5.2. Conclusion

It can be concluded that the best K for KNN is equal to 3 for Polarity v2.0 dataset. KNN has been be compared to another machine learning method such as NB, SVM and RF. Comparison between KNN, NB, SVM and RF without IG and with IG were done by using 10 fold Cross validation to know the performance. In this research, the more relevance features will get the better accuracy of NB and KNN. IG can helps RF and SVM to improve the accuracy only if the threshold is equal to 0.1. It show that SVM and RF only can build a good model to classify if the feature have a good relevance and the amount of features is not too much or too little. Without feature selection KNN only achieved the performance equal to 60% with a value of K=3. After using Information gain, KNN has a better performance which also the highest performance compared to other methods with 96.8% at K=3. It can be concluded that the reduction of irrelevant features has a greater effect on KNN method than other methods. Feature selection with IG improve all machine learning methods performance. It's because IG can reduce features that are less relevant to the class. However, the results of the comparison of machine learning may differ in the case of different datasets. For further research, author recommend to find the optimal threshold by a method. It's because in this paper the optimal threshold is set manually.

References

[1] Sebastiani F *Machine learning in automated text categorization* 2002 CSUR **34(1)** 1 47

[2] Jiang S, Pang G, Wu M and Kuang L *An improved K-nearest-neighbor algorithm for text categorization* 2012 Expert Systems with Applications **39(1)** 1503 09

[3] Pratiwi A I *On the feature selection and classification based on information gain for document sentiment analysis* 2018 Applied Computational Intelligence and Soft Computing

[4] Pang B, Lee L, and Vaithyanathan S *Thumbs up?: sentiment classification using machine learning techniques* 2002 July EMNLP ACL **10** 79 86

[5] Tan S and Zhang J *An empirical study of sentiment analysis for chinese documents* 2008 Expert Systems with applications **34(4)** 2622 29

[6] Basari A S H, Hussin B, Ananta I G P and Zeniarja J *Opinion mining of movie review using hybrid method of support vector machine and particle swarm optimization* 2013 Procedia Engineering **53** 453 62

[7] OKeefe T and Koprinska I *Feature selection and weighting methods in sentiment analysis* 2009 December In Proceedings of the 14th ADCS 67 74

[8] Sahami M and Koller D *Using machine learning to improve information access* 1998 (Doctoral dissertation, Stanford University, Department of Computer Science).

[9] Chaovalit P and Zhou L *Movie review mining: A comparison between supervised and unsupervised classification approaches* 2005 January HICSS 112c- 112c IEEE.

[10] Lee C and Lee G G *Information gain and divergence-based feature selection for machine learning-based text categorization* 2006 Information processing and management **42(1)** 155 65.

[11] Soucy P and Mineau G W *A simple KNN algorithm for text categorization* 2001 ICDM 2001 647 48)

[12] Pang B and Lee L *A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts* 2004 July ACL 271

[13] Yang Y and Pedersen J O *A comparative study on feature selection in text categorization* 1997 July Icml **97** 412 20

[14] Cha S H *Comprehensive survey on distance/similarity measures between probability density functions* 2007 **1(2)** 1

[15] Na J C, Sui H, Khoo C S and Zhou Y *Effectiveness of simple linguistic processing in automatic sentiment classification of product reviews* 2004

[16] Elmurngi E, and Gherbi A *An empirical study on detecting fake reviews using machine learning techniques* 2017 In International Conference on Innovative Computing Technology (pp. 107-114).

[17] Thapa L B , and Bal B K *Classifying sentiments in Nepali subjective texts In Information Intelligence, Systems & Applications* 2006 IISA.

[18] Xu Guo X and Ye Y and Cheng, J *An Improved Random Forest Classifier for Text Categorization* 2012 JCP **7(12)** 2913 2920.

[19] Gupte A, Joshi S, Gadgul P and Kadam A *Comparative study of classification algorithms used in sentiment analysis* JCSIT **5(5)** 6261 64