are used to evaluate the models.
  + Weka
    - Weka is used in the project, which is a commonly used tool providing a number of machine learning algorithms.
  + Classifiers
    - In this project, four classifiers are chosen to compare with each other. The four classifiers are Naïve Bayes, Decision Tree, Random Forest, and KNN. The classifiers are chosen from the ones introduced in the subject due to the interpretability.
* Feature engineering
  + Twitter features
  + However, the data is still not only a plain text message, but also contains features other than words which may contribute to its sentiment.
  + Negation words
    - The present of the negation word will change the opinion of the tweet entirely. For example, not good means bad. But the negation words could various representations, such as not, didn’t, doesn’t, which means that for each of the negation words, the frequency could be relatively low, and to capture all of them may increase the dimension of the data significantly.
    - Another property of the negations words is that the existence of double negations. For example, “I couldn’t not help her” actually means “I felt I should help her”.
    - To solve the problems mentioned above, a feature named “NEGATIONWORD” is constructed. First, the system was given a collection of negation words. Then, instead of using the number of appearance of the negation words in the tweet, the modulo operation was applied to handle the double negation. The attribute NEGATIONWORD is calculated out of the following formula.
    - NEGATIONWORD(t) = c mode 2
    - c indicating the number of negative words in the tweet.
    - Emoticons
    - Emoji and other symbol expressions such as “:)”, “:(”, are widely used now to represent emotions. A good news for sentiment analysis is that some of tend to indicate obvious positive or negative emotions and the negative words will not be applied on the emoji and facial expressions, which means that these emoticons usually express strong and clear opinions than normal words.
  + Preprocessing

The tweets used in the project has been preprocessed to remove less informative contents such as author, time stamp, etc. The preprocessing is still needed for the project.

In the preprocessing stage, the URLs and the mentions is the tweets will be removed. And most of the stop words such as ‘the’, ‘a’ will also be removed due to limited information they provide. Another critical step for preprocessing is stemming. Stemming is an approach to assemble the strongly related tokens to the same type of token. For example, the present tense and the past tense of the same word usually represents the similar opinion.

* Feature selection
  + To select features that are more related to the sentiment of the tweets
* Results and effectiveness analysis
  + The results
  + Interpreting this data
    - Native Bayes
    - Tree
    - Forest
    - KNN
  + Comparing the data
  + Some knowledge

* Limitation

* Conclusion

Rosenthal, Sara, Noura Farra, and Preslav Nakov (2017). SemEval-2017 Task 4: Senti- ment Analysis in Twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval ’17)*. Vancouver, Canada.