International Journal of Applied Engineering Research ISSN 0973-4562 Volume 14, Number 11 (2019) pp. 2689-2697
© Research India Publications. http://www.ripublication.com

Critical Factors that Impact Performance of Machine Learning Algorithms for Polarity Determination of Movie Reviews

[1]Shashank Shekhar Sharma

PhD Scholar, Indian Institute of Foreign Trade
IIFT Bhawan, B-21, NRPC Colony, Block B,
Qutab Institutional Area, New Delhi, Delhi 110016, India

(Corresponding Author)

[2]Gautam Dutta

Assistant Professor, Indian Institute of Foreign Trade
IIFT Bhawan, B-21, NRPC Colony, Block B,
Qutab Institutional Area, New Delhi, Delhi 110016, India

Abstract

Sentiment Analysis (SA) of movie reviews has been shown to be extremely sensitive to the inherent review specific factors. The researches mostly focus on feature selection options and algorithms but do not dive deeper into other factors that may affect the performance of algorithms in prediction of sentiment polarity. Machine learning based supervised methods can be affected by factors such as the rating cut off point selected as label marker, the impact of the ratio of positive to negative reviews in the sample set or the number of reviews taken for training. In particular, the impact of sub-domains with regard to country specific factors is also not well understood and it can be interesting to study its impact by comparing the SA performances for the two largest country specific sub-domains within the domain of movie reviews: Bollywood and Hollywood. The study shows that the machine learning models perform differently on these sub-domains due to inherent differences in the nature of reviews. Also the study highlights that the impact of the number of training samples, selection of rating cut-off point and the positive to negative skew within the sample set. The study also concludes that while SVM (Support Vector Machine) delivers best performance in most cases, there are certain situations in which other machine learning algorithms may prove more useful.

Keywords: Polarity determination; sentiment analysis; movie reviews; natural language processing; classification of reviews; feature selection; classification methods; machine learning.

1. INTRODUCTION

A range of efforts aimed at decoding consumer reaction for a given product and service have attempted to utilize the fingerprints left by the consumer in the digital media. A particularly useful ‘fingerprint’ is the digital text that is getting generated at a phenomenal rate everyday by users across the world on various digital platforms. In case of several product categories like movies, a huge amount of electronic word of mouth (eWOM) gets generated every time a new movie gets released. More and more consumers use online discussion forums, consumer review sites, weblogs, social network sites etc. to exchange product information[1]. Various researches have established the high correlation between online reviews and product sales[2-5].

The effect of online reviews is also much more exaggerated in case of experience goods like movies which have a very short life and one doesn’t have much insight into how good or bad her transaction is going to turn out before she actually makes the transaction and watches the movie[6]. One has to make the purchase going not by his own evaluation but mostly on evaluation of others[3]. Several studies have established the high degree of causative role that online reviews and comments play on the outcome of a movie[2, 4, 7, 8-10]. Beyond the movie specific factors like the actor, director, production house etc., the movie’s reviews, and in particular the polarity of the reviews in terms of overall sentiment, has been found to be the single largest external factor that can help predict the total box office potential of a movie[8].

However given the abundance of reviews generated in a very short time after a movie’s release it is impossible to manually label the reviews based on the sentiment polarity of the review (positive or negative) and model a movie’s prediction in a practical time period to actually maximize its’ utility. In recent years, machine learning based algorithms have been shown to be quite efficient at this labeling task[11]. The automation makes it practical to quickly sieve through millions of reviews and determine the polarity (positive or negative) of individual reviews to predict the success of the movie. This automation for extracting sentiment expressed in a text is called Sentiment Analysis (SA).

Most of the researches in the domain of sentiment analysis of movie reviews are aimed at increasing the accuracy of sentiment classification based on input data, mostly in the form of text. Many of the NLP methods of feature extraction and engineering have been coupled with a host of Machine Learning (ML) techniques to increase accuracy of SA. Many other hybrid approaches such as using ML algorithms with existing or built-for-purpose lexicons and corpora have been used to enhance the accuracy[11]. The researches mostly focus on feature selection options and algorithms but do not dive deeper into other factors affecting the prediction of polarity such as the number of reviews taken for training, the rating cut off point selected as label marker, or the impact of the ratio of positive to negative reviews in the sample set. In particular the impact of sub-domain with regard to country specific factors is not well understood and it remains an unexplored territory to robustly compare the SA performances for two separate sub-domains such as Bollywood and Hollywood. In fact, very little research has been done for the domain of Bollywood movie reviews despite the fact the Bollywood produces more movies each year compared to Hollywood. Different ML algorithms have been shown to be effective for polarity determination task[12,13]. Support Vector Machine (SVM), Naïve Bayes (NB) and Maximum Entropy (ME) are three particularly popular techniques which have shown consistent performance in terms of accuracy in many of the researches[11-14]. While SVM has generally delivered a slightly better performance than the other two, there are cases when ME and NB have in fact outperformed SVM[11]. There is no clear understanding of circumstances when a certain algorithm can be picked as the preferred option. This paper tests all three algorithms to understand their differential behaviour in different circumstances emerging from choice of number of reviews, domain of interest, rating cut-off point and polarity skew in the training and test sample sets.

2. LITERATURE REVIEW

The first landmark paper on the subject by Pang, Lee and Sivakumar[12] used a data set of 1400 Hollywood movie reviews from IMDB which is a popular movie database website and one of the most widely consulted movie review website for user reviews and ratings. They used SVM algorithm with unigrams as features and were able to achieve accuracy of 86% with their methodology. Consequently several papers have been published over the years but many of them used the same data set as Pang, Lee and Sivakumar[12] for making models. Many others have used a bigger dataset of movie reviews which was used by Pang and Lee[13] in their 2004 paper and have developed models with varying degrees of accuracy. Among the various algorithms Support Vector Machine (SVM), Naïve Bayes (NB) and Maximum Entropy (ME) are the most favoured[11].

The data is in the form of text which must be prepared so that it can be used for model learning or classification task. The unstructured text needs to be modified into a structured format that can be further processed in meaningful manner. In data mining in general and text mining specifically the pre-processing phase is of significant impact on the overall outcomes[15-16]. This step is important not only to put the data in a structured format but also to select and represent the features in the most optimum way. Bag of words (BoW) is the most commonly used feature in which the words or phrases are simply represented as a multiset based on their presence in a document and the frequency disregarding grammar. Pang, Lee and Sivakumar[12] used bag of features framework with unigrams alone with using both unigrams and bigrams together and found unigrams alone to be slightly more accurate. While term frequency – inverse document frequency (TF-IDF) had been very successful in other domains like topic classification, Pang et al.[12] pointed out that frequency of words may not be a good predictor for sentiment. Instead binary representation of each feature (presence or absence) can lead to more accurate prediction.

More nuanced approaches such as use of part of speech information or sentiment orientations of the words have been shown to improve accuracy as well. Opinion words are words that are commonly used to express positive or negative sentiments[17]. For example, adjectives like nice, fantastic, good, and great are positive opinion words while ugly, bad, and boring are negative opinion words. Semantic Orientation (SO) is a real number measure of the positive or negative sentiment expressed by a word of [18]. Mullen and Collier[19] selected and defined the SO scores of the adjectives using ‘Point Mutual Information’ method with the seed words like “excellent” and “poor” and then denoting a SO value to each objective using below formula for calculating PMI. Pang and Lee[13] used SVM and only subjective words as features and achieved accuracy of 87% with SVM.

As the simplest method used for classification, use of n-grams with any of the ML classification methods among SVM, NB and ME deliver a decent performance with accuracies above 80% and reaching as high as 86% with SVM. These baseline methods can be used a good starting point to explore other important factors beyond feature selection.

3.1 Research Gaps

Most studies use common datasets[12, 13, 17, 20-22] which have an equal distribution of positive and negative. Also most studies use only accuracy for performance assessment. Both these aspects limit the real life applicability since a good proportion of reviews for a given movie may be have significantly higher percentage of positive or negative review which may get exaggerated unless we use models that are applicable for biased data set as well which perform reasonably well on recall and precision both along with accuracy. The most well accepted and best performing methods discussed in above section can be tested on larger and more contemporary data sets with different clusters of biased and unbiased dataset to compare performances on each of the three key aspects: accuracy, recall and precision. Also, there is scope of establishing correlation between performance and no of reviews (or training samples) for different methods.

These methods have not been robustly tested for a different domain like Bollywood either. Viewers in different countries show differences in their movie consumption habits and have different expectations[23]. Reviews may also vary in the level of objectivity or how extreme the ratings tend to be based on the domain they belong to[24]. In the Pang and Lee[13] dataset the ratings had been used for labeling the reviews as positive or negative, however, the cut-off rating point had been arbitrarily chosen. There is a need for deeper analysis to narrow down the neutral rating point as well for each sub-domain so as to enable a more accurate classification of labels.

The studies reviewed earlier preferred different algorithms or models although SVM was the most preferred model. But these studies didn’t compare the performance of different algorithms under special situations since there can be a lot of variance in available movie reviews’ data for training in terms of number of samples available or the skew of ratings in a given sample.

3.2 Research Questions

- Is performance of ML algorithms for polarity determination different for different domains like

Bollywood and Hollywood reviews?

- What is the impact of the number of reviews and the ratio of positive to negative reviews on performance accuracy?
- What should be selected as the ideal cut off point of rating for positive and negative classification?
- Which ML algorithm delivers the most optimum performance consistently across domains? Are there special conditions under which one may prefer a different ML algorithm?

3. METHODOLOGY

IMDB (Internet Movie Database) is one of the most popular websites where hundreds of audience reviews can be found for movies released across the world. This source of movie reviews has been used across several key researches on sentiment analysis of movie reviews. For the purpose of this study, a total of 83,500 reviews were extracted from IMDB using Python library 'Scrapy' for web scraping. Many of the reviews were found to contain less than 5 words and were dropped from the sample set. Also some of the reviews did not contain rating and had to be dropped as well as they could not have been labeled. Finally 40,000 reviews each for Bollywood and Hollywood were selected for experiment.

Python provided some excellent libraries like 'nltk' for natural language processing tasks and 'scikit-learn' for building machine learning models. These libraries were predominantly used for the experiment for text processing, model learning and model testing. The reviews were preprocessed using 'nltk' modules and a pipeline was executed for each sample set on order to implement the feature annotations in order to optimize and extract useful information and finally arrive at a feature vector for each review which can be run on an ML algorithm.

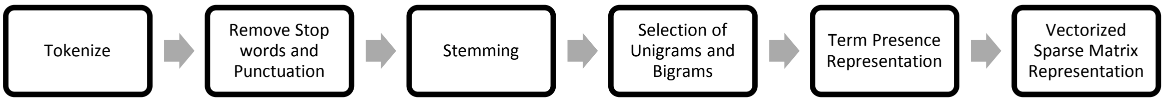
Tokenize
Remove Stop words and Punctuation
Stemming
Selection of Unigrams and Bigrams
Term Presence Representation
Vectorized Sparse Matrix Representation

Figure 1. Steps of feature processing and representation

Following the steps given in Figure 1, each review was first tokenized and tokens were removed if they belonged to stop words or were a punctuation mark. Stemming was then performed to reduce each token to its root word. Both unigrams and bigrams were both selected as this approach has been shown to contain more word relations thus positively impacting the classification accuracy. Each n-gram was then represented by a binary value '1'or '0' which denoted the presence or absence of that n-gram in that review respectively.

Rating cut-off was chosen basis a survey done on 112 participants who regularly used IMDB to decide on whether or not they will watch a particular movie. Each participant was asked to label each rating point as either positive or negative basis how these rating points impact their perception of the movie. Also result from this survey was compared to average sentiment scores of reviews across each rating point for both domains. Sentiment scores were calculated using SentiWordNet[25] which is a commonly used lexicon. SentiWordNet provides information about polarity identification as well as for subjectivity detection. It provides discrete values for sentiment score where 0.25, 0.5, 0.75 and 1.0 show positive polarity with varying degree of positivity with 1 being highest. The case is exactly reverse in case of negative scores. Using sentiment lexicons, the sentiment orientation scores can be calculated at a phrase level, sentence level or document level. Common way to aggregate the scores at any level is to average out the sentiment score values of all the words present in the text. If the final score is positive, the text is considered as positively oriented or else negative if the average score comes out to be negative [26].

Several experiments were run for both sample sets (Hollywood and Bollywood) using combination of variables including the number of selected reviews and positive-negative rating ratios and each combination was tested with the three most preferred ML algorithm (SVM, ME and NB) in the SA domain along with parameter tuning, done with help of 'Grid Search CV' module provided in 'scikit-learn' library, to arrive at the most optimized models for each case. The results of these experiments are described below with emphasis on highlighting the impact of the key variables being studied. In total 72 experiments were run initially for with below combinations using all three ML algorithms. Table 1 lists the values for each of the factors that were tested. The scores presented in the results section are all basis 10 fold cross validation results for each of the model.

Table 1. Selected variables and their values for various iterations

No of Reviews (6 iterations)	Rating Cut-off Point (4 iterations)	Positive – Negative Ratio
1000, 2000, 5000, 10000, 15000, 20000	5, 6, 7, 8	0.5, 1.0, 2.0

4. RESULTS

The graph below shows how the ratings were dispersed across the selected 20,000 movie reviews for each of the sub-domains. It can be clearly seen that Bollywood reviews tend to have a much more skewed ratings with almost 40% reviews rated as either '1' or '10' whereas Hollywood reviews have less skew with only 22% reviews rated as either '1' or '10'.

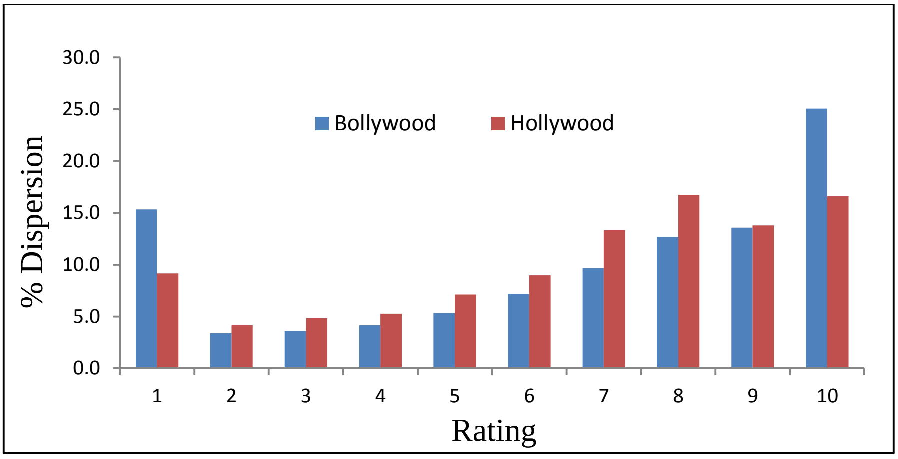

Figure 2. Rating Dispersion of 20,000 movie reviews each for Bollywood and Hollywood

Whether it is due to the nature of Bollywood movies themselves that they tend to be either too good or too bad or due to a more objective or nuanced opinion of Hollywood movie reviewers, it can be concluded on the basis of extremity of ratings that Bollywood reviews are more extreme. This may have an impact of how much more accurately the algorithms are able to classify the polarity of Bollywood movie reviews versus Hollywood movie reviews. Figure 3 shows that increase in sentiment scores across rating points is steeper for Bollywood reviews. This may be because they are probably more opinionated in nature. Also Figure 4 shows that Hollywood reviews in fact contain more number of words especially for extreme reviews compared to Bollywood reviews. Again this indicates that Hollywood reviews are more nuanced in nature and not as extreme as Bollywood reviews. This difference in the nature of domains can directly impact performance in prediction of polarity and this can be tested with the above described data set.

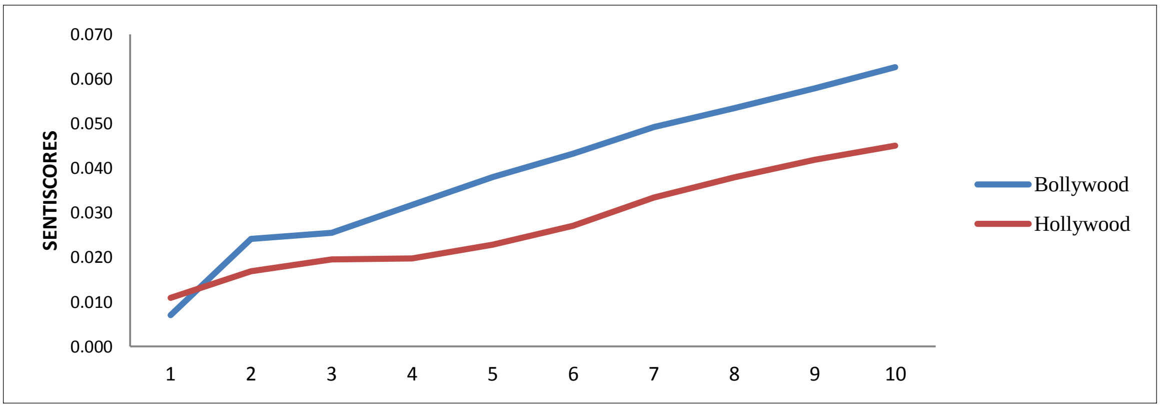

Figure 3. Average sentiment scores across rating

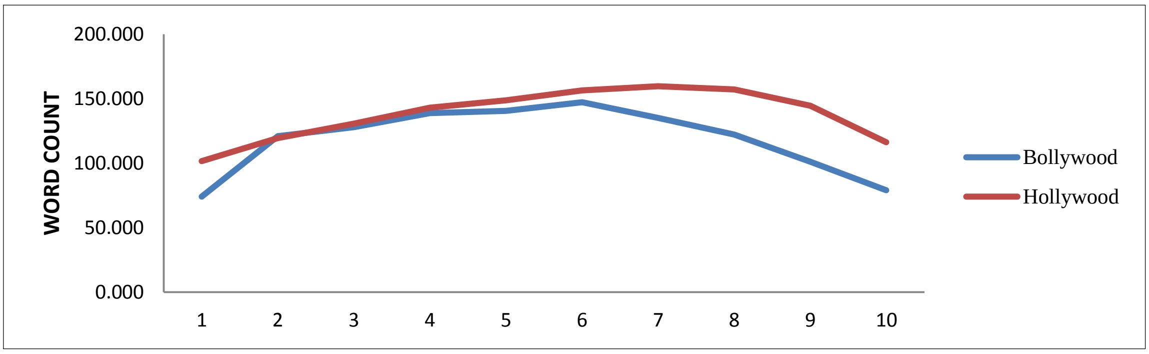

Figure 4. No of words/review across rating

4.1 Impact of Sub-domain on classification

An average of results across experiments clearly shows that independent of the number of reviews or other factors like rating cut-off point and sample polarity skew (positive-negative ratio), all metrics used for classification model evaluation show a significantly better result in case of Bollywood movie reviews.

Table 2 shows the average metric scores across 72 experiments run on each of the sub-domain. Bollywood movie reviews have 2.7% higher accuracy and 2.9% higher f1-score compared to Hollywood movie reviews. As stated earlier this may be explained by the skew in extreme rating scores observed for Bollywood reviews.

Table 2. Average scores of classification across key metrics for 72 experiments

	Average of Accuracy	Average of f1-score	Average of Precision	Average of Recall
Bollywood Reviews	85.3%	83.8%	84.2%	83.6%
Hollywood Reviews	82.6%	80.9%	81.9%	80.2%

Since the above variation in score could be due to skew in extreme ratings and not due to the inherent nature of reviews, the same can be tested by selecting reviews from both domain in a manner that both sample set have the same rating dispersion and then running the models again for both to compare the results. This experiment was carried out across all combinations as done in the earlier case. Table 4 shows the average scores of classification across key metrics for 72 experiments with same rating dispersion in both domains.

Table 4. Average scores of classification across key metrics for 72 experiments with same rating dispersion in both domains

	Average of Accuracy	Average of f1-score	Average of Precision	Average of Recall
Bollywood Reviews	86.3%	85.8%	84.8%	86.6%
Hollywood Reviews	84.7%	84.2%	84.5%	83.9%

The results show that while the difference between the scores for both domains has narrowed down there is still a significant difference and classification scores across metrics are better for Bollywood movie reviews.

4.2 Impact of rating cut-off point:

The result of survey on rating cut-off point is provided in Table 5. Any rating under ‘5’ was seen as negative by an overwhelming majority. Any rating above ‘6’ was seen as positive by almost the same percentage of participants. This result indicated that choosing ‘5’ or ‘6’ is a valid rating cut-off point as has been done previously by several researches.

Table 5. Results of survey on rating cut-off point

Rating	Postive	Negative
1	0%	100%
2	0%	100%
3	0%	100%
4	2%	98%
5	33%	67%
6	56%	44%
7	70%	30%
8	3%	97%
9	97%	0%
10	97%	0%

To test this further, in this study, four different iterations were tried for 20,000 samples each from Hollywood and Bollywood sample set. In each of the iteration, reviews with a certain rating point were treated as neutral and excluded from the sample set. Reviews with ratings higher than this rating point were labeled as positive and the others were labeled as negative. The four rating points excluded in each of the iterations were 5, 6, 7 and 8 respectively. The experiments were run with positive to negative ratio for each set maintained at 1 so as to control for this variable. The results are listed in table 6.

Table 6. Results of classification for each iteration where for each sample set listed, the corresponding rating mentioned in the table has been treated as neutral point and the corresponding reviews have been removed

Domain	Rating	Accuracy	f1-score	Precision	Recall
Bollywood Reviews	5	89.8%	90.0%	90.3%	89.8%
	6	89.3%	89.3%	88.5%	90.1%
	7	86.3%	86.4%	86.5%	86.3%
	8	85.2%	85.3%	85.1%	85.6%
Hollywood Reviews	5	88.1%	88.0%	88.2%	87.8%
	6	88.0%	88.2%	87.8%	88.5%
	7	85.5%	85.5%	84.5%	86.4%
	8	84.0%	84.0%	84.5%	84.4%

The accuracy scores for both domains are best for rating point ‘5’ although only very slightly better than ‘6’. In fact f1-score for ‘6’ is 0.2% better in case of Hollywood reviews. There is a substantial drop across scores when rating point ‘7’ is selected as neutral and excluded. This is a further drop for rating point ‘8’ as expected but again the drop is not too exaggerated and the model still returns a fairly high accuracy score of 85.2% and 84.0% for Bollywood and Hollywood respectively.

4.3 Impact of number of reviews (sample size) and ratio of positive to negative reviews:

The model was trained on sample sizes 500 and 1000 and then incremental of 1000 samples till 20,000 samples (movie reviews). This experiment was done only for Bollywood reviews to assess the impact of number of reviews on model accuracy. Positive to negative ratio for the samples was again maintained at 1 control for this variable. Rating cut-off was taken as ‘6’. The previous experiment showed that there was not a significant difference between rating of ‘5’ and ‘6’ so an arbitrary choice between the two was made in this case in favor of ‘6’. The graph below maps the best accuracy and best f1-score obtained for each sample set using all three ML algorithms against the sample size. The graph in Figure 5 shows a substantial increase in accuracy and f1-score between sample size of 500 and 3000 when both these scores reached around 87% but both plateau out after that and there is no clear trend of any gain in accuracy or f1-score with size as per the results obtained for sample size beyond 3000.

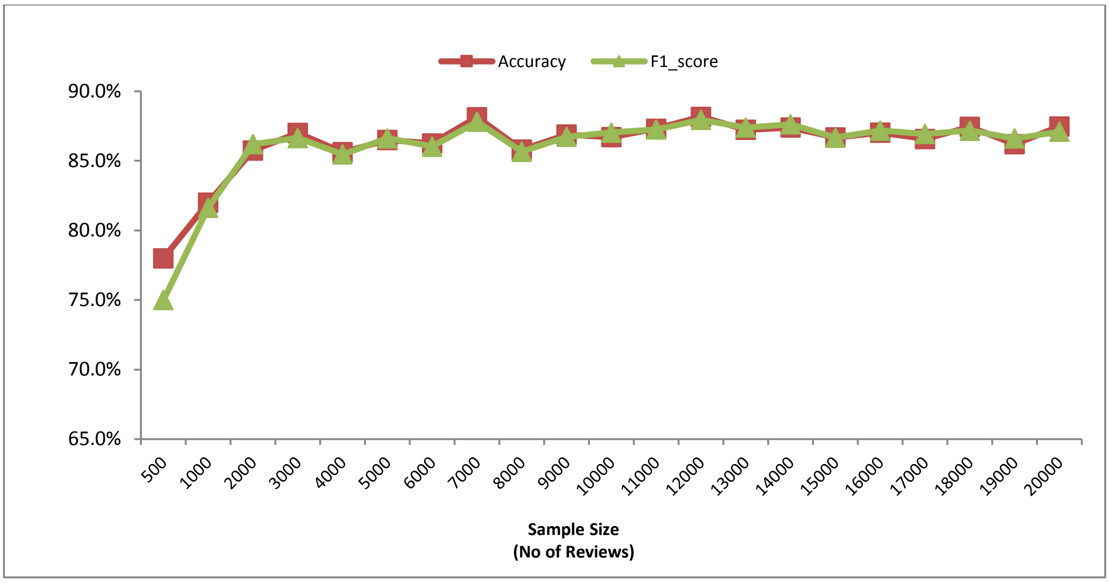

Figure 5. Line graph of accuracy and f1-score obtained for each sample set against the sample size

The above experiment was repeated for different sample sizes between 1000 and 20,000 and trained on sample set with different positive to negative ratio as well. While there was no moderating impact of ratio on accuracy which again plateaued after sample size of 3000 but there was a significant impact on f-score which is illustrated in the Figure 6.

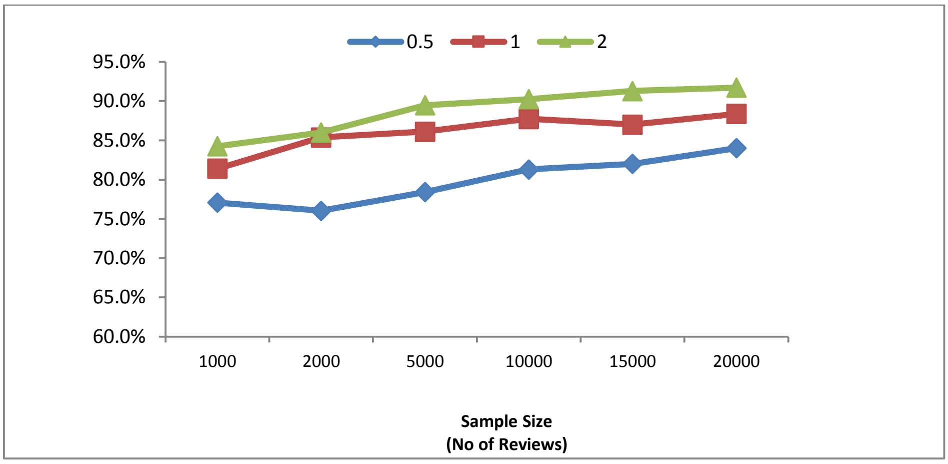

Figure 6. Line of f-score obtained for each sample set for all three ratio against the sample size

Since the f-score was calculated for positive reviews, as expected, in sample set with lower number of positive reviews the f-scores were very low as compared to samples with higher number of positive reviews. Interestingly though, it can be observed that for skewed samples there was a gain in f-score as the number of samples increase. For sample set with positive to negative ratio at 50%, the f-score increase from a low 77.1% for 1000 samples to 84.0% in case of sample set with 20,000 training samples. Similarly the impact of number of reviews can be seen in case of sample sets with positive to negative ratio of 2 as well. There was a gain of 7.5% in f1-score from 1000 training samples to 20,000 training samples.

Ratio and sample size also impacted the choice of ML algorithm. The table below lists the best performing ML algorithm in each case for the two variables basis accuracy score. SVM was the most consistent in terms of performance on accuracy. It delivered the best performance across all experiments when the ratio of positive to negative reviews is 1. Also, for smaller sample sets SVM delivered the best performance even for skewed samples. Different values of the most important parameters for each of the algorithms were also tried for training of each model. The best parameter value for each case is listed in the Table 7.

Table 7. Best performing ML algorithm basis accuracy score for each sample set with different combination of sample size and positive to negative ratio.

Ratio	Size	Best Performing Classifier	Parameter
0.5	1000	SVM	C=0.01, loss='squared hinge'
0.5	2000	Multinomial NB	alpha=1.0
0.5	5000	ME	C=1000.0
0.5	10000	ME	c=1.0
0.5	15000	Multinomial NB	alpha=1.0
0.5	20000	SVM	C=0.01, loss='squared hinge'
1	1000	SVM	C=0.01, loss='squared hinge'
1	2000	SVM	C=0.01, loss='squared hinge'
1	5000	SVM	C=0.01, loss='squared hinge'
1	10000	SVM	C=0.01, loss='squared hinge'
1	15000	SVM	C=0.01, loss='squared hinge'
1	20000	SVM	C=0.01, loss='squared hinge'
2	1000	SVM	C=0.01, loss='squared hinge'
2	2000	SVM	C=0.01, loss='squared hinge'
2	5000	SVM	C=0.01, loss='squared hinge'
2	10000	Multinomial NB	alpha=1.0
2	15000	SVM	C=0.01, loss='squared hinge'
2	20000	Multinomial NB	alpha=1.0

5. DISCUSSION

There are several factors that impact the robustness of ML models when it comes to the task of sentiment analysis. Domains have been known to have a huge impact on classification accuracy of reviews. From the results obtained in this study, it was evident that even within a particular domain like movie reviews, the ML models do not produce similar results for its sub-domains like Hollywood and Bollywood reviews. In the particular case of Hollywood and Bollywood reviews it was found that the ratings for Bollywood reviews tend to be more extreme and models trained on Bollywood reviews deliver 2.7% higher accuracy and 2.9% higher f1-score compared to Hollywood movie reviews. Even after controlling for ratings dispersion similar results were obtained although the gap between scores for both sub domains got reduced. This could be due to a more extreme nature of Bollywood reviews themselves. This was tested by calculating sentiment value dispersion of reviews by adding up the Sentiment Orientation (SO) scores of opinion words in the reviews using a lexicon like SentiWordNet and those scores indicated that Bollywood reviews were more opinionated and extreme compared to Hollywood revoews. It could also be that Hollywood reviews are more nuanced and therefore use more varied set of adjectives as compared to Bollywood reviews in which case probably same adjectives get repeated more often and this helps machine learning models to weigh them more and produce better results in polarity determination. It can be worthwhile to check this by looking at the dispersion of term frequencies of adjectives across both sub-domains.

Earlier researches have generally considered reviews with rating of 5 or 6 as neutral and for training the researchers tend to manually label the reviews as positive or negative based on their corresponding rating points. This assumption was found to be a robust assumption in this study. The survey results indicate that most participants see either ‘5’ or ‘6’ as the neutral point and there is little to choose between the two. Even the models trained on training set with neutral rating set as ‘5’ or ‘6’ easily outperformed the other models with a higher or lower rating cut-off point. Surprisingly, it was also observed that unlike most other ML tasks, there was very little impact of number of samples on the accuracy of prediction for sentiment analysis of movie reviews. There was a significant rise in prediction scores till about 3000 training samples. Beyond 3000, increasing number of training samples resulted in little extra advantage. In fact based on some earlier studies, it appears that choice of features and better feature representations along with use of hybrid methods for classification can have a much more significant impact. Also, a larger skew in training samples in terms of polarity where the number positive and negative reviews are not equal did not have much of an impact on the accuracy of the model either. However, skew in positive to negative ratio can substantially impact the f-score of the model output. Interestingly though, it was observed that for skewed samples there was a gain in f1-score as the number of samples increased. So it is suggested that in case of skewed training samples, it is advisable not to restrict to only 3000 samples. In fact, study showed that for the more skewed sample sets there is merit in going up to 20,000 samples as well for a higher f-score.

The results also indicate that both, the ratio and sample size, impacted the choice of ML algorithm. While SVM was found to produce best performing models in the majority of cases, it lagged behind in prediction accuracy when it came to larger samples sets with skewed positive to negative ratio compared to other algorithms which delivered better performance. Multinomial NB performed best in cases where the positive reviews were twice the number of negative reviews while ME seemed to have more advantage in the reverse case. The values of chosen model’s parameters also didn’t vary much with change in sample size, positive to negative skew or rating cut-off point. They didn’t vary even in case of training with different sub-domains either. This finding can have a huge impact on the computational efficiency as it obviates the need for training and comparing models with a range of value for model parameters which can be extremely intensive and take much longer. Further validation of this can be done through more robust testing of models across different sample sets in different domains.

REFERENCES

[1] Cheung, C. M. K., & Lee, M. K. O. (2008). Online consumer reviews: Does negative electronic word-of-mouth hurt more? Proceedings from the AMCIS 2008. A review for the influential factors in e-WoM research

[2] Dellarocas, C., Awad, N., & Zhang, X. (2004).Exploring the value of online reviews to organizations: Implications for revenue forecasting and planning. Proceedings from the ICIS 2004.

[3] Dellarocas, C., Zhang, X. & Awad, N. (2007).Exploring the value of online product reviews in forecasting sales: The case of motion pictures. Journal of Interactive Marketing, 21(4), 23

[4] Liu, Y. (2006). Word of mouth for movies: Its dynamics and impact on box office revenue. Journal of marketing, 70(3), 74-89.

[5] Godes, D., & Mayzlin, D. (2004).Using Online Conversations to Study Word-of-Mouth Communication.Marketing Science, 23, 545-560.

[6] Reinstein, D. A., & Snyder, C. M. (2005). The influence of expert reviews on consumer demand for experience goods: A case study of movie critics. The journal of industrial economics, 53(1), 27-51.

[7] Eliashberg, J. & Shugan, S. (1997). Film critics: Influencers or predictors?.JournalofMarketing, 61, 68-78.

[8] Duan, W., Gu, B.& Whinston, A. (2008). "The dynamics of online word-of-mouth and product sales—An empirical investigation of the movie industry." Journal of retailing 84.2 , 233-242.

[9] Niraj, R., & Singh, J. (2015). Impact of user-generated and professional critics reviews on Bollywood movie success. Australasian Marketing Journal (AMJ), 23(3), 179-187.

[10] Rui, H., Liu, Y., & Whinston, A. (2013). Whose and what chatter matters? The effect of tweets on movie sales. Decision Support Systems, 55(4), 863-870.

[11] Sharma, S. S., & Dutta, G. (2018). Polarity Determination of Movie Reviews: A Systematic Literature Review. *International Journal of Innovative Knowledge Concepts*, *6*(12), 43-55.

[12] Pang, B., & Lee, L. (2004, July). A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In Proceedings of the 42nd annual meeting on Association for Computational Linguistics (p. 271). Association for Computational Linguistics.

[13] Pang, B., Lee, L., & Vaithyanathan, S. (2002, July). Thumbs up?: sentiment classification using machine learning techniques. In Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10 (pp. 79-86). Association for Computational Linguistics.

[14] Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. Ain Shams Engineering Journal, 5(4), 1093-1113.

[15] Uysal, A. K., & Gunal, S. (2014). The impact of preprocessing on text classification. Information Processing & Management, 50(1), 104-112.

[16] Abbasi, A., France, S., Zhang, Z., & Chen, H. (2011). Selecting attributes for sentiment classification using feature relation networks. IEEE Transactions on Knowledge and Data Engineering, 23(3), 447-462.

[17] Hung, C., & Lin, H. K. (2013). Using objective words in SentiWordNet to improve sentiment classification for word of mouth. IEEE Intelligent Systems, 1.

[18] Hatzivassiloglou, V., & McKeown, K. R. (1997, July). Predicting the semantic orientation of adjectives. In Proceedings of the eighth conference on European chapter of the Association for Computational

Linguistics (pp. 174-181). Association for Computational Linguistics.

[19] Mullen, T., & Collier, N. (2004, July). Sentiment Analysis using Support Vector Machines with Diverse Information Sources. In EMNLP (Vol. 4, pp. 412-418).

[20] Abbasi, A., Chen, H., & Salem, A. (2008). Sentiment analysis in multiple languages: Feature selection for opinion classification in Web forums. *ACM Transactions on Information Systems (TOIS)*, *26*(3), 12.

[21] Xia, R., Zong, C., & Li, S. (2011). Ensemble of feature sets and classification algorithms for sentiment classification. Information Sciences, 181(6), 1138-1152.

[22] Moraes, R., Valiati, J. F., & Neto, W. P. G. (2013). Document-level sentiment classification: An empirical comparison between SVM and ANN. Expert Systems with Applications, 40(2), 621-633.

[23] Astous, A., Colbert, F., & Nobert, V. (2007). Effects of country-genre congruence on the evaluation of movies: The moderating role of critical reviews and moviegoers' prior knowledge. *International Journal of Arts Management*, *10*(1), 45.

[24] Koh, N. S., Hu, N., & Clemons, E. K. (2010). Do online reviews reflect a product’s true perceived quality? An investigation of online movie reviews across cultures. *Electronic Commerce Research and Applications*, *9*(5), 374-385.

[25] Andrea Esuli Baccianella, Stefano and Fabrizio Sebastiani. (2010). SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In Proceedings of LREC, volume 10, pages 200-204.

[26] Musto, Cataldo, Giovanni Semeraro, and Marco Polignano.(2014). "A comparison of lexicon-based approaches for sentiment analysis of microblog posts." *Information Filtering and Retrieval* 59