**Sentiment Analysis on IMDB movie**

kING kHALID uNIVERSITY |Abha,

Introduction to Data science

AMAL ALSHAHRANI

‏2020

**Abstract:**

People can think, comment and express their emotions by challenging subject in machine learning which is called Sentiment analysis. People able to give a movie review which it can be automatically classified in categories. These categories can be user defined (positive, negative) or whichever classes you want. We consider the problem of classifying documents by overall sentiment, specify review is positive or negative. Using movie reviews as data, we find that standard machine learning techniques definitively outperform human-produced baselines. However, the three machine learning methods we employed (Feature engineering, Dimension reduction , Classification algorithms) and achieving both the lowest error rate of 0.7% and the highest F1 measure of 0.994. Some exploration of data and discussion about this, for error analysis and tuning, is also included through the work. Future work of this model may be focusing on some new approaches like deep learning, and introduction of embedded context information in the feature space.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**INTRODUCTION**

A lot of people prefer to entertainment by watching movies**.** we have experienced a growing in internet users along with increasing power of online review sites and social media has given birth to Sentiment analysis.

. Also, analyze news articles, blog posts, forum discussions, and other texts on the internet over a period of time to see sentiment of a particular audience. Therefore, feature of a product has got a major role in sentiment analysis. The goal is to identify overall people experience, and find ways to elevate all customers to “promoter” a specific movie, where they theoretically will watch more, stay longer, and refer other customers.

Sentiment analysis, also called mining opinion, is The field of study that analyzes people's opinions and feelings and appraisals and bring them and their attitudes and emotions towards entities such as products, services, organizations and individuals, issues, events and themes and Their qualities [3]. In applications such as Systems recommendation, 'what others think' has It was always important information during the decision-making process [4] such as The recommendation of the clothing in the Amazon region based on Sentiment analysis for reviews users.

Detected the application of sentiment analysis In multiple regions. In social media such as Facebook And Snapchat, there are gold mines to bombard the consumer and the statements of opinion (Lexalytics, and 2019), which can be Can be used for advertising and marketing recommendation Sign of a new friendship, etc.

Filled with competence, knowledge, abbreviations, and emoticons, a problem has non-trivial water Space in machine learning models. The sheer size Is the problem also. Hopefully, the model can successful Saved some hours conversion value Mountains of social data manually [5] You can apply these things also on the models of business intelligence. For example, sentiment analysis Statistics can be used to assess the rate of warming Of customers in a new product, modify marketing situation current and trying to please Customers in a better manner [1]. Like this Enabled approach to surveillance companies Air conditioning plan their work in real time, which can lead To assess the potential cost of marketing.

For the analysis a review of the film, the sentiment analysis means finding the mood The public about how to judge a certain movie [2]. For more details, the classification of documents from the user reviews based on More messing about. Some of the tasks of sentiment analysis based classification of binary such as positive and Negative, some may say the other classification of multi-level Like positive, somewhat positive, neutral, negative somehow.

**1.1 The Problem:**

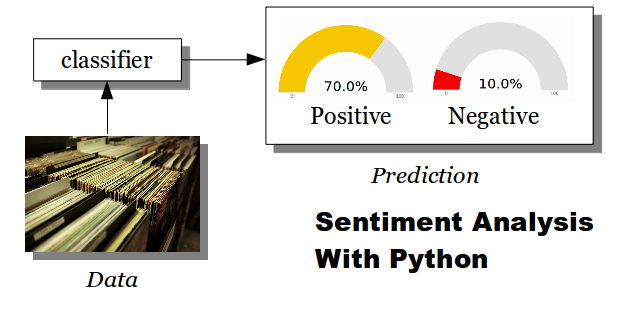
The Large Movie Review Dataset (often referred to as the IMDB dataset) contains 25,000 moving reviews (good or bad) for training and the same amount again for testing. The problem is to determine whether a given moving review has a positive or negative sentiment.

Sentiment analysis performed on the data to see the statistics of what type of movie do users like. The role of Sentiment analysis involves analyzing the textual data by identifying the emotion of the user, Positive or Negative.

**Literature view:**

This study's goal is to create a model of sentiment analysis on a 2000 rows IMDB movie comments by using machine learning; positive or negative preliminary information about the text is to provide.

For a long time, been dealing with sentiment analysis As the task of natural language processing at many levels of precision [7]. The first form of feelings Analysis on the document level. In the work of Turney [8], the learning approach is controlled The classification of the reviews into recommended (positive) or not Recommended (negative). There was also used to mark the PoS (part of speech) to determine Expressions with attributes or conditions. There are, however, attributes to them .In applications such as Systems recommendation, 'what others think' has It was always important information during the decision-making process, such as The recommendation of the clothing in the Amazon region based on Sentiment analysis for reviews users.

****

Detected the application of sentiment analysis In multiple regions. In social media such as Facebook And Snapchat, there are gold mines to bombard the consumer and the statements of opinion , which can be Can be used for advertising and marketing recommendation Sign of a new friendship, etc. Filled with competence, knowledge, abbreviations, and emoticons, a problem has non-trivial water Space in machine learning models. The sheer size Is the problem also. Hopefully, the model can successful Saved some hours conversion value Mountains of social data manually . You can apply these things also on the models of business intelligence. For example, sentiment analysis Statistics can be used to assess the rate of warming Of customers in a new product, modify marketing situation current and trying to please Customers in a better way. Like this Enabled approach to surveillance companies Air conditioning plan their work in real time, which can lead To assess the potential cost of marketing.

**Data Collection**

**Data Set** This is a dataset for binary sentiment classification containing substantially more data than previous benchmark datasets. We provide a set of 25,000 highly polar movie reviews for training, and 25,000 for testing. There is additional unlabeled data for use as well. Raw text and already processed bag of words formats are provided. See the README file contained in the release for more details.

* **Pre-processing:** for text consists of removing non-alphabetic characters, stop words (a set of very common words like the, a, and, etc.) and changing all words to lowercase. In this instance, we also need to remove HTML tags from the movie reviews. These steps can be packaged into the following function. Text is messy, people love to throw in attempts at expressing themselves more clearly by adding extravagant punctuation and spelling words incorrectly. However, machine learning models can’t cope with text as input, so we need to map the characters and words to numerical representations. (scaling)

**Data Preparation**

We will assume that we will use to model bag of words or maybe include the word don't require a lot of preparation.

Split into tokens

First, let me Protect Document one-take a look at the tokens to the raw material divided by vacuum. We'll use the function load\_doc () that have been developed in the previous section. We can use the function split () to submit the document to tokens separated by spaces white. Just look to icons raw can give us a lot of ideas for things you can try such as:

* Remove the punctuation from words (such as 'what').
* Remove icons that are just punctuation (for example '-').
* Remove the tokens that contain numbers (such as '10/10 '').
* Remove tokens that are not single letter (like 'a').
* Remove symbols that do not make much sense (such as 'and')

Some ideas:

* We can filter out punctuation from tokens using a series string ().
* We can remove the tokens that are just punctuation, or contain numbers using isalpha () on each token.
* We can remove stop words of English using the list that was downloaded using NLTK.
* Can we filter the tokens are short of during the control of the length.
* The following is an updated version of the organization of this review.

**METHODOLOGY**

In this section we present methodology adopted for conducting literature review, steps taken for this purpose include Phases of a Data Science Project, expected outcomes and summary.

We use two sets of data sets review of neo-realist films to stimulate emotions and emotional Rating. Difficult to use the reviews directly. As such, we need to apply engineering features Reduce the dimensions to create a feature suitable for resorts movies. After pre-processing, we The introduction of different classification algorithms to solve the problem of classification with the two classifications The problem of classification of multi-tags.

**Feature engineering**

First, we create a matrix to review the terms and words that have a frequency less than a The minimum specified and larger than the maximum specified. We believe such words such as Does not contain effective information of emotion, and handle the remaining words as the original Features all reviews.

Secondly, instead of using single words (unigrams) as features, the one way in which we live by building a water space is Based on features of 2 grams, which water is considered as a sequence contiguous from the two components of the selected item Revisions. For example, we might consider 'not good' as a term one rather than the 'No' and 'good' As two different, which will give the meaning more precise. There is another way is to provide a matrix of term frequency inverse (TF-IDF). The Going back to the number of words in each review in reverse through a number of audits which show where the dog In this way, words that appear frequently, which may not have significant contribution in the information, It will be no small weight.

**Dimension reduction**

Because each review contains hundreds of different words, the dictionary that covers all words great Matrix review term consistent. There is a lot of noise and redundancy in the Matrix, review the terms of the original, Matrix 2 gram matrix of TFIDF. We have implemented many of reduce the various dimensions of the Algorithms to reduce noise and redundancy in the original data. The roads considered in our project

Includes principal components analysis (PCA), and Principal Component Analysis Kernel (KPCA), the Multidimensional scaling (MDS) embedding of the linear locally (LLE). For PCA and KPCA Algorithms, the data is displayed in the original dimensions that have the greatest disparities, while Will KPCA projection data to the high-dimensional space with kernel functions. MDS computes the low Embed the dimensions who is trying to better the distances, the marriage between data points. LLE improves Decorative low dimensions and preserve the neighborhood for the input high-dimensional.

**Classification algorithms**

Our rating of the bipolar electrode sentiment classification and multi-classification for sentimentality Water associated with the reviews of the film. For the problem of classification of two categories, we use my algorithm rating classic of the two, the Naive Bayes Algorithms and logistic regression area. Algorithm Naive Bayes calculates the background Probability of each category for each review and picks the class with respect top. As for the Logistic regression regular, we apply methods of different organizations, such as the L1 and L2 network flexible, Based on logistic regression.

In the problem of classification multi-classifications, can realize the prediction in three ways, the first is Are the classic styles. In the first method, is the use of disaggregated multiple needs (such as Naive Bayes) To obtain the likelihood for each label. Is determined the decision on the registration when specific threshold of Likely. The second method is method One-vs.-Rest (OvR). In the OvR, the source of independent binary They are trained to distinguish one class from all other classes, and establishment of such Bank of each category. The truth is strictly the classification of posters on quality. One supplement as possible to classify multi-brand Are the groups RAndom K-labelss (printing Rakel). Consists printing Rakel of a set of workbooks for each of them Rated a random subset of small adjustments you think one category (PowerSet) and classified binary I learned how to distinguish between the labels in the PowerSet versus outside the PowerSet. Is to achieve the prediction by a The classification system across a range of works.

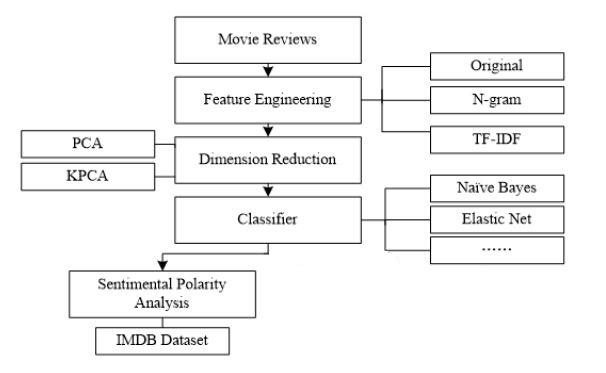


Figure 1 : Methodology

our Sentiment Analysis problem and Emotion Classification problem can be seen in the above figure. The first step is to perform feature engineering by generating a set of features using the movie review text data. This is done in one of three ways (single-words, 2-grams, and TFIDF). The feature set size is as follows:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Data Set | Sample Size | Size Single- Word Features | 2-Gram Features | TF-IDF Features |
| IMDB | 629 (sentences) | 728 | 1,205 | 728 |

The second step is to perform a dimensionality reduction on the feature sets that were created previously. We Proceeded to end up with three ways (there is no reduction, PCA, and Kernel PCA). In the heart of the nuclear Ingredients, the use of a function basis of radiosurgery as a core, and kept all components non-zero. In general, the number of components saved data IMDB 2000 and 626 Respectively. Between the first and second step, we created a total of 9 different sets of features. Finally, in the third step we apply many different. The methods that we followed are Naive Bayes and logistic regression regular (L1 and L2 network flexible).

**Evaluation**

The assessment methods our classification on the test data using the control intersecting 5-fold. Used in primarily two ways for evaluation. The first is the error rate, which is the ratio of the predictions are accurate to the size of the overall sample. While this simple procedure is widely used, it is not the only measure of the effectiveness of the source. Thus we also calculate the F1, which is expected to compromise accuracy and precision. For multiple tags, the production of a set of predictions (one per emotion) for each review, and sensitive assessment measures using all the emotions.

**Results**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | | | Training  Error | Testing  Error | Precision | Recall | F1 |
| Naïve Bayes | Original | | 0.030 | 0.337 | 0.655 | 0.690 | 0.672 |
| 2-gram | | 0.001 | 0.267 | 0.695 | 0.832 | 0.757 |
| TFIDF | | 0.008 | 0.351 | 0.638 | 0.690 | 0.663 |
| KPCA | Original | 0.067 | 0.082 | 0.920 | 0.916 | 0.918 |
| 2-gram | 0.121 | 0.187 | 0.785 | 0.862 | 0.822 |
| TFIDF | 0.254 | 0.361 | 0.731 | 0.442 | 0.551 |
| L2 Regularized Logistic  Regression | Original | | 0.000 | 0.160 | 0.843 | 0.835 | 0.839 |
| 2-gram | | 0.000 | 0.148 | 0.855 | 0.848 | 0.851 |
| TFIDF | | 0.028 | 0.160 | 0.831 | 0.853 | 0.842 |
| KPCA | Original | 0.000 | 0.499 | 0.501 | 0.800 | 0.616 |
| 2-gram | 0.000 | 0.499 | 0.501 | 0.800 | 0.616 |
| TFIDF | 0.000 | 0.486 | 0.517 | 0.432 | 0.471 |
| L1 Regularized Logistic  Regression | Original | | 0.000 | 0.186 | 0.816 | 0.812 | 0.814 |
| 2-gram | | 0.000 | 0.174 | 0.824 | 0.829 | 0.827 |
| TFIDF | | 0.209 | 0.247 | 0.726 | 0.812 | 0.767 |
| KPCA | Original | 0.005 | 0.007 | 0.987 | 1.000 | 0.994 |
| 2-gram | 0.041 | 0.041 | 0.987 | 0.930 | 0.958 |
| TFIDF | 0.333 | 0.428 | 0.571 | 0.577 | 0.574 |
| Elastic Net | KPCA | Original | 0.000 | 0.044 | 0.946 | 0.968 | 0.957 |

Figure 3: multi-label classification problem

In the multi-label emotion classification, we were not able to determine any one method as the best method. In general we were able to achieve an error rate around 17-20%, but our F1 measure indicate poor performance of our classifiers in identifying the true positives (i.e. predicting the presence of emotions correctly). Additionally, we saw that there was a trade off between error rate and F1 measure: in general, a decrease in error rate is correlated with a decrease in the F1 measure. However out of our best results, the Naive Bayes classifier was the most common classifier.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Problem | Classifier | Dimension  Reduction | Feature  Engineering | Training  Error | Testing  Error | Precision | Recall | F1 |
| Sentiment  Analysis | Naive Bayes | Kernel  PCA | Single-words | 0.067 | 0.082 | 0.920 | 0.916 | 0.918 |
| L1 Logistic  Regression | Kernel  PCA | Single-words | 0.005 | 0.007 | 0.987 | 1.000 | 0.994 |
| Elastic Net (l1 ratio = 0.85) | Kernel  PCA | Single-words | 0.000 | 0.044 | 0.946 | 0.968 | 0.957 |
| Naive Bayes | None | Single-words | 0.104 | 0.219 | 0.321 | 0.380 | 0.348 |
| Naive Bayes | PCA | Single-words | 0.179 | 0.197 | 0.351 | 0.330 | 0.340 |
| Elastic Net (l1 ratio = 0.15) | None | TF-IDF | 0.186 | 0.177 | 0.375 | 0.232 | 0.286 |

The table shows the best three results from Sentiment Analysis and Emotion Classification. In Sentiment Analysis, we saw that the use of KPCA had the largest increase in the performance of the classifier among the dimension reduction methods. The effect of KPCA on the performance was also the largest on the single-word features. Among the candidates of classifiers, the L1 Logistic Regression had the best performance, achieving both the lowest error rate of 0.7% and the highest F1 measure of 0.994. In general, we were able to achieve an almost perfect classification of the sentiment in movie reviews.

Conclusions

We found that the classification of the emotional much more difficult than analysis of emotions. There are some possible reasons. The first is the relative size of the data sets. For the group of data of the IMDB, because the sample was very small, I didn't have a source that we have trained the ability to distinguish between different emotions. The second related to the inherent nature of the problems. Describe in words generally either a Negative or positive emotions, but they can describe many emotions, and often these feelings are tied to significantly.

**References**

[1]Buitinck, Lars, et al. "Multi-Emotion Detection in User-Generated Reviews." European Conference on Information Retrieval. Springer International Publishing, 2015.

[2] Kim, Seungyeon, et al. "Beyond Sentiment: The Manifold of Human Emotions." AISTATS. 2013.

[3]Kouloumpis, Efthymios and Wilson, Theresa, and Moore, Johanna, “Twitter Sentiment Analysis: The Good the Bad and the OMG!” Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media, 2011

[4] Maas, Andrew L., et al. "Learning word vectors for sentiment analysis."Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1. Association for Computational Linguistics, 2011.

[5]Ng, Andrew and Jordan, Michael. “On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes”. Advances in neural information processing systems, 2002.

[6] Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan. "Thumbs up?: sentiment classification using machine learning techniques." Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10. Association for Computational Linguistics, 2002.

[7] Pang, Bo and Lillian Lee. “Opinion Mining and Sentiment Analysis”. Foundations and Trends in Information Retrieval Vol. 2, No 1-2 (2008) 1–135

[8] Roweis, Sam T., and Lawrence K. Saul. "Nonlinear dimensionality reduction by locally linear embedding." Science 290.5500 (2000): 2323-2326.

[9] Saul, Lawrence K., and Sam T. Roweis. "An introduction to locally linear embedding." unpublished. Available at: http://www. cs. toronto. edu/~ roweis/lle/publications. html (2000).

IMBD dataset: (<http://ai.stanford.edu/~amaas/data/sentiment/>