The University of Queensland

Project Proposal

Classification of News Headlines

March 29, 2017

By,

Arjun C (s4463233@student.uq.edu.au)

Supervisor,

Prof. Neil Bergmann

Contents

Table of Figures

1. Introduction

News, its digital data is being created in a flashing rate. Most often, we humans can identify the category the news belongs to by just reading the headlines. This capability is inherent to us humans. An intriguing question has been if computers can think or if they can be made to think. Given the learning ability to classify data, computers can extend the same for such similar requirements and tend to be smarter.

Classification is one of the most commonly used machine learning techniques in categorizing the available data. Classification of news based on just the headlines using the bag of words model will serve as a stepping stone in understanding the model. The use of machine learning techniques like Support Vector Machines, Naïve Bayes and Logistic Regression for the classification will provide an interpretation of the complexity, accuracy, efficiency and feasibility of the technique for this type of dataset and help us in deciding technique to be used. This approach could further be implemented for classifying documents, images with appropriate design changes.

Objectives

The main objectives of this project are:

- To learn and implement accurate data mining techniques and the bag of words approach in preprocessing the data.
- To implement different classifiers to perceive the mathematical approach behind the algorithms.
- To evaluate their complexity, effects of complexity of the model, performance and feasibility.

2. Background

News is a report of current affairs. Thousands of news articles are published every day. They are made available globally to the people in numerous ways: media, broadcasting and electronic availability being the major sources. With today's advancement in technology, a noticeable increase in the number of websites and the applications that can be used to subscribe for news can be seen. Similarly, social networking sites have been a platform for the diverse media organizations to publish variety of information. One such platform is twitter, which serves as one of the easiest sources to refer to because of the word limit for posts. However, the sites publishing RSS feeds are something worth using as they are customizable to suit our preference of news type thus saving our time scrolling through everything. This makes the user experience better. To customize subscriptions, firstly, one needs to know the user's preferences. Secondly, the system should be aware of the categories of the existing articles. Categorizing news manually is one of the ways achieving the second requirement. However, machines that can learn from the existing pattern saves a lot of manual effort.

UCI machine learning repository [1] is a collection of datasets for machine learning enthusiasts to evaluate the learning algorithms. This project is based on the news aggregator data set provided by UCI machine learning repository which has around 400,000 instances that are grouped appropriately. The dataset contains a large collection of data that can be categorized to business, entertainment, health and science & technology.

Machine learning is a field in computer science, using which computers can be made to learn without being programmed to do normal sequence and conditional operations. This is one of the major building blocks of the intelligence in the intelligent and the smart machines. The machine analyzes the data and recognizes a pattern in the data. This pattern is later used to fit new data. The types of problems machine learning can handle are supervised learning, unsupervised learning and reinforcement learning. In supervised learning the machine is given input and the respective desired output using which the pattern mapping is done. In unsupervised learning the computer is given just the inputs and the computer has to find a pattern by its own. In reinforcement learning, the computer is given a goal to achieve like defeating an opponent in a game of chess.

In this project we make use of supervised learning as we have both the input features and the respective desired output. The input would be the headlines with other relevant features and the desired output would be the class the news headlines are categorized to.

Classification is a type of supervised learning technique. It can be extended to classify multiple classes. There are a number of classifiers like Support Vector Machines (SVM), Logistic Regression, and Naïve Bayes etc. available to solve this problem. The technique should be decided considering the data complexity, model complexity, and computational power of the system and prior usage of these algorithms for solving similar classification datasets.

Scope

The scope of this project is to consider data belonging to two of the categories in the news aggregator dataset for the complete implementation of the project. The data from two classes are separated for the preprocessing step and then when the final inputs are prepared after data mining, they are fed into the different learning algorithms to get models and later predictions are verified. Multi class classification is not considered at this moment as the implementation is just an extension of two class classification.

The reason for choosing this project is the vastness of the problem-set the approach could be used to target in future and to apply the learning algorithms learnt in the machine learning course to build classifiers.

3. Literature Review

The key elements to look for when we see how to go ahead solving such a text based classification problem are a few simple yet crucial questions. This is because calculations with numbers is what is easier with the equations than text. The research presented in this field are apt for any beginner to understand the approach and try to tackle the issues that arise in the process.

The data has to be cleaned before it could be used for any processing. The use of bag of words approach is more evident for text to numeric interpretation in the preprocessing phase. The extended use of this approach in image classification is worth noting. The performance of the techniques used seem to be more dependent on the dataset we are training on and the complexity and many such factors.

Below are some of the questions to consider to understand the previously used and existing approaches and methodologies to solve this problem.

a. Should we differentiate the letter casing and how do we handle punctuations?

The headline that is the text to be considered for the classification can have punctuations or the same words could be in upper case or lower case in different headlines. This makes it difficult for the computer to understand that they are the same. This is what makes data noisy and training models less accurate.

The authors Fageeri et al. [2] put forward a discussion that the missing data or any unnecessary features of data that appears as noise on analysis should be disregarded and removed as being part of data. Considering this, it is reasonable to remove the punctuations and change all the text to same case so that it will be easier to identify the pattern in the available headlines.

In [3], Dilrukshi et al. use an approach where the most used and the least used words are removed as they do not carry much information. Their statement seems very thoughtful as there is a very high chance of the most repeated words like “the” to be present in most of the sentences and do not carry any required information for the learning algorithm or some very rarely used words like “Draconian” which are hardly seen in any messages.

These provide an understanding on how the elimination of words should be handled.

b. What are the approaches to handle text in preprocessing?

The commonly used approach is Bag of Words model for text classification.

As seen from the implementation by Dilrukshi et al. [3], an approach called the Bag of words in which all the words are concatenated into a vector and the frequency of the words in the vector is calculated for each news message thus providing the input features for that instance.

Their remark on the increase in the dimensionality with considering n-gram approach thereby resulting in an increase in the complexity is something to be employed to solve this problem. If there is a possibility to solve the problem in a simple way, it is always better to keep it simple, bearing in mind the accuracy and efficiency.

The calculation of term weightage and not just term frequency as implemented by Hyun and Minyoung [4], also similar to what was discussed by Zhen Niu et al. [5] gives a very practical approach to solve the problem. This helps to reduce the noise and increase the probability of getting a more efficient hypothesis.

The implementation of twitter sentiment analysis by NanWang et al. [6] gives a clear picture of using n-gram model. However, their statement on not using dictionary based approach where the analysis is made on just the term frequency is what raises a question. Sentiments are either positive or negative and since only few terms could be categorized into each group, it should be possible to classify the messages as positive sentiment or negative sentiment with very less computation.

The data mining process implemented by Ajeng and Kusprasapta [7] gives a very practical approach to solve this problem. The sequential steps case folding, tokenizing, filtering, stop word removal and stemming similar to ideas presented in other papers gives a strong confirmation on using this approach.

In the article by Figueiredo et al. [8], their idea on considering co-occurrence of non-adjacent words gives a new contemplation for implementing the word frequency approach.

From all the implementations verified it is evident that the bag of words approach is the method to tackle the problem of using the text for classification. However, the way bag of words is used is also important. This could be decided by the type of data that is being taken into account for classification, the required complexity for training and the computation capability.

c. Should we train the model with the complete data available?

In any type of learning it is always necessary that, be it humans or machines, they are to be taught the method and not the solutions for all the problems. The problems are to be solved by the methods learnt. If solutions for all the problems are taught, it is more of memorizing the solution than understanding the method. In that case when a new problem is encountered the error rate would be very high. As put forward by Alpaydin [9], it is always important to divide the data to train, validate and test the hypothesis learnt.

d. What classifiers can be used for training?

There are a number of classifiers that could be used for such a binary class classification or the multiple class classification. From the implementations and research considered [10, 11], it is evident that Support Vector Machines, Naïve Bayes, Logistic regression and Random forests

are more suitable to solve this type of classification problem. There are advantages and limitations for every technique over the other.

The discussions on complexity, performance and computational requirement of Support Vector Machines in [8], their advantages and disadvantages provides a deep insight on how they are dependent and how it may apply to other techniques as well. Also, it is evident from the discussions from the papers that some techniques outperform the other for specific dataset and the complexity. It is always better to train a few models that are suitable to solve this issue and they choose the best one by comparing them.

4. Implementation

The implementation is divided into individual steps which put together form the whole classification process. This gives a rather parallel approach to carry out the implementation thus providing an opportunity to save time required to complete the independent modules.

The input for this project will be the headlines and the associated features as will be discussed later. This data will be processed and the resulting feature input for the training models will be the bag of words with appropriate frequencies and term weightage. The techniques to be compared will be Naïve Bayesian estimation, Logistic Regression and Support Vector Machines and the time taken to generate the hypothesis will be noted. Once we train the data and we have a model, the validation step is carried out. Appropriate errors like training error, validation error and test error are also noted. Validation dataset is used to fine tune the model parameters if required. Then testing is done on the remaining test dataset and necessary parameters noted down. The time taken and the accuracies serve as the criteria to compare these techniques and evaluate the feasibility of the model for this problem set.

The project would be carried out using Python due to the ease of coding, availability of machine learning library and the extensive amount of support from the python machine learning community.

4.1 Data retrieval and selection

The required data is downloaded from the UCI machine learning repository [1]. Any instances with missing data is removed. This data is further divided to get just the entries that belong to two pre-selected categories. Further, this separated data is divided 80% into training set, 10% into validation dataset and 10% into test dataset.

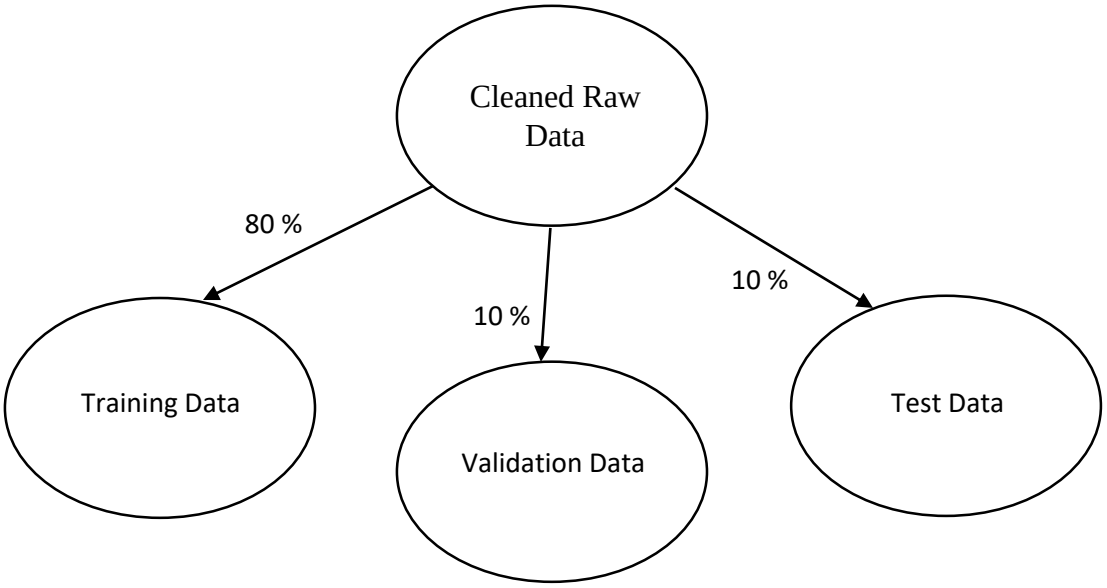

Figure 1. Splitting data for Training, Validation and Testing [7]

The training dataset is used to train the different classifier models considered. The same training dataset cannot be used to validate the generalization and performance. The validation dataset is used to fine tune the model parameters and to select the best suitable classifier for this problem set. Then the test dataset is used to get the categories and the error/accuracy of the model. This test data need not be from the same dataset, the model should be able to classify any current real-time data provided.

The dataset consists of the following features.

Features	Description	Sample
ID:	Numeric ID of the article.	1
TITLE:	Headline of the article	"Fed official says weak data caused by weather, should not slow taper"
URL:	URL of the article	"http://www.latimes.com/business/money/la-fi-mo-federal-reserve-plosser-stimulus-economy-20140310,0,1312750.story\?track=rss"
PUBLISHER:	Publisher of the article	Los Angeles Times
CATEGORY:	The category of the news item. *b* for business, *t* for science and technology, *e* for entertainment and *m* for health.	*B*
STORY:	Alphanumeric ID of the news story that the article discusses	ddUyU0VZz0BRneMioxUPQVP6sIxvM
HOSTNAME:	Hostname where the article was posted	www.latimes.com
TIMESTAMP:	Approximate timestamp of the article's publication, given in Unix time (seconds since midnight on Jan 1, 1970)	1394470370698

Table 1. Data description with sample [1]

The important feature that has to be considered for the classification is the “Title” for every instance. The “URL” and the “Hostname” could also play a role in simplifying the training and prediction process and to give a better accuracy. The “Category” field is the desired output that the classifiers use to map the input features and generate a hypothesis. “Story” and “ID” fields do not include any important information as they are just the IDs. “Timestamp” does not seem to play a major role in the classification process.

By eliminating the unnecessary fields by inspection, required fields are retained and further processing is carried out on the data.

4.2 Data Preprocessing

The separated and selected data is then required to be preprocessed to the desired format, in this case the bag of words. The data preprocessing is divided into five main steps, case folding, tokenizing, filtering, stop word removal and stemming [7]. All the punctuations and numbers are removed before going ahead with the below sequence.

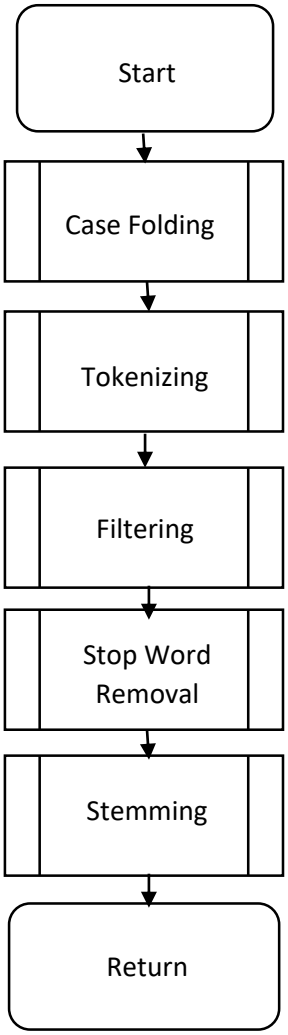

Figure 2. Data Preprocessing [7]

Case folding: The headlines may contain same words in different casing. To avoid the confusion for the system, all words are converted to same casing, usually lower case.

Tokenizing: Words are separated in this step to get a vector of tokens using which the word frequency can be calculated for each news headlines.

Filtering: Since the data we are considering is from the digital media, there are chances of the headlines having the URL to any relevant sources. These have to be filtered out as they can act as noise in the input features.

Stop word removal: The removal of the most occurring and the least occurring words are done in this step. The threshold for this removal is decided by analysis of the data.

Stemming: In this step, the derived words are replaced by the parent word to reduce the chances of noise in that data. E.g. The words serious, seriously, seriousness are all changed to the parent word serious to avoid confusions for the system.

Once we get all the tokens with the term frequency for each headlines, the bag of words, this has to be fed into the next process. Before going ahead with that, the word weightage has to be taken care of for the words have a very high probability of appearing in a specific category of news and not in the others. E.g. "Cricket" has a very high probability of appearing in the sports category and not very much in any other category, similarly "Economy" appearing in business category. This weightage can make the training more precise and results in better hypothesis and thus better prediction.

By the end of this step we will have the bag of words with word frequency for each headlines which is the input instance. This will have the weightage manipulation incorporated as found advantageous. Next we go ahead with the training process.

4.3 Training

Once the feature extraction is complete and the inputs for the classifiers are final, these are fed into the different models chosen, Support Vector Machines, Naïve Bayesian classifier and Logistic Regression. The time taken to train each model is noted for comparison purpose.

Bayesian Classification

This technique is based on the probabilistic approach. The probability of an instance "X" belonging to class "C_i" given probability of the class and probability of "X" can be calculated using the below equation [9].

$$P(C_i|X) = P(X| C_i) * P(C_i) / P(X)$$

This can be implemented in many ways. Using the above equation, the probability of an instance belonging to both the classes is calculated and the decision is made by the one with the greater probability [9].

Logistic Regression

Logistic regression is a simple approach in which the categorization is made using logistic function [7]. It is a statistical model that has been used by many for similar applications. Logistic regression uses the sigmoid function to calculate the probability by fitting a linear model to the data [9]. The independent variables like headlines, the information from the host name act as predictors for the category that we want to estimate.

Support Vector Machines

This is yet another effective model to tackle the classification problems. This technique finds the optimum hyper plane separating the classes [12]. This is a non-probabilistic approach. In this technique, the classes are separated with a plane which is at a maximum distance from instances of each class. When new instances are given to predict, the algorithm groups them to either of these separate spaces of the two categories thus giving us a category it belongs to [13].

These techniques are used to train the data respectively and to build the hypothesis for each technique. Once the hypotheses are generated, they can be used to validate the generalization, performance and accuracy. Validation gives a better idea on deciding the better model.

Validation is carried out using the validation split of the dataset. The same procedure as testing is followed for validation. Even though this may seem unnecessary, this can be used to fine tune the parameters like kernel, type of decision boundary to be fit etc. (e.g. fine tuning number of hidden layers in a deep neural network).

4.4 Testing

Once the hypotheses are ready for each technique, they are tested with the testing split of the dataset. The test dataset is fed to the hypotheses and the outputs are collected for each technique. These outputs are then cross verified with the desired output from the test dataset. The error percentage is calculated.

Also, to monitor the performance of the models, the training time and the testing time are noted down. Once we have the predictions, we can go ahead and plot the error matrix as shown below.

Error Matrix		Predicted Class	
		Class 1	Class 2
Actual Class	Class 1	1000	20
	Class 2	40	780

Figure 3. Error matrix [3]

Error matrix gives a visual representation of the number of instances that were classified correctly and the number of instances that weren't classified as in the desired output. Using this for all the three methods, the errors/accuracy of each method can be calculated.

Using these error/accuracy we can plot a graph comparing their errors and also the time taken for training and testing.

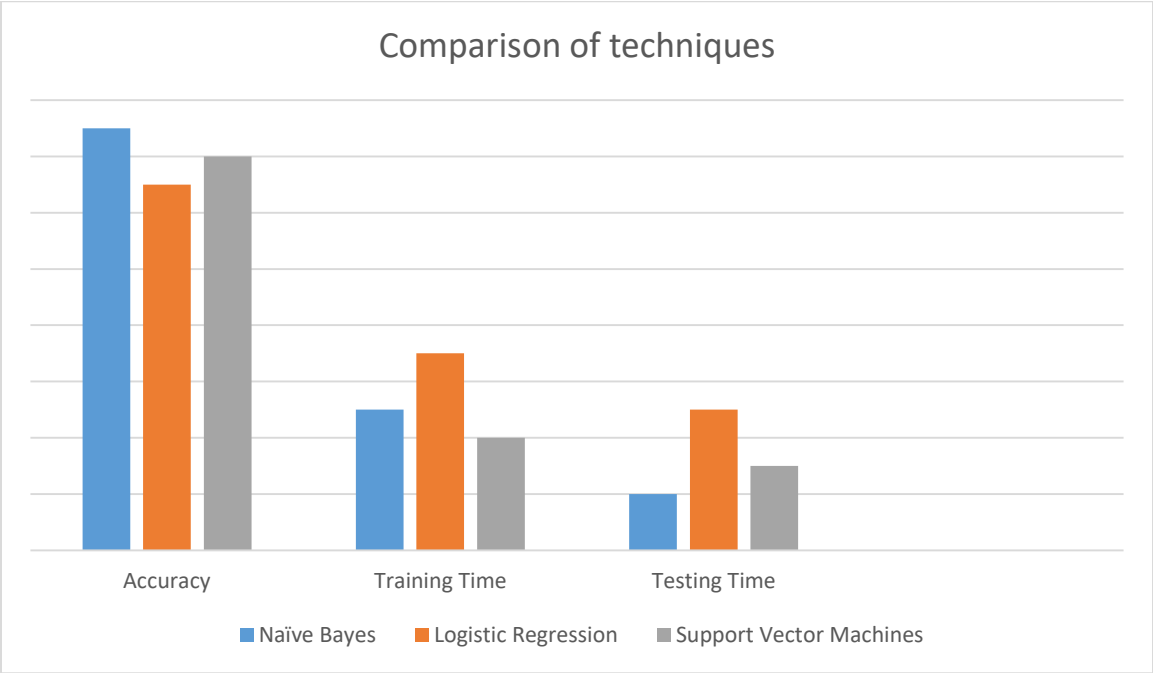

Figure 4. Sample comparison plot

Using this data it can be decisively concluded as to which of the classifiers is more feasible for this dataset, the advantages and the disadvantages over the others which can be the parameters for the users to decide which method to use. Eg. In case the model with better complexity takes more time but gives better performance with respect to accuracy, the time taken can be reduced by using systems with better configuration and performance.

5. Project Plan

The completion of the project requires targeting completion of modules in time. The timeline is equally distributed for the decided tasks. There is always a probability of some tasks exceeding the targets and some finishing even before the targets. These are handled with proper planning to meet the targets. Below is a plan for completion of the yearlong project.

Plan Item	Target
Semester 1	
Academic Integrity Tutorial, Start Learning basic machine learning techniques and Python.	Week 1 03-03-2017
Select a project topic and search for related published references.	Week 2 10-03-2017
Analyze the research references and start preparing annotated bibliography.	Week 3 17-03-2017
Submit annotated bibliography. Start writing Project Proposal.	Week 4 24-03-2017
Submit Project Proposal. Start working on the project.	Week 5 31-03-2017
Collect the required data for the project.	Week 6 07-04-2017
Learn the required data mining technique to process the text data.	Week 7 14-04-2017
Data collection, search for alternate data just in case there is a requirement.	Week 8 28-04-2017
Prepare for Seminar.	Week 9 05-05-2017
Learn different approaches of weightage calculation and implementation as seen in research papers.	Week 10 12-05-2017
Seminar Presentation	Week 11 19-05-2017
Learn the techniques to be implemented	Week 12 26-05-2017
Implement Data preprocessing and obtain the bag of words.	Week 13 02-06-2017
Semester 2	
Start understanding the requirements for the conference paper and start working on the paper.	Week 1 28-07-2017
Gather required resources for the conference paper.	Week 2

	04-08-2017
Implement Naïve Bayes and Logistic Regression.	Week 3 11-08-2017
Implement Support Vector Machines.	Week 4 18-08-2017
Fine tune the parameters as required and validate the models.	Week 5 25-05-2017
Start writing the Conference paper.	Week 6 02-09-2017
Test the models.	Week 7 08-09-2017
Start working on Thesis report.	Week 8 15-09-2017
Compare the models and evaluate them.	Week 9 22-09-2017
Complete the conference paper and verify the project.	Week 10 06-10-2017
Conference Paper Submission	Week 11 13-10-2017
Project Demo	Week 12 20-10-2017
Thesis Report Submission	Week 13 27-10-2017

Table 2. Project Plan

Certain aspects that involve learning the related concepts may take time which is taken care by allotting equal distribution of time and using some days specifically for the learning activity. Time assigned for the modules of the project that include data collection, preprocessing, training the models, validation and testing are as seen in the table. The academic targets are also handled and appropriate preparation time is planned as found necessary. These are an estimate time requirement, however, there is always a possibility that the progress may vary at times.

6. Risk Assessment

A proactive method to manage risks is always a better option than to react to risks which results in not meeting the deadlines due to many associated disturbances. Since the project is carried out on our personal laptops, the risk of the effects of external entities disturbing the setup is eliminated. We could categorize the risks into different categories like health & security risks, general risks, project specific risk and environment specific risks and so on.

Taking into account the Occupational Health and Safety assessment, this project could be considered as a low risk project. This project involves working in a computer lab or at home with just the computer and does not involve any hazardous substances or machinery that requires vigilance.

Below are some of the risks identified and the planned techniques to overcome them. Probabilities and severities are just an estimate. Higher the severity and higher the probability of the risks, they have to be monitored with utmost care. Lower the risk and lower the severity, they could be ignored.

Risk Items	Planned management technique	Probability	Severity
Occupational Health and Safety Assessment			
a. Stress due to many factors like delay in the progress, monotonous methodology etc.	Efficient planning on handling the progress and selecting interesting methodologies than just following what every other research suggests.	High	High
b. Muscle strain due to prolonged use of keyboard.	Plan breaks and exercises to avoid muscle stress.	Medium	Low
c. Stress on eyes due to unpleasant lighting in the room we are working in.	Adjust the screen brightness as required and use matt finish and anti-glare add-ons to the computer screen	High	Medium
d. Irritation due to noise.	Use ear plugs.	Medium	Low
e. Fire emergency.	Be aware of the emergency exits and keep in mind the guidelines for fire safety and evacuation procedures. Always keep a backup of the data being worked on.	Low	High

General and Project specific risk assessment			
Risk Item	Planned management technique	Probability	Severity
a. Loss/damage of laptop or corruption of the laptop's storage.	Keep a backup of the files being worked on in repositories like Google drive or Github.	Medium	High
b. Writing code and reaching a point where everything falls apart.	Version the files before working on them and keep a backup all the versions.	Low	Medium
c. Planning the implementation eats up more time than necessary.	Avoid more detailing and jump into the implementation.	High	Low
d. Limited understanding of the concepts and technologies to be used.	Refer easily understandable tutorials on the topic of interest.	Low	Low
e. Dataset being corrupt.	Make sure there is an availability of similar datasets.	Low	Low
f. Unrealistic planning of targets.	Estimate the targets keeping in mind all the possible risks.	Medium	Medium
g. Libraries, if used, not performing the task we need to achieve	Understand the usage of the selected libraries and their intended purpose, search for alternatives just in case.	Low	Medium
h. System performance shortfall.	Know your system. Try and use as much data the system can handle.	Low	Low
i. Conclusion that some method may never work for the project.	Make sure an alternative is considered.	Low	Low
j. Classes and other assignments taking more time than expected.	Sketch out effective ways to manage time.	High	Medium
k. The dimension of the features reaching more than the number of features.	Handle data effectively. Text mining should be implemented as required.	High	Medium

Table 3. Risk assessment

7. Conclusion

This project focuses on classifying news utilizing headlines from multiple publishers. The main idea of the project is to implement simple text mining approach, to validate and understand the bag of words approach and its variations as researched by many data scientists. The machine learning approach is a very reliable way to carry out this classification. The project serves as basis for understanding the thorough implementation of a selective set of machine learning techniques, understanding the mathematics behind the implementation and the way this could be extended or modified as required to solve any other similar problem.

The key characteristics like performance, accuracy and other related properties of the models are important to consider the model for such classification. Even though this project targets binary class classification, the same concepts could be used for implementing multi class classification. Bag of words approach is not just a textual data mining technique, the same could be implemented for image data to form a bag of features to classify images later with a better understanding. The vastness of the fields of application this project could hold well for and the way they can be extended is sure motivating.

8. References

[1] M. Lichman. (2013). *{UCI} Machine Learning Repository*. Available: http://archive.ics.uci.edu/ml

[2] S. O. Fageeri, S. M. M. Ahmed, S. A. Almubarak, and A. A. Mu'azu, "Eye refractive error classification using machine learning techniques," in *2017 International Conference on Communication, Control, Computing and Electronics Engineering (ICCCCEE)*, 2017, pp. 1-6.

[3] I. Dilrukshi, K. D. Zoysa, and A. Caldera, "Twitter news classification using SVM," in *2013 8th International Conference on Computer Science & Education*, 2013, pp. 287-291.

[4] H. Kim and M. Kim, "Model-induced term-weighting schemes for text classification," *The International Journal of Artificial Intelligence, Neural Networks, and Complex Problem-Solving Technologies,* vol. 45, no. 1, pp. 30-43, 2016.

[5] Z. Niu, Z. Yin, and X. Kong, "Sentiment Classification for Microblog by Machine Learning," in *2012 Fourth International Conference on Computational and Information Sciences*, 2012, pp. 286-289.

[6] N. Wang, B. Varghese, and P. D. Donnelly, "A machine learning analysis of Twitter sentiment to the Sandy Hook shootings," in *2016 IEEE 12th International Conference on e-Science (e-Science)*, 2016, pp. 303-312.

[7] A. A. Turdjai and K. Mutijarsa, "Simulation of marketplace customer satisfaction analysis based on machine learning algorithms," in *2016 International Seminar on Application for Technology of Information and Communication (ISemantic)*, 2016, pp. 157-162.

[8] F. Figueiredo, L. Rocha, T. Couto, T. Salles, M. A. Gonçalves, and W. Meira Jr, "Word co-occurrence features for text classification," *Information Systems,* vol. 36, no. 5, pp. 843-858, 2011.

[9] E. Alpaydin, *Introduction to Machine Learning* (Adaptive Computation and Machine Learning). Cambridge, MA: The MIT Press, 2014.

[10] R. Khanna and M. Awad, *Efficient Learning Machines Theories, Concepts, and Applications for Engineers and System Designers* (The expert's voice in machine learning). Berkeley, CA : Apress : Imprint: Apress, 2015.

[11] P. Tsangaratos and I. Ilia, "Comparison of a logistic regression and Naïve Bayes classifier in landslide susceptibility assessments: The influence of models complexity and training dataset size," *Catena,* vol. 145, pp. 164-179, 2016.

[12] Y. Ma and G. Guo, *Support Vector Machines Applications*. Cham: Cham : Springer International Publishing, 2014.

[13] S. Hans Georg, "Support Vector Machines," in *Machine Learning in Image Steganalysis*: Wiley-IEEE Press, 2012, pp. 179-196.