FINAL PROJECT DATA MINING

TEXT CLASSIFICATION

Dosen Pengampu : Al Ustadz Oddy Virgantara Putra, S.Kom., M.T,

DISUSUN OLEH :

FATHIN MUHAMMAD WASMANSON

NIM: 372016611508

PROGRAM STUDI TEKNIK INFORMATIKA

FAKULTAS SAINS DAN TEKNOLOGI

UNIVERSITAS DARUSSALAM GONTOR

2019

1. Latar Belakang Masalah

Klasifikasi teks merupakan sebuah pemisahan data pada sebuah teks yang atau artikel yang berbentuk teks. Ini dimaksudkan untuk mempermudah kita dalam menganalisis informasi yang ada dalam suatu teks baik itu jenis, sifat dan lain lain. Semua itu dapat dengan mudah kita pilah hanya dengan menggunakan metode klasifikasi ini. Khususnya pada sistem yang memiliki aliran data yang besar, metode ini sangatlah amat berguna, untuk mempelajari pola data atau informasi yang ada.

Metode ini berjalan diawali oleh text processing dan kemudian diklasifikasikan. Dari sini penulis ingin mengklasifikasikan berdasarkan review sebuah film terbaru yang berjudul shazam. Dengan mengambil 20 sampel dari website imdb yang diambil secara acak. Film ini menuai kontroversi karena bersifat anti mainstream membuat banyak penilaian beragam terhadap film ini.

2. Rumusan Masalah

Rumusan masalah yang dapat diambil dari latar belakang yang diangkat adalah kriteria film shazam ini berdasarkan klasifikasi review.

3. Metode

Dalam final project ini penulis menggunakan metode Text classification using Java dan algoritma Naive Bayes dengan tahapan :

Preprocessing

1. Tokenization
2. Stemming
3. Removing special characters for instance: (),.!?/'"!@#$%^&*
4. Bag of Words

Text Classification steps:

1. Data Learning
2. Data Testing

4. Hasil

a. Hasil Learning Naïve Bayes

```
run:
===== Loaded dataset: data/review.small.arff =====

Correctly Classified Instances          14               70      %
Incorrectly Classified Instances         6               30      %
Kappa statistic                          0.4
Mean absolute error                      0.3002
Root mean squared error                  0.5477
Relative absolute error                 59.3334 %
Root relative squared error            108.0955 %
Total Number of Instances               20

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0.900     0.500     0.643       0.900   0.750      0.436    0.525     0.491     positive
               0.500     0.100     0.833       0.500   0.625      0.436    0.540     0.678     negative
Weighted Avg.  0.700     0.300     0.738       0.700   0.688      0.436    0.533     0.584

===== Evaluating on filtered (training) dataset done =====
===== Training on filtered (training) dataset done =====
===== Saved model: data/wow.model.dat =====
BUILD SUCCESSFUL (total time: 0 seconds)
```

b. Hasil Klasifikasi

```
run:
===== Loaded text data: data/smstest.txt =====
 this is positive or negative ?
===== Loaded model: data/wow.model.dat =====
===== Instance created with reference dataset =====
@relation 'Test relation'

@attribute class {positive,negative}
@attribute text string

@data
?,' this is positive or negative ?'
===== Classified instance =====
Class predicted: positive
BUILD SUCCESSFUL (total time: 0 seconds)
```

c. Kesimpulan

Kesimpulan yang dapat diambil dari permasalahan yang telah dibahas dan dijelaskan diatas adalah bahwa kriteria film shazam berdasarkan review acak adalah berkriteria **Positive.**