Project Report

Classification of Customer Product Reviews Using Sentiment Analysis.

Mentor: Prof. Srikumar Krishnamoorthy
Student: Shaimak Reddy

Date: 08/07/2013

Abstract

We consider the problem of finding out the best techniques in classifying documents not by topic, but by overall sentiment, e.g., determining whether a review is positive or negative. Using novel reviews as data, we try to find out the combination of feature extraction, feature selection and machine learning methods which results in maximum accuracy. We also look at some new weighing schemes which we later find out, are more accurate than our traditional schemes. We conclude by apparently suggesting the combination of tools which is likely responsible for optimum accuracy considering sentiment classification only.

1. Introduction

Today, very large amounts of information are available in on-line documents. We often deal with websites like *Amazon.com* which have huge inflows of customer reviews loaded onto server round the clock. As part of the effort to better organize this information for users, researchers have been actively investigating the problem of automatic text categorization. The bulk of such work has focused on topical categorization, attempting to sort documents according to their subject matter (e.g., sports vs. politics). However, recent years have seen rapid growth in on-line discussion groups and review sites where a crucial characteristic of the posted articles is their sentiment, or overall opinion towards the subject matter. For example, whether a product review is positive or negative. Labeling these articles with their sentiment would provide succinct summaries to readers.

The rise of social media such as blogs and social networks has fueled interest in sentiment analysis. With the proliferation of reviews, ratings, recommendations and other forms of online expression, online opinion has turned into a kind of virtual currency for businesses looking to market their products, identify new opportunities and manage their reputations. As businesses look to automate the process of filtering out the noise, understanding the conversations, identifying the relevant content and dividing it appropriately, many are now looking to the field of sentiment analysis. If web 2.0 was all about democratizing publishing, then the next stage of the web may well be based on democratizing data mining of all the content that is getting published

2. Problem Statement

Sentiment analysis refers to the application of natural language processing, computational linguistics, and text analytics to identify and extract subjective information in source materials.

Generally speaking, sentiment analysis aims to determine the attitude of a speaker or a writer with respect to some topic or the overall contextual polarity of a document. The attitude may be his or her judgment or evaluation, affective state (that is to say, the emotional state of the author when writing), or the intended emotional communication (that is to say, the emotional effect the author wishes to have on the reader).

A basic task in sentiment analysis is classifying the *polarity* of a given text at the document, sentence, or feature/aspect level — whether the expressed opinion in a document, a sentence or an entity feature/aspect is positive, negative, or neutral.

Intuitions seem to differ as to the difficulty of the sentiment detection problem. An expert on using machine learning for text categorization predicted relatively low performance for automatic methods. On the other hand, it seems that distinguishing positive from negative reviews is relatively easy for humans, especially in comparison to the standard text categorization problem, where topics can be closely related.We further proceed to find out which feature extraction methods, feature selection methods and prediction methods outperform both independently and as a combination. Feature extraction methods we employed are: lexical features (fs1), n-gram features (fs2), sentiword features (fs3), aspect extraction (fs4), a combination of aforementioned, also some tf-idf features. All the coding is done in Python 2.7.3

Feature Selection methods employed are: information gain(IG), gain ratio(GR), Chi-square statistic(CS), point wise mutual information(PMI), categorical proportional difference(CPD), KL Divergence(JSD), while the Prediction methods employed are: Gaussian naive bays(NB), Logistic Regression(LR), Random forest Classifier(RFC), Support Vector Classifier(linear kernel), SVC (radial kernel).

As mention earlier, our aim is to implement these above mentioned functions to find out the best possible combination in terms of higher *accuracy*, f-value. More insights regarding the feature extractions and feature selections are described below.

2.1 Challenges:

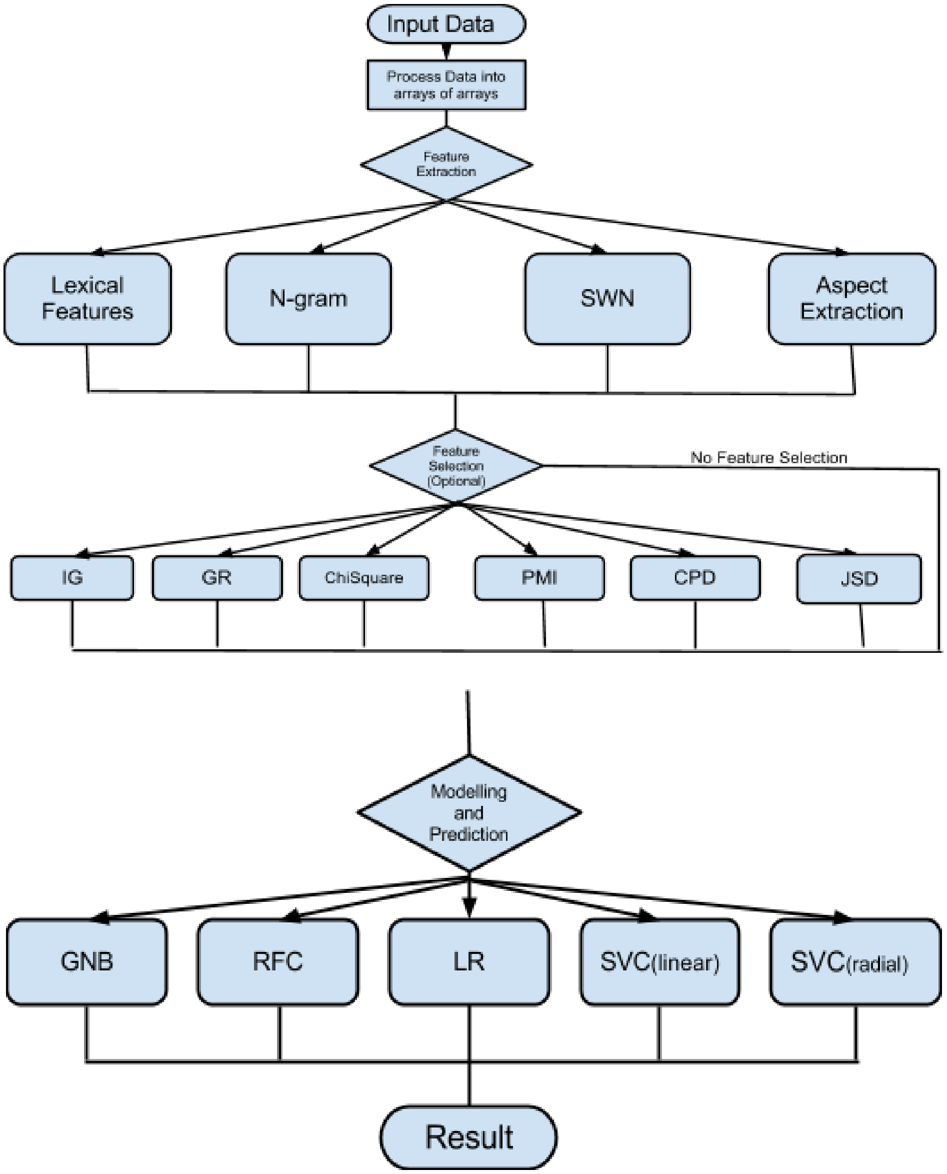
Like most scientific methods sentiment analysis is not without its problems. Sentiment analysis is a very subjective method of classification and if there are more than one observer to the test, there will more than likely be differences in opinion.

This is actually problem most often encountered with sentiment analysis. Interpreting the mood of a subject may vary from one person to another; a problem made even harder by the format the subject may be analyzed in.

The research in the field started with sentiment and subjectivity classification, which treated the problem as a text classification problem. Sentiment classification classifies whether an opinionated document (e.g. Product reviews) or sentence expresses a positive or negative opinion. Subjectivity classification determines whether a sentence is subjective or objective. Many real-life applications, however, require more detailed analysis because the user often wants to know what the opinions have been expressed on. For example, from the review of a product, one wants to know what features of the product have been praised and criticized by consumers.

3. Methodology

The below flow chart outlines the various steps involved in the methodology of arriving with the conclusion. This involves the preprocessing, different extracting and predicting methods and the order of performing.



3.1 Preprocessing:

For our experiments, we chose to work with novel reviews. This domain is experimentally convenient because there are large on-line collections of such reviews (eg: Amazon.com), and because reviewers often summarize their overall sentiment with a machine-extractable rating indicator, such as a number of stars; hence, we did not need to hand-label the data for supervised learning or evaluation purposes. It can quite conveniently be extended to any kind of product reviews.

Our test case consists 2000 reviews from amazon.com of various novels, 1000 positive reviews and 1000 negative reviews, with ratings of 4, 5 considered positive and 1,2 from considered negative. For now, we are not considering reviews with the specific rating of 3. This would reduce the ambiguity of the user's rating, i.e. if the user wanted to rate is as useful or not. Although these reviews are not specific to any book or author, we still find common words like *book*, *novel*, *author* and many more *stop-words* which help us in no way to find out the overall sentiment. We filter stop words in our first level of processing. Although the former kind of words mentioned are harder to remove in some cases, we definitely can overcome them in second level of filtering. However, since they are equally likely to occur in both positive and negative cases, we only lose on some processing time.

3.2 Feature extraction:

When the input data to an algorithm is too large to be processed and it is suspected to be notoriously redundant then the input data will be transformed into a reduced representation set of features (also named features vector). Transforming the input data into the set of features is called *feature extraction*. If the features extracted are carefully chosen it is expected that the features set will extract the relevant information from the input data in order to perform the desired task using this reduced representation instead of the full size input.

Generally speaking, Feature extraction involves simplifying the amount of resources required to describe a large set of data accurately. When performing analysis of complex data one of the major problems stems from the number of variables involved. Analysis with a large number of variables generally requires a large amount of memory and computation power or a classification algorithm which over fits the training sample and generalizes poorly to new samples. Feature extraction is a general term for methods of constructing combinations of the variables to get around these problems while still describing the data with sufficient accuracy.

Feature Extraction (FE) is perhaps the most difficult task in SA. The following are some of the feature extraction methods we shall be using are mentioned below:

3.2.1 Lexical Features (FE1)**:**

This consists of a set of 250 lexical (character-based + word-based + syntactic + Structural) features including five vocabulary richness measures.

Character based features:

1.	Total number of characters(C)
2.	Total number of alphabetic characters/C
3.	Total number of uppercase characters/C
4.	Total number of digit characters/C
5.	Total number of white-space characters/C
6.	Total number of tab spaces/C
7-32	Frequency of letters(26 features)
33-53	Frequency of special characters(21 features) ~,@,#,$,%,^,&,*,-,_,=,+,>,<,[,],{,},/,\,|

Word-based Features:

54.	Total number of words(M)
55.	Total number of short words (less than 4 characters)/M
56.	Total number of characters in words/C
57.	Average word length
58.	Average sentence length in terms of character
59.	Average sentence length in terms of word
60.	Total different words/M
61.	Hapax legomena - Frequency of once-occurring words
62.	Hapax dislegomena - Frequency of twice-occurring words
63.	Yule’s K measure
64.	Simpson’s D measure
65.	Sichel’s S measure
66.	Brunet’s W measure
67.	Honore’s R measure
68-87	Word length frequency distribution/M (20 features) - frequency of words of different length

Syntactic features:

88-95	Frequency of punctuations (8 features) “.”,“,”,“?”,“!”,“:”,“;”,““”,“””
96-245	Frequency of function words(150 features)

['a','about','above','after','all','although','am','among','an','and','another','any','anybody','anyone','anything','are','around','as','at','be','because','before','behind','below','beside','between','both','but','by','can','cos','do','down','each','either','enough','every','everybody','everyone','everything','few','following','for','from','have','he','her','him','i','if','in','including','inside','into','is','it','its','latter','less','like','little','lots','many','me','more','most','much','must','my','near','need','neither','no','nobody','none','nor','nothing','of','off','on','once','one','onto','opposite','or','our','outside','over','own','past','per','plenty','plus','regarding','same','several','she','should','since','so','some','somebody','someone','something','such','than','that','the','their','them','these','they','this','those','though','through','till','to','toward','towards','under','unless','unlike','until','up','upon','us','used','via','we','what','whatever','when','where','whether','which','while','who','whoever','whom','whose','will','with','within','without','worth','would','yes','you','your']

Structural Features:

246.	Total number of lines
247.	Total number of sentences
248.	Total number of paragraphs
249.	Total number of sentences per paragraph
250.	Total number of characters per paragraph

3.2.2 N-gram features (FE2):

Here we firstly extract n-grams (preferably unigrams or bigrams) as potential features, initially filtered out based on minimum count considering total number of occurrences in all reviews combined.

3.2.3 Sentiment Subjectivity Scoring (FE3):

Features here are selected firstly based on parts of speech, mainly Adjectives, adverbs and verbs. Only those features which have significant effect in determining the class of review, i.e. have positive weighting significantly higher than negative or vice versa. There is no other cutting down of words based on weights.

SentiWordNet (SWN) assigns three scores (positive, negative, objective) to each synset in WorldNet. The scores are also provided for different adjective, adverb, verb and noun. In this study, we extract the sentiment scores for adjective, adverb and verbs which normally have sentiment orientation.

$$\mathit{Score}\ (f{=}POS)_i = \sum_{\mathrm{k \epsilon SWN(f=\ POS\ and\ polarity=i)}} \frac{SWN(k)}{|\mathrm{synsets(f{=}POS)}|}$$

For each feature, up to 9 scores are calculated (3 based on POS - adjective, adverb, verb x 3 based on polarity - positive, negative, objective). The features are weighted and initially selected using the following strategy:

```
if Score(f = POS)obj > 0:5, discard the feature
Else if Score(f = POS)pos > Score(f = POS)neg
    add feature, fwithscore = | Score(f = POS)pos |
if Score(f = POS)pos < Score(f = POS)neg
    add feature, fwithscore = | Score(f = POS)pos |
else discard the feature
```

3.2.4 Aspect Extraction (FE4)**:**

Our features here are nouns, weight simply by the number of times they occur in our review set. A parameter is set to exclude the aspects which occur lesser number of times then the value of itself. Around 2000 features would be optimum in terms of speed and accuracy for a train + test set of 2000 reviews.

3.3 TF-IDF Weighing:

Another kind of feature weighing scheme which can overcome the redundant processing time and memory is *tf-idf* weighting scheme, which is unlike any other traditional weighing schemes discussed above. The **tf–idf**, term frequency–inverse document frequency, is a numerical statistic which reflects how important a word is to a document in a collection. The tf-idf value increases proportionally to the number of times a word appears in the document, but is offset by the frequency of the word in the corpus, which helps to control for the fact that some words are generally more common than others.tf–idf is the product of two statistics, term frequency and inverse document frequency. Various ways for determining the exact values of both statistics exist. In the case of the **term frequency** tf (*t*,*d*), the simplest choice is to use the *raw frequency* of a term in a document, i.e. the number of times that term *t* occurs in document *d*. The **inverse document frequency** is a measure of whether the term is common or rare across all documents. It is obtained by dividing the total number of documents by the number of documents containing the term, and then taking the logarithm of that quotient.

$$\text{TFIDF (f)} = tf*log_2\,(N/df)$$

N is the total number of documents/reviews, tf is the frequency of occurrence of feature f in the collection, and df is the total number of reviews that contain feature f.

3.3.1 TF weighting:

B (Boolean) = 1, TFIDF>0,

=0, otherwise

A (augmented) = 0.5+ 0.5*tf/max$_f$(tf)

O (bm25) = (k+1)*tf /(k*((1-b)+b*no_of_features/avg_word_count) + tf)

Where tf is the term frequency of feature in the particular review and max$_f$(tf) is the maximum frequency of any feature in the review collection, k=1.2, b=0.95.

3.3.2 IDF weighting:

***Default** =1*

Delta (smoothed IDF) $= log_2((\mathrm{N1} * \mathrm{df2}\ + 0.5) \div (\mathrm{N2} * \mathrm{df1} + 0.5))$

Delta (SmoothedProbIDF) $= log2((\mathrm{N1} * \mathrm{df2}\ + 0.5) \div (\mathrm{N2} * \mathrm{df1} + 0.5))$

A high weight in tf–idf is reached by a high term frequency (in the given document) and a low document frequency of the term in the whole collection of documents; the weights hence tend to filter out common terms. Since the ratio inside the idf's log function is always greater than or equal to 1, the value of idf (and tf-idf) is greater than or equal to 0. As a term appears in more documents, the ratio inside the logarithm approaches 1, bringing the idf and tf-idf closer to 0.Various (mathematical) forms of the tf–idf term weight can be derived from a probabilistic retrieval model that mimics human relevance decision making.

3.4 Feature Selection:

As the dimensionality of the data increases, many types of data analysis and classification problems become significantly harder. Sometimes the data also becomes increasingly sparse in the space it occupies. This can lead to big problems for both supervised and unsupervised learning. On the one hand, in the case of supervised learning or classification the available training data may be too small, i. e, there may be too few data objects to allow the creation of a reliable model for assigning a class to all possible objects. On the other hand, for unsupervised learning methods or clustering algorithms, various vitally important definitions like density or distance between points may become less convincing (as more dimensions tend to make the proximity between points more uniform). As a result, a high number of features can lead to lower classification accuracy and clusters of poor quality.

3.4.1 Information Gain (IG):

Given a training review example of the form: (X, C), where C is the class label, X is represented as (f1, f2...fn), and f1, f2 and so on are the features of the review.

$$Entropy(C) = -\sum_{i=1}^{m} \mathrm{Pi} * \log(\mathrm{Pi})$$

Where m is the number of classes (eg.2-C1, C2 here since it is a binary review classification) Pi is the proportion of tuples in each class calculated as|CϵCi|/|C|

$$Information\ Gain(C,f) = Entropy(C) - Entropy(C|f)$$

3.4.2 Gain ratio (GR):

Gain ratio is an extension of the Information gain measure. This means normalizes the information gain scores based on the number of partitions of a feature.

$$GainRatio(C,f)=InformationGain(C,f) \div \mathrm{Entropy(f)}$$

3.4.3 Chi-Square Statistic (CS):

Chi-squared statistic measures the association between the feature and the corresponding class. The chi-square value is zero when the feature fj and class Ci are independent of each other. The contingency table that follows provides the number of training instances where the feature f is found / not found in class C.

Contingency Table:

	C	~C
f	a	b
~f	c	d

$$X^2(f,C_i) = \mathrm{N} * (\mathrm{a} * \mathrm{d} - \mathrm{b} * \mathrm{c})\ {}^2 \div ((\mathrm{a} + \mathrm{c}) * (\mathrm{b} + \mathrm{d}) * (\mathrm{a} + \mathrm{b}) * (\mathrm{c} + \mathrm{d})\ ),\ N=a+b+c+d$$

$$X^2(f)=max_i\ X^2(f,C_i)$$

3.4.4 Point wise Mutual Information (PMI):

PMI measures the mutual dependence of a feature and a class. It is zero when the feature and class are independent.

$$PMI(f,C_i) = P(f,C_i) \div\ (\ P(f) * P(Ci)\ )$$

$$= a*N/\ (\ (a+b)*(a+c)\ )$$

$$PMI(f)=max_i\ PMI(f,C_i)$$

3.4.5 Categorical Proportional Difference (CPD)

CPD is used to measure the degree of dependence of a feature to one class over the other.

$$\mathrm{CPD}(\mathrm{f},\mathrm{C}_\mathrm{i})= (\mathrm{a}-\mathrm{b})\ /\ (\mathrm{a}+\mathrm{b})$$

$$\mathrm{CPD}(\mathrm{f})=\mathrm{max}_\mathrm{i}\ \mathrm{CPD}(\mathrm{f},\mathrm{C}_\mathrm{i})$$

3.4.6 KL Divergence - Jensen Shannon Divergence

Measures the difference in probability distributions of feature f and class C.

$$\begin{aligned} D_{JS}(C, f) &= 0.5*D_{KL}(C||f) + 0.5*D_{KL}(f||C) \\ &= 0.5*(Entropy(C, f) - Entropy(C)) \\ &+0:5\ (Entropy(f,C) - Entropy(f)\ ) \end{aligned}$$

3.5 Model building

As mentioned earlier, after the pre-processing being done, we begin with our base case considering n-gram features for basic feature extraction and comparing the results using different prediction methods(Gaussian naive Bayes(NB), LR, Random forest Classifier, Support Vector Classifier(linear kernel), SVC (radial kernel)). Considering prediction methods which are giving better accuracies, we proceed to test the results on other feature extraction methods using the successful prediction methods and hence come up with one of the combination of feature extraction + selection with highest accuracy. We further try these with a combination of feature extraction methods to see if can improve our accuracy. Later we apply feature selection methods on our feature extraction methods to find out the best selection method which preserves /increases the accuracy of our afore-deduced best case.

On the other hand, we also try to investigate feature extraction by term frequency (tf) and also by multiplying by idf (inverse document factor) and predicting using the same methods. Similarly we also use feature selection to find out the best combination of tf-idf with feature selection and prediction methods in terms of accuracy.

Towards the end, we use the best feature extraction, feature selection(if any) and prediction to determine the highest accuracy and later see the influence of change of parameters(will be mentioned subsequently) on the accuracy to determine the maximum possible accuracy. We conclude by the former and also by throwing some light upon the times of execution of programs.

SWN	Sentiment subjectivity scoring	PMI	Pointwise mutual Information	LR	LR
IG	Information Gain	CPD	Categorical Proportional Difference	GNB	Gaussian Naive Bayes
GR	Gain Ratio	JSD	JSD - Jensen Shannon Divergence	RFC	Random Forest Classifier

3.6 Prediction:

3.6.1 Gaussian Naive Bayes:

A naive Bayes classifier is a simple probabilistic classifier based on applying Bayes' theorem with strong (naive) independence assumptions. A more descriptive term for the underlying probability model would be "independent feature model". A naive Bayes classifier assumes that the presence or absence of a particular feature is unrelated to the presence or absence of any other feature, given the class variable.

Using Bayes' theorem, this can be written

$$p(C|F_1,\dots,F_n) = \frac{p(C)\ p(F_1,\dots,F_n|C)}{p(F_1,\dots,F_n)}.$$

And the conditional distribution over the class variable C is:

$$p(C|F_1,\dots,F_n) = \frac{1}{Z}p(C)\prod_{i=1}^{n} p(F_i|C)$$

Where Z (the evidence) is a scaling factor dependent only on $F_1, \dots, F_n$, that is, a constant if the values of the feature variables are known.

3.6.2 Logistic Regression:

Logistic regression is used for predicting the outcome of a categorical dependent variable based on one or more predictor variables. That is, it is used in estimating empirical values of the parameters in a qualitative response model. The probabilities describing the possible outcomes of a single trial are modeled, as a function of the explanatory (predictor) variables, using a logistic function. Logistic regression measures the relationship between a categorical dependent variable and one or more independent variables, which are usually (but not necessarily) continuous, by using probability scores as the predicted values of the dependent variable.

3.6.3 Random Forest Classifier:

Random forests are an ensemble learning method for classification (and regression) that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes output by individual trees. The method combines idea and the random selection of features, in order to construct a collection of decision trees with controlled variation.

3.6.4 Support Vector Classifier:

In machine learning, support vector machines (SVM) are supervised learning models with associated learning algorithms that analyze data and recognize patterns, used for classification and regression analysis. The basic SVM takes a set of input data and predicts, for each given input, which of two possible classes forms the output, making it a non-probabilistic binary linear classifier. Given a set of training examples, each marked as belonging to one of two categories, an SVM training algorithm builds a model that assigns new examples into one category or the other.

An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall on.

In addition to performing linear classification, SVMs can efficiently perform non-linear classification using what is called the kernel trick, implicitly mapping their inputs into high-dimensional feature spaces. We deal with two kinds of SVC's, namely SVC with linear kernel and SVC with radial kernel

4. Experiments

4.1 Design

Design refers to our selection of data and the procedure of pre-processing and evaluation we consider to arrive at our result. We also mention the details of certain parameters we've encountered which are variable and how we chose those values which affect our results to some extent. The values chosen serve as a trade-off between accuracy and running time. More insight on how we proceed selectively testing and parameters we chose to compare the results are described below:

4.1.1 Dataset

The Dataset generated from the preprocessing done from a set of 2000 book reviews with 50% odds(1000 positive reviews and 1000 negative reviews) is an array of arrays. Where each element in a row (or major array) corresponds to a review bag and each such bag has two elements, the first stores the rating (0 or 1 based on negative or positive review), second sores the review in form of string. Thus making a 2000x2 matrix. The matrices after different stages are different in size, also vary with the review type. They also vary with the parameters used. More will be discussed later.

4.1.2 Evaluation

Evaluation being concerned, we primarily refer to four terms namely Precision, recall, f-measure and accuracy. We also do a cross fold validation to average out the values and get fairly consistent results.

Precision:

The precision of a measurement system, also called reproducibility or repeatability, is the degree to which repeated measurements under unchanged conditions show the same results. In other words, Precision is the probability that a (randomly selected) retrieved document is relevant.

$$\text{Precision} = \frac{tp}{tp + fp}$$

Recall:

Recall in this context is defined as the number of true positives divided by the total number of elements that actually belong to the positive class (i.e. the sum of true positives and false negatives, which are items which were not labeled as belonging to the positive class but should have been). In other words, Recall is the probability that a (randomly selected) relevant document is retrieved in a search.

$$\text{Recall} = \frac{tp}{tp + fn}$$

F-measure:

Often, there is an inverse relationship between precision and recall, where it is possible to increase one at the cost of reducing the other. Usually, precision and recall scores are not discussed in isolation. Instead, either values for one measure are compared for a fixed level at the other measure (e.g. precision at a recall level of 0.75) or both are combined into a single measure, such as their harmonic mean the F-measure, which is the weighted harmonic mean of precision and recall

$$F = 2 \cdot \frac{\mathrm{precision} \cdot \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}}$$

Accuracy:

Accuracy is the proportion of true results (both true positives and true negatives) in the population. It is a parameter of the test. An accuracy of 100% means that the measured values are exactly the same as the given values.

$$\mathrm{accuracy} = \frac{\text{number of true positives} + \text{number of true negatives}}{\text{number of true positives} + \text{false positives} + \text{false negatives} + \text{true negatives}}$$

Cross validation:

Cross-validation, sometimes called rotation estimation, is a model validation technique for assessing how the results of a statistical analysis will generalize to an independent data set. It is mainly used in settings where the goal is prediction, and one wants to estimate how accurately a predictive model will perform in practice. One round of cross-validation involves partitioning a sample of data into complementary subsets, performing the analysis on one subset (called the training set), and validating the analysis on the other subset (called the validation set or testing set). To reduce variability, multiple rounds of cross-validation are performed using different partitions, and the validation results are averaged over the rounds. We are currently using a 3-fold validation with train-test data sets randomly generated.

4.1.3 Process

Here we describe the methodology in steps to arrive at the best possible classifier measured on accuracy. We proceed by following the below steps, where each step has (apparently) significant effect on the resulting ranking. It is to be noted that we are ranking these different algorithms based on (increased) accuracy and secondly, if needed by f-measure. Precision and recall are more variant in compared to the former.

A. **Classification Evaluation:**

Firstly, the dataset extracted and preprocessed in for of a matrix is passed through N-gram feature extraction, consistently treating this as a base case for comparing different classifiers. Procedurally, the new matrix obtained from the n-gram feature extraction is our ground level input to test and evaluate our prediction methods. We compare their accuracy to find out the best two classifiers which will be carried forward into subsequent steps.

B. **Feature Extraction:**

Once we get the top two classifiers, we can now go back to test the best feature extraction method based on the results our selected prediction methods show. The extraction methods compared are Logistic regression, N-gram feature extraction, SWN sentiment subjectivity scoring, and Aspect extraction. Their accuracies are compared and the feature extraction method which yields the highest accuracy is carried forward for feature selection.

C. **Feature Selection:**

Feature selection algorithms basically weighs features based on apparent importance assigned from some definite characteristics. It as chops down various insignificant features thus saving precious runtime. But there is a tradeoff between the former and the number of features we actually select because we may be losing on important features thus in a way limiting our knowledge obtained from the dataset. As a result, feature selection may act as a boon or bane. In Our results, we select about 500 features, thus resulting in a 2000x500 matrix which is a fair deal to predict with. Also, since we are relative in predicting, the sense of losing out on number of features cannot be so harmful are projected.

D. Weighting Schemes (TF-IDF):

Another kind of feature extraction + weighting schemes are tf-idf schemes, based on n-grams (more aptly unigrams and bi-grams) which are also predominant in text classification. The term frequency – Inverse Document frequency collectively acts as feature extraction and feature selection combined. The features are weighted based on some schemes but all those features are selected, i.e. none are chopped off for prediction. The results are compared with that of best former feature extraction algorithms to find out the better of the two. We can also apply feature selection algorithms on the resultant already *weighed* features to see if there might be an increase.

Depending on the results produced, we can arrive with the best prediction match. It is to note that a minimum of 50% accuracy is the minimum possible rationally (with probability 0, predicting a constant value throughout) and around 80% was believed to be achieved earlier using the traditional (former) feature selection algorithms.

4.2 Analysis and Results:

Parameters used:

Parameter	Value Used	Description of parameter
N	1	Describes the n-gram used for feature extraction
Wt_ngram	15	The minimum no. of times a particular n-gram is to be present to be counted as a feature
No_of_features	500	States the number of features which are selected in the feature selection algorithm after they are scaled based on their respective traits
Noun_count	100	The number of noun-features used per review. The preference is based on the frequency of occurrence.

Confusion matrix:

	Actual [**Yes**]	Actual [**No**]
Predicted [**Yes**]	True positive [**tp**]	False positive [**fp**]
Predicted [**No**]	False negative [**fn**]	True negative [**tn**]

A. Classification Evaluation:

As mentioned earlier, we first look at results produced from FE2 (N-gram) feature extraction using various prediction methods. Note that no feature selection methods are employed. Here a parameter wt. =15 is used, which is the minimum count of n-gram to be counted as a feature. We thus get around 2000 features for the 2000 review set we are experimenting on.

Prediction	Precision	Recall	Accuracy	F-measure
Gaussian NB	0.6473	0.766	**0.674**	0.701
LR	0.7371	0.734	**0.736**	0.735
RFC	0.6870	0.574	**0.656**	0.625
SVC (linear)	0.7152	0.710	**0.713**	0.712
SVC (radial)	0.5653	0.853	**0.598**	0.680

From the above results we can conclude that LR (LR) and SVC (linear) have higher accuracies when compared to the rest. We hence proceed with the results predicted by SVC (linear kernel) and LR for the remaining feature extraction methods:

B. Feature Extraction Evaluation:

Lexical Features:

Prediction	Precision	Recall	**Accuracy**	F-Measure
LR	0.6063	0.6660	**0.616**	0.6342
SVC (linear)	0.6013	0.6990	**0.617**	0.6462

Sentiment Subjectivity scoring:

Prediction	Precision	Recall	Accuracy	F-Measure
LR	0.6391	0.6990	**0.651**	0.6677
SVC (linear)	0.6427	0.6990	**0.654**	0.6695

Aspect Extraction:

Prediction	Precision	Recall	Accuracy	F-Measure
LR	0.6034	0.5890	**0.601**	0.5964
SVC (linear)	0.6037	0.5850	**0.600**	0.5940

The results are directing towards n-gram being in the lead with an accuracy of *73.6% (LR) and 71.3% (SVC)* .We now see how the results are with combination of sum of above features:

FE3+FE4 (SWN+ Aspect Extraction):

Prediction	Precision	Recall	**Accuracy**	F-Measure
LR	0.6773	0.6700	**0.6755**	0.6733
SVC (linear)	0.6681	0.6530	**0.6635**	0.6600

FE2+FE3+FE4 (N-gram +SWN + Aspect Extraction):

Prediction	Precision	Recall	**Accuracy**	F-Measure
LR	0.7392	0.7340	**0.7370**	0.7361
SVC (linear)	0.7083	0.7000	**0.7055**	0.7039

FE1+FE2+FE3+FE4 (Lexical + N-gram +SWN + Aspect Extraction):

Prediction	Precision	Recall	**Accuracy**	F-Measure
LR	0.7562	0.7480	**0.7525**	0.7513
SVC (linear)	0.7300	0.7170	**0.7255**	0.7230

Results above suggest that the combination of all four feature extractions leads to a slightly better accuracy **(75.2%)** as compared to using only n-gram features **(73.7%)**. We further try to investigate using some feature selection methods, as described below.

C. Feature Selection evaluation:

We hereby test if the accuracy of prediction using N-gram feature selection as improved on applying feature selection algorithms. We continue to compare the accuracies measured by LR and SVC (linear kernel). Note that a parameter here which we are using is the number of features selected and it is 500 now for the moment to optimize the time in comparison to codes running excluding feature selection algorithms.

Results for n-gram feature selection:

Information Gain:

Prediction	**Accuracy**	F-measure
LR	**0.7925**	0.7910
SVC (linear)	**0.7750**	0.7745

Gain ratio:

Prediction	**Accuracy**	F-measure
LR	**0.7925**	0.7910
SVC (linear)	**0.7750**	0.7745

Chi-Square Statistic:

Prediction	**Accuracy**	F-measure
LR	**0.7940**	0.7927
SVC (linear)	**0.7815**	0.7807

Point wise Mutual Information (PMI):

Prediction	**Accuracy**	F-measure
LR	**0.7960**	0.7994
SVC (linear)	**0.7725**	0.7750

Categorical Proportional Difference (CPD):

Prediction	**Accuracy**	F-measure
LR	**0.7960**	0.7995
SVC (linear)	**0.7725**	0.7750

KL Divergence - Jensen Shannon Divergence:

Prediction	**Accuracy**	F-measure
LR	**0.7945**	0.7936
SVC (linear)	**0.7745**	0.7735

It is evident from the results that the accuracy has definitely increased by our feature selection schemes but the accuracies seem to be quite close. I.e. are equally good. But from the results of Tf-idf which we will cover later, Chi-Square Statistic, Information gain and JSD seem to have an appreciable edge over the remaining three feature selection methods with Chi-Square leading by a very small margin.(**~0.1%**) . We thus can safely assume *N-gram* feature extraction with *Chi-Square* feature selection and predicting using *LR* gives a very high accuracy in sentiment classification.

D. Weighting Schemes:

Similarly, we first compare different prediction methods, firstly without any feature selection algorithms involved. Our base case will be for Boolean tf weighing with idf weighting default at 1.

Prediction	Precision	Recall	**Accuracy**	F-Measure
GNB	0.6575	0.5670	**0.6360**	0.6066
LR	0.7567	0.7510	**0.7545**	0.7534
RFC	0.6932	0.6450	**0.6805**	0.6676
SVC (linear)	0.7318	0.7140	**0.7260**	0.7227
SVC (radial)	0.3538	0.6326	**0.5325**	0.4524

The above results are similar with the ones we got in case of n-gram feature extraction. We hence carry forward LR and SVC (linear) for predicting the results. Below are results of other tf-idf schemes implemented without any feature selections, predicted using LR, SVC(linear):

Results for TF-IDF:

Tf=b (Boolean), idf=s(SmoothedIDF):

Prediction	**Accuracy**	F-measure
LR	**0.8920**	0.8941
SVC (linear)	**0.7965**	0.7939

tf=b, idf=sp(SmoothedProbIDF):

Prediction	**Accuracy**	F-measure
LR	**0.8825**	0.8830
SVC (linear)	**0.7865**	0.7821

Tf=a, idf=d (Default) (Binned data):

Prediction	**Accuracy**	F-measure
LR	**0.6570**	0.6704
SVC (linear)	**0.6310**	0.6515

tf=a (augmented), idf=s(SmoothedIDF):

Prediction	**Accuracy**	F-measure
LR	**0.8335**	0.8580
Gaussian NB	**0.8695**	0.8610

Tf=a (augmented), idf=sp (SmoothedProbIDF):

Prediction	**Accuracy**	F-measure
LR	**0.7285**	0.7538
Gaussian NB	**0.8695**	0.8610

tf=o (bm25), idf=default:

Prediction	**Accuracy**	F-measure
LR	**0.7615**	0.7594
SVC (linear)	**0.7220**	0.7195

Tf=o (bm25), idf=s(SmoothedIDF):

Prediction	**Accuracy**	F-measure
LR	**0.9080**	0.9092
SVC (linear)	**0.8285**	0.8306

tf=o(bm25), idf=sp(SmoothedProbIDF):

Prediction	**Accuracy**	F-measure
LR	**0.8955**	0.8957
SVC (linear)	**0.8155**	0.8163

Clearly, TF-IDF Schemes have a very high accuracy (tf=o, idf=s i.e. [o,s] at **90.8%** being the highest) reaching up to 90% of accuracies. Some of them are: [b,s] at 89.2%, [b,sp] at 88.2%, [a,sp] at 86.9%, [o,s] at 90.8% and [o,sp] at 89.5%. These results confirm the point that tf-idf feature extraction processes are superior to n-gram feature extraction and aspect extraction processes.

We move forward to test the effect of Feature selection on our best case, i.e. obm25 with idf=s (SmoothedProbIDF):

Results for [O, S]:

Information Gain:

Prediction	**Accuracy**	F-measure
LR	**0.8595**	0.8600
SVC (linear)	**0.8065**	0.8093

Gain ratio:

Prediction	**Accuracy**	F-measure
LR	**0.6195**	0.6986
SVC (linear)	**0.5925**	0.6822

Chi-Square Statistic:

Prediction	**Accuracy**	F-measure
LR	**0.8540**	0.8549
SVC (linear)	**0.7985**	0.8003

Point wise Mutual Information (PMI):

Prediction	**Accuracy**	F-measure
LR	**0.7535**	0.8022
SVC (linear)	**0.7100**	0.7752

Categorical Proportional Difference (CPD):

Prediction	**Accuracy**	F-measure
LR	**0.7535**	0.8023
SVC (linear)	**0.7100**	0.7752

KL Divergence - Jensen Shannon Divergence:

Prediction	**Accuracy**	F-measure
LR	**0.8480**	0.8493
SVC (linear)	**0.7940**	0.7980

Not so surprisingly, the above results suggest that applying feature selection algorithms onto features weighed by tf-idf algorithms, the accuracy tends to decrease all the feature selection algorithms. Binning the data before predicting increases the accuracy when predicted by SVC(linear) but still remains below the original accuracy using all the features whereas in case of

LR, the accuracy further decreases. Thus, the best possible accuracy is attained for *tf=o, idf=s* (o,s) without any feature selection, predicted using *Logistic Regression*.

5. Conclusion:

We started by off by comparing reviews to determine their overall sentiment in order to effectively classify them. As per the tests we've performed, TF-idf feature extraction methods are performing significantly better than the traditional N-gram feature extraction methods (**90.8%** as compared to **79.5%**). When it comes to feature selection algorithms, we can also state that they are beneficial only for traditional feature selection algorithms and not for tf-idf. Logistic Regression proved to have the highest accuracy in prediction for both kinds of feature extraction models. Thus, with *okapi (bm25)* tf weighing with *smoothed idf* as feature extraction, and with *logistic regression* as prediction, we can expect very high accuracies.

Limitations: We've come up with a fairly reasonable method which does have a very good accuracy but we are still open to more explorations. These mainly correspond to changing the parameters and observing the effect of them on our primary results. Starting with the traditional methods where feature selection was a boon, two things could have been altered to probably get better accuracy. These being the 'n-gram' (which is unigram in our experiments), and the count of features we select from feature selection algorithms.

The former may have better results considering the bi-grams like 'not pleasing', 'not bad' as compared to unigram based evaluation. The later may not be of any importance since feature selection didn't have any benefits in case of tf-idf weighing schemes. However, reducing around 20,000 features to 500 would definitely reduce the accuracy by quite some bit. Hence is still worth the effort.

6. Learnings:

This section of the report describes the benefits of pursuing this project and how this project plays an important role in my career. This also describes the level of detail with which the topics have been covered, which would eventually score me some more points in the long run.

To begin with, the topic itself, with some background information was intriguing enough to fuel interest throughout the term. Having done only mini- projects in *python* before, I can safely say that I have improved significantly that I now prefer doing all my future coding in *python*. Regarding more about the topic, though *Sentiment analysis* and machine learning techniques were new to me, it was an interesting enough topic to work on, I got well acquainted with the topic.

This project also served as a major leap forward in work concerned with my minor field (Computer Science), and Since my assignment was two month long, I am now well off with my coding skills on a larger scale. Besides the above mentioned, I also learnt quite a few tactics such as I've gained the ability to work independently, more organized in a manner. I've gained some experience on practical applications of my minor courses since I've been dealing with very large amounts of data and processing time up to 60hrs at peaks.

Not only has my internship contributed to my long term skills and interests besides providing me with adequate experience, It also threw some light upon my future career plans and my department orientation. All in all, it has been a great learning and also a memorable experience.

7. References:

- Thumbs up? “Sentiment Classification using machine learning techniques", by Bo Pang, L. Lee, S. Vaithyanathan, click here.
- ‘Sentiment Analysis: A Multi-Faceted Problem’, by Bing Liu; Handbook of Natural Language Processing, Second Edition, (editors: N. Indurkhya and F. J. Damerau)
- A Feature Extraction Process for Sentiment Analysis of Opinions on Services, Henrique Siqueira and Flavia Barros
- A study of Information Retrieval weighting schemes for sentiment analysis, by Georgios Paltoglou, Mike Thelwall.
- Wikipedia. click here

APPENDIX:

GNB	Gaussian Naive Bayes
LR	Logistic Regression
RFC	Random Forest Classifier
SVC (linear)	Support Vector Classifier with linear kernel
SVC(radial)	Support Vector Classifier with radial kernel
IG	Information Gain Selection
GR	Gain Ratio Selection
CS	Chi-Square Statistic
PMI	Point wise Mutual Information
CPD	Categorical Proportional Difference
JSD	KL Divergence - Jensen Shannon Divergence

Lexical features (FE1)

FS Type	Prediction	Precision	Recall	Accuracy	F-Measure
-	GNB	54.63%	72.60%	56.00%	62.26%
-	LR	60.64%	66.60%	61.65%	63.42%
-	RFC	61.10%	49.70%	59.05%	54.79%
-	SVC (linear)	60.13%	69.90%	61.75%	64.62%

N-gram features (FE2)

FS Type	Prediction	Precision	Recall	Accuracy	F-measure
-	GNB	64.74%	76.70%	67.45%	70.17%
-	LR	73.71%	73.50%	73.60%	73.55%
-	RFC	68.71%	57.50%	65.65%	62.52%
-	SVC (linear)	71.52%	71.10%	71.35%	71.28%

-	SVC (radial)	56.54%	85.40%	59.85%	68.03%
IG	GNB	72.76%	86.90%	77.05%	79.12%
IG	LR	79.65%	78.60%	79.25%	79.10%
IG	RFC	71.76%	62.80%	69.00%	66.87%
IG	SVC (linear)	77.56%	77.40%	77.50%	77.45%
IG	SVC (radial)	68.27%	78.10%	70.90%	72.85%
GR	GNB	72.76%	86.90%	77.05%	79.12%
GR	LR	79.65%	78.60%	79.25%	79.10%
GR	RFC	72.17%	64.50%	69.80%	68.11%
GR	SVC (linear)	77.56%	77.40%	77.50%	77.45%
GR	SVC (radial)	68.27%	78.10%	70.90%	72.85%
CS	GNB	72.71%	86.70%	76.90%	78.97%
CS	LR	79.68%	78.90%	79.40%	79.27%
CS	RFC	71.40%	63.10%	68.80%	66.78%
CS	SVC (linear)	78.31%	77.90%	78.15%	78.07%
CS	SVC (radial)	68.22%	78.10%	70.85%	72.81%
PMI	GNB	68.30%	86.50%	73.10%	76.28%
PMI	LR	78.58%	81.40%	79.60%	79.95%
PMI	RFC	75.06%	78.40%	76.15%	76.65%
PMI	SVC (linear)	76.52%	78.60%	77.25%	77.50%
PMI	SVC (radial)	66.16%	90.20%	71.75%	76.20%
CPD	GNB	68.30%	86.50%	73.10%	76.28%
CPD	LR	78.58%	81.40%	79.60%	79.95%
CPD	RFC	74.08%	77.30%	75.10%	75.62%
CPD	SVC (linear)	76.52%	78.60%	77.25%	77.50%
CPD	SVC (radial)	66.16%	90.20%	71.75%	76.20%
JSD	GNB	72.04%	86.30%	76.30%	78.44%
JSD	LR	79.65%	79.10%	79.45%	79.36%
JSD	RFC	71.65%	62.40%	68.85%	66.68%
JSD	SVC (linear)	77.64%	77.10%	77.45%	77.35%
JSD	SVC (radial)	68.26%	78.50%	71.00%	73.02%

Sentiment Subjectivity Scoring (FE3 A)

FS Type	Prediction	Precision	Recall	Accuracy	F-measure
-	GNB	52.42%	64.78%	53.20%	53.51%
-	LR	63.91%	69.90%	65.20%	66.77%
-	RFC	63.02%	69.30%	64.30%	66.00%
-	SVC (linear)	64.27%	69.90%	65.50%	66.95%
-	SVC (radial)	35.29%	62.56%	53.05%	44.97%
IG	GNB	65.13%	71.98%	60.45%	60.33%
IG	LR	67.56%	75.60%	69.60%	71.33%
IG	RFC	65.57%	72.10%	67.10%	68.67%
IG	SVC (linear)	66.63%	76.90%	69.20%	71.39%
IG	SVC (radial)	35.21%	63.16%	53.00%	45.09%
GR	GNB	65.13%	71.98%	60.45%	60.33%
GR	LR	67.56%	75.60%	69.60%	71.33%
GR	RFC	64.80%	73.70%	66.75%	68.92%
GR	SVC (linear)	66.63%	76.90%	69.20%	71.39%
GR	SVC (radial)	35.21%	63.16%	53.00%	45.09%
CS	GNB	57.71%	91.50%	61.90%	70.68%
CS	LR	66.34%	73.40%	68.05%	69.69%
CS	RFC	64.30%	72.60%	66.10%	68.18%
CS	SVC (linear)	65.65%	72.50%	67.30%	68.89%
CS	SVC (radial)	35.62%	61.66%	53.40%	44.93%
PMI	GNB	66.91%	69.67%	59.05%	56.77%
PMI	LR	59.77%	90.00%	64.70%	71.83%
PMI	RFC	60.18%	89.40%	65.10%	71.92%
PMI	SVC (linear)	57.57%	90.90%	61.95%	70.49%
PMI	SVC (radial)	33.54%	66.67%	50.40%	44.62%
CPD	GNB	66.91%	69.67%	59.05%	56.77%
CPD	LR	59.77%	90.00%	64.70%	71.83%

CPD	RFC	60.48%	89.20%	65.45%	72.08%
CPD	SVC (linear)	57.57%	90.90%	61.95%	70.49%
CPD	SVC (radial)	33.54%	66.67%	50.40%	44.62%
JSD	GNB	53.70%	95.00%	56.55%	68.62%
JSD	LR	66.40%	73.60%	68.15%	69.81%
JSD	RFC	63.72%	71.10%	65.30%	67.20%
JSD	SVC (linear)	65.48%	72.70%	67.20%	68.88%
JSD	SVC (radial)	35.46%	62.26%	53.25%	45.00%

Aspect Extraction (FE4)

FS Type	Prediction	Precision	Recall	Accuracy	F-Measure
-	GNB	59.76%	63.60%	60.40%	61.59%
-	LR	60.34%	59.00%	60.10%	59.64%
-	RFC	58.88%	56.10%	58.45%	57.45%
-	SVC (linear)	60.37%	58.50%	60.05%	59.40%
-	SVC (radial)	59.49%	70.89%	61.25%	64.60%
IG	GNB	59.76%	63.60%	60.40%	61.59%
IG	LR	60.34%	59.00%	60.10%	59.64%
IG	RFC	59.66%	56.70%	59.15%	58.11%
IG	SVC (linear)	60.37%	58.50%	60.05%	59.40%
IG	SVC (radial)	59.49%	70.89%	61.25%	64.60%
GR	GNB	59.76%	63.60%	60.40%	61.59%
GR	LR	60.34%	59.00%	60.10%	59.64%
GR	RFC	57.02%	56.40%	56.95%	56.71%
GR	SVC (linear)	60.37%	58.50%	60.05%	59.40%
GR	SVC (radial)	59.49%	70.89%	61.25%	64.60%
CS	GNB	59.76%	63.60%	60.40%	61.59%
CS	LR	60.34%	59.00%	60.10%	59.64%
CS	RFC	57.71%	57.40%	57.70%	57.54%
CS	SVC (linear)	60.37%	58.50%	60.05%	59.40%

CS	SVC (radial)	59.49%	70.89%	61.25%	64.60%
PMI	GNB	59.76%	63.60%	60.40%	61.59%
PMI	LR	60.34%	59.00%	60.10%	59.64%
PMI	RFC	58.50%	56.20%	58.15%	57.32%
PMI	SVC (linear)	60.37%	58.50%	60.05%	59.40%
PMI	SVC (radial)	59.49%	70.89%	61.25%	64.60%
CPD	GNB	59.76%	63.60%	60.40%	61.59%
CPD	LR	60.34%	59.00%	60.10%	59.64%
CPD	RFC	57.75%	53.50%	57.20%	55.52%
CPD	SVC (linear)	60.37%	58.50%	60.05%	59.40%
CPD	SVC (radial)	59.49%	70.89%	61.25%	64.60%
JSD	GNB	59.76%	63.60%	60.40%	61.59%
JSD	LR	60.34%	59.00%	60.10%	59.64%
JSD	RFC	58.30%	56.50%	58.05%	57.35%
JSD	SVC (linear)	60.37%	58.50%	60.05%	59.40%
JSD	SVC (radial)	59.49%	70.89%	61.25%	64.60%

FE3 (A) +FE4

FS Type	Prediction	Precision	Recall	Accuracy	F-Measure
-	GNB	52.89%	65.18%	53.40%	54.07%
-	LR	67.73%	67.00%	67.55%	67.33%
-	RFC	64.49%	64.50%	64.45%	64.47%
-	SVC (linear)	66.81%	65.30%	66.35%	66.00%
-	SVC (radial)	34.92%	63.46%	52.60%	44.95%
IG	GNB	63.66%	83.59%	64.65%	70.12%
IG	LR	68.91%	69.40%	69.05%	69.15%
IG	RFC	66.21%	66.19%	66.25%	66.20%
IG	SVC (linear)	69.00%	69.30%	69.10%	69.14%
IG	SVC (radial)	56.28%	91.11%	59.70%	69.38%
GR	GNB	63.66%	83.59%	64.65%	70.12%
GR	LR	68.91%	69.40%	69.05%	69.15%
GR	RFC	64.62%	67.60%	65.30%	66.07%

GR	SVC (linear)	69.00%	69.30%	69.10%	69.14%
GR	SVC (radial)	56.28%	91.11%	59.70%	69.38%
CS	GNB	58.55%	91.10%	63.05%	71.20%
CS	LR	69.17%	68.70%	69.05%	68.91%
CS	RFC	63.67%	65.10%	63.90%	64.34%
CS	SVC (linear)	67.91%	67.90%	67.90%	67.90%
CS	SVC (radial)	56.49%	90.21%	59.85%	69.26%
PMI	GNB	57.61%	94.30%	62.15%	71.43%
PMI	LR	60.42%	90.80%	65.65%	72.55%
PMI	RFC	60.37%	90.00%	65.45%	72.26%
PMI	SVC (linear)	58.33%	90.70%	62.95%	71.00%
PMI	SVC (radial)	34.28%	65.17%	51.70%	44.90%
CPD	GNB	57.61%	94.30%	62.15%	71.43%
CPD	LR	60.42%	90.80%	65.65%	72.55%
CPD	RFC	60.26%	89.30%	65.20%	71.96%
CPD	SVC (linear)	58.33%	90.70%	62.95%	71.00%
CPD	SVC (radial)	34.28%	65.17%	51.70%	44.90%
JSD	GNB	53.67%	95.00%	56.50%	68.59%
JSD	LR	69.48%	68.80%	69.30%	69.11%
JSD	RFC	65.43%	67.70%	65.95%	66.54%
JSD	SVC (linear)	68.03%	68.50%	68.15%	68.26%
JSD	SVC (radial)	56.35%	90.31%	59.70%	69.19%

FE2+FE3 (A) +FE4

FS Type	Prediction	Precision	Recall	Accuracy	F-Measure
-	GNB	60.65%	74.89%	62.15%	66.11%
-	LR	73.92%	73.40%	73.70%	73.61%
-	RFC	68.75%	59.90%	66.35%	63.96%
-	SVC (linear)	70.83%	70.00%	70.55%	70.39%
-	SVC (radial)	55.33%	87.60%	58.35%	67.80%
IG	GNB	73.51%	87.70%	77.95%	79.92%
IG	LR	80.63%	81.10%	80.80%	80.85%
IG	RFC	70.86%	61.30%	68.00%	65.66%
IG	SVC (linear)	79.35%	80.70%	79.85%	80.00%

IG	SVC (radial)	69.05%	79.40%	71.90%	73.85%
GR	GNB	73.51%	87.70%	77.95%	79.92%
GR	LR	80.63%	81.10%	80.80%	80.85%
GR	RFC	73.46%	67.10%	71.40%	70.11%
GR	SVC (linear)	79.35%	80.70%	79.85%	80.00%
GR	SVC (radial)	69.05%	79.40%	71.90%	73.85%
CS	GNB	72.18%	88.10%	77.00%	79.31%
CS	LR	80.94%	81.50%	81.15%	81.21%
CS	RFC	72.18%	64.30%	69.75%	68.01%
CS	SVC (linear)	78.41%	80.60%	79.20%	79.48%
CS	SVC (radial)	69.14%	79.30%	71.95%	73.85%
PMI	GNB	60.86%	97.00%	67.00%	74.71%
PMI	LR	62.34%	95.70%	68.85%	75.47%
PMI	RFC	62.63%	94.90%	69.05%	75.43%
PMI	SVC (linear)	61.00%	94.90%	67.05%	74.24%
PMI	SVC (radial)	33.72%	66.57%	50.75%	44.76%
CPD	GNB	60.86%	97.00%	67.00%	74.71%
CPD	LR	62.34%	95.70%	68.85%	75.47%
CPD	RFC	62.91%	94.90%	69.40%	75.64%
CPD	SVC (linear)	61.00%	94.90%	67.05%	74.24%
CPD	SVC (radial)	33.72%	66.57%	50.75%	44.76%
JSD	GNB	71.49%	88.70%	76.60%	79.13%
JSD	LR	81.01%	81.40%	81.15%	81.19%
JSD	RFC	74.35%	66.59%	71.80%	70.18%
JSD	SVC (linear)	78.79%	81.00%	79.60%	79.87%
JSD	SVC (radial)	69.09%	79.30%	71.90%	73.82%

FE1+FE2+FE3 (A) +FE4

FS Type	Prediction	Precision	Recall	Accuracy	F-Measure
-	GNB	60.53%	73.89%	62.00%	65.92%
-	Logistic Regression	75.62%	74.80%	75.25%	75.13%
-	RFC	64.83%	51.40%	61.75%	57.33%
-	SVC (linear)	72.99%	71.70%	72.55%	72.31%
-	SVC (radial)	57.50%	82.80%	60.80%	67.87%
IG	GNB	72.38%	87.30%	76.95%	79.10%
IG	Logistic Regression	80.07%	80.90%	80.35%	80.46%
IG	RFC	72.36%	63.10%	69.50%	67.40%
IG	SVC (linear)	78.44%	80.70%	79.25%	79.54%
IG	SVC (radial)	66.34%	78.10%	69.20%	71.73%
GR	GNB	72.38%	87.30%	76.95%	79.10%
GR	Logistic Regression	80.07%	80.90%	80.35%	80.46%
GR	RFC	73.70%	62.00%	69.85%	67.29%
GR	SVC (linear)	78.44%	80.70%	79.25%	79.54%
GR	SVC (radial)	66.34%	78.10%	69.20%	71.73%
CS	GNB	71.47%	87.30%	76.20%	78.56%
CS	Logistic Regression	79.88%	81.50%	80.45%	80.66%
CS	RFC	74.22%	63.30%	70.65%	68.32%
CS	SVC (linear)	78.90%	80.90%	79.60%	79.85%

CS	SVC (radial)	66.70%	77.80%	69.45%	71.81%
PMI	GNB	60.76%	97.30%	66.95%	74.73%
PMI	Logistic Regression	62.27%	95.80%	68.80%	75.46%
PMI	RFC	62.91%	94.80%	69.40%	75.62%
PMI	SVC (linear)	60.97%	95.10%	67.05%	74.28%
PMI	SVC (radial)	34.62%	66.17%	52.35%	45.44%
CPD	GNB	60.76%	97.30%	66.95%	74.73%
CPD	Logistic Regression	62.27%	95.80%	68.80%	75.46%
CPD	RFC	62.76%	95.20%	69.30%	75.64%
CPD	SVC (linear)	60.97%	95.10%	67.05%	74.28%
CPD	SVC (radial)	34.62%	66.17%	52.35%	45.44%
JSD	GNB	72.28%	87.40%	76.90%	79.09%
JSD	Logistic Regression	79.78%	80.90%	80.15%	80.31%
JSD	RFC	70.79%	62.80%	68.40%	66.53%
JSD	SVC (linear)	78.29%	80.70%	79.15%	79.47%
JSD	SVC (radial)	66.47%	78.00%	69.30%	71.77%

TF=b, IDF=Default (=1)

FS Type	Prediction	Precision	Recall	Accuracy	F-Measure
-	GNB	65.75%	56.70%	63.60%	60.67%
-	Logistic Regression	75.67%	75.10%	75.45%	75.34%
-	RFC	69.32%	64.50%	68.05%	66.76%
-	SVC (linear)	73.18%	71.40%	72.60%	72.27%
-	SVC (radial)	35.38%	63.26%	53.25%	45.24%
IG	Logistic Regression	82.48%	81.90%	82.25%	82.18%
IG	SVC (linear)	81.14%	81.70%	81.35%	81.41%
GR	Logistic Regression	57.70%	86.90%	61.60%	69.35%
GR	SVC (linear)	56.41%	88.50%	60.05%	68.90%
CS	Logistic Regression	82.53%	81.70%	82.20%	82.11%
CS	SVC (linear)	81.01%	81.40%	81.15%	81.18%
PMI	Logistic Regression	67.48%	100.00%	75.90%	80.58%
PMI	SVC (linear)	64.52%	100.00%	72.50%	78.43%
CPD	Logistic Regression	67.48%	100.00%	75.90%	80.58%
CPD	SVC (linear)	64.52%	100.00%	72.50%	78.43%
JSD	Logistic	82.16%	82.00%	82.10%	82.07%

	Regression				
JSD	SVC (linear)	81.24%	82.10%	81.55%	81.64%

TF=b, IDF=s (SmoothedIDF)

FS Type	Bin	Prediction	Precision	Recall	Accuracy	F-Measure
-	0	GNB	91.49%	78.30%	85.50%	84.37%
-	1	GNB	82.55%	65.50%	75.85%	73.03%
-	0	Logistic Regression	87.69%	91.20%	89.20%	89.41%
-	1	Logistic Regression	71.70%	70.20%	71.20%	70.90%
-	0	RFC	67.84%	62.00%	66.35%	64.70%
-	1	RFC	67.56%	61.80%	66.10%	64.51%
-	0	SVC (linear)	80.42%	78.40%	79.65%	79.39%
-	1	SVC (linear)	74.05%	72.40%	73.50%	73.21%
-	0	SVC (radial)	78.75%	37.03%	52.15%	29.22%
-	1	SVC (radial)	35.45%	63.26%	53.35%	45.30%
IG	0	Logistic Regression	84.62%	85.50%	84.95%	85.02%
IG	1	Logistic Regression	77.57%	76.20%	77.05%	76.86%
IG	0	SVC (linear)	79.06%	82.50%	80.30%	80.70%
IG	1	SVC (linear)	81.14%	81.70%	81.35%	81.41%
GR	0	Logistic Regression	57.60%	88.40%	61.65%	69.75%
GR	1	Logistic Regression	50.30%	65.80%	50.40%	57.02%

GR	0	SVC (linear)	55.44%	88.20%	58.65%	68.08%
GR	1	SVC (linear)	56.40%	88.60%	60.05%	68.92%
CS	0	Logistic Regression	85.22%	85.40%	85.25%	85.26%
CS	1	Logistic Regression	76.61%	76.90%	76.70%	76.75%
CS	0	SVC (linear)	79.80%	80.50%	80.00%	80.08%
CS	1	SVC (linear)	80.85%	81.80%	81.20%	81.30%
PMI	0	Logistic Regression	66.41%	100.00%	74.70%	79.81%
PMI	1	Logistic Regression	66.54%	100.00%	74.85%	79.91%
PMI	0	SVC (linear)	63.02%	100.00%	70.65%	77.31%
PMI	1	SVC (linear)	64.15%	100.00%	72.05%	78.16%
CPD	0	Logistic Regression	66.41%	100.00%	74.70%	79.81%
CPD	1	Logistic Regression	66.54%	100.00%	74.85%	79.91%
CPD	0	SVC (linear)	63.02%	100.00%	70.65%	77.31%
CPD	1	SVC (linear)	64.15%	100.00%	72.05%	78.16%
JSD	0	Logistic Regression	84.33%	85.80%	84.80%	84.95%
JSD	1	Logistic Regression	77.14%	76.80%	77.00%	76.95%
JSD	0	SVC (linear)	79.80%	82.50%	80.70%	81.07%
JSD	1	SVC (linear)	81.24%	82.10%	81.55%	81.64%

TF=b(Boolean) IDF=sp (SmoothedProbIDF)

FS Type	Bin	Prediction	Precision	Recall	Accuracy	F-Measure
-	0	GNB	91.49%	78.30%	85.50%	84.37%
-	1	GNB	82.55%	65.50%	75.85%	73.03%
-	0	Logistic Regression	87.83%	88.80%	88.25%	88.31%
-	1	Logistic Regression	71.70%	70.20%	71.20%	70.90%
-	0	RFC	65.64%	61.40%	64.65%	63.41%
-	1	RFC	68.29%	67.00%	67.95%	67.60%
-	0	SVC (linear)	79.92%	76.60%	78.65%	78.21%
-	1	SVC (linear)	74.05%	72.40%	73.50%	73.21%
-	0	SVC (radial)	78.75%	37.53%	52.40%	30.04%
-	1	SVC (radial)	35.45%	63.26%	53.35%	45.30%
IG	0	Logistic Regression	85.06%	85.80%	85.35%	85.41%
IG	1	Logistic Regression	77.57%	76.20%	77.05%	76.86%
IG	0	SVC (linear)	79.62%	82.60%	80.70%	81.05%
IG	1	SVC (linear)	81.14%	81.70%	81.35%	81.41%
GR	0	Logistic Regression	57.60%	88.40%	61.65%	69.75%
GR	1	Logistic Regression	50.30%	65.80%	50.40%	57.02%
GR	0	SVC (linear)	55.44%	88.20%	58.65%	68.08%
GR	1	SVC (linear)	56.40%	88.60%	60.05%	68.92%

CS	0	Logistic Regression	85.66%	85.20%	85.45%	85.40%
CS	1	Logistic Regression	76.61%	76.90%	76.70%	76.75%
CS	0	SVC (linear)	79.84%	80.10%	79.90%	79.93%
CS	1	SVC (linear)	80.85%	81.80%	81.20%	81.30%
PMI	0	Logistic Regression	66.41%	100.00%	74.70%	79.81%
PMI	1	Logistic Regression	66.54%	100.00%	74.85%	79.91%
PMI	0	SVC (linear)	63.02%	100.00%	70.65%	77.31%
PMI	1	SVC (linear)	64.15%	100.00%	72.05%	78.16%
CPD	0	Logistic Regression	66.41%	100.00%	74.70%	79.81%
CPD	1	Logistic Regression	66.54%	100.00%	74.85%	79.91%
CPD	0	SVC (linear)	63.02%	100.00%	70.65%	77.31%
CPD	1	SVC (linear)	64.15%	100.00%	72.05%	78.16%
JSD	0	Logistic Regression	84.38%	85.00%	84.55%	84.62%
JSD	1	Logistic Regression	77.14%	76.80%	77.00%	76.95%
JSD	0	SVC (linear)	79.77%	81.80%	80.45%	80.72%
JSD	1	SVC (linear)	81.24%	82.10%	81.55%	81.64%

TF= a (Augmented), IDF=d(Default = 1)

FS Type	Bin	Prediction	Precision	Recall	Accuracy	F-Measure
-	0	GNB	63.78%	59.30%	62.80%	61.31%
-	1	GNB	59.47%	61.49%	59.55%	59.84%
-	0	Logistic Regression	56.08%	57.49%	56.20%	55.14%

-	1	Logistic Regression	64.53%	69.80%	65.70%	67.04%
-	0	RFC	68.03%	65.10%	67.25%	66.46%
-	1	RFC	58.42%	71.30%	60.15%	64.18%
-	0	SVC (linear)	30.93%	33.93%	49.85%	23.35%
-	1	SVC (linear)	61.73%	69.00%	63.10%	65.15%
-	0	SVC (radial)	29.46%	33.83%	49.80%	23.16%
-	1	SVC (radial)	33.43%	66.67%	50.20%	44.53%
IG	0	Logistic Regression	33.38%	66.47%	50.10%	44.44%
IG	1	Logistic Regression	69.20%	93.60%	75.95%	79.56%
IG	0	SVC (linear)	33.31%	66.67%	49.95%	44.42%
IG	1	SVC (linear)	75.51%	78.78%	73.80%	73.54%
GR	0	Logistic Regression	33.38%	66.47%	50.10%	44.44%
GR	1	Logistic Regression	65.22%	49.26%	58.45%	50.93%
GR	0	SVC (linear)	33.31%	66.67%	49.95%	44.42%
GR	1	SVC (linear)	64.12%	48.46%	57.55%	49.50%
CS	0	Logistic Regression	33.38%	66.47%	50.10%	44.44%
CS	1	Logistic Regression	70.40%	91.10%	76.40%	79.42%
CS	0	SVC (linear)	33.31%	66.67%	49.95%	44.42%
CS	1	SVC (linear)	68.68%	91.00%	74.75%	78.27%
PMI	0	Logistic Regression	33.38%	66.47%	50.10%	44.44%
PMI	1	Logistic Regression	65.75%	100.00%	73.95%	79.34%

PMI	0	SVC (linear)	33.31%	66.67%	49.95%	44.42%
PMI	1	SVC (linear)	62.78%	100.00%	70.35%	77.13%
CPD	0	Logistic Regression	33.38%	66.47%	50.10%	44.44%
CPD	1	Logistic Regression	65.75%	100.00%	73.95%	79.34%
CPD	0	SVC (linear)	33.31%	66.67%	49.95%	44.42%
CPD	1	SVC (linear)	62.78%	100.00%	70.35%	77.13%
JSD	0	Logistic Regression	56.74%	57.29%	56.75%	55.37%
JSD	1	Logistic Regression	69.00%	93.60%	75.75%	79.43%
JSD	0	SVC (linear)	29.46%	33.83%	49.80%	23.16%
JSD	1	SVC (linear)	66.52%	93.10%	73.10%	77.58%

TF=a(Augmented), IDF=s (SmoothedIDF):

FS Type	Bin	Prediction	Precision	Recall	Accuracy	F-Measure
-	0	GNB	91.92%	81.10%	86.95%	86.10%
-	1	GNB	67.87%	53.70%	64.15%	59.93%
-	0	Logistic Regression	78.62%	95.90%	83.35%	85.80%
-	1	Logistic Regression	55.79%	82.40%	58.55%	66.53%
-	0	RFC	66.85%	63.00%	65.90%	64.84%
-	1	RFC	56.63%	84.90%	59.95%	67.94%
-	0	SVC (linear)	50.00%	100.00%	50.00%	66.67%
-	1	SVC (linear)	56.46%	84.70%	59.70%	67.76%

-	0	SVC (radial)	33.64%	66.67%	50.60%	44.72%
-	1	SVC (radial)	33.36%	66.67%	50.05%	44.47%
IG	0	Logistic Regression	16.64%	33.33%	49.95%	22.20%
IG	1	Logistic Regression	66.44%	89.50%	72.15%	76.26%
IG	0	SVC (linear)	16.64%	33.33%	49.95%	22.20%
IG	1	SVC (linear)	64.02%	92.20%	70.20%	75.56%
GR	0	Logistic Regression	16.64%	33.33%	49.95%	22.20%
GR	1	Logistic Regression	55.73%	77.40%	57.95%	64.80%
GR	0	SVC (linear)	16.64%	33.33%	49.95%	22.20%
GR	1	SVC (linear)	54.85%	84.30%	57.45%	66.46%
CS	0	Logistic Regression	16.64%	33.33%	49.95%	22.20%
CS	1	Logistic Regression	66.11%	88.40%	71.55%	75.64%
CS	0	SVC (linear)	16.64%	33.33%	49.95%	22.20%
CS	1	SVC (linear)	64.07%	92.40%	70.30%	75.67%
PMI	0	Logistic Regression	16.64%	33.33%	49.95%	22.20%
PMI	1	Logistic Regression	65.63%	100.00%	73.80%	79.24%
PMI	0	SVC (linear)	16.64%	33.33%	49.95%	22.20%
PMI	1	SVC (linear)	62.74%	100.00%	70.30%	77.10%
CPD	0	Logistic Regression	16.64%	33.33%	49.95%	22.20%
CPD	1	Logistic	65.63%	100.00%	73.80%	79.24%

		Regression				
CPD	0	SVC (linear)	16.64%	33.33%	49.95%	22.20%
CPD	1	SVC (linear)	62.74%	100.00%	70.30%	77.10%
JSD	0	Logistic Regression	73.42%	60.65%	60.11%	52.37%
JSD	1	Logistic Regression	66.54%	89.50%	72.25%	76.33%
JSD	0	SVC (linear)	33.67%	66.67%	50.65%	44.74%
JSD	1	SVC (linear)	63.92%	92.50%	70.15%	75.59%

TF=a(Augmented), IDF=sp(SmoothedProbIDF):

FS Type	Bin	Prediction	Precision	Recall	Accuracy	F-Measure
-	0	GNB	91.92%	81.10%	86.95%	86.10%
-	1	GNB	67.21%	53.50%	63.75%	59.55%
-	0	Logistic Regression	69.65%	83.00%	72.85%	75.38%
-	1	Logistic Regression	58.23%	83.50%	61.80%	68.61%
-	0	RFC	66.67%	66.20%	66.55%	66.43%
-	1	RFC	57.76%	82.50%	61.10%	67.94%
-	0	SVC (linear)	50.15%	100.00%	50.30%	66.80%
-	1	SVC (linear)	57.36%	83.50%	60.70%	68.00%
-	0	SVC (radial)	33.74%	66.67%	50.80%	44.81%
-	1	SVC (radial)	33.36%	66.67%	50.05%	44.47%
IG	0	Logistic Regression	16.64%	33.33%	49.95%	22.20%
IG	1	Logistic	67.20%	88.10%	72.55%	76.24%

		Regression				
IG	0	SVC (linear)	16.64%	33.33%	49.95%	22.20%
IG	1	SVC (linear)	64.71%	90.40%	70.55%	75.43%
GR	0	Logistic Regression	16.64%	33.33%	49.95%	22.20%
GR	1	Logistic Regression	55.57%	77.30%	57.75%	64.66%
GR	0	SVC (linear)	16.64%	33.33%	49.95%	22.20%
GR	1	SVC (linear)	54.81%	84.90%	57.45%	66.62%
CS	0	Logistic Regression	16.64%	33.33%	49.95%	22.20%
CS	1	Logistic Regression	67.10%	86.70%	72.10%	75.65%
CS	0	SVC (linear)	16.64%	33.33%	49.95%	22.20%
CS	1	SVC (linear)	65.36%	89.40%	71.00%	75.51%
PMI	0	Logistic Regression	16.64%	33.33%	49.95%	22.20%
PMI	1	Logistic Regression	65.55%	100.00%	73.70%	79.18%
PMI	0	SVC (linear)	16.64%	33.33%	49.95%	22.20%
PMI	1	SVC (linear)	62.90%	100.00%	70.50%	77.22%
CPD	0	Logistic Regression	16.64%	33.33%	49.95%	22.20%
CPD	1	Logistic Regression	65.55%	100.00%	73.70%	79.18%
CPD	0	SVC (linear)	16.64%	33.33%	49.95%	22.20%
CPD	1	SVC (linear)	62.90%	100.00%	70.50%	77.22%
JSD	0	Logistic Regression	58.33%	69.11%	59.45%	62.83%
JSD	1	Logistic Regression	66.47%	88.40%	71.90%	75.88%

JSD	0	SVC (linear)	33.43%	66.57%	50.20%	44.51%
JSD	1	SVC (linear)	64.23%	90.30%	70.00%	75.06%

TF=o(bm25), IDF=d (Default = 1)

FS Type	Bin	Prediction	Precision	Recall	Accuracy	F-Measure
-	0	GNB	64.43%	58.00%	63.00%	61.04%
-	1	GNB	65.92%	57.20%	63.80%	61.23%
-	0	LR	76.59%	75.40%	76.15%	75.94%
-	1	LR	74.90%	75.20%	75.00%	75.04%
-	0	RFC	65.96%	63.20%	65.25%	64.46%
-	1	RFC	67.21%	63.70%	66.30%	65.39%
-	0	SVC (linear)	74.73%	73.20%	74.20%	73.93%
-	1	SVC (linear)	72.63%	71.30%	72.20%	71.95%
-	0	SVC (radial)	39.85%	57.66%	56.81%	45.95%
-	1	SVC (radial)	36.40%	63.06%	54.60%	45.94%
IG	0	LR	82.78%	82.70%	82.75%	82.73%
IG	1	LR	82.04%	83.50%	82.60%	82.76%
IG	0	SVC (linear)	80.38%	82.30%	81.10%	81.31%
IG	1	SVC (linear)	82.30%	84.10%	83.00%	83.19%
GR	0	LR	57.69%	86.70%	61.55%	69.28%

GR	1	LR	68.35%	55.37%	63.00%	57.46%
GR	0	SVC (linear)	56.55%	87.20%	60.10%	68.61%
GR	1	SVC (linear)	65.74%	52.66%	60.65%	54.22%
CS	0	LR	82.30%	82.70%	82.45%	82.48%
CS	1	LR	81.25%	82.80%	81.85%	82.01%
CS	0	SVC (linear)	80.03%	82.20%	80.85%	81.08%
CS	1	SVC (linear)	81.84%	81.40%	81.65%	81.60%
PMI	0	LR	67.48%	100.00%	75.90%	80.58%
PMI	1	LR	70.93%	100.00%	79.50%	82.99%
PMI	0	SVC (linear)	65.71%	100.00%	73.90%	79.31%
PMI	1	SVC (linear)	67.66%	100.00%	76.10%	80.71%
CPD	0	Logistic Regression	67.48%	100.00%	75.90%	80.58%
CPD	1	Logistic Regression	70.93%	100.00%	79.50%	82.99%
CPD	0	SVC (linear)	65.71%	100.00%	73.90%	79.31%
CPD	1	SVC (linear)	67.66%	100.00%	76.10%	80.71%
JSD	0	Logistic Regression	82.93%	83.10%	83.00%	83.01%
JSD	1	Logistic Regression	82.16%	83.30%	82.60%	82.72%
JSD	0	SVC (linear)	80.53%	82.60%	81.30%	81.54%
JSD	1	SVC (linear)	82.49%	83.80%	83.00%	83.14%

TF=o(bm25), IDF=s(SmoothedIDF)

FS Type	Bin	Prediction	Precision	Recall	Accuracy	F-Measure
-	0	GNB	90.96%	81.10%	86.50%	85.73%
-	1	GNB	75.90%	65.10%	72.20%	70.05%
-	0	Logistic Regression	89.78%	92.10%	90.80%	90.92%
-	1	Logistic Regression	69.43%	68.50%	69.15%	68.95%
-	0	RFC	65.63%	67.80%	66.15%	66.68%
-	1	RFC	67.75%	65.50%	67.15%	66.60%
-	0	SVC (linear)	81.98%	84.20%	82.85%	83.07%
-	1	SVC (linear)	73.28%	72.10%	72.90%	72.68%
-	0	SVC (radial)	38.17%	66.57%	57.46%	48.33%
-	1	SVC (radial)	36.43%	63.16%	54.65%	45.99%
IG	0	Logistic Regression	85.65%	86.49%	85.95%	85.99%
IG	1	Logistic Regression	74.79%	77.40%	75.65%	76.05%
IG	0	SVC (linear)	79.84%	82.20%	80.65%	80.93%
IG	1	SVC (linear)	82.19%	83.90%	82.85%	83.03%
GR	0	Logistic Regression	57.84%	88.20%	61.95%	69.86%
GR	1	Logistic Regression	51.91%	64.50%	52.40%	57.52%
GR	0	SVC (linear)	55.91%	87.50%	59.25%	68.23%
GR	1	SVC (linear)	65.71%	52.66%	60.70%	54.29%
CS	0	Logistic Regression	85.08%	86.00%	85.40%	85.49%

CS	1	Logistic Regression	73.95%	75.20%	74.35%	74.57%
CS	0	SVC (linear)	79.37%	80.80%	79.85%	80.03%
CS	1	SVC (linear)	81.90%	80.80%	81.45%	81.32%
PMI	0	Logistic Regression	66.98%	100.00%	75.35%	80.23%
PMI	1	Logistic Regression	69.74%	100.00%	78.25%	82.16%
PMI	0	SVC (linear)	63.30%	100.00%	71.00%	77.52%
PMI	1	SVC (linear)	67.39%	100.00%	75.80%	80.52%
CPD	0	Logistic Regression	66.98%	100.00%	75.35%	80.23%
CPD	1	Logistic Regression	69.74%	100.00%	78.25%	82.16%
CPD	0	SVC (linear)	63.30%	100.00%	71.00%	77.52%
CPD	1	SVC (linear)	67.39%	100.00%	75.80%	80.52%
JSD	0	Logistic Regression	84.32%	85.70%	84.80%	84.93%
JSD	1	Logistic Regression	74.35%	76.50%	75.05%	75.40%
JSD	0	SVC (linear)	78.45%	81.40%	79.40%	79.80%
JSD	1	SVC (linear)	82.06%	83.60%	82.65%	82.82%

TF=o(bm25), IDF=sp(SmoothedProbIDF)

FS Type	Bin	Prediction	Precision	Recall	Accuracy	F-Measure
-	0	GNB	90.96%	81.10%	86.50%	85.73%
-	1	GNB	75.90%	65.10%	72.20%	70.05%
-	0	Logistic	89.35%	89.80%	89.55%	89.57%

		Regression				
-	1	Logistic Regression	69.43%	68.50%	69.15%	68.95%
-	0	RFC	66.41%	62.20%	65.30%	64.20%
-	1	RFC	67.31%	67.29%	67.30%	67.21%
-	0	SVC (linear)	81.28%	82.00%	81.55%	81.63%
-	1	SVC (linear)	73.28%	72.10%	72.90%	72.68%
-	0	SVC (radial)	39.70%	66.37%	59.11%	49.36%
-	1	SVC (radial)	36.43%	63.16%	54.65%	45.99%
IG	0	Logistic Regression	85.66%	85.99%	85.75%	85.76%
IG	1	Logistic Regression	74.79%	77.40%	75.65%	76.05%
IG	0	SVC (linear)	80.60%	80.90%	80.65%	80.70%
IG	1	SVC (linear)	82.19%	83.90%	82.85%	83.03%
GR	0	Logistic Regression	57.80%	88.20%	61.90%	69.83%
GR	1	Logistic Regression	51.91%	64.50%	52.40%	57.52%
GR	0	SVC (linear)	55.91%	87.50%	59.25%	68.23%
GR	1	SVC (linear)	65.71%	52.66%	60.70%	54.29%
CS	0	Logistic Regression	85.32%	85.30%	85.25%	85.26%
CS	1	Logistic Regression	73.95%	75.20%	74.35%	74.57%
CS	0	SVC (linear)	79.73%	80.80%	80.05%	80.19%
CS	1	SVC (linear)	81.90%	80.80%	81.45%	81.32%
PMI	0	Logistic Regression	66.98%	100.00%	75.35%	80.23%
PMI	1	Logistic	69.74%	100.00%	78.25%	82.16%

		Regression				
PMI	0	SVC (linear)	63.30%	100.00%	71.00%	77.52%
PMI	1	SVC (linear)	67.39%	100.00%	75.80%	80.52%
CPD	0	Logistic Regression	66.98%	100.00%	75.35%	80.23%
CPD	1	Logistic Regression	69.74%	100.00%	78.25%	82.16%
CPD	0	SVC (linear)	63.30%	100.00%	71.00%	77.52%
CPD	1	SVC (linear)	67.39%	100.00%	75.80%	80.52%
JSD	0	Logistic Regression	84.48%	85.90%	85.00%	85.13%
JSD	1	Logistic Regression	74.35%	76.50%	75.05%	75.40%
JSD	0	SVC (linear)	78.04%	80.80%	78.90%	79.31%
JSD	1	SVC (linear)	82.06%	83.60%	82.65%	82.82%