Mining TED for understanding viewers preferences and behavior overtime

Dong Ziyuan (CSCI5502)
Department of Computer Science
University of Colorado
Boulder, CO USA
zido6913@colorado.edu

Ali Raza (CSCI5502)
Department of Computer Science
University of Colorado
Boulder, CO USA
a.raza@colorado.edu

ABSTRACT

Public speaking is an art, if given as an effective one, it can have an impact on the life of the listeners. Akin, TED Talks are famous for having a long-lasting impact on its listeners. So, our aim was to understand what makes TED talks so interesting in terms of the features and speakers of the talks. To better answer our questions, we preprocessed a data file which contains the information about the TED talks till 2017. We had views, comments, speaker's information, tags, languages, duration, events, and transcripts as some of the main attributes. After doing data preprocessing, we analyzed the data thoroughly and highlighted some insights from the Ted Talks and used different visualization techniques such as heat map, line graph, bar graphs and word cloud, for projecting our results. We also analyzed the important topics that were extracted based on each talk. For modeling, we implemented three algorithms to Ridge regression, Lasso and Elastic nets for modeling about Views and Comments. Elastic nets performed best in providing more accurate modeling to the Views and Comments. Whereas we also identified important features (Ted NYC, Ted annual, Ted Global, TedX) while modeling these attributes. We used the Mean square error as the performance metric for these three algorithms. We also did modeling for predicting Languages, Views and number of speaker for talks using three algorithms (ZeroR, Bagging, Random Forest) using WEKA [15]. We used Root mean square as an evaluation metric for the performance measurement of these algorithms. Random Forest performed best in predicting Languages and Comments. Whereas our baseline algorithm, ZeroR performed best for modeling the number of speakers. The features that are yielded as important in making the prediction from these three algorithms are: Duration and the Views of the videos.

KEYWORDS

TED Talks, algorithms, analysis, modeling

[*]Article Title Footnote needs to be captured as Title Note
[†]Author Footnote to be captured as Author Note

WOODSTOCK'18, June, 2018, El Paso, Texas USA

ACM Reference format:

FirstName Surname, FirstName Surname and FirstName Surname. 2018. Insert Your Title Here: Insert Subtitle Here. In *Proceedings of ACM Woodstock conference (WOODSTOCK'18). ACM, New York, NY, USA, 2 pages.* https://doi.org/10.1145/1234567890

1 Introduction

Data is generated in our lives every day whether it is in our regular routine or through some kind of action involving computer device. It is becoming more important than ever to mine the knowledge hidden in the enormous data that we are producing. When we say about mining then it does not mean that we need to focus on some subset of the data instead the idea is to generalize our findings based on an overall picture of the data. This can help us to understand the ratio of different patterns occurring together and at the same time providing us with the capacity of making a prediction [1]. For our problem, we have selected the TED dataset that consists of text and numeric data containing hundreds of tuples. Ted or TedX is one of the most popular shows that inspires many people around the globe. It brings speakers from all parts of the world explaining their thoughts, sharing experiences, and providing thought provoking directions for the viewers. But we want to understand which speakers were really influential and what was there topic that attracted so many people. More specifically, we would like to mine interesting patterns in the numeric and text data and also making sense out of data with the help of the graphical visualizations. Some of the things that we are interested in mining out of the TED datasets are:
Which themes are the most common among the TEDsters? Which speakers were most popular, in other terms what were the relationship between view, comments and tags/duration? Which features are most effective in predicting how much views a particular video is going to receive? Also, which features would be able to tell, how many comments did a video is going to receive? From the text transcripts of the talks, which topics are the talks most related to? How accurately we can predict the views, comments, language, speakers?
Mining the information out of this dataset will provide us with some interesting insights about the behavior of the viewers and about presenter's selections about a particular topic. Gaining knowledge about these key aspects will guide us in exploring

some interesting facts, relationships, and patterns in different attributes.

2 Related Work

In the section, we are going to provide an overview of the data mining techniques and its applications that are currently being utilized in different fields.

Data preparation is one of the most important steps for effectively applying the data mining techniques because of the following reason: i) Real world data is not clean, ii) High-quality mining algorithms requires high quality data, iii) Quality data will provide quality patterns [4].

In a research study, the top ten algorithms for the data mining are indicated with their significance impact of each algorithm with the future research on the algorithm that is being done. The list of the algorithms includes: C4.5, K-means, SVM, Apriori, EM, PageRank, AdaBoost, kNN, Naïve Bayes, and CART [2]. CLIQUE is proposed as a clustering algorithm that can tackle with the challenges like high dimensional data, scalability, insensitivity of the order of the input record. This method creates a cluster of data in the DNF expressions which are decreased for getting comprehension whereas the results of the algorithms suggest that it accurately finds clusters in high dimensional large datasets [3]. Ratanamahatana & Gunopulos [5] proposed a Selective Bayesian Classifier that selects only those features which are selected by the C4.5 decision trees when it is learning a subset example of the training set. The results show that the combination of Naïve Bayesian classifier and C4.5 decision trees outperform the Naïve Bayesian and C4.5 decision tree on all the ten datasets on which experiments were conducted. Whereas this hybrid approach also learns faster than both the algorithms requiring few examples to acquire the most accurate rate of classification.

Li & Huan [6] proposed a new learning model called constructivism learning with applying the Bayesian nonparametric model which can dynamically handle the creation of new learning tasks. Two key processes were hypnotized for transparent machine learning; a holding process where we enhance our existing learning of models, a second process where we create new learning models. The model was studied on the synthetic and real data and it performed well enough to provide high quality models and also providing the transparency of the learning process. Mautz et al. [7] expanded the K-means algorithm and proposed a new algorithm called SubKmeans which contains two goals in it; create a k-means style clustering partitioning and then move all the clusters into a particular common subspace. One of the most prominent advantages of this newly proposed algorithm is that it deals with the curse of high dimensionality. SubKmeans is also compatible with many extensions of the k-means algorithm.

Sathe & Aggarwal [8] expanded an application of the random forest algorithm to handle the arbitrary data. Whereas this approach also surpasses the perform of random forest on the multidimensional data. The authors named this expansion of the random forest as Similarity forest. In order to work with the arbitrary set of data objects, similarity must be computed between the data objects and also this method computes a very small amount of the $O(n^2)$ pair similarity for the construction of forest between the objects. Wu et al. [9] improved the time complexity of running the pairwise preference between the items by proposing the Primal-CR and Primal-CR++ algorithms. Both algorithms are an expansion of the collaborative ranking system. It is the only collaborative ranking algorithms capable of working with the full Netflix dataset consisting of 20 billion rating pairs and also provides a better recommendation as compared to simple collaborative ranking. When considering the ranking loss while making the recommendations, the algorithms outperformed the classical matrix factorization in terms of the efficiency. Currently, that's the only algorithm of making better recommendations by ranking the un-rated items. Li et al. [10] used the actigraphy data to predict the Alzheimer's disease in the old age people with the application of multivariate time series classification method TATC. Modeling effect for Circadian rhythm is used with Neural deep learning approach for the time aware attention data. Whereas this approach also provides an analysis of the daily activities of the participants. TATC was compared with four other algorithms and it out performed them in recognizing the disease. This can facilitate doctors in providing timely feedback and guidance to the old patients. Rather than waiting for the symptoms to occurs this approach will provide an analysis of all the patient's data by monitoring their daily activities. Beeck et al. [11] calculated the fatigue prediction in the runners in the outdoor environment and intimating runners accurately when a fatigue injury is going to take place. Three models for learning were used: All runners model, other runners only model, and individual model. The models were evaluated using the four regression techniques; Gradient boosted regression trees, artificial neural network, linear regression with elastic net regularization, and linear regression with least absolute shrinkage and selection operator regularization. The results produced successful prediction of the runner's fatigue with respect to placing the inertia sensors on different parts of the body. Silvis et al. [12] implemented a mobile based notification system to select and notify the users intelligently for donating the excessive food to the food banks. Reinforcement learning was used to determine how many notifications to send to which particular users. A dataset consisting of 1000 users were used. Whereas these baseline algorithms were used Frequent First, Round Robin, Pantry First. Whereas FIPS linear value model was used to judge user on the basis of metrics and booQ algorithm at selecting the user for notification. Finally, a comparison between booQ at correcting user probability score bias to the baseline algorithms is made. The results show that the implemented technique will facilitate the reduction of the food waste. Liu et al. [13] did a study on finding the similar exercises in the online education system with proposing a new Multi model attention based Neural Network (MANN). In the first step, convolutional neural network is utilized for the image representations. Then an Attention based Long Term Short

Term is created to learn each representation of the exercise having both text and images. Further, a similarity attention was designed to measure the similar parts in the exercise. Finally, to return similar exercise a pair wise strategy was used. Three evaluation metrics were used; Precision, Recall, and F1 measure to evaluate the performance of MANN across with the different other algorithms. The results showed that using MANN approach in all three-metrics used, it was highly effective in returning pairwise exercises. Cao et al. [14] proposed a prediction method for the early detection of the Bipolar affective disorder using the mobile phone text entry. Participants of the study were provided with a text phone in which the customized keyboard was installed for recording of the input text of the user. DeepMood architecture with a multi view machine layer for data fusion was the best predictor algorithm with an accuracy prediction rate of 90.3% when compared with other predictors algorithms. With the use of this application, an early intervention for the Bipolar affective disorder can be made and can be helpful for the patients. SVM and Random forest algorithm performed best in making prediction about the fire structures in the Arizona. When compared with the other algorithms Logistic regression and Gradient Boosting. SVM reached a performance of 71% as true positive and Random Forest performed at 69% with the true positive. Using an interactive map for displaying results Arizona department of fire was able to identify almost six thousand structures which were at the risk of fire [23].

3 Proposed Work

In this project, we aimed to mine text and numeric data consisting of different large files. We wanted to highlight significant patterns and meanings out of our work that can help us find the meaning of our research questions highlighted in section 1. After finding answers to our questions we would like to generalize our findings based on how to be a successful speaker or which topics are appraised by the audience mostly. From the implementation point of view, we will preprocess our data by recovering incomplete data, purifying data, resolving data conflicts. Whereas while preprocessing of our data, to name a few we worked the with following preprocessing functions; attribute selection, sampling or instance selection, handling missing values. Further, after the preprocessing of the data, we apply classification algorithms to identify the patterns in the data. After applying these algorithms then we aim to show our results using visualizations that can show the effectiveness of the different algorithms that ran on it. Whereas we also evaluate our algorithms using the Root mean square error as the metric.

4 Data analysis TED Talk

For the data analysis purpose, we cleaned the data and handled the missing values. Our main data file consists of the main description about the speakers and the talks they gave including date of the talks and when they published but all the data was in Linux timestamp so we converted the time stamp into human readable form.

4.1 Views and comments

From our analysis of the top ten viewed and commented talks, we found that they all have been released after the year 2006. We used the Pearson correlation coefficient to find the relationship between the views and the comments; we get a value of 0.531 which led us to understand that there is a positive-moderate correlation between these two attributes.

4.2 Popularity of TED Talk by month and year

We used several visualizations; bar charts, heat maps, and running line graph over time to explore the information for the popularity of the TedTalk overtime. In the figure 1 below, we show a heat map to produce show the popularity of TedTalks by month and time. We were able to find out that since the February 2009 the TedTalk reached the new peak of popularity each year since that mainly because every February new TedTalks are released and with the passage of time, more talks have been produced. We figured out the reason why there was a boost in the popularity of the TED talk and it was because in 2009 two new talks were added into the show which led to massive popularity.

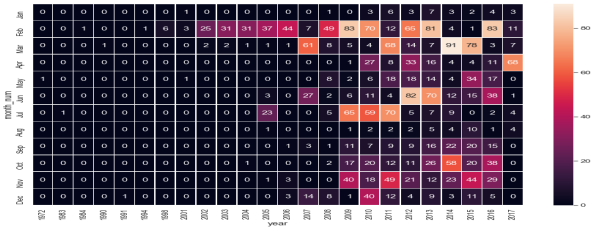

Figure 1: Popularity of the talks since its inception

4.3 Popular speakers and Talks tags

We did analysis to recognize which were the most popular speakers by recognizing how many time one particular speaker is invited for the talk. Hans Rosling, Juan Enriquez, Rives, Marcos Tempest and Clay Shirky were the most invited speakers with the subsequent number of talks: 9,7,6,6,5. Each talk have different tag associated to them and we consider these tags as the themes for each talk. The most talks given to top five tags/themes are Technology, science, global issues, culture, design, business and entertainment. With the talks given about each theme and tag to be: 561,455,393,339, 299,263,221.

Another aspect that we tried to understand is the trend about these tags/themes overtime and how are they getting enhanced or decreased overtime. So, figure 2 below provides an overview of the trends that are significantly enhanced or decreased overtime. We calculate the trends from the 2009 since the popularity of the TedTalks reached the peak; we found the technology tags are increasing every year. Whereas science tags provide a sudden growth and decline each alternative year and very interestingly global issues tags are on the decline since the 2009. We infer that technology related tags are on the rise since the 2009 because of the inception of new technologies; block chain, augmented and virtual reality, 3D printing, more engaging gaming experiences to the global audience all over the world.

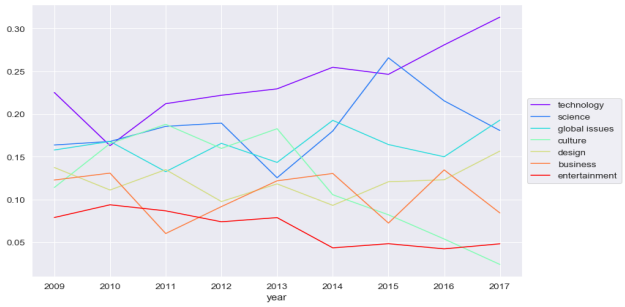

Figure 2: Trends of top tags from 2009

Another interesting question that we were trying to explore was which particular theme will harness more views. We explored an interesting phenomenon using the boxplots and found out that culture is one of the most important tag that have received most acclamation in terms of the view.

4.4 Correlation between duration and views, speed and views

We were interested in understanding the relationship between the duration of the videos talks and the number of views that each video gets. So, we used Pearson correlation methodology to find the correlation between two attributes and found it to be 0.049 which shows there is little or no correlation between the two values so duration of the video has number impact on the number of views of that video.

Also, we were interested in understanding if there is any relationship between the speed of a talk (Quantity of the words spoken by the speaker) and the views for the talk. We count the numbers of words for all transcripts from the other dataset file, and divided them by corresponding durations to show the speed of each talk. So, we found a correlation between the two attributes by 0.066, which shows that both of them are not significantly correlated to each other.

4.5 Correlation between rating, views and comments

As rating is one of the most important attribute because it is used as a parameter to help understand the perspective of the viewers using the fourteen defined tags to express themselves. So, we grouped these ratings into the positive and negative ratings to help us understand the relationship of rating with the views and comments. From the positive rating scale, we selected the 'inspiring' as the value to calculate the correlation with View and Comments. So, we calculated the Pearson Correlation Coefficient between the Rating and Views which we found out to be 0.775 and then we found the correlation between the Rating and comments, which comes to be 0.559. It provides us with this discussion that there is a very strong relationship with the value 'inspiring' and the views. Whereas also there is a positive correlation between the 'inspiring' and the comments.

4.6 Related Videos Network

We construct a related videos network where a node represents a talk and an edge means one talk is on the list of recommended watches of another talk. Then we calculated the degree, closeness, commonness, Eigen vector for centrality for each node and sort all the talks by each centrality score. And the talk "12 truths I learned from life and writing" given by Anne Lamott has the highest score on all four centralities. So, this talk plays the most important role to attract users watch other videos by browse the recommended list. This talk involved 12 themes and its views is in the top quarter.

4.7 Word analysis

For our transcripts data file, we were curious on which words were spoken or repeated in the transcript of each speaker. So, we analyzed the transcripts and found the words which are most popular in the Ted talks in the figure 3 below.

Figure 3: Most popular words in the transcripts of TED talk

4.8 Popular Ratings

We analyzed the ratings which were rated by the different viewers for understanding the perspective about each video from the user point of view. So, we divided the ratings into positive and negative for better understanding this phenomenon. We found that Inspiring, jaw-dropping, informative, fascinating, Funny and beautiful as the most positive rated. Whereas Obnoxious, ok, confusing, unconvincing and Longwinded are the most negative rating tags assigned to a video by the viewers.

The videos that have the duration more than the twenty-five minutes are mostly rated as positive by the viewers as a percentage of 92 percent. In total, there were 14 videos that were greater than the twenty-five minutes.

4.9 Arbitrary Findings

In this section, we will talk about the arbitrary findings that we explored using the WEKA analysis tool [15]. These findings highlight the maximum, minimum and some patterns based on the binning of the different attributes based on taking event as the main class for implementing the classification. For the attribute view, these are the findings: the minimum number of views that a video talk received took place at the university of California with 50443. The most views received as an event program were by the TED global in 2012 with 43155405. Almost forty-two of the events received the views in this range of 64 to 124 thousand. This finding provides us with the generalizing that most of the TED Talks event generally receive a viewership of greater than fifty thousand.

Some of the interesting arbitrary findings from the language attribute are: The videos which was translated into most language was by Matt Cunt have inn TED global 2011 by the name 'Try something new for 30 days'. There are 401 video talks which are translated mostly in the range of 28-31 in different languages which represent sixteen percent of the data. To understand the distributions of the languages over all

the TED talks, the distribution of the languages over the binning was made up to 30 bins and almost eighty percent of the videos were translated in the languages between these six ranges. The table 1 below shows the distribution of the languages based on the videos.

Language	Major Talks
33-36	224
31-33	227
29-31	401
26-28	271
24-26	260
21-24	366
19-21	169

Table 1: Language range of the most talks

Most videos released: TED 2009 and 2014 have the most videos released in these two years with the numbers 83 and 84. A couple of noticeable events which released subsequent great number of videos during these years apart from the mainstream TED talks are TED India and TED women with 34 and 35 videos each. This finding helps us understand about the importance of the individual TED talks that are organized apart from the mainstream TED Global. Time duration of the overall videos from the TED talk: The table 2 below shows the range of the more than half of the TED talks into the six bins. Whereas in total there were almost twenty bins. The talks that fall in these categories are almost fifty seven percent. So, there is chance that any new video which is made will fall under these five ranges with almost fifty-five to sixty percent of the chance.

Video Time Range	Major Talks
19-18 min	240
18-16.5 min	325
16.3 -14.8 min	290
14.8-13 min	254
13.3-12 min	250

Table 2: Time range of the most talks

We also found interesting patterns based on the comments by understanding the differentiation among the videos in five binning. As ninety two percent of the videos fell into the range of these patterns. Table 3 below shows the distribution of these comments in the five categories, as it covers ninety two percent of the videos. The most commented video received sixty-four hundred and four comments by the Richard Dawkins video on the militant atheism.

Comments Range	Major Talks
2-70	750
70-140	695
140-208	409
208-277	234
277-346	143

Table 3: Comments range of most talks

5 Data Preprocessing:

In this section, we explain our process of Data preprocessing and also detail about the dataset that we used for analysis and modeling purposes.

1. Data Cleaning & Data Integration

Our dataset includes some unofficial talks held a long time ago. So, we removed them. The dataset consisted of two documents. One is the basic information of all the talks and another one is the transcripts of these talks. Two documents have only one common attribute named “url” of which values are non-identical to each other. So, we use it as the identification to combine a transcript with an information tuple of the corresponding talk.

2. Data Reduction

Attributes “description” and “title” are textual descriptions about the talks. Attributes “name”, “main speaker” are irrelevant to the mining task. Besides, the values of attribute “speaker occupation” and “tags” are a completely confusing. So, we have to remove these attributes.

3. Data transformation

The format of date in this dataset is UNIX timestamp which isn't readable and computable. So, we transform the date to "day-month-year" format. Then we calculated the length of time of each talk from published date to the present to replace the attribute “published date”. Because we thought the former was more related to the views and comments. We also counted the average "views" of all the related videos of each talk to replace the attribute “related videos”. Moreover, the values of attribute “event” include about four hundred kinds of different events which required concept hierarchy generation. So, we categorize these events into 6 classes and transform the attribute “event” to 5 binary attributes representing 5 classes respectively. When the value of these 5 attributes are all equal to 0, this talk belongs to the sixth classes. TED provides fourteen positive or negative adjectives for audience ratings. The attribute “ratings” contains the votes for each adjective of all the talks. Ratings shows how popular a talk is. So, it’s meaningful to predict the ratings of a talk. But it is too difficult to predict specific votes for each adjective of a talk. We classified these 14 adjectives into 5 classes (“very bad”, “bad”, “moderate”, “good”, “very good”, each class contains 2-3 adjectives) based on their meanings. Then we added five new attributes to represent the fractions of votes for each class to the total votes for a talk.

4. Other tasks

Whereas we also handled missing values, changing nominal to numeric data, applying normalization for understanding the

results, understanding which models will suit best based on the data since we mostly had numeric data, using binning to range the data attributes in different categorize in order to better understand the data and relationship between attributes. Assigning topics to all the transcripts, eradicated the unwanted attributes by removing them.
In addition, we removed the attribute "tags" before. So, we need a new similar attribute that can highly summarize the contents of the talks, which is useful for prediction. Therefore, we extracted 15 topics from the transcripts with the help of non-negative matrix factorization (NMF). By this method, we could know how relevant to each topic a talk is (a score from 0 to 1) and use 15 new attributes to record the results. Each topic with top 5 most related words extracted from transcripts and figure about this is provided in the supporting material of the project. Finally, we standardized all the attributes for a more accurate result. After the data pre-processing, the remaining 26 attributes are list in table 4.

Attribute	Format
Duration	Numerical value (unit: sec)
Events (5 binary attributes)	Boolean
Languages	Numerical value
Length of time	Numerical value (unit: day)
Fractions of votes for each new rating class (5 attributes)	Numerical value from 0 to 1
Degrees of membership for each topic (15 attributes)	Numerical value from 0 to 1
Views	Numerical value
Comments	Numerical value
Speakers	Numerical value

Table 4: Data attributes used for modeling

6 Data Mining Algorithms for modeling

To help answer our questions about the different predictions that we want to make in terms of the success rate about the particular video, the rating or popularity in terms of being positive for the released video.
Here's a list of algorithms that we used to work with for answering our questions: Ridge Regression (We labeled the rating as positive and negative), Lasso regression, Elastic nets. We implemented these three algorithms for making the predictions for the views and the comments. In the subsections below, the details about implementation of these algorithms is explained which also details about why the implementation of each particular algorithm has taken place. Whereas we also utilized the Weka data analysis tool [15] for applying different classification algorithms.

6.1 Ridge Regression

Ridge regression is a biased estimate that improves the least square method. When the dimension of a dataset is higher than its size, multicollinearity of the features will lead to the overfitting and the cost will be sensitive to input errors. To solve this problem, Ridge regression adds a penalty term. This process is called regularization. Ridge regression set penalty term at a L2-norm. So, it's difficult to shrink the regression coefficients to 0. It means that all the features have influence on the prediction result. After the data pre-processing, the dataset included 26 attributes and 2250 tuples. So, Ridge regression was suitable for our task.
For making the prediction on the views and the comments based on which features are the most important for the prediction. We used the Ridge Regression since it uses a linear model for making the predictions. We opted for this type of regression because we had less then hundred thousand samples and we wanted to take into account most of the features for making the prediction and
how Ridge results would be impacted because it does not make a selection based on the feature [16].

6.2 Lasso Regression

Lasso regression is very similar to Ridge regression but differs in few important aspects. The only difference is that Lasso use a L1-norm as the penalty term. We can use this property to remove the influence of these irrelevant attributes.
Lasso uses the property of the feature selection, which is not present in the case of the Ridge regression. In Lasso, some of the coefficients can be reduced to zero or shrinked. Lasso is used when we have more number of features and it automatically detects the feature [17]. In our case, we provided Lasso with all the features and wanted to check the impact on the performance. It made Lasso regression also suitable for our task.

6.3 Elastic nets

Elastic nets are the hybrid of both Ridge regression and the Lasso. As it handles the penalty of both the regression types discussed previously. Elastic nets linearly combine the penalties of the Lasso and Ridge (L1, L2). It generally gives good performance in case of large datasets but it can't be always true especially when the dataset is small. In such cases, the performance of the Elastic net may not be significant [18,19].

6.4 Using WEKA for implementing algorithms

Our major aim to use WEKA was based on understanding how different algorithms work and how the performance of each algorithm can vary based on different parameters. We first wanted to utilize the basic algorithms that we have studied in the class to understand the performance difference of those algorithms. Because this can help us understand the underlying functions of each algorithm. We had to do a rigorous amount of data cleaning because WEKA does not allow to input any data file which has any special characters and other warning in it which require the data preprocessing. After the completion of the data preprocessing, we applied different supervised and unsupervised attribute filters (handling the missing values, add

classification, add clustering, converting nominal to numeric). The reason behind applying the filters is to open different algorithms in the Weka because when we input the file in the Weka then it already recognizes which algorithms can be applied on this dataset. For instance, to apply the Linear Regression algorithm we had to convert all the nominal data into the numeric. So, we converted the attribute event into distinct numeric attributes. This helped us to unlock the option of Linear Regression algorithm, albeit we did not consider extending the use of this algorithm. Akin reason was also that we wanted to try out simple algorithms first that we understand the underlying process and the performance first. In the next subsections, we explain the use of each algorithm and why it was considered as a choice.

6.4.1 Holistic view of the modeling attributes

In this section, we are going to visualize the relationships of the attributes (Languages, Speakers, Comments) that we have selected for the modeling using the Weka. In the figure 4 below, we are showing a visualization (Scatter plot) where each point represents an instance. The visualization shows the relationship between the language (x-axis) and the comments (y-axis). The majority of the videos are translated into the language range from eighteen till forty-three. We also recognized an outlier, with the most comments sixty-four hundred and translated into the forty-two languages.

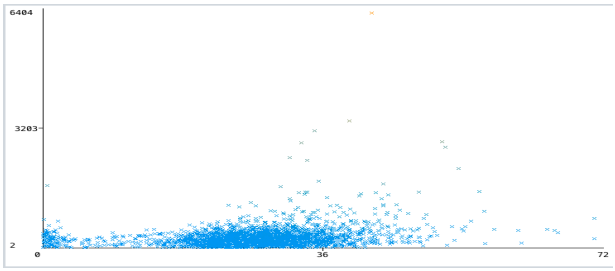

Figure 4: Relationship between Language and Comments

Akin, we also analyzed relationship between other attribute Speakers with the Language. The main findings from these relationships are: Most of the Ted talk events have mostly the one speaker in the event. The video that was most translated into the different languages; seventy-two times. That particular talk was also given by the one speaker only. Albeit we detected an outlier, at most there have been five speakers in an event (TED 2017) and it was translated into the twelve languages. The details about this and other such analysis is submitted in the material submission for the course.

6.4.2: Algorithms for prediction

We were interested in modeling three attributes (Language, Speaker and Comments) and understanding their performance on different algorithms. In this section below, we define why we selected each algorithm and a brief introduction about basic concept of each algorithm. The performance for each algorithm is explained in the next section 5 (Evaluation).

ZeroR:

We selected the ZeroR algorithm as a baseline algorithm since it is one of the simplest and straight forward algorithm. By simply predicting the majority class in the nominal data and average of the data in the numerical data, it provides a baseline performance for any dataset provided [20]. It constructs a frequency table and select the majority value from that table. This helps us to set a benchmark in terms of the performance for other classifiers. Albeit, it is important to remember that in some cases ZeroR can have a better performance than the other algorithms whom it is compared with, this also indicate us towards the problem of overfitting. We ran the algorithm based on the three attributes (Languages, comments, number of speaker) that we were interested to explore.

Random Forest:

Random forest was selected as an algorithm because it is an meta ensemble method and constructs a multitude decision tree. Based on this idea that Ensemble methods would generally perform better then standalone methods. Random forests utilize a combination of the tree predicators and the value of each tree is dependent on the value of random vector which is sampled independently having equal distribution for every tree in the forest. It creates a diversity by alternating branches in the tree. Similarly, this method utilizes the tree function so we opted for the random forest. It uses the averaging for controlling the prediction accuracy and also handles the overfitting of the data [21].

Bagging:

We opted for using the Bagging predictor since we had the most important attributes as the numerical. Another reason to use this method was because it is one of the most widely used ensemble method. It provides with the facility of classification and regression depending on the base learner. Bagging predictor algorithm generates many versions of a predictor and then aggregate the best predictor by averaging and use the voting system for predicting about a class [22]. Bootstrap copies the versions of the learning set and then making new learning sets from these one's.

7 Evaluation

For conducting process of evaluation on each algorithm. We used Root mean square error (R^2) as the metric since the modeling that we are applying for the attributes are numeric. We used 10 folds cross validation for the creating a balance between the training and test data set. Which means that our datasets would be divided into ten parts and then randomly each part would be used for the training and test purpose. We adopted this methodology in both the implementation using Python and Weka.

7.1 Evaluation for algorithms (Ridge Regression, Lasso, Elastic nets):

We implemented three algorithms from the regression family for modeling about the attributes Views and the Comments. For the first implementation of the algorithms, we did not perceive a highly correlated model for the attribute. So, we had to categorize the event values into six class labels for achieving high accuracy. Whereas we also defined classes for the attribute rating into six categorizes, based on the fourteen adjective words which a Ted viewer can assign to a particular video. The table 5 below shows the result of the three regression algorithms.

	Ridge	Lasso	Elastic net
Views	0.0175	0.2928	**0.0918**
Comments	0.0234	0.2642	**0.0290**

Table 5: Evaluation results for Ridge, Lasso and Elastic nets

Based on the accuracy of the modeling of these algorithms based on the attributes Views and Comments. Elastic nets performance surpassed the other two algorithms. We perceive the results of the algorithm because Elastic net usually performs since they utilize a hybrid approach in shrinking the data and selecting the important features. Whereas the result about the coefficient of these three algorithms are not that different. For understanding which features mostly contribute to the process of modeling the Views, we found out that based on the events; Ted NYC, Ted annual, Ted Global, TedX are the one which impact the process of modeling.

7.2 Evaluation for algorithms on Weka:

In order to better comprehend our results from algorithms, we tried to understand the role of the overfitting of the data also. Using seed value from (1-5), each algorithm had to reapply the model for five times for achieving a result based on the mean of the five settings. This helped us explore, is our data is overfitting or not. The table 6 below shows the information about the five runs on each algorithm (ZeroR, Random Forests, Bagging) based on R^2 value. The value outside the brackets informs about the accuracy using R^2 with seed value of one. The values inside the brackets are the means after five rounds of iteration showing accuracy based on R^2 with different seed values. We also normalized our data so we can easily interpret the differences in performance. For the performance measure, we take the mean values of the random seeds. As we can comprehend from the results that, Random Forest predicted best in terms of the Language, as there is the least different in the data points (Predicted and original). Similarly, ZeroR produces the best output in terms of the number of speaker for the prediction of the speakers. And finally, Random Forest provides the best output for the comments attribute as compared to the rest of the algorithms.

	ZeroR R^2	Bagging R^2	Random Forest (R^2)
Language	0.1322(0.1321)	0.1033(0.1033)	0.0952(**0.0952**)
Speaker	0.052(**0.052**)	0.0533(0.0525)	0.055(0.0545)
Comment	0.042(0.042)	0.0389(0.0388)	0.0376(**0.0379**)

Table 6: Evaluation results for ZeroR, Bagging, Random Forest

By taking means using different seeds values, we understand this that there is no overfitting of the data. As the difference between the default seed value result and the mean seed value result is very minimal.

For the accuracy that we have achieved, we used the six attributes after doing the data pre-processing on it. We got interested that which attributes are most useful in providing this accuracy. So, we utilized the functionality wrapper subset evaluation for understanding the importance of the different attributes. Duration, Languages, and the Views are the most important attributes based on the Random Forest algorithm. Whereas Duration, Comments and View attribute are most important based on the Bagging algorithm.

Project progress breakdown:

For the course project, as a team we followed these different milestones that helped us to identify the progress of the project or foresee any new challenge. We list our milestone below for each week:

i) Week 1-2: Data preprocessing, exploring toolsets and libraries
ii) Week 3-4: Data preprocessing, experimenting or learning the libraries, WEKA
iii) Week 5-6: Applying algorithms on our dataset with some more data preprocessing
iv) Week 7-8: Using different algorithms to compare and contrast the results,
v) Week 9 onwards: Summarizing the results, trying to make meaningful interpretation out of data by doing analysis and documenting the results. Gaining reasoning knowledge about the results and algorithms.

8 Discussion

With the review of the literature and also along with the discussion between the team. We have largely realized the importance of the Data processing technique and how it can have an impact on the accuracy of your results. So, it is important to establish the data cleaning parameters first at the start of the project. We also recognized the difference between data analysis and data modeling. Considering data analysis as an integral part for better understanding our data in terms of depth. By understanding our data which was mostly numeric, we came to recognize that which algorithms would work best on it. So, we selected regression based algorithms (Ridge regression, Lasso, and Elastic nets) for the implementation. We used 10-folds validation for separating the data into the test and training set. After the selection of these algorithms, the next challenge was to select the evaluation measure that is going to help us provide the comparative differences in performance. Since we were utilizing mostly the numeric data so we selected the means square error as the metric for all the six algorithms that we implemented. We found out that Elastic nets performed most accurate in modeling the attributes Views and Comments. We relate this success of Elastic nets with the hybrid approach it uses for shrinking and selecting the attributes. Whereas labeling the data and reducing the dimensionality also helped improve the performance of the algorithms. As we wanted to diversify our skill set, we also explored and learned WEKA [15] for modeling the attributes Languages, Comments and Speakers. So, we preprocessed and cleaned the dataset again using different filters and techniques.

We selected ZeroR as the baseline algorithm as it is simple and provides with a straightforward model. Whereas our other options for applying algorithms were: Bagging and Random Forest. Since we idealized that using and ensemble method would probably provide us with better performance. As our most of the data we were going to made our model was numeric so we used Root Mean Squared error as the evaluation metric. Random Forest algorithm provided better performance for modeling two attributes and ZeroR performed better performance for one attribute. For understanding that are our algorithms overfitting the data; we wanted to understand the estimate in variation of the results so we set the random seed and repeated the experiment with an increasing value of random seed. We found the results to be minor different in the case of the Bagging and Random Forest. But the results in the case of the ZeroR stayed the same at each seed value. So, we infer that since ZeroR always select the class with the highest probabilities that's why it is giving the same values regardless of changing the random seed.

9 Conclusion:

With the exploration of this dataset using different tools and techniques, it has provided us with immense insights on how to plan for making a talk successful. It is likely that if a video has views there would be more comments on it. We also found out that since February 2009 TED talk have reached a new peak of popularity every year. We infer, it is because since TED released its more videos every year since then. Hans Roling was the most invited speaker in TED events nine times. Whereas most talks are given on these themes: Technology, Science, Global issues, Culture, Design. Since 2009, the popularity of the Technology talks increases every year due to the emergence of new technologies like Bitcoin, AR, VR and applied Machine Learning. We also recognized that if a video is made based on the Culture as a tag then it is going to harness more view as compared to others. There is no relationship between the frequency of the words spoken in a talk and the views it receives. Inspiring tag has a very strong relationship with the views and the comments a video receives. The words most commonly spoken in a talk are: 'One, Now, People, Know, Think, Will'. Each Ted Talk would generally receive more than fifty thousand views. Almost eighty percent of the Ted Talks have been translated into languages from a range of 19-33. Only two events in 2013 and 2014 released more videos after the global events are Ted India with 34 videos and Ted Women with 35 videos. Almost sixty percent of the videos fall under the time range of 13 minutes to 19 minutes. Also, ninety two percent of the videos have comments in the range from 2 till 346. Based on the talk transcripts, these topics (Women, Music, Brain, Power, Water, Country, Design, Data, Universe, Energy) were the one's o which most videos were made.

We applied six algorithms in total for modeling the prediction on the attributes like Languages, Comments, number of Speakers, Views. Firstly, we implemented Ridge regression, Lasso, Elastic nets from the regression family since we mostly had the numeric data for working. The root mean square method (R^2) was used as an evaluation metric for these algorithms. Elastic nets outperformed and made accurate predictions then the other two algorithms based on making prediction for views and comments. For understanding the features which are most crucial in making the predictions, we recognized in terms of the Ted talks (Ted NYC, Ted annual, Ted Global and TedX), as the most important ones.

For modeling, we also used WEKA [15] for applying three algorithms (ZeroR, Bagging, Random Forest). The attributes we predicted about were languages, number of speakers, and comments. ZeroR was selected as the baseline algorithm because of its simplicity. After five rounds of repeated experimentation with different seed values, Random Forest predicted the most accurately about the Language and the Comment. Whereas ZeroR predicted most accurately for the number of speakers. We found the top most important features in predicting are: Duration and Views of the videos. These features were rated most important by both Bagging and Random Forest algorithm.

After doing the analysis about Ted, we have identified what will constitutes a Ted Talk or any informative talk for being successful**:**

Based on profession: If you are a writer, psychologist, and journalist then your chances of popular talks are extended.

Based on time: Talk should be between 13-18 minutes

Based on themes: Technology, Science, Global issues, Culture can gain most popularity

Getting more Views: Try to add some element of 'inspiration' in talk, this will garner more views.

Receiving more Comments: More comments would be received if you talk about some element of psychology or world affairs in your talk.

With the analysis and modeling of this rich dataset, we have gained valuable insights and we have identified major themes that are important and can contribute in identifying what can constitutes a successful talk. Whereas we also learned how we can use modeling for predicting different numeric data and what are the most important features used in the prediction.

10 Contributions by Team

We both worked equally on the project on all parts starting from start till end. Since we sat together throughout the semester when we implemented and wrote the results in the report. We both set exact roles for each day when we started learning and implementing the algorithms. Then we shared and this allowed us to trouble shoot each other. For a detailed breakthrough of the activities, we have listed them by week in the section 5 of the evaluation.

ACKNOWLEDGMENTS

We would like to thank our Professor for providing us with a strong starting point in data mining and allowing us freedom to explore the tools for data mining.

REFERENCES

[1] Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. AI magazine, 17(3), 37.

[2] Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., ... & Zhou, Z. H. (2008). Top 10 algorithms in data mining. Knowledge and information systems, 14(1), 1-37.

[3] Agrawal, R., Gehrke, J., Gunopulos, D., & Raghavan, P. (1998). Automatic subspace clustering of high dimensional data for data mining applications (Vol. 27, No. 2, pp. 94-105). ACM.

[4] Zhang, S., Zhang, C., & Yang, Q. (2003). Data preparation for data mining. Applied artificial intelligence, 17(5-6), 375-381.

[5] Ratanamahatana, C. A., & Gunopulos, D. (2003). Feature selection for the naive bayesian classifier using decision trees. Applied artificial intelligence, 17(5-6), 475-487.

[6] Li, X., & Huan, J. (2017, August). Constructivism Learning: A Learning Paradigm for Transparent Predictive Analytics. In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 285-294). ACM.

[7] Mautz, D., Ye, W., Plant, C., & Böhm, C. (2017, August). Towards an Optimal Subspace for K-Means. In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 365-373). ACM.

[8] Sathe, S., & Aggarwal, C. C. (2017, August). Similarity forests. In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 395-403). ACM.

[9] Wu, L., Hsieh, C. J., & Sharpnack, J. (2017, August). Large-scale Collaborative Ranking in Near-Linear Time. In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 515-524). ACM.

[10] Li, J., Rong, Y., Meng, H., Lu, Z., Kwok, T., & Cheng, H. (2018, July). TATC: Predicting Alzheimer's Disease with Actigraphy Data. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (pp. 509-518). ACM.

[11] Op De Beéck, T., Meert, W., Schütte, K., Vanwanseele, B., & Davis, J. (2018, July). Fatigue Prediction in Outdoor Runners Via Machine Learning and Sensor Fusion. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (pp. 606-615). ACM.

[12] Silvis, M., Sicilia, A., & Labrinidis, A. (2018, July). PittGrub: A Frustration-Free System to Reduce Food Waste by Notifying Hungry College Students. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (pp. 754-763). ACM.

[13] Liu, Q., Huang, Z., Huang, Z., Liu, C., Chen, E., Su, Y., & Hu, G. (2018, July). Finding Similar Exercises in Online Education Systems. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (pp. 1821-1830). ACM.

[14] Cao, B., Zheng, L., Zhang, C., Yu, P. S., Piscitello, A., Zulueta, J., ... & Leow, A. D. (2017, August). Deepmood: modeling mobile phone typing dynamics for mood detection. In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 747-755). ACM.

[15] WEKA: Data mining software, available at: https://www.cs.waikato.ac.nz/ml/weka/

[16] Model selection, available at: http://peekaboo-vision.blogspot.com/2013/01/

[17] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological), 267-288.

[18] Zou, H., & Hastie, T. (2003). Regression shrinkage and selection via the elastic net, with applications to microarrays. JR Stat Soc Ser B, 67, 301-20.

[19] Elastic nets, available at: https://www.analyticsvidhya.com/blog/2017/06/a-comprehensive-guide-for-linear-ridge-and-lasso-regression/

[20] Witten, I. H., Frank, E., Trigg, L. E., Hall, M. A., Holmes, G., & Cunningham, S. J. (1999). Weka: Practical machine learning tools and techniques with Java implementations.

[21] Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

[22] Breiman, L. (1996). Bagging predictors. Machine learning, 24(2), 123-140.

[23] Madaio, M., Chen, S. T., Haimson, O. L., Zhang, W., Cheng, X., Hinds-Aldrich, M., ... & Dilkina, B. (2016, August). Firebird: Predicting fire risk and prioritizing fire inspections in atlanta. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 185-194). ACM.