COMP 3503 – Project Progress Report

Axel Eschholz 100161044

March 1st, 2024

**Introduction:**

This project is centered around text mining and sentiment classification, specifically in the realm of customer reviews. The dataset in question is an unlabelled corpus of Trustpilot reviews for the company Asos. The goal of the project is to train a model to classify consumer reviews as ‘positive’, ‘negative’, or ‘neutral’. Given the raw nature of the data, text mining techniques will need to be implemented to enable effective analysis. Furthermore, an effective labelling method will need to be applied to produce a meaningful model.

**Model Choice**

The classification model specified by the project is a Decision Tree. In pursuit of implementing an effective decision tree, this project will leverage Weka’s J48 flavor of the C4.5 classification algorithm to build the decision tree. J48 was initially chosen for its relative performance compared to its ease of use and understanding. It was further justified by a recent study on the accuracy of this algorithm in the specific context of classifying e-commerce reviews.

For attribute selection, the CfsSubsetEval evaluator model was used in conjunction with the BestFirst search method. The justification for this was to find a subset of attributes highly correlated to the sentiment value, but with low intercorrelation.

**Related Works**

The 2020 article by Kasthuri & Jebaseeli presents a sentiment analysis approach for Twitter data, employing a decision tree classifier. It includes preprocessing steps like stop-word removal and feature extraction using TF-IDF to determine the importance of words within a dataset. This method aims to classify tweet sentiments as ‘positive’, ‘negative’, and ‘neutral’. The use of TF-IDF is important in identifying key features from tweets for further classification, contributing to improved accuracy in sentiment analysis by 6-20% compared to existing systems.

The 2021 article by Ikhsanti et al evaluates the accuracy of the C4.5 Decision Tree algorithm for sentiment analysis of e-commerce application reviews. The study emphasizes the efficacy of the C4.5 algorithm for sentiment analysis, specifically in categorizing user reviews into ‘positive’, ‘negative’, and ‘neutral’ sentiments. With an average accuracy of 92.63% on a very similar dataset, this article provides ample justification for the use of the C4.5 algorithm in this context.

**Work Done**

Preprocessing

The raw data has been cleaned, tokenized, and vectorized using Term Frequency – Inverse Document Frequent (TF-IDF). This translates a dataset of text blocks into a numeric dataset across 1000 attributes that can be properly fed into Weka.

Labelling

Since the raw data was unlabelled, this project labelled initial sentiment using the Affin Sentiment Lexicon. This assigned blocks of text a numerical score denoting the relative positive or negative sentiment expressed within, as defined by the manually rated lexicon. (Nielsen, 2011)

Attribute Selection

Once the preprocessed and labelled data was loaded into Weka, the evaluation and search methods were applied via a filter to select the most meaningful attributes.

**Future Work**

Now that the general structure and pipeline for this project has been developed, there are some refinements that need to happen before the model can be applied. Specifically, the labelling output needs to be properly bounded to produce a nominal class for the decision tree to be built on. This includes searching for and eliminating outliers, assigning logical boundaries for the ‘positive’, ‘neutral, and ‘negative’ classifications, then iterating through the dataset to apply these classifications. The attribute selection will still be conducted on the numerical score so as not to lose accuracy because of the loss of information.

Once the labelling is complete, the data will need to be split into training and test sets. The J48 model will be run using cross-validation on the training set and then evaluated using the test set. Based on these results the model will be refined and/or replaced to achieve the highest possible accuracy. Certain preprocessing steps or attribute selection may also need to be rerun if an accuracy plateau is reached.

**References**

Finn Årup Nielsen (2011). A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. *Proceedings of the ESWC2011 Workshop on 'Making Sense of Microposts': Big things come in small packages 718 in CEUR Workshop Proceedings* 93-98. <http://arxiv.org/abs/1103.2903>.

Ikhsanti, A. Y., Fauziah, Y., & Perwira, R. I. (2021). Implementation of the c4. 5 decision tree learning algorithm for sentiment analysis in e-commerce application reviews on google play store. *Computing and Information Processing Letters*, *1*(1), 25-30. <http://jurnal.upnyk.ac.id/index.php/cip/article/viewFile/6128/3956>

Kasthuri, S., & Jebaseeli, A. N. (2020). An efficient decision tree algorithm for analyzing the twitter sentiment analysis. *Journal of Critical Reviews*, *7*(4), 1010-1018. <https://www.jcreview.com/admin/Uploads/Files/61adfe235a6725.71379760.pdf>