**Syracuse University**

**IST-736 Assignment 7**

ThulasiRam RuppaKrishnan

IST 736

Professor Ami Gates

Contents

[Introduction 3](#_Toc18191913)

[Analysis and Models 6](#_Toc18191914)

[**About the data** 6](#_Toc18191915)

[**Models** 8](#_Toc18191916)

[Results 22](#_Toc18191917)

[Conclusion 24](#_Toc18191918)

## 

## **Introduction**

A review is an evaluation of a publication, service, or company such as a movie (a movie review), video game (video game review), musical composition (music review of a composition or recording), book (book review); a piece of hardware like a car, home appliance, or computer; or an event or performance, such as a live music concert, play, musical theater show, dance show, or art exhibition. In addition to a critical evaluation, the review's author may assign the work a rating to indicate its relative merit. More loosely, an author may review current events, trends, or items in the news. A compilation of reviews may itself be called a review. The New York Review of Books, for instance, is a collection of essays on literature, culture, and current affairs.

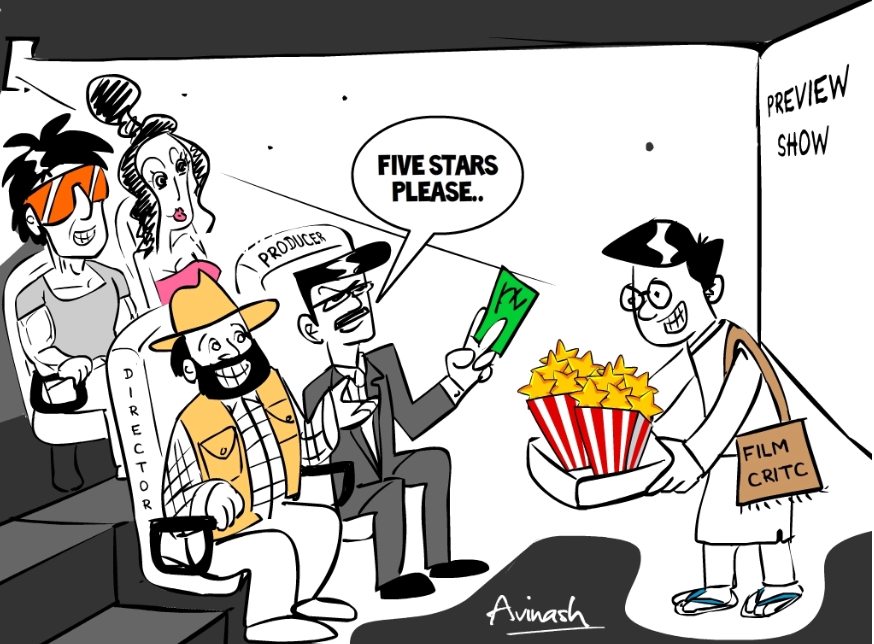


A user review refers to a review written by a user or consumer for a product or a service based on her experience as a user of the reviewed product. Popular sources for consumer reviews are e-commerce sites like Amazon.com, Zappos or lately in the Yoga field for schools such as Banjaara Yoga and Ayurveda, and social media sites like TripAdvisor and Yelp. E-commerce sites often have consumer reviews for products and sellers separately. Usually, consumer reviews are in the form of several lines of texts accompanied by a numerical rating. This text is meant to aid in shopping decision of a prospective buyer. A consumer review of a product usually comments on how well the product measures up to expectations based on the specifications provided by the manufacturer or seller. It talks about performance, reliability, quality defects, if any, and value for money. Consumer review, also called 'word of mouth' and 'user generated content' differs from 'marketer generated content' in its evaluation from consumer or user point of view. Often it includes comparative evaluations against competing products. Observations are factual as well as subjective in nature. Consumer review of sellers usually comment on service experienced, and dependability or trustworthiness of the seller. Usually, it comments on factors such as timeliness of delivery, packaging, and correctness of delivered items, shipping charges, return services against promises made, and so on.

**Movie Review**

Film was introduced in the late 19th century. The earliest artistic criticism of film emerged in the early 1900s. The first paper to serve as a critique of film came out of The Optical Lantern and Cinematograph Journal, followed by the Bioscope in 1908.Film is a relatively new form of art, in comparison to music, literature and painting which have existed since ancient times. Early writing on film sought to argue that films could also be considered a form of art. In 1911, Ricciotto Canudo wrote a manifesto proclaiming cinema to be the "Sixth Art" (later "Seventh Art"). For many decades after, film was still being treated with less prestige than longer-established art forms.

By the 1920s, critics were analyzing film for its merit and value as more than just entertainment. The growing popularity of the medium caused major newspapers to start hiring film critics. In the 1930s, the film industry developed concepts of stardom and celebrity in relation to actors, which led to a rise in obsession with critics as well, to the point that they were often seen on "red carpet" and at major events with the actors. It was in the 1940s that new forms of criticism emerged. Essays analyzing films with a distinctive charm and style to persuade the reader of the critic's argument. It was the emergence of these styles that brought film criticism to the mainstream, gaining the attention of many popular magazines; this made film reviews and critiques an eventual staple among most print media. As the decades passed, the fame for critics grew and gave rise to household names among the craft like James Agee, Andrew Sarris, Pauline Kael



**Online film critics**

Blogging has also introduced opportunities for a new wave of amateur film critics to have their opinions heard. These review blogs may focus on one genre, director or actor, or encompass a much wider variety of films. Friends, friends of friends, or strangers can visit these blogsites, and can often leave their own comments about the movie and/or the author's review. Although much less frequented than their professional counterparts, these sites can gather a following of like-minded people who look to specific bloggers for reviews as they have found that the critic consistently exhibits an outlook very similar to their own. YouTube has also served as a platform for amateur film critics.

Some websites specialize in narrow aspects of film reviewing. For instance, there are sites that focus on specific content advisories for parents to judge a film's suitability for children. Others focus on a religious perspective (e.g. CAP Alert). Still others highlight more esoteric subjects such as the depiction of science in fiction films. One such example is Insultingly Stupid Movie Physics by Intuitor. Some online niche websites provide comprehensive coverage of the independent sector; usually adopting a style closer to print journalism. They tend to prohibit adverts and offer uncompromising opinions free of any commercial interest. Their film critics normally have an academic film background.

The Online Film Critics Society, an international professional association of Internet-based cinema reviewers, consists of writers from all over the world, while New York Film Critics Online members handle reviews in the New York tri-state area.

**User-submitted reviews**

Several websites allow Internet users to submit movie reviews and aggregate them into an average. Community-driven review sites have allowed the common movie goer to express their opinion on films. Many of these sites allow users to rate films on a 0 to 10 scale, while some rely on the star rating system of 1–5, 0–5 or 0–4 stars. The votes are then culled into an overall rating and ranking for any film. Some of these community driven review sites include Reviewer, Movie Attractions, Flixster, FilmCrave, Flickchart and Everyone's a Critic. Rotten Tomatoes and Metacritic aggregate both scores from accredited critics and those submitted by users.



On these online review sites, users generally only must register with the site in order to submit reviews. This means that they are a form of open access poll, and have the same advantages and disadvantages; notably, there is no guarantee that they will be a representative sample of the film's audience. In some cases, online review sites have produced wildly differing results to scientific polling of audiences.

## **Analysis and Models**

### **About the data**

The dataset is comprised of tab-separated files with phrases from the Rotten Tomatoes dataset. The train/test split has been preserved for the purposes of benchmarking, but the sentences have been shuffled from their original order. Each Sentence has been parsed into many phrases by the Stanford parser. Each phrase has a PhraseId. Each sentence has a SentenceId. Phrases that are repeated (such as short/common words) are only included once in the data.

**train.tsv** contains the phrases and their associated sentiment labels. We have additionally provided a SentenceId so that you can track which phrases belong to a single sentence.

**test.tsv** contains just phrases. You must assign a sentiment label to each phrase.

The sentiment labels are:

0 - negative

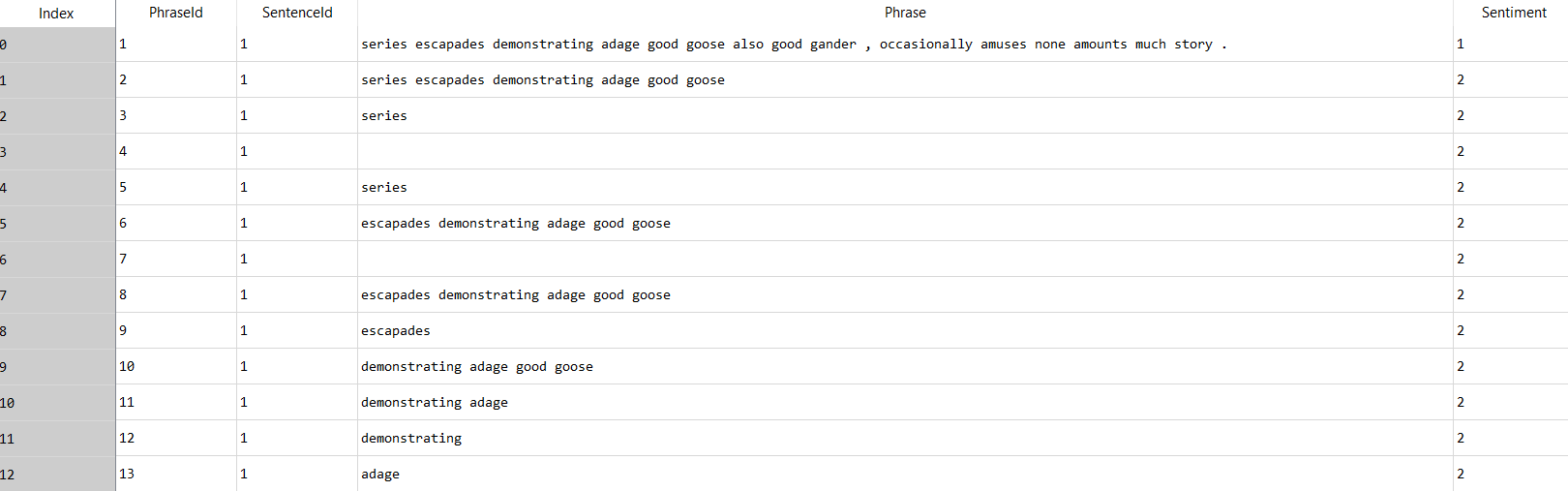
1 - somewhat negative

2 - neutral

3 - somewhat positive

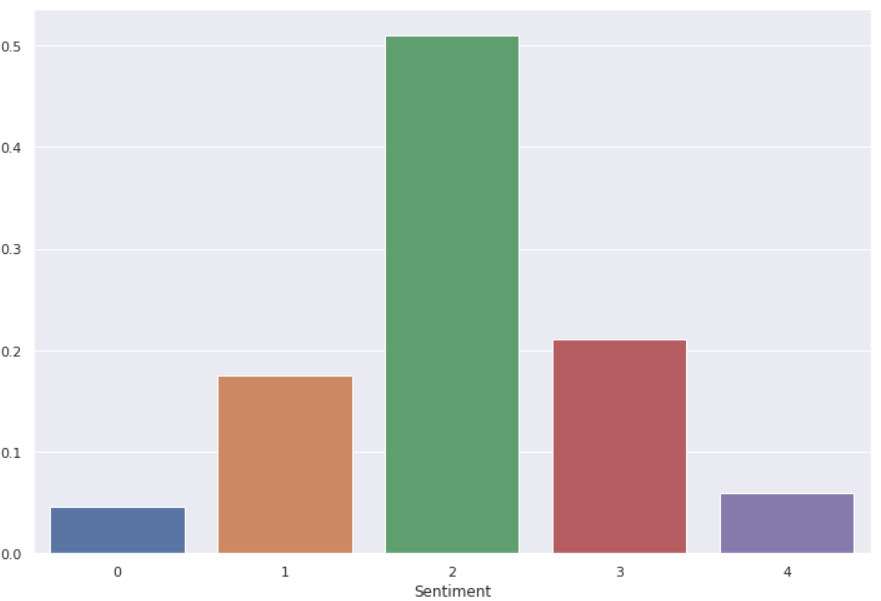
4 – positive

**Table 1.1** given below shows the sample dataset from the Rotten Tomatoes dataset.



**Table 1.1 Sample movie review dataset**

In the exploratory data analysis, word clouds are generated for every review and the positivity and negativity is tagged by using Sentiment Intensity analyzer. **Figure 1.2** shows word cloud and tagged sentiment intensity by each review and for the entire dataset. Also, **Figure 1.1** shows the sentiment distribution in the dataset and the labels are differentiated by color.

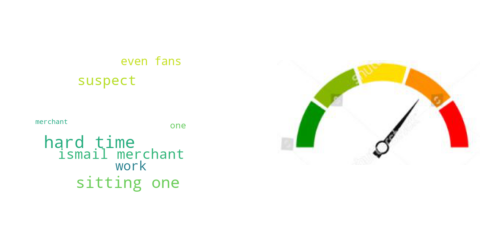


**Figure 1.1 Sentiment distribution in the dataset**

**SentimentId:1 SentimentId:2**



**SentimentId:3 SentimentId:4**





**Figure 1.2 Word Clouds for each review and the entire dataset**

### **Models**

In this exercise, models are developed using Naïve Bayes and SVM to compare their efficiency and accuracy in identifying sentiments on Rotten Tomatoes movie review dataset.

#### **Naïve Bayes Classification**

It is a classification technique based on Bayes’ Theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a specific feature in a class is unrelated to the presence of any other feature. For example, a fruit may be an apple if it is red, round, and about 3 inches in diameter. Even if these features depend on each other or upon the existence of the other features, all of these properties independently contribute to the probability that this fruit is an apple and that is why it is known as ‘Naive’.

Naive Bayes model is easy to build and particularly useful for very large data sets. Along with simplicity, Naive Bayes is known to outperform even highly sophisticated classification methods.

**Bayes theorem**

Bayes theorem provides a way of calculating posterior probability P(c|x) from P(c), P(x) and P(x|c). Look at the equation below:



* P(c|x) is the posterior probability of class (c, target) given predictor (x, attributes).
* P(c) is the prior probability of class.
* P(x|c) is the likelihood which is the probability of predictor given class.
* P(x) is the prior probability of predictor.

**Classification based on conditional probability**

To classify whether players will play or not based on weather condition using Naïve Bayes classification approach

Likelihood table Frequency Table are derived by finding the probabilities like Overcast probability = 0.29 and probability of playing is 0.64.

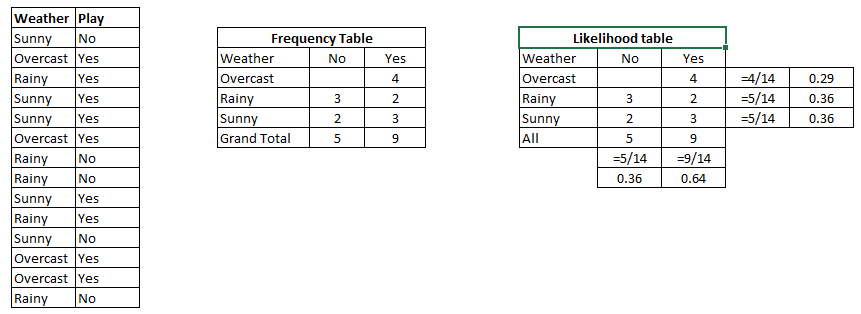
[](https://www.analyticsvidhya.com/wp-content/uploads/2015/08/Bayes_41.png)

Table 2.1

Using Naive Bayesian equation, the posterior probability for each class is calculated. The class with the highest posterior probability is the outcome of prediction.

Say if we want to find out if the Players will play when the weather is sunny?

To solve the above discussed method of posterior probability.

P (Yes | Sunny) = P (Sunny | Yes) \* P(Yes) / P (Sunny)

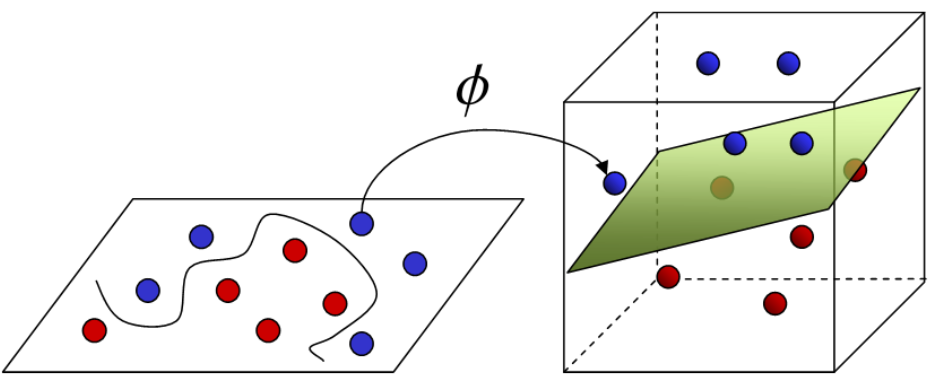
Here we have P (Sunny |Yes) = 3/9 = 0.33, P(Sunny) = 5/14 = 0.36, P(Yes)= 9/14 = 0.64

Now, P (Yes | Sunny) = 0.33 \* 0.64 / 0.36 = 0.60, which has higher probability.

Naive Bayes uses a similar method to predict the probability of different class based on various attributes.

#### **SVM (Support Vector Machine)**

A Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyperplane. In other words, given labeled training data (supervised learning), the algorithm outputs an optimal hyperplane which categorizes new examples. In two-dimensional space this hyperplane is a line dividing a plane in two parts where in each class lay in either side.



**Kernel**

The learning of the hyperplane in linear SVM is done by transforming the problem using some linear algebra. This is where the kernel plays role. For linear kernel the equation for prediction for a new input using the dot product between the input (x) and each support vector (xi) is calculated as follows:

f(x) = B (0) + sum (ai \* (x, xi))

This is an equation that involves calculating the inner products of a new input vector (x) with all support vectors in training data. The coefficients B0 and ai (for each input) must be estimated from the training data by the learning algorithm. The polynomial kernel can be written as K(x,xi) = 1 + sum(x \* xi)^d and exponential as K(x,xi) = exp(-gamma \* sum((x — xi²)).

Polynomial and exponential kernels calculate separation line in higher dimension. This is called kernel trick

**Regularization**

The Regularization parameter (often termed as C parameter in python’s sklearn library) tells the SVM optimization how much you want to avoid misclassifying each training example. For large values of C, the optimization will choose a smaller-margin hyperplane if that hyperplane does a better job of getting all the training points classified correctly. Conversely, a very small value of C will cause the optimizer to look for a larger-margin separating hyperplane, even if that hyperplane misclassifies more points.

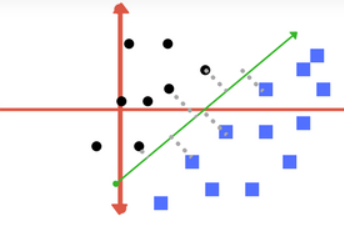
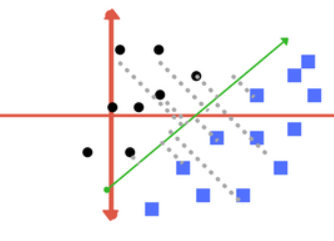
The images below (same as image 1 and image 2 in section 2) are example of two different regularization parameter. Left one has some misclassification due to lower regularization value. Higher value leads to results like right one.



**Left: low regularization value, right: high regularization value**

**Gamma**

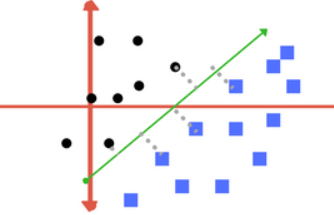
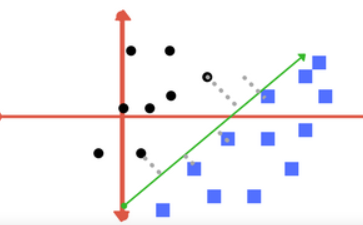
The gamma parameter defines how far the influence of a single training example reaches, with low values meaning ‘far’ and high values meaning ‘close’. In other words, with low gamma, points far away from plausible separation line are considered in calculation for the separation line. Whereas high gamma means the points close to plausible line are considered in calculation.

**High Gamma Low Gamma**

**Margin**

And finally, last but very important characteristic of SVM classifier. SVM to core tries to achieve a good margin. A margin is a separation of line to the closest class points. A good margin is one where this separation is larger for both the classes. Images below gives to visual example of good and bad margin. A good margin allows the points to be in their respective classes without crossing to other class.

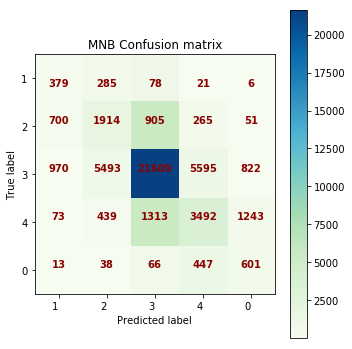
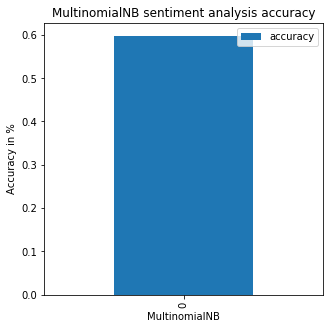
 

**Good Margin Bad Margin**

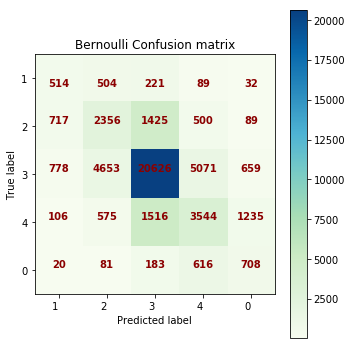
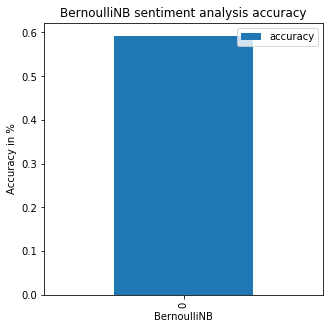
**Model 1.1 Unigram Boolean vectorizer on MultinomialNB, Bernoulli and SVM**

**Figure 2.1, 2.2 and 2.3** shows the confusion matrix and the accuracy report of the MultinomialNB, Bernoulli and SVM algorithm.

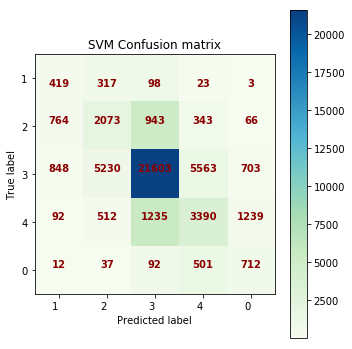
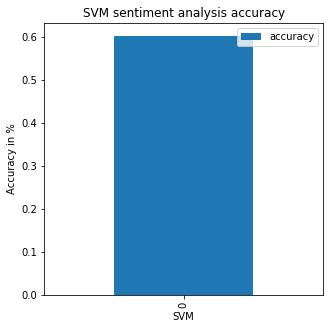
**Figure 2.4** is the comparison chart of accuracy for the above models

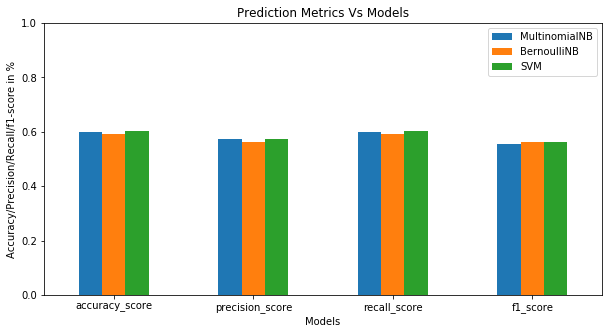
**Figure 2.1 Confusion matrix and accuracy for MNB**

**Figure 2.2 Confusion matrix and accuracy for Bernoulli**

**Figure 2.3 Confusion matrix and accuracy for SVM**

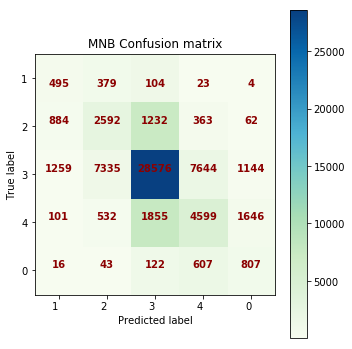
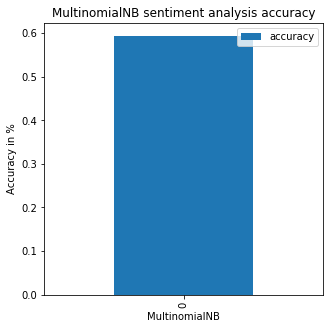


**Figure 2.4 Accuracy by model**

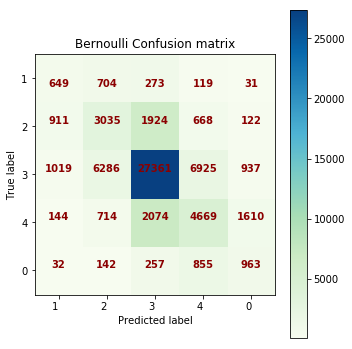
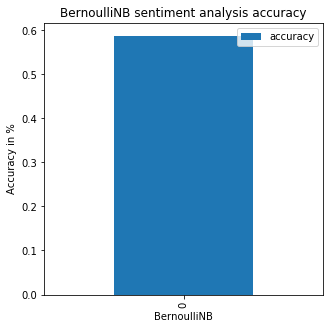
**Model 1.2 Unigram Frequency vectorizer on MultinomialNB, Bernoulli and SVM**

**Figure 2.5, 2.6 and 2.7** shows the confusion matrix and the accuracy report of the MultinomialNB, Bernoulli and SVM algorithm.

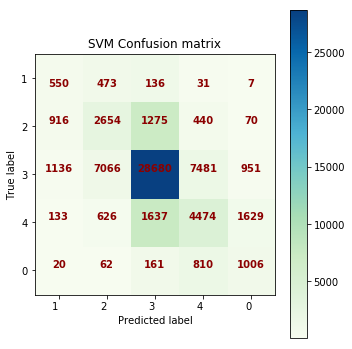
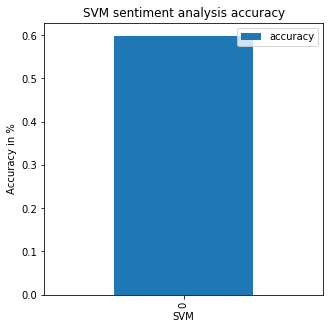
**Figure 2.8** is the comparison chart of accuracy for the above models

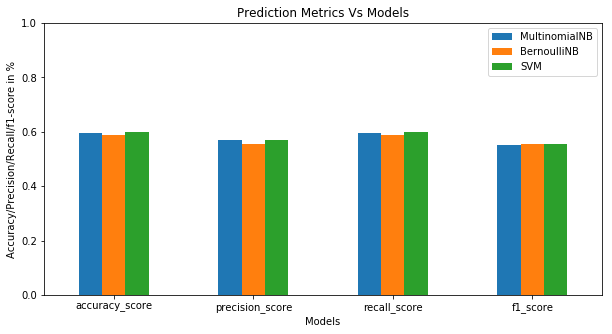
**Figure 2.5 Confusion matrix and accuracy for MNB**

**Figure 2.6 Confusion matrix and accuracy for Bernoulli**

**Figure 2.7 Confusion matrix and accuracy for SVM**

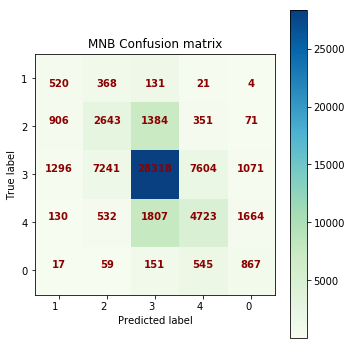
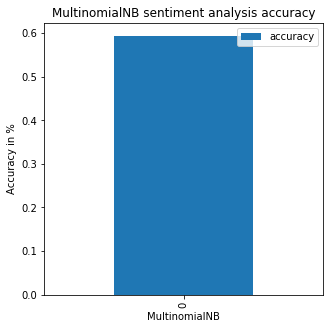


**Figure 2.8 Accuracy by model**

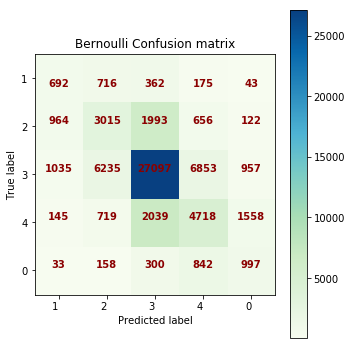
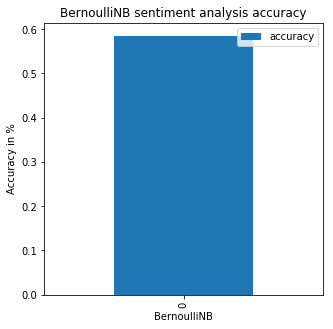
**Model 1.3 ngram(1,2) Boolean vectorizer on MultinomialNB, Bernoulli and SVM**

**Figure 2.9, 2.10 and 2.11** shows the confusion matrix and the accuracy report of the MultinomialNB, Bernoulli and SVM algorithm.

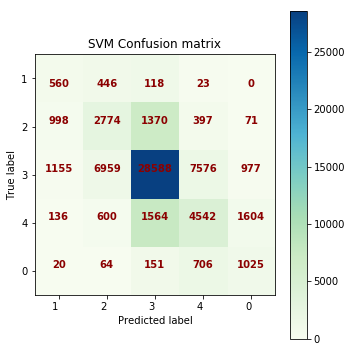
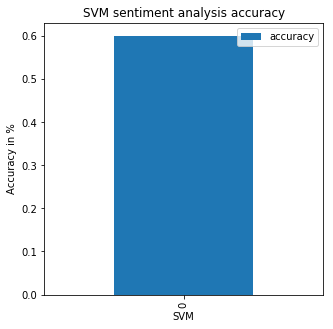
**Figure 2.12** is the comparison chart of accuracy for the above models

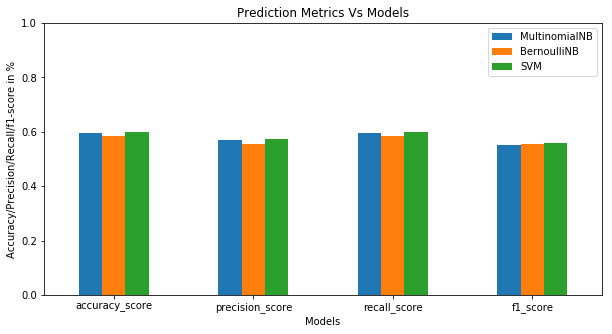
**Figure 2.9 Confusion matrix and accuracy for MNB**

**Figure 2.10 Confusion matrix and accuracy for MNB**

**Figure 2.11 Confusion matrix and accuracy for MNB**

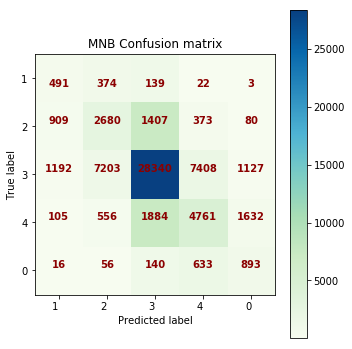
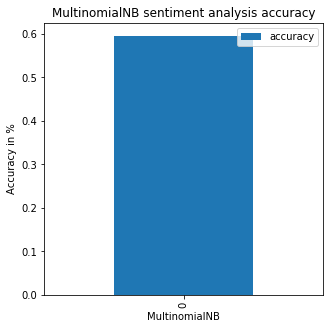


**Figure 2.12 Accuracy by model**

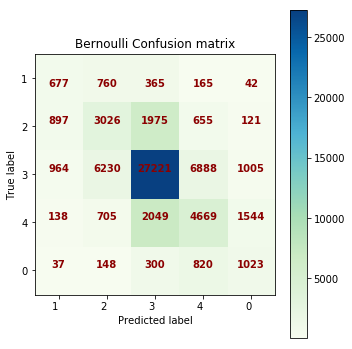
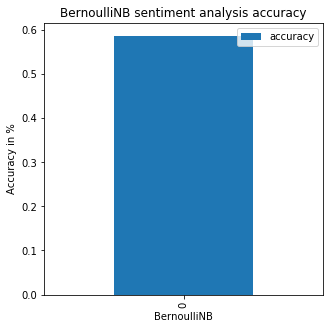
**Model 1.4 ngram(1,2) Frequency vectorizer on MultinomialNB, Bernoulli and SVM**

**Figure 2.13, 2.14 and 2.15** shows the confusion matrix and the accuracy report of the MultinomialNB, Bernoulli and SVM algorithm.

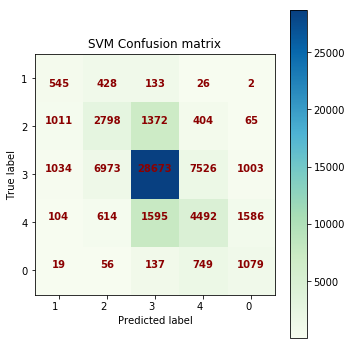
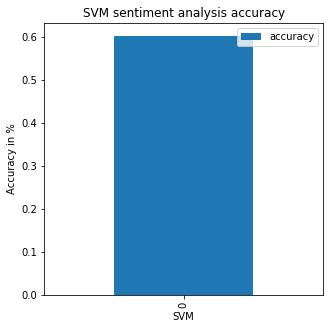
**Figure 2.16** is the comparison chart of accuracy for the above models

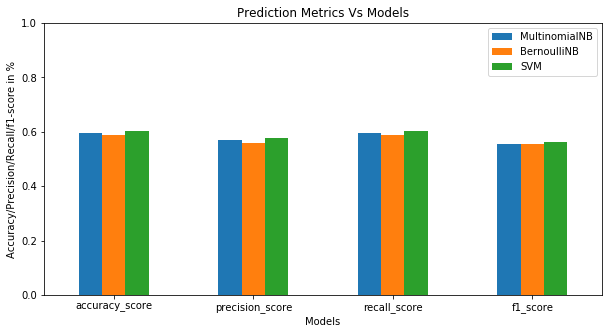
**Figure 2.13 Confusion matrix and accuracy for MNB**

**Figure 2.14 Confusion matrix and accuracy for Bernoulli**

**Figure 2.15 Confusion matrix and accuracy for SVM**



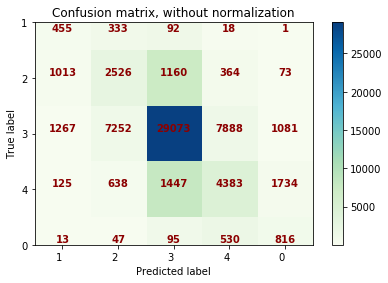
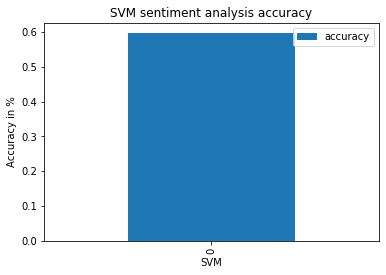
**Figure 2.16 Accuracy by model**

**Model 1.4 ngram(1,2) Frequency vectorizer on SVM using C as 0.1,1,10,100,1000**

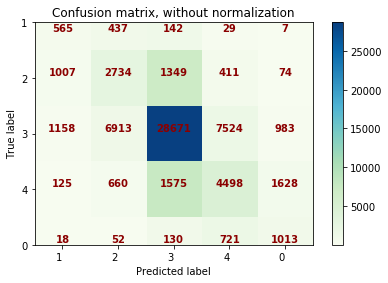
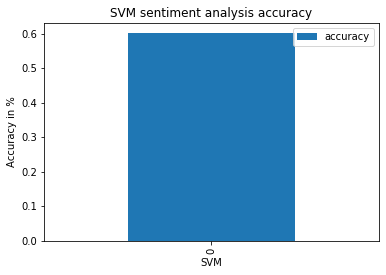
C is the penalty parameter of the error term. It controls the tradeoff between smooth decision boundary and classifying the training points correctly.

Below figure shows the accuracy and confusion matrix for different values of C

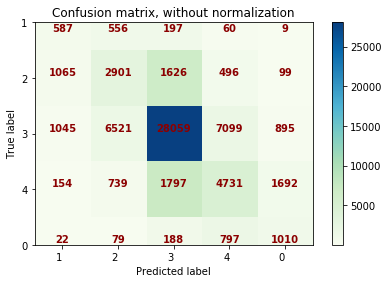
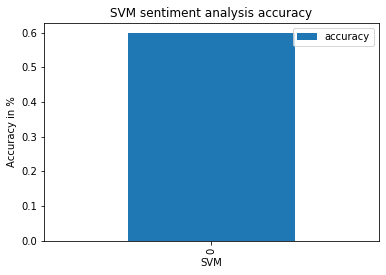
**When C=0.1**

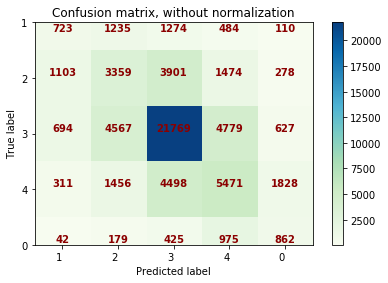
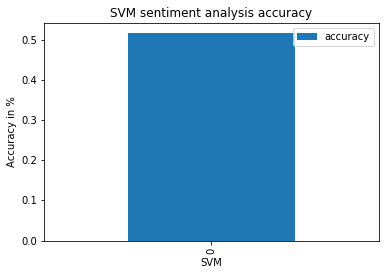
**When C=1**

**When C=100**

**When C=1000**

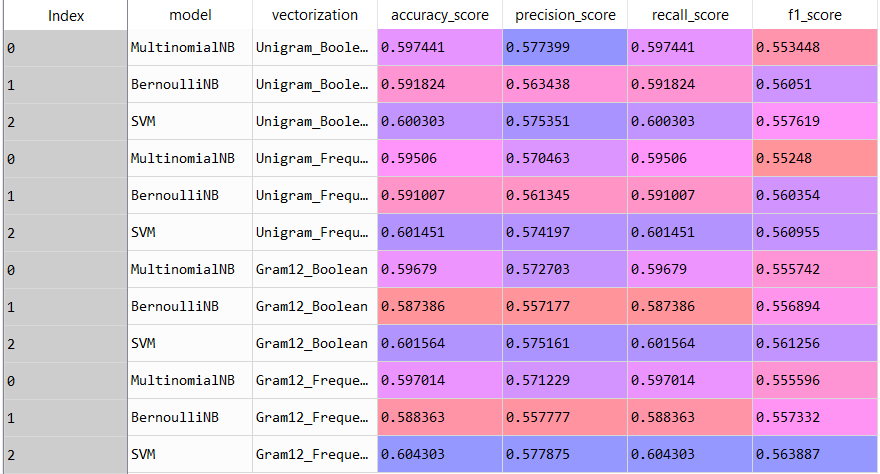
 

## **Results**

**Sentiment polarity prediction**

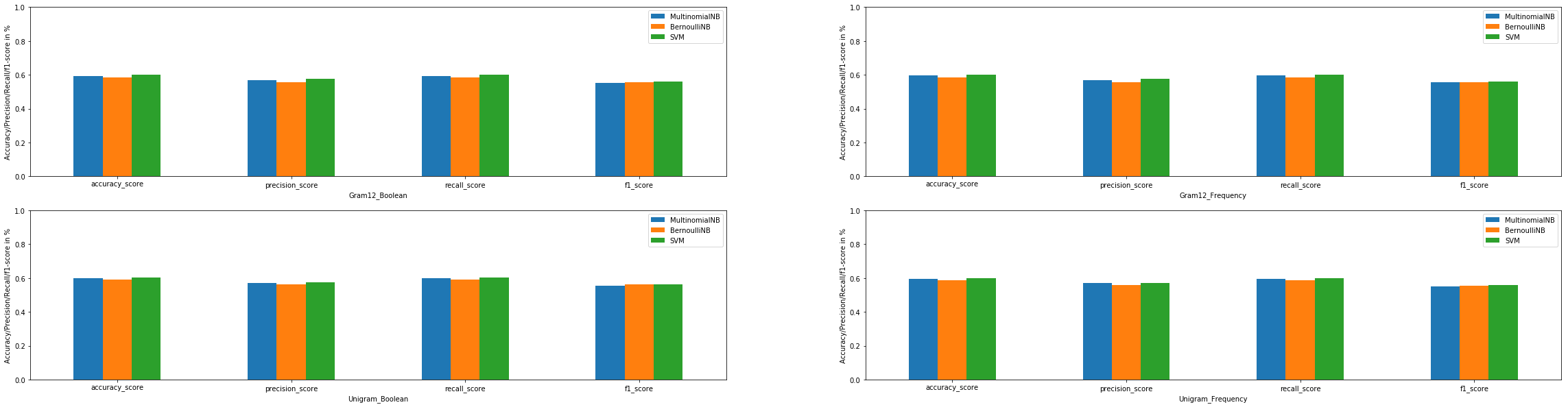
Based on the experimental results, it appears that the SVM model generated using ngram(1,2) frequency vectorization is more accurate when compared to another vectorization technique. Also, there are not much variance in the accuracy across other models (**Figure 3.1**).

Accuracy, precision and recall using MultinomialNB , Bernoulli and SVM are tabulated in **Table 3.1**

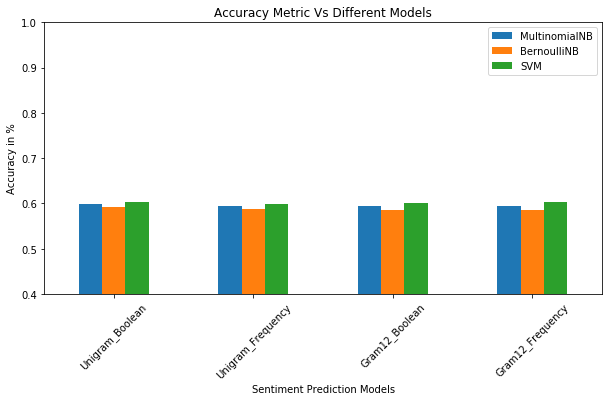


**Table 3.1 Model performance comparison for sentiment analysis**

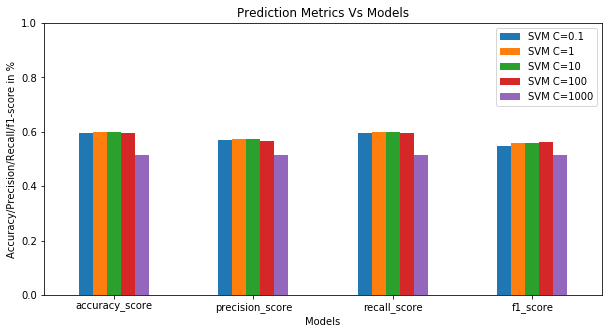
SVM is the one which has highest accuracy of more than 60% is the best model and the escalation parameter should 1,10,100



**Figure 3.1 Accuracy,Precision, Recall and F1 Score**



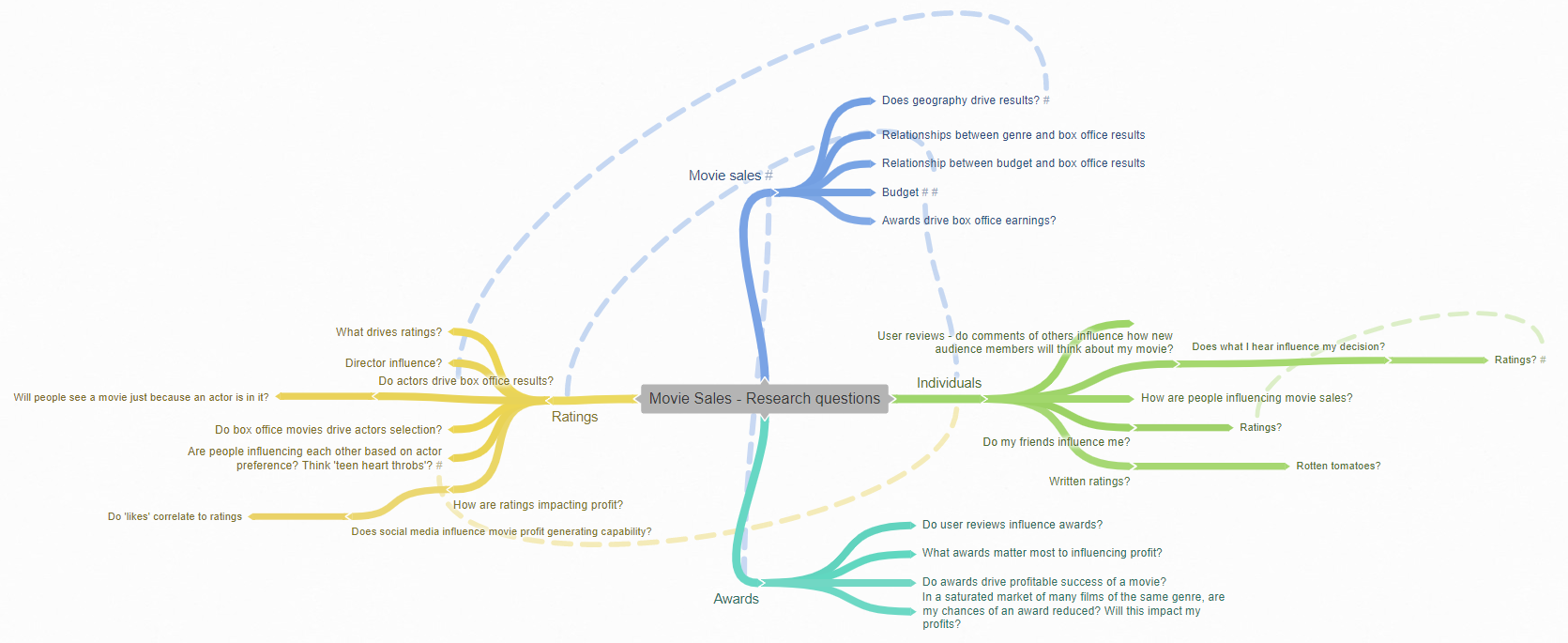
**Figure 3.2 Accuracy Vs vectorization and Models**



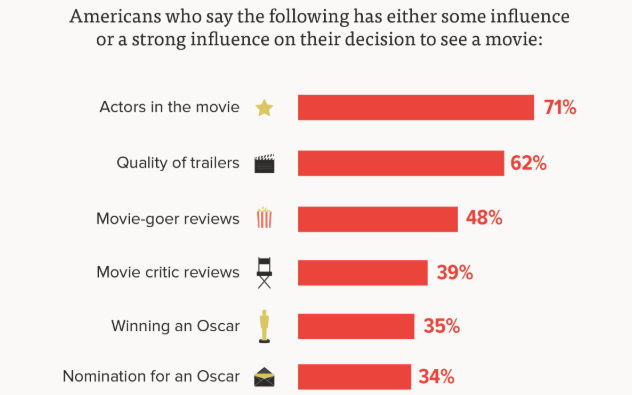
**Figure 3.3 Accuracy Vs penalty parameters**

## **Conclusion**

There are a few conclusions that can be drawn here. The first is that people with a high Need for Cognitive Closure prefer reviews written by consumers. This means that they trust consumer reviews to give them more accurate information than critics. This could be because they see consumers more like themselves, in terms of taste and knowledge. They may feel that critics write longer and more complex reviews that are too inconvenient to read. Additionally, people with a high Need for Cognitive Closure prefer to read positive reviews over others. This also has to do with the ease at which they feel positive reviews allow them to decide. They are more inclined to trust the reviewer’s opinion, and a mixed review simply makes it more difficult and time consuming to reach a conclusion. To back this up, people who showed a low Need for Cognitive Closure preferred mixed reviews because they are less concerned with the effort it takes to find an answer but would rather find the best answer. A mixed review offers more 2-dimensional information which thus services to provide more well-rounded information.



However, in recent years, there has been a growing belief in the film industry that critic aggregators (especially Rotten Tomatoes) are increasing the collective influence of film critics. The underperformance of several films in 2017 was blamed on their low scores on Rotten Tomatoes. This has led to studies such as one commissioned by 20th Century Fox claiming that younger viewers give the website more credibility than the major studio marketing, which undercuts its effectiveness.



Today, fan-run film analysis websites like Box Office Prophets, CineBee and Box Office Guru routinely factor more into the opinions of the general public on films produced.