Anomaly Detection Model training for New York Times Bestseller Fiction Novels

M. Madhuram, Chinmaya Joshi, Ananthajith TCA, Anoushka Dutta

Abstract: The New York Times bestsellers list is one of the ultimate authority when it comes to ranking the best selling novels in the world. Their list is globally recognized and respected by everyone when it comes to which books are among the best. We propose an idea to use machine learning to predict in the early stages of a book's lifecycle whether they have a chance of becoming a NY Times Bestseller or not. The idea is to take the bestselling fiction novels dataset and train a machine learning model to teach it to try and recognize a pattern in all the novels which are in the list in between the years 2008-2018. The features used to train the model are the publisher's name, author's name, date of publication, title and description of the novel. This would be an unsupervised learning problem due to the reason that the dataset only contains the list of bestseller novels and not those which can be used to train the model to recognize what a non-bestseller novel looks like. This would be an example of a one-class classification. These type of machine learning problems are categorized as "anomaly/novelty detection" problems. In these problems, we train the model to associate the data with the bestseller novels, and when we check the model for some new data it checks whether that data is the same as those which the model is trained to recognize as bestseller and depending on that it classifies that specific tuple of data/novel as a bestseller or not (anomaly or not).

Index Terms: New York Times Bestseller, Novels, Fiction, Data Anomaly, Unsupervised, bag-of-words, One Class SVC

I. INTRODUCTION

Books have always been useful when it comes to pleasure reading or gaining information regarding a specific subject. They are irreplaceable no matter how many other types of entertainment and infotainment source are found. Novels allow the reader to understand and enjoy stories that may or may not seem plausible but accomplish their objective of providing entertainment. But in the vast and endless collection of novels available for the readers, they usually are unsure regarding which novels would be "worth the time". To give an idea for what novels are worth looking into, there are authorities and departments that collect data regarding books and based upon the figures and the factors collected recommend some novels specifically that are not to be missed. The New York Times Bestsellers list or "NYTBS" as they are colloquially called are one such example of list of such books that are used as a basis of classification for the best novels on the market from the rest. They collect data from the sales figures while keeping in mind some of their internal judging guidelines. The NYTBS listing has been categorized into multiple sub-lists spawning different categories, broken down by fiction and non-fiction; hardcover, paperback and e-books and so on. It is based on weekly sales reports that is obtained from specific samples of independent retailers or chain bookstores and wholesale throughout the globe. The sales figures represent books that have actually been sold at retail particularly as the Times surveys the booksellers in an attempt to better track the sales of each individual copy of that specific novel. However, the building of the list and the method that goes into inclusion of books in the list has always been criticized for being too easy to push through. The above mentioned sales records can be used as features for training the machine learning model for the said prediction.[1] Our aim is to analyse the data from previously declared bestseller fiction novels and get an intuition for what it means for a novel of becoming a New York Times Bestseller fiction novel. Ultimately we plan to train a machine learning model which can make predictions, with reasonably good accuracy and precision, the chance of a newly released novel of becoming a NYTBS in their near future.[2]

Manuscript published on 30 April 2019.
***** Correspondence Author (s)

M. Madhuram*, Assistant Professor, Department of Computer Science and Engineering, SRM Institute of Science and Technology, Chennai, India.

Chinmaya Joshi, Department of Computer Science and Engineering, SRM Institute of Science and Technology, Chennai, India

Ananthajith TCA, Department of Computer Science and Engineering, SRM Institute of Science and Technology, Chennai, India

Anoushka Dutta, Department of Computer Science and Engineering, SRM Institute of Science and Technology, Chennai, India

II. CURRENT EVALUATION SYSTEM

A. Working

The existing system that dictates as to which novels are compiled for the bestseller list make little to no use of algorithms or computational assistance whatsoever. The list is compiled by the editors of the "News Surveys" department but is published by the Book review department A novel is added to the list on the basis of its weekly sales reports obtained from particular samples of independent and chain bookstores and wholesalers throughout the globe. The sales figures are widely taken into account to represent the books that have actually been sold at such standalone shops or retail, rather than wholesale, as the Times' survey tries to include books in such a way the list emphasizes the books that are mainly purchased and read by individual buyers. At times a book does not need to be studied for weeks for it to be included in the list. Some books are shipped and have significant amount of retail orders that allow the book a pre-emptive spot on the list.

The books are compiled according to the highest sales estimate as well. The system has evolved compiling in such a way that a single list is now subdivided into multiple lists in different categories, broken down by fiction and non-fiction, paperback and hardcover, and various different genres. The list is mainly divided on the basis of fiction and non-fiction with each list containing 15 to 20 titles each week.

B. Drawbacks

However robust the shortlisting procedure might be, it has some downsides that can easily allow books to be included regardless of that book's genuine claim for the bestseller title. Some of these problems are:

a) Short boom of sales:
While the NYTBS (New York Times BestSeller) list reflects sales in a given week, it does not however take into account total sales. Thus, one book may sell heavily in a given week, making the list, whereas another may sell at a slower rate, never making the list, but latter given more time could easily be deserving of the opportunity to be put on the bestsellers list.

b) False sales count:
Wholesalers report how much they sell to retailer authorities, and they in turn report how many copies are sold to the customers. At this point there can be an overlap with the same book being sold multiple number of times within a given time period. In addition, retailers may return books to wholesalers at a later period in case they never sell, thus resulting in a false "sale" being reported.

c) Manipulation by authors/publishers:
Since the list judges the books mainly based on the number of copies of that novel being sold, there may be a possibility wherein the author/publisher themselves may arrange for a large chunk of the books to be bought by themselves such that while the actual sales haven't started yet, the book may already be termed as a "bestseller".

d) The "Bestseller" phenomenon:
When a book is termed a bestseller despite being a bestseller for a single week, the term itself causes an increase in sales as people tend to buy books for the apparent guarantee that the said book is highly recommended .

Thus while the current system has been using the fore-mentioned ideals and methods, it has a lot of problems that may easily be exploited causing a huge amount of errors to take place.

III. PROPOSED MODEL PIPELINE

The proposed idea for the pipeline shown in Figure 1.1 portrays the lifecycle of the system from the creation of the system to the use-case for the said. Our aim is to create a model which can look at some of the details of a fiction novel and predict whether it is NYTBS material or not.

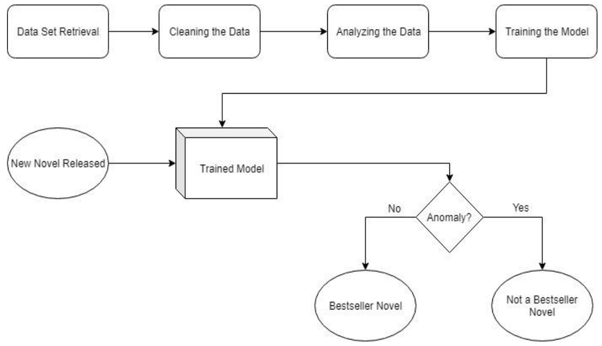

Fig. 1 System Model Architecture Flowchart

The first step for achieving this system is to feed the model a large datasets of New York Times Bestseller Fiction Novels[3] and then train the model to predict if a newly released fiction novel with its own characteristics can become a bestseller or not. After retrieving the dataset, it needs to be cleaned. For cleaning the dataset, all incomplete, inaccurate, incorrect or irrelevant data is identified and replaced, modified or removed. The actual process consists of data auditing, workflow specification, workflow execution, post processing and controlling. After the data is cleaned and is ready for usage, the data is then analyzed. The process of analyzing gives us a goal of discovering useful information, informing conclusions and supporting decision making by getting an intuition for the correlation in the data. The next step is the process of training the machine model. In this process, machine learning algorithms may be applied to data to identify relationships among the variables, such as correlation or causation. This allows the algorithm to form a pattern of sorts and look for anomalies. After the model is trained, it checks for new novels, collects their data and detects if any anomaly occurs or not. If an anomaly is detected, then the model predicts against the new novel becoming a best seller whereas if the data is not an anomaly, then the model predicts for a good chance of it becoming a best seller.

IV. DATA INTERPRETATION AND EXTRAPOLATION

A. Dataset Description

The provided dataset was initially in the JSON (JavaScript Object Notation) file format. For an easier working environment and handling of the data, the dataset was transformed into a '.csv' file format. It contained the relevant information of NYTBS Fiction Novels from June 7, 2008 to July 22, 2018. The total number of rows of data present in the dataset was 10,195. Each data tuple consisted of 12 columns though we utilized only 8 of them while training the model. This was mainly due to the discrepancy occurring in various feature-columns of the dataset. Some of these disparities of the data occurred in the form of missing values for a majority of the dataset for a specific column. One such example of this would be the "Price" column which had the values of 6184 novels as 0. That rendered 60.6% of the dataset worthless for use. Another disparity which led to the use of reduced number of columns was the presence of insignificant column data with respect to the aim of training a machine learning model.

The profound example of that would be the JSON ids assigned to each individual row of data which carried no significance in terms of a bestseller novel. Another such example would be the ‘weeks_on_list’ column. Though it carried a profound meaning when it came to the novel’s bestseller information, in our model’s prediction-state our aim is to give the data for a newly published non-bestseller fiction novel which would not have a feature such as that. Hence the ‘weeks_on_list’ feature was also not taken into consideration while the training of the model.

We also used feature engineering[4] to create a new feature called ‘title_word_count’ which gives the number of words present in the NYTBS novel. After the removal of the unusable columns, the following features remain which are used to train the model.

Table 1: Features used for training the model

S.No.	Feature Name	Feature Description
1	“description”	The description/synopsis of the published fiction novel.
2	“title”	The official title (name) of the published fiction novel.
3	“published_year”	The year of publication of the fiction novel.
4	“published_month”	The month of publication of the fiction novel.
5	“published_day”	The day of publication of the fiction novel.
6	“publisher_name”	The name of the publisher of the fiction novel.
7	“author_name”	The name of the author of the fiction novel.
8	“title_word_count”	The number of words present in the title of the fiction novel.

Another hurdle that arose while conquering this problem was the occurrence of redundant values in the dataset. In other words, a single novel’s entry was repeated multiple times in the dataset. This repeated occurrence was due to the nature of that specific novel being a bestseller for more than a week, but since our model need to classify whether the fiction novel would become a bestseller once, this redundancy was also omitted during the data analysis and training process.

B. Preprocessing

The preprocessing consisted of majorly two parts: one being the cleaning of the said data, second being the preprocessing for the non-numeric features in the dataset. The cleaning process of the data has been explained extensively in the last section. At the end of the data cleaning portion, the shape of the dataset had become 2328x8. The preprocessing section would mainly focus on the non-numeric feature handling portion. One of the biggest issue while working with machine learning algorithms and their implementation on a computer is the computer’s lack of ability of understanding non-numeric data. Almost all the machine learning algorithms rely on heavy statistical and mathematical application on the dataset for the desired output or result. The computer does not possess the innate ability of performing these complex mathematical functions on a string of data. As a result, we cannot train a machine learning model as long as the dataset still contained non-numeric values.

The obvious solution to this problem is to represent the initial non-numeric features into a format which is understandable by the machines. There exists 4 features that need these transformation for them be usable; namely “description”, “title”, “publisher_name” and the “author_name”. All these column’s data are present in the form of alphanumeric data-type and need to be represented in a purely numeric form. This area of working with textual data for machine learning purposes is categorized as “Natural Language Processing” (NLP).

We can consider a single column of data as an example since the same methodology would be used for the transformation of rest of the text-based feature sets. Let’s take the “description” feature column. Each individual tuple data for this column contains a description of the bestselling novel. In mathematical terminologies, we can say that the column is just a vector of descriptions where each individual vector value is a corpus of words. Our initial aim was to represent each individual words present in the corpus in terms of a number. Instead of doing so, the final solution to this step was found in the methodology of portraying each corpus of words in the form of a sparse matrix where the 0s would signify the absence of the word associated with that specific column of the matrix. This methodology is better know as “bags-of-words”.[5] After the above transformation is successfully applied onto each of the four non-numeric columns, the dataset reaches it’s final form where any mathematical operations can be applied on the dataset.

C. Data Analysis

The dataset is analyzed thoroughly for the objective of achieving an intuition for where the machine learning (ML) model being trained in the later stages may detect a pattern. This step is a crucial stage as this is the process of finding a correlation between the various features available within the dataset.

The graph in Fig 2 gives us an idea about the number of words used in the title of books that makes it to NYTBS list of fiction novels. In the depicted graph, the y-axis represents the count of books with the x-axis representing the number of words in the title of those books. The graph provides an intuition that the model may recognize a pattern which illustrates a margin between bestseller books with fewer words in their title and bestseller books with more than four words in the title being extremely rare. This graph conclusively represents the relation of bestselling novels with the number of words available in the titles. As there is a growth in the number of words used in title, we see an exponential decrease in the number of books becoming Best sellers. This gives us an vague idea of how titles of books contribute to their best seller’s position.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication (BEIESP)
© Copyright: All rights reserved.

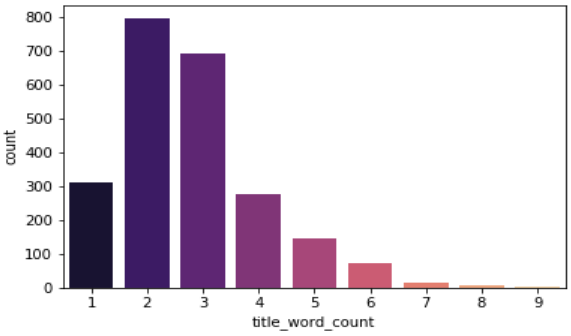

Fig. 2 Number of words in the Title vs Novel Count

The graph in Fig 3 provides the data of the number of books becoming a best seller feature on the New York Times Fiction bestseller list. This graph with its analysis gives us an idea of the distribution of published books becoming a bestseller after their release. The data doesn't show any major fluctuation apart from its initial and last month. The graph results in a balance among bestsellers over the course of ten years.

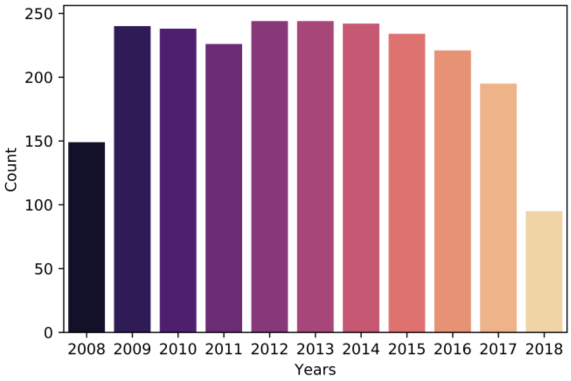

Fig. 3 Countplot for novel count each year

The graph in the Fig 4 represents the number of books sold in a span of 10 years that begin at the year 2008 and go up to 2018 with an emphasis on the individual month counts. The graph takes the months of the years in the x axis and take the number of books sold as y axis coordinate values. Each month has been numerically presented with 1 representing January and 12 representing December. From the graph we can observe that the amount of novels sold the highest was in the months of April, May and October while it is comparatively lower for other months. The lowest selling point is found to be at the months of January and December. It is also inferred that the books are preferred to be sold during the months of summer i.e. April and May. It is safe to assume that the selling of books during the span of these 10 years is maximum in the months of April , May and October while it is lowest during the months of January, February and December i.e. during the winter season.

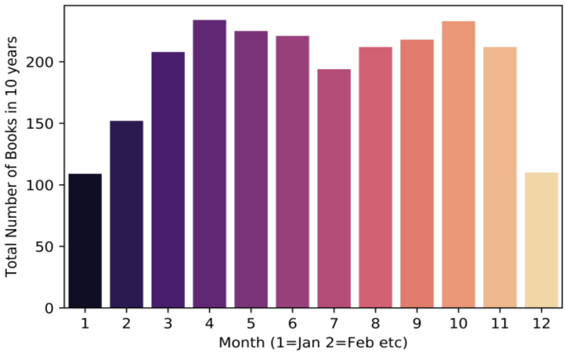

Fig. 4 Countplot for novels count each month

The graph (Fig 5) details the number of books published in each year during the period of 2008-2018 and also goes on to specify the amount of books published in each month during that specific year. The graph has the x axis representing the years between 2008-2018 and also has each month represented by color gradient bars. The y-axis specifies the amount of books published with a scale of 5 books each step. The year 2008 had the highest number of novels published through the entire graph with more than 35 books published in the single month. During the year of 2009, the highest number of books published is found to have been done in May while the least number belongs to that of the month of January and December. This trend can be seen throughout the entire graph as the amount of novels sold has a uniform distribution all around the years for each individual months. This concurs with the intuition we got from the Fig 4 of bestsellers count being higher in summer and fall seasons, and lower during the winter season.

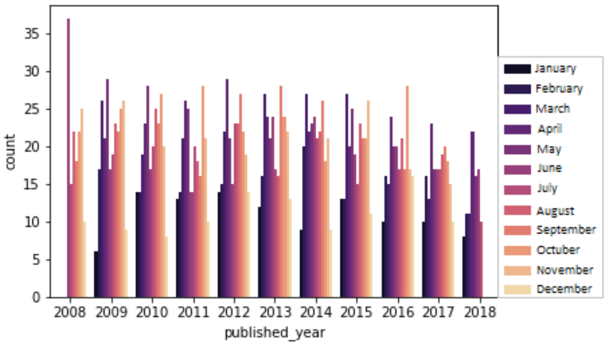

Fig. 5 Detailed Countplot for novel published

The below presented figure (Fig 6) gives us the visualization of words associated with the title of the bestseller novels of our dataset. The terms have been highlighted based on their frequency of appearing on titles of bestseller novels. Hence a wordcloud has been generated using the most used words in the titles of bestselling novels. There is a very subtle point to notice when examining this particular wordcloud. The dataset used consists of over two thousand books, but only a handful of words are used or rather chosen for title. This leads to a slight understanding of the disposition of words used in the title of the book in the fiction genre to attract the target customer base. The wordcloud achieves a pattern of words which are most used in bestselling novels. This process further helps to understand the anomaly within them.

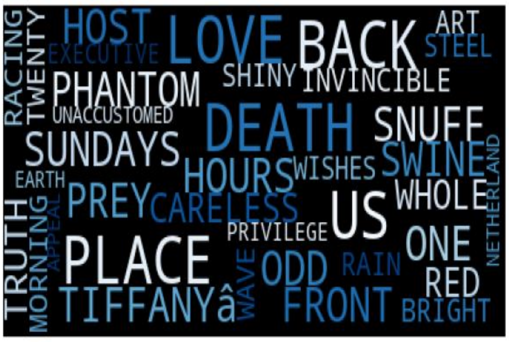

Fig. 6 Wordcloud for Title

The given wordcloud in the Fig. 7 diagram showcases the frequency of the particular words used in the tittles of the NYTBS fiction novels. In the dataset the occurrences of the words were counted after the cleaning of the datasets and it was observed that the words "man" and "woman" had the highest occurrences with 218 and 190 occurrences throughout the titles present within the titles of the novels in the dataset. The words "when", "life" and "detective" also had a total high count values of 119, 112 and 109. The lowest number of occurrences is given by the words such as "artificial" ,"silent" and "sapiens" with 3 ,2 and 1 occurrence throughout the 10 year span. Thus we can safely presume that novels with the words "man and "woman" etcetera have a greater probability of a New York Times Bestseller while the book titles with words from the bottom have a lesser chance of being one.

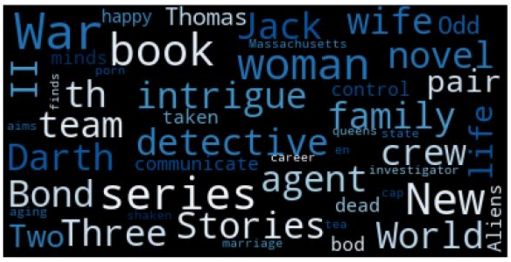

Fig. 7 Wordcloud for Description

D. Model Training

When dealing with datasets where the data stored is unlabeled, we usually use the unsupervised learning techniques for handling such a problem. The term “unlabeled” here signifies the lack of a “label” or “category” column for each tuple value. In the NYTBS dataset’s case, each novel data present in the dataset is of a bestseller fiction novel i.e. there are no tuple entry for a model to recognize a non-bestseller novel data. As a result, the infamous supervised classification algorithms such as Logistic Regression, Linear Discriminant Analysis (LDA), Support Vector Classification (SCV)[6], Classification and Random Forests (CART), use of Neural Networks etcetera cannot be implemented with this specific dataset. The unsupervised algorithm being used for the training of the model is “Anomaly/Novelty Detection” or “OneClassSVM” (OCSVM). In this type of machine learning problems, we train the model to look at a single class of data which in our case would be the NYTBS novels. The machine learning model tries to recognize the patterns within this data and associate the meaning behind them to the bestsellers. When a previously unseen data for a fiction novel is given to the model to make a prediction, it checks the data for the patterns that it had associated with the bestseller and if the pattern does not exist, it classifies this data as non-bestseller or “anomaly”. This is achieved by calculating the probability of the previously unseen fiction novel data of becoming a bestseller. Then that probability is compared with a previously determined epsilon value. Each individual feature probability is calculated and a cumulative probability is generated by taking the product of all the individual probabilities.

$P(x) = P_1(x_1)*P_2(x_2)*P_3(x_3)*\ldots\ldots\ldots\ldots\ldots\ldots\ldots*P_8(x_8)$

Here, $x = [x_1,x_2,x_3,x_4,x_5,x_6,x_7,x_8]$,

where x is the newly tested fiction novel tuple data and the x_1 to x_8 are the individual feature values for that tuple entry. The probability of individual feature is calculated using the Gaussian Distribution formula given in the below.

$$f_g(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{\frac{-(x-a)^2}{2\sigma^2}}$$

Symbol	Meaning
a	Mean of the column-g
sigma	Standard Deviation of the column-g
$f_g(x)$	Function to predict the probability of the column-g
x	Current testing value for column-g

Table 2: Gaussian Formula Variable

Once we calculate the probability of each individual feature, the product of all those 8 probabilities (one for each of the eight features) is compared with the previously determined epsilon value. If the probability is more than epsilon, the tuple is classified as a bestseller; else the tuple is classified as an anomaly.

V. CONCLUSION

We propose an anomaly detection model for New York Times Bestseller Fiction novels based on unsupervised machine learning, using which we can detect abnormal patterns in the behavior among the data which in turn allows us to suggest the behaviour of the various parameters of the New York times bestseller fiction novel and build a theoretical blueprint of novels that becomes the bestseller every year. Traditionally there is no method to predict if a book can become a bestseller novel or not, but research on this topic has been going on for a long time. This model offers to investigate the behaviour of New York Times Bestseller fiction Novel and then understand its pattern , analyse the growth and parameters of various books and then build a model that detects any book showing variation from the characteristic of the bestselling books. The main contribution of this model lies in its effectiveness and stability. This model can learn the far and near behavioral patterns of any subject provided.

VI. FUTURE WORK

The dataset taken for predictive modelling was solved with an unsupervised problem statement due to the unlabeled form factor of the data. If a dataset is created with both the label classes present, i.e. novels that are bestsellers and non-bestsellers, then a more efficient supervised algorithm such as LR, LDA, SVM or Neural Nets could be utilized for a better accuracy and precision generating model. Also, in the present dataset there were many redundant or unusable data which were not useful when analyzing the data. The "Price" column had majority of their entries as empty leading to the drop of that column for data analysis and model training all together. Getting the prices of them and using that data for the model training would also, intuitively, increase the efficiency of the model.

REFERENCES

1. Zhen Zhu, Jing-Yan Wang; "Book Recommendation Service By Improved Association Rule Mining Algorithm"; Sixth International Conference on Machine Learning and Cybernetics, Hong Kong, 19-22 August 2007
2. Keita Tsujia, Nobuya Takizawab, Sho Satoc, Ui Ikeuchid, Atsushi Ikeuchia, Fuyuki Yoshikanea and Hiroshi Itsumura; "Book Recommendation Based on Library Loan Records and Bibliographic Information "; 2014 IIAI 3rd International Conference on Advanced Applied Informatics; August 2014
3. "New York Times Best Sellers Hardcover Fiction Best Sellers from 2008 to 2018" dataset uploaded by Carlo "https://www.kaggle.com/cmenca/new-york-times-hardcover-fiction-best-sellers"
4. Suraj Maharjan, John Arevalo and Fabio A. Gonzalez, Manuel Montes-y-Gome, Thamar Solorio; "A Multi-task Approach to Predict Likability of Books"; 15th Conference of the European Chapter of the Association for Computational Linguistics; Volume 1, pp. 1217–1227
5. Jeffrey Pennington, Richard Socher, Christopher D. Manning; "GloVe: Global Vectors for Word Representation"; 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1532–1543
6. Pranam Kolari, Akshay Java, Tim Finin, Tim Oates, Anupam Joshi; "Detecting Spam Blogs: A Machine Learning Approach"; 21st National Conference on Association for Advancement of Artificial Intelligence (AAAI2006), July, pp. 1351-1356, 2006