

TABLE OF CONTENTS

FOREWORD

ACKNOWLEDGEMENTS

LIST OF FIGURES

LIST OF TABLES

1. TECHNICAL CONTRIBUTIONS (INDIVIDUAL)

1. Introduction
2. Project Description
3. Work Breakdown Structure
4. Machine Learning Classifier Development
 - 4.1. Data collection
 - 4.2. Data cleaning and Feature extraction
 - 4.2.1. Total number of claims
 - a. Description of claims
 - b. Process of extraction
 - 4.2.2. Number of independent and dependent claims
 - 4.2.3. Length of first independent claim
 - 4.2.4. Written description alignment
5. Machine Learning Classifier and Results
 - 5.1. Logistic regression classifier
 - 5.2. Decision tree classifier
6. Conclusion

2. ENGINEERING LEADERSHIP (TEAM WRITTEN)

1. Executive Summary
2. Introduction
3. Market Demand Assessment
 - 3.1. Patent service industry
 - 3.2. Major Players
 - 3.3. Porter's five forces analysis
 - 3.4. SWOT analysis
 - 3.5. Stakeholders
4. Strategic business plan
 - 4.1. Technical strategies
 - 4.2. Marketing strategies
5. Conclusion

APPENDIX A: Code for claims data processing

APPENDIX B: Code for cosine similarity metric

APPENDIX C: Code for merging claims data

APPENDIX D: Code for building classifiers

BIBLIOGRAPHY

FOREWORD

This work represents the final report accounting for my work as well as my teammates' work on the Capstone Project titled 'Machine Learning Classifier for Patent Grant Prediction'. This report consists of the combination of two papers. Firstly, my personal contribution to the project with my technical contributions (Chapter 1). Secondly, a team-written paper focusing on the Market demand assessment, Industry analysis, Porter's five forces, SWOT analysis and Strategic business plan which includes technical and marketing strategies (Chapter 2).

ACKNOWLEDGEMENTS

Working on the Machine Learning Classifier for Patent Grant Prediction project has been a great experience for me, and I would like to express my sincere gratitude to all of the following persons: To Prof. Dr. Lee Fleming, our Capstone advisor and Chair of the reviewing committee, for having accepted me to work on this project, for his availability any time I needed him and for the valuable advice that he has been giving us throughout the project. To Mr. Guan Cheng, who supervised and supported our team closely, who took the time to answer our queries, and who always ensured that we were able to make progress. To Prof. Dr. Paul Grigas for having kindly accepted to be a member of the reviewing committee and to represent the Department of Industrial Engineering and Operations Research within it. Particularly, to Saketa Lakshmi Bhojanapally, Yen An Chen and Yujia Xu for having been such fantastic capstone teammates, for the many hours of work and numerous experiences that we shared, as well as for the mutual support we gave each other. Furthermore, to Dr. Alexandre Beliaev and Ms. Amy Lee, respectively our lecturer and GSI for the Capstone Integration class, who closely followed our endeavors and decisively helped us to improve our writing and presentation skills. And finally, to all my friends (near and far) as well as to my family for all their love, patience, and for their irreplaceable support throughout this year.

LIST OF FIGURES

Figure No.	Description	Page No.
1	Work Breakdown Structure	
2	Average total number of claims for granted and not granted patents	
3	Boxplot of total number of claims for granted and not granted patents	
4	Independent claims and dependent claims	
5	Average number of dependent claims for granted and not granted patents	
6	Boxplot of number of dependent claims for granted and not granted patents	
7	Average length of first independent claim	
8	Boxplot of length of first independent claim	
9	Importance of features	
10	Accuracy of different classifiers	
11	Accuracy improvement with addition of new features	
12	Market Share of Patent Grant for Top 100 Ranked Patent Law Firms	
13	Porter's Five Forces Analysis	
14	SWOT Analysis	

LIST OF TABLES

Figure No.	Description	Page No.
1	Description of data sources	
2	Descriptive statistics of total number of claims for granted and not granted patents	
3	Descriptive statistics of number of dependent claims for granted and not granted patents	
4	Descriptive statistics of length of first independent claim	
5	Confusion matrix of logistic regression classifier	
6	Confusion matrix of decision tree classifier	

CHAPTER 1

1. Introduction

The goal of our project is to build a machine learning classifier using patent-relevant data to predict whether a particular patent application will be granted or not. Our algorithm can suggest areas of improvement for increasing the probability of patent grant for a particular application, thereby saving time, energy and cost required for filing a patent application. The patent application process is not an easy process due to the following reasons. First, according to USPTO Economic Working Paper Series (Joan Farre-Mensa et al. 2015: 7), “On average, it takes the USPTO 1.75 years to make a preliminary decision on the patent applications, and a full 3.2 years to make a final decision”. Furthermore, the total number of patent applications filed in U.S. is increasing year-by-year. For instance, in the past five years from 2011-2015, there has been an 18% increase in the number of applications being filed at USPTO (United States Patent and Trademark Office) (“U.S. Patent Statistics Chart”, 2016). This further causes an additional delay on an already lengthy process. Second, the probability of patent approval in its first examination is very low. For years 1996-2005, 11.4% of applications were granted in first examination of the application while 86.4% of applications received a non-final rejection decision and remaining 2.3% of applications were abandoned prior to the first examination of the application. Third, The average cost of filing a patent is \$17,078 and the cost of obtaining a patent in multiple countries could exceed \$400,000 (Walter G.Park 2010:42).

Our algorithm can help all the stakeholders involved in the patent application process. In this paper, we have focused on the process of extraction of textual features namely, claims, abstract and title of a patent application and the results of two machine learning classifiers built on the dataset have been discussed. We have also discussed about different data sources used for feature extraction and the procedure of data cleaning in the following sections.

2. Project Description

Our algorithm is trained on USPTO dataset from years 2001 - 2011. This is because the meta data for all applications is available only from year 2001. We have not considered patent applications from the year 2012 for analysis because there may be some applications which are still being prosecuted and their decision of being granted or not granted is not yet available. Hence, we have considered all patent applications from year 2001 to 2011 for analysis. The algorithm takes different features (columns) as input and it returns the probability of patent approval. Our aim is to train different algorithms on this dataset and measure their accuracy on test data and select the best performing algorithm. Our algorithm can act as a guide to all the stakeholders involved in the patent application process such as patent evaluation firms, startups

and medium-sized corporations to evaluate their current patent application thereby improving the odds of patent approval.

3. Work Breakdown Structure

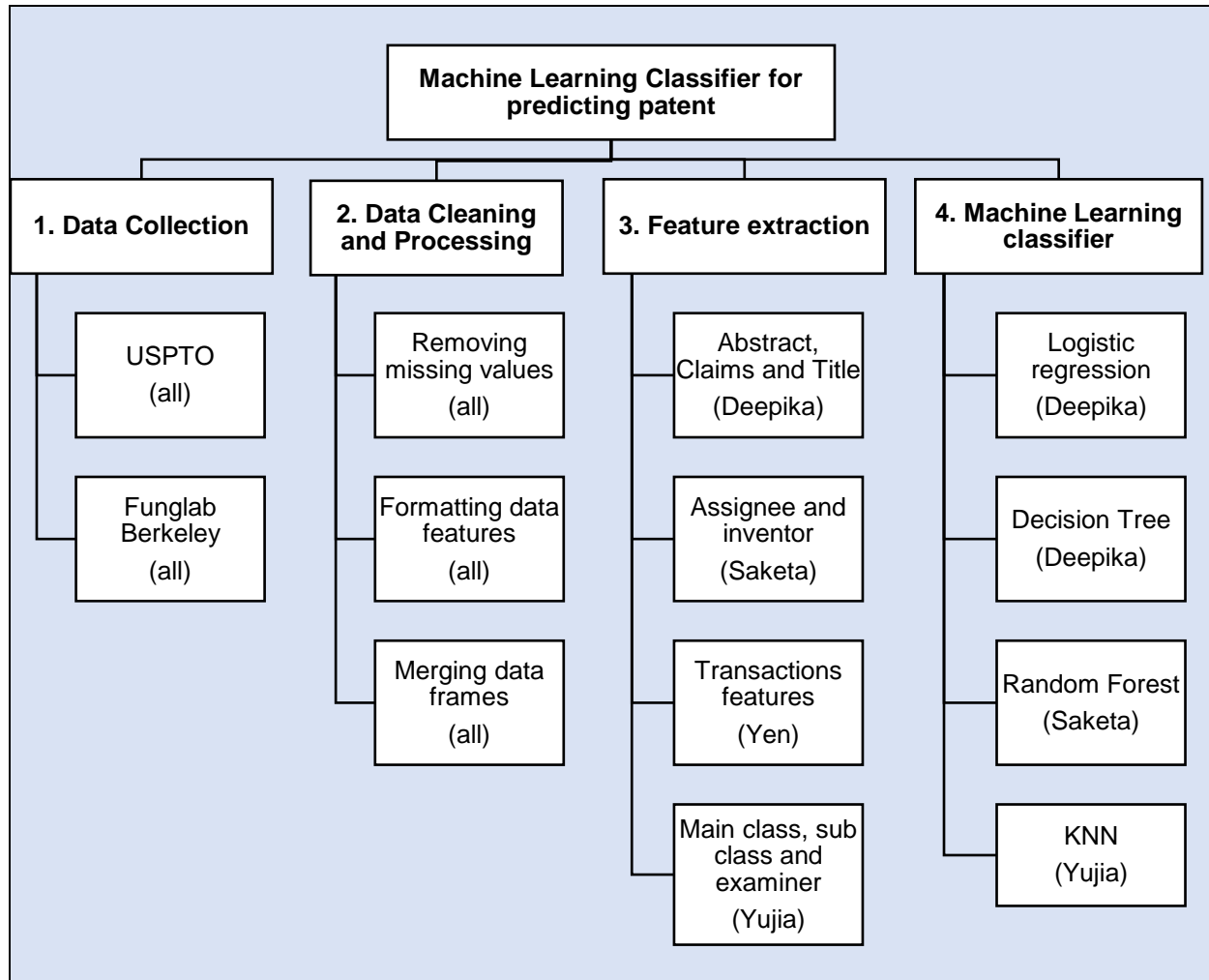


Figure 1: Work Breakdown Structure

We have four important phases in the development of our machine learning classifier.

- 1. Data Collection:** This phase emphasizes different sources of our data and different data files we used in our project and relation between different files
- 2. Data cleaning and processing:** In this stage, exploratory data analysis is carried out. This phase is about analyzing different attributes present in the data like finding datatype of the column and number of missing values and outliers in the column. All missing values which cannot be imputed are removed and the data which is not in proper format is brought into a proper format. Also, the data from different databases is merged in this step. The details about number of missing values removed, the number of observations that are used for

building classifier and whether selection bias was present, are discussed in the following sections.

- 3. Feature extraction:** The quantitative features are extracted from textual and non-textual columns which can be used in the stage of machine learning classifier.
- 4. Developing machine learning classifier and selecting the best classifier:** Various machine learning classifiers are built on the features extracted in above step and the output is analyzed to derive insights on data

During Fall semester, we performed all the above four steps and we achieved an accuracy of 73% on test dataset from our preliminary results. During the Spring semester, we have been trying to add some more features to already existing features like examiner information, art class unit and similarity index between abstract and claims for patents. After this, we will tune the classifiers on the training dataset using cross-validation approach in order to further improve our accuracy.

This paper will focus on data cleaning and processing of textual features in the data, exploratory analysis and results of two machine learning classifiers. The description of each of the textual feature extracted, hypothesis of their effect on outcome, their process of extraction and analysis of the output from the classifiers has been discussed in detail. Yen-An-Chen's paper will focus on extracting the information about the types of transactions that have occurred between USPTO and the patent applicant. Yujia's paper will focus on the patent attributes like examiner information and the art unit to which the patent belongs to. Saketa's paper explains attributes like information about assignee and inventor, main class, sub class and technology relevance of the patent.

4. Machine Learning Classifier Development

4.1 Data Collection

Our primary data sources for the project are USPTO (Patent Examination Research Dataset) and data records from Funlab Berkeley. The programming language used for data analysis in the project is Python which consists of libraries regex, nltk, pandas. The raw data for analysis is available in .csv , .tsv, .txt, http formats.

Our data is extracted from the following four different files. This table which describes in detail about all the datasets has been taken from Saketa's paper.

No .	DATA SOURCE	AVAILABILITY YEAR	DATA SIZE (rows*columns)	DESCRIPTION	PRIMARY KEY
1.	https://www.uspto.gov/learning-and-resources/electronic-data-products/patent-examination-research-dataset-public-pair	1976-2015	3.5 billion * 4	This dataset contains all the information about the transactions that have taken place between USPTO and patent applicant.	Application id
2.	http://portal.uspto.gov/pair/PublicPair	1976-2016	6 million*23	This dataset contains the information of all patent applications which have been granted. It contains date of filing patent, application number, technology class, examiner, art-unit, inventor and assignee information of a patent application (these are meta data and technical features).	Patent number, Application id
3.	http://funlab.berkeley.edu/publications	2001-2016	5.5 million*23	This dataset contains the meta data and technical information of all published patent applications.	Publication number, Application id
4.	http://funlab.berkeley.edu/publications/records_0714.tsv	1976-2016	9 billion*2	This dataset contains information about claims, title and abstract of a patent application	Patent number, Publication number

Table 1: Description of data sources (Source: Saketa Lakshmi Bhojanapally's paper)

The terms used in the above table are explained in Saketa's paper as follows.

- Application Id in the above table is specific ID given to a patent application at the time of filing by USPTO (source: Saketa's paper)

- Publication number is assigned to a patent application if it has survived 18 months after the filing application date (source: Saketa's paper)
- Patent number is assigned to patent applications which have been granted by USPTO
- Examiner information about a patent is about the examiner who analyzes the subject matter of a patent application and determines if an application should be patented or not ("United States Patent and Trademark Office, Patent Examiner Positions", 2017)
- Art unit is assigned to a group of related patent arts. Each art unit consists of one supervisory patent examiner and a number of patent examiners who determine if an application has to be patent or not. They are identified by a four digit number for e.g., 1642
- The technology class of a patent represents the technology group to which the patent has been classified into according to its technological content ("U.S. PATENT AND TRADEMARK OFFICE, Patent Technology Monitoring Team (PTMT)", 2017)
- Assignee of a patent is a business or a person who receives the ownership of patent by assignment from inventor of the patent. ("uspto.GOV, 301 Ownership/Assignability of Patents and Applications [R- 07.2015]", 2017)

There are three different types of features that are extracted from the data sources.

- Meta data features : These features describe the details of a patent application like assignee and inventor invention, time of filing of the patent and transactions
- Technical features : These features describe the examiner and art unit information, technology class to which the patent belongs to
- Lexical features : These are the textual features which contain the textual information about claims, abstract and title of the patent

The information from all these data sources is merged into a single data frame consisting of all attributes for patent applications. After merging all datasets on relevant columns, we have 763,968 patent applications or records for which all the information is present. The final dataset that we are using for analysis consists of patent applications from the year 2001 to 2011.

The response variable that we are trying to predict is whether a particular patent application will be granted or not. In the final dataset, we have 57 % observations for which patent has not been granted while there are 43 % applications which have been patented. The distribution of variables before merging with the final dataset is not significantly different from the distribution of variables after merging with the final dataset and after removing the missing values. For example, the average number of total claims for all patent applications in the dataset from years 1976-2016 is 21 while the final dataset that we are using for analysis from years 2001-2011 has average number of total claims as 18. Moreover, the observations have been removed randomly and the specific information of variables haven't been removed i.e., it is ensured while

dealing with the inventor and assignee location variable that all observations from a particular country or region have not been wiped out. Therefore, as there is a fair distribution of the classes being predicted, there is no skewness in data and the final data is representative of the target population and hence, there is no selection bias.

4.2 Data Cleaning and Feature Extraction

I have performed lexical feature analysis on the dataset from 'http://funglab.berkeley.edu/pub2/records_0714.tsv'. This dataset consists of 9 billion rows and 2 columns. First column consists of two types of data. These two types of data are patent number and publication number of the application, which are present in different formats. This means that some of the rows in column have patent number as their primary key while other rows have publication number as their primary key. The challenge is to identify which rows correspond to the patent number format and which ones correspond to publication number format. For example, the dataset consisted of 2 rows as follows:

Number	Text data
9086608	Laser probe with light beam. ABSTRACT. A laser...CLAIMS. 1. A system...
20040261315	Floral grouping wrapperof use. ABSTRACT. A method forCLAIMS. 1/ A method....

In the above table, the 7-digit number '9086608' corresponds to the application which has been patented (this is patent number) while the 11-digit number '20040261315' corresponds to the application which has not yet been patented (this is publication number). So, the task here is to identify which rows correspond to the patented applications and which rows correspond to applications which haven't been patented. After identifying these rows, this dataset will be merged with the final dataset for further analysis.

The second column consists of text about title, abstract and claim of the application. The abstract is the short summary of the idea or invention which is being patented. The abstract in an application filed may not exceed 150 words("uspto.GOV, 1826 The Abstract [R-07.2015]", 2017) whereas claims specify the limits of what is covered and not covered in the patent. The patent applicant has an exclusive right to make, use or sell only the things that are mentioned in the claims (Brown & Michaels: How do I read a patent?, 2017). The information about abstract, claims and title is present in the dataset in the following format.

"Welding shirt. ABSTRACT. This invention is an article of protective apparel having pliable, insulative panels integral to a sleeve or sleeves, CLAIMS. 1. An article for the protection of a person's arm from work-surface transfer heat, said article comprising: a front panel, a back panel and at....."

The algorithm is developed to parse this single column containing all the information about 'title, abstract and claims' and extract this information into different columns containing textual data about 'title', 'abstract' and 'claims'. The algorithm has been successfully implemented on the whole dataset using regex and pandas libraries from python.

The major challenge in handling this dataset is the time the algorithm takes to execute on whole dataset. In order to overcome this, the whole dataset has been divided into 60 chunks and every algorithm is executed on each chunk.

The following features have been derived from the columns 'title', 'abstract' and 'claims':

- Total number of claims
- Number of independent claims
- Number of dependent claims
- Total length of the first claim without stop words
- Written description alignment

The procedure for extraction of each of these features mentioned above is discussed as follows.

4.2.1. Total number of claims

In this section, the importance of claims feature, its description and the process of extraction of this feature has been discussed in detail.

Patent scope is one of the important parameters in measuring "patent quality". Patent quality represents the measure of strength of a patent that it could withstand rigorous examination if it is challenged in a court of law (Yan Liu et al. 2011: 1145). The metrics used for measuring patent scope are total number of claims, length of independent claim, number of independent and dependent claims (the difference between independent claims and dependent claims has been described in detail in the following sections). It is important to derive these features from the data available to improve the accuracy of our predictions about the patent. Our hypothesis is that more the total number of claims, the lesser the grant probability for a patent. (Alan C. Marco et al. 2016) (Elizabeth Webster et al. 2014) . Our results show an increase in accuracy of classifier by around 10% after the addition of these features in the algorithm.

a. Description of claims:

Claims of a patent application represent the legal bounds of the invention or discovery. It is very common to make changes to claims of a patent during its prosecution. Claims can be either described in a broader way incorporating larger set of technologies which the owner can exclude others from using or they can be described in a narrower way so that they are not so broad as to

overlap with any prior art. Here, we show the relation between the claims features and the probability of granting a patent.

b. Process of extraction:

The major challenge in extracting the feature, total number of claims, is that the claims feature is present in different formats for all patent applications. The majority of patent applications which have already been granted have their claim features in the following format:
" 1. A multi-purpose of a user. 2. A multi-purpose garment as in claim 1 , wherein garment. 3. A multi-purpose garment as recited in claim 1 , wherein ends. 4. A multi-purpose garment"

We have to write code for an algorithm such that it can handle all the exceptions. (The code for the feature extraction has been attached in the Appendix section). The progression of the code and the method of dealing with the exceptions and cleaning of the file is being described as follows:

For instance, in the above example, the total number of claims in the application is 4. Here, in order to count the total number of claims we can select all the numbers which are followed by '.' and count all such numbers. This would give us the number of total claims. In this case we would capture [1. , 2. , 3. , 4.] and it will give us the output as 4. But there are two exceptions to this rule.

1. Some of the applications are in the following format: *"1. (canceled) 2. (canceled) 3. (canceled) 4. (canceled) 5. (canceled) 6. (canceled) 7. (canceled) 8. (canceled) 9. (canceled) 10. (canceled) 11. A kit comprising twoacid of formula I. 12. A kit of claim 11,"*

In this case, the code will capture [1. , 2. , 3. , 4. , 5. , ..., 12.] and it will give us the output of 12 which is wrong since there are only two claims which are not cancelled and are active.

2. Another exception is that there are multiple formats in which claims are written like,
 - *"1 - 7. (canceled) 8. a refrigerating appliance...."*
 - *"1 : A multi-band antenna formed 2 : On the major surface 3 : a first part resonating at a"*
 - Instead of all claims being written in the standard format like '1. Claim 2. Claim 3. Claim', the different formats that have been encountered in all the files are '1- Claim 2-Claim 3-Claim' or '1) Claim 2) Claim 3) Claim' or '1/ Claim 2/ Claim 3/ Claim' or '1.) Claim 2.) Claim 3.) Claim' or '1.- Claim 2.- Claim 3.- Claim' or '1). Claim 2). Claim 3). Claim' or '1 Claim 2 Claim 3 Claim' or '1 . Claim 2 . Claim 3 . Claim'

There are close to 1 million files in the file containing all these formats apart from the standard format. The code has been written such that it captures the majority of the claims falling under a standard format to the maximum extent and it can handle them with greater accuracy.

All the exceptions which cannot be handled by this code have been removed from the data manually. This data has been cleaned so that there is no noise or incorrect values in the final data, this will help us in developing a better classifier with accurate data values.

In order to handle all the above discussed exceptions, the code has been written to handle the three most commonly occurring formats which are of the form '1. Claim 2. Claim 3. Claim', '1. (cancelled) 2. Claim' and '1- 94 (cancelled) 95. Claim'.

In the code, the text is parsed in order to get all the numbers found in the patent of the form '1.' and we select the last number present. For example, in this case, the numbers [1., 2., 3., 4.] present in the claim are captured and the length of this list is selected as the total number of claims. But there are few exceptions to following this logic; the procedure of handling these exceptions has been described as follows. Some of the claims end with a numerical in their last claim, like for patent application number 6836770, "1. A computer..... expressions are simplified. 2. The computer-implemented method of claim 1method of claim 2. 31. A computer method of claim 3. 32. A computer programsteps of the method of claim 4.." For such formats, the code would capture [1., 2., 2., 31., 3., 32., 4.] and it would return 4 as the total number of claims since this is the last number present. In order to deal with this exception, all the numbers which are followed by an alphabet are selected. Hence, [1., 2., 3., ...32.] would be captured and length of the list 32 will be taken as the total number of claims. This logic would work even for the claims with 'cancelled' word in them. The total number of claims which are active and not cancelled are counted for these claims which have 'cancelled' word.

On average, the total number of claims in a patent application which is granted is 16 while the average number of claims in a patent application which is not granted is 19. This shows that there is a significant difference in this feature between patented and non-patented applications. Our results after running the machine classifier also validate our hypothesis that the probability of granting for a patent application with fewer number of total claims is more than that of an application with higher number of total claims. (Alan C. Marco et al. 2016) (Elizabeth Webster et al.2014)

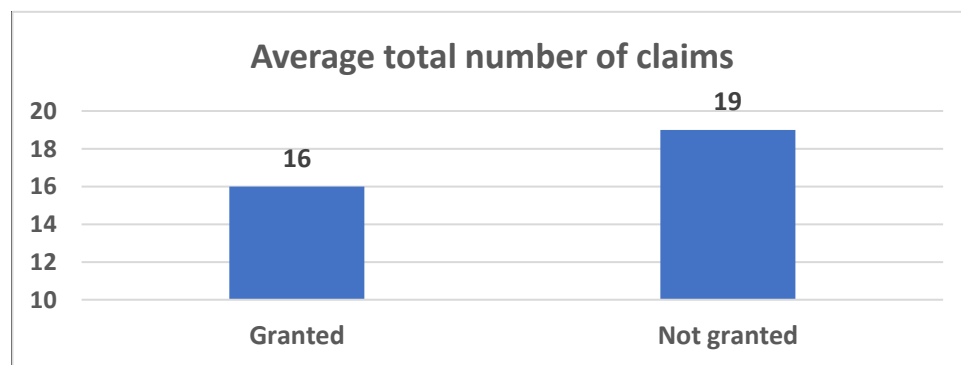


Figure 2: Average total number of claims for granted and not granted patents

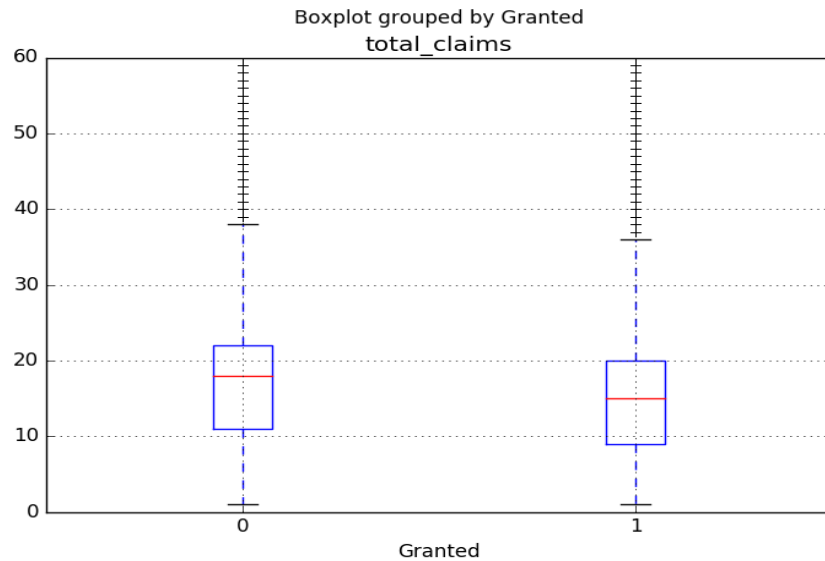


Figure 3: Boxplot of total number of claims for granted and not granted patents

In the above boxplot, the distribution of total number of claims for both patented and not patented applications is shown. The applications which are granted have the response variable as 1 while which are not granted have the response variable equal to 0. We can see that the median number of total claims for applications which are not granted is more than the median of total number of claims for applications which are granted.

DESCRIPTIVE STATISTICS	GRANTED	NOT GRANTED
mean	16	19
standard deviation	10.59	13.12
minimum	1	1
25%	9	11
50%	15	18
75%	20	22
maximum	372	565

Table 2: Descriptive statistics of total number of claims for granted and not granted patents

We can see from table 2 that all the descriptive statistics of total claims for applications which are not granted is more than the descriptive statistics of total number of claims for applications which are granted.

4.2.2. Number of independent and dependent claims

Our hypothesis is that increase in the number of dependent claims will lead to a decrease in the probability of patent grant (Alan C. Marco et al. 2016) (Yoshifumi Nakata et al. 2011). Keeping the total number of claims as constant, increase in number of dependent claims means an increase in the number of independent claims. According to (Alan C. Marco et al. 2016: 11), an increase in the number of independent claims leads to a broader patent scope. Also, he states that applications for which patent has not been granted tend to have narrow scope as compared to patented applications. A possible explanation according to Alan C. Marco is that, for applications with one independent claim, the application is more likely to get rejected soon if this one independent claim is rejected while for applications with multiple independent claims, they are more likely to continue prosecution and be in patenting process if one independent claim is rejected. In this section, the description of independent and dependent claims, and procedure for extraction of these features has been discussed.

Generally, claims in a patent consists of two parts: independent and dependent claims. According to WIPO Journal by Thomas Ewing et. al. 2017 : 78-80, independent claims are the broadest claims which stand alone and does not refer to another claim to be complete. Every claim has at least one independent claim and there may be more than one independent claims. An independent claim includes the essential features necessary to define an invention excluding the generic terms implied by the invention. For example, if we are claiming a patent for a bicycle then it does not typically need to mention the wheels.

Dependent claims, on the other hand, make a reference to either independent or other dependent claims. There may be zero or more number of dependent claims in any patent application. (U.S. Department of Commerce, Patent and Trademark Office, (Washington, DC 20231), 20,21)

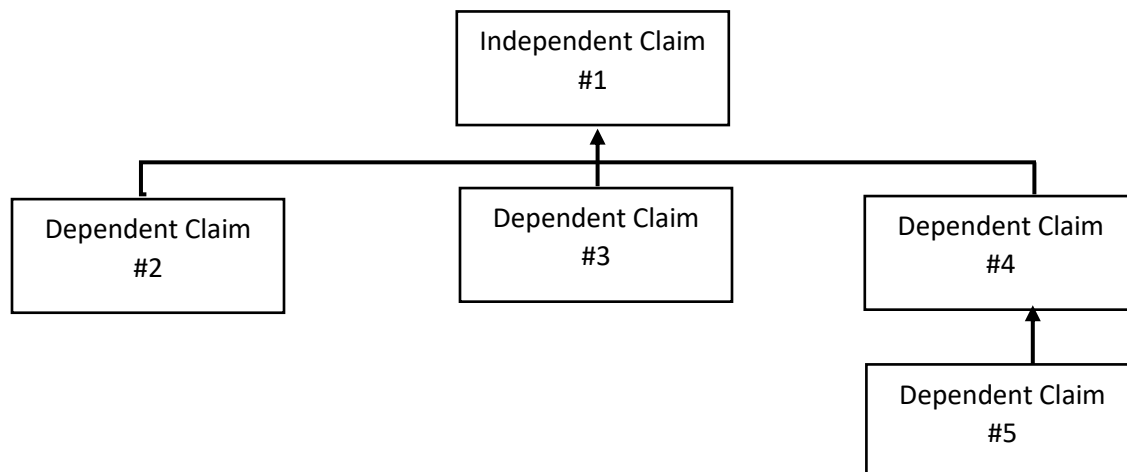


Figure 4: Independent claims and dependent claims (Stephen Yelderman 2014: 116)

Below is an example for a claim which consists of both independent and dependent claims. This example has been taken from Santa Clara High Technology Law Journal Volume 28, Issue 3 (Kristen Osenga, 2012: 626)

1. A chair comprising: a seat, having a top and a bottom; and a plurality of leg members, extending downwards from and connected to the bottom of the seat. [Independent claim]
2. The chair of claim 1, where the seat is made of walnut wood. [Dependent claim, refining the independent claim]

In a patent, the first claim should always be independent. There are two major disadvantages to having many number of independent claims. The first one is the cost of filing. The standard filing fee with USPTO for a utility patent will include only 3 independent claims and up to 20 total claims and any additional claim to be filed above these numbers will result in additional fee. The additional fee is 100\$ per independent claim and 50\$ per claim, if greater than 20 claims in total. This fee structure will discourage many firms from filing a patent with more number of claims, while for big corporations who can afford huge amounts and willing to file a strong patent, this fee structure may not hinder them. (Robert D. Fish Esq. 2014: Chapter 5)

Another disadvantage is that USPTO may sometimes force the applicant to prosecute each independent claim (including its dependencies) in a separate patent application, by issuing restriction requirements. Also, patent examiners typically do not like large number of independent claims due to an increase in caseload, as opposed to an application which could have been well drafted with just 1 or 2 independent claims and the remainder as dependent claims. A case with huge number of claims can be looked at unfavorably from the beginning of the application. (Robert D. Fish Esq. 2014: Chapter 5)

Owing to the above mentioned characteristics of independent and dependent claims, their importance as a feature for prediction cannot be overstated.

The procedure for extraction of dependent claims is as follows:

A typical example of a claim which consists of an independent claim as well as multiple dependent claims is as follows:

"17. An isolated, synthetic, or recombinant polypeptide having phytase activity made by a method comprising: (a) providing an exogenous nucleic acid encoding the polypeptide of claim 1, claim 2, claim 3, claim 4, claim 5, claim 6, claim 7, claim 8, claim 9, claim 10, claim 11, claim 12, claim 13, claim 14, claim 15, or claim 16; and (b) culturing the cell of (a).... 18. The isolated, synthetic, or recombinant polypeptide of claim 17, wherein the cell is a bacterial cell. 19. A foodstuff, a byproduct of a foodstuff, a feed, an animal feed, a food, a feed supplement, or a food supplement comprising the isolated, synthetic, or recombinant polypeptide claim 1, claim 2, claim 3, claim 4, claim 5, claim 6, claim 7, claim 8, claim 9, claim 10, claim 11, claim 12, claim 13, claim 14, claim 15, or claim 16"

Here, the number of times text appears between word 'claim' is counted using regular expressions and this is equal to the number of dependent claims. Thereby, we can derive the number of independent claims by subtracting the number of dependent claims from total number of claims. The average number of dependent claims for applications which are granted is around 13 while for applications which are not granted is 16. This gives us evidence that this feature will be helpful in predicting the probability of grant. Our results described in the section 5 also validate that our hypothesis is true.

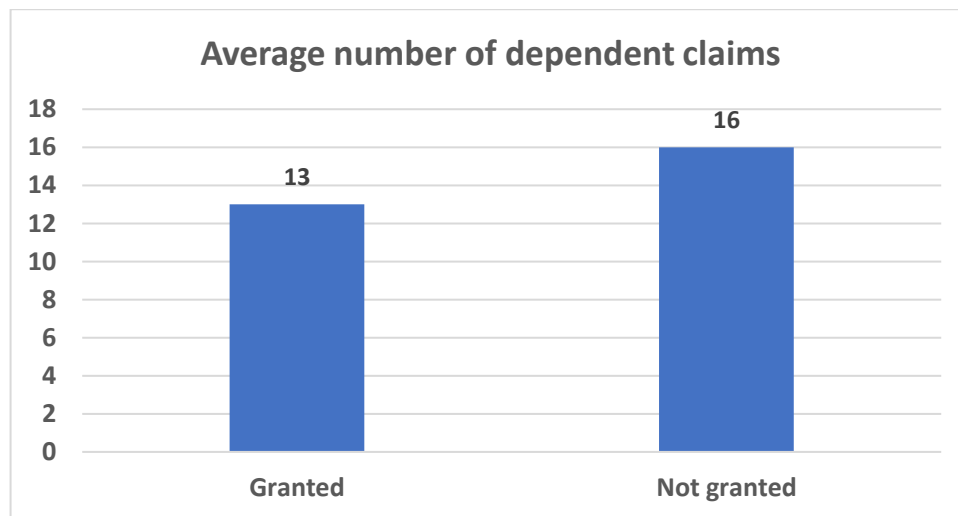


Figure 5: Average number of dependent claims for granted and not granted patents

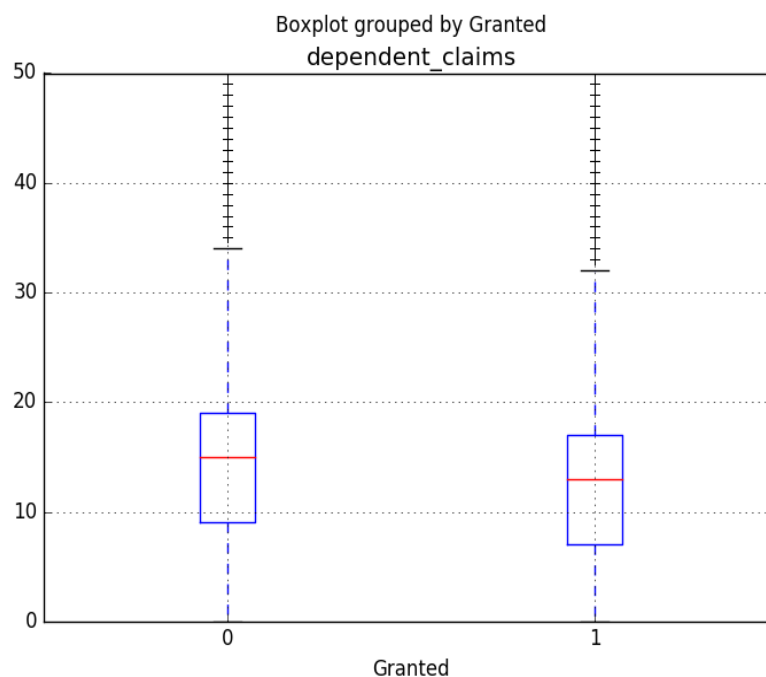


Figure 6: Boxplot of number of dependent claims for granted and not granted patents

DESCRIPTIVE STATISTICS	GRANTED	NOT GRANTED
mean	13	16
standard deviation	9.9	12
minimum	1	1
25%	7	9
50%	13	15
75%	17	19
maximum	361	404

Table 3: Descriptive statistics of number of dependent claims for granted and not granted patents

In the above boxplot, the distribution of dependent number of claims for both patented and not patented applications is shown. The applications which are granted have the response variable as 1 while which are not granted have the response variable equal to 0. We can see that the median number of dependent claims for applications which are not granted is more than the median of dependent number of claims for applications which are granted.

We can see from table 3 that all the descriptive statistics of total claims for applications which are not granted is more than the descriptive statistics of total number of claims for applications which are granted.

4.2.3. Length of first independent claim

Our hypothesis is that the number of words used in the first independent claim is an important feature in predicting the probability with which a patent can be granted. This is because the length of independent claim and number of independent claims together describe the breadth of claims and scope of patent applications during prosecution (Alan C. Marco et al. 2016: 50) and as mentioned above, patent scope is important to analyze the quality of a patent. The applications which have narrower claims (in terms of length of independent claim) have more probability to get granted than those with broader claims. (Alan C. Marco et al. 2016: 7,8)

The process of extraction of this feature is as follows:

- The first claim present between '1. ' and '2. ' is extracted and data cleaning is performed
- The punctuations are removed from the textual feature thus extracted (M.Ravichandran et. al. 2015: 4)
- Stop words are the most commonly used words which do not add value while retrieving information from the textual data (Rachel Tsz-Wai Lo et. al.: 1). The noise from the textual data is reduced by removing the stop words using a pre-compiled list of stop words. Some examples of obvious stopwords are 'the', 'for', 'is', 'and', 'it',

etc. Removing these words which have low discrimination power will result in greater accuracy of the algorithm (Saif Hassan et. al. 2014: 1,2)

- In addition to the above steps, letters and words which have a length of less than or equal to 2 have been removed to reduce noise in the data. This has been done because there are many letters in patents pertaining to chemical compounds which consist of formula of compounds and letters like 'xy', 'cf', 'ch', 'x', 'y'. Removing these words resulted in effectively capturing the accurate length of the first claim

On an average, the patent applications which are granted have length of their first independent claim as 109 while the applications which are not granted have the length as 65. The output of our machine learning classifier has validated the hypothesis that the length of the first independent claim is a significant predictor in predicting the probability of grant. (Alan C. Marco et al. 2016: 7,8)

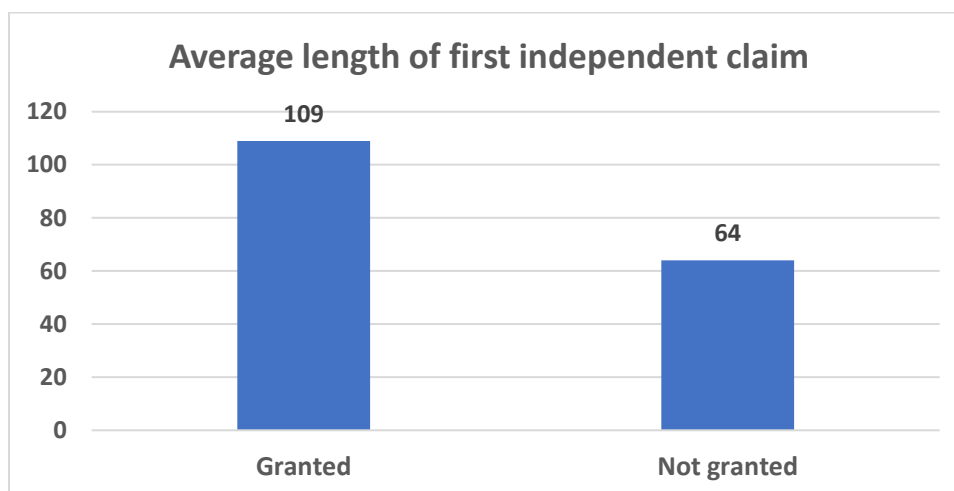


Figure 7: Average length of first independent claim

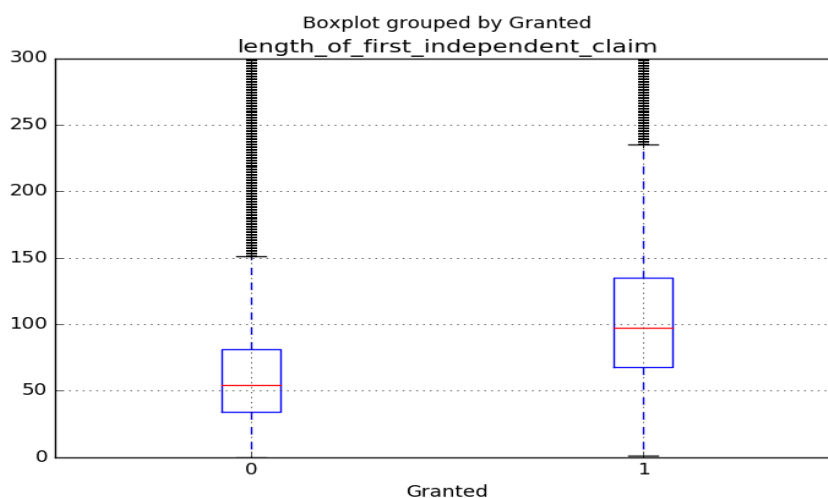


Figure 8: Boxplot of length of first independent claim

DESCRIPTIVE STATISTICS	GRANTED	NOT GRANTED
mean	109	65
standard deviation	73.17	64.68
minimum	20	15
25%	68	34
50%	97	54
75%	135	81
maximum	10103	6670

Table 4: Descriptive statistics of length of first independent claim

In the above boxplot, the distribution of length of first independent claim for both patented and not patented applications is shown. The applications which are granted have the response variable as 1 while which are not granted have the response variable equal to 0. We can see that the median length of first independent claim for applications which are not granted is more than the median length of first independent claim for applications which are granted. We can see from table 4 that all the descriptive statistics of total claims for applications which are not granted is less than the descriptive statistics of total number of claims for applications which are granted.

4.2.4. Written description alignment

Another important feature which can measure the quality of a patent application is the degree of alignment between the abstract and patent claims. This is captured by calculating the similarity metrics between pairs of word vectors representing abstract and claims. This measure is the distance between those parts of the patent, which will increase if there is poor alignment between abstract and claims and thereby, decreasing the probability of patent grant. (Yan Liu et al. 2011: 1149). This feature is extracted by performing following steps:

- Stop words and words which have length less than 3 are removed
- Stemming operation is performed on abstracts and claims which results in capturing stems of the words by clearing the affixes that contain grammatical information about the word. This results in aggregation of similar words and thereby resulting in accurate measure of similarity metric. (Cristian Moral et. al. 2014: 1,2) . PorterStemmer() function available in python has been used for this operation. ("Porter Stemming algorithm" from "Snowball", 1980). For example, the words 'cats', 'troubled', 'failing', 'feudalism' after performing stemming operation will result in 'cat', 'trouble', 'fail', 'feud' respectively
- Lemmatization has been performed on the output from the previous operation. Lemmatization is used as a normalization technique used to reduce derivational form of a word to its root word. (Riddhi Dave et. al. 2015: 1-2). For example, the words 'car',

- 'mouse' would be matched with the words 'automobile', 'mice' respectively as they refer to same things
- We have performed parts-of-speech tagging which would help in capturing only the relevant important information by tagging the words with their parts-of-speech. The written description alignment between abstract and claims is calculated using cosine similarity distance metric.(Anna Huang 2008: 51 - 53)

We have observed that there is no significant difference in the average cosine similarity metric between the patented applications (0.645) and applications which haven't been patented (0.642). This feature turns out to be the sixth most important predictor according to the output of decision tree classifier. It explains 7% of the variation in the data.

In addition to the above discussed features, we have added additional features like examiner information, transactions and art class of a patent. All the features from different databases are merged and the columns with missing data which cannot be imputed with values are removed. The final dataset consists of 763,968 observations upon which machine learning classifier is trained.

5. Machine Learning classifier and results

In this section, the two machine learning classifiers that we have trained on the data and their results have been discussed. We have tried two machine learning classifiers on the dataset namely, logistic regression and decision tree classifiers. The patent applications which have been filed until the year 2010 have been taken as the training dataset while the observations from the year 2011 have been taken as testing dataset. splitting it this way results in 80% of the observations being in the training dataset and 20% of the observations in the testing dataset.

In our dataset, there are 43% of the applications which are granted and 57% of the applications that are not granted. If we build a very simple algorithm which predicts that all the patents are not granted, then the accuracy of this algorithm will be 57%. This is known as the baseline accuracy. Our objective is to build an algorithm or classifier which outperforms this baseline accuracy.

5.1. Logistic Regression classifier

Logistic regression classifier is used to predict a binary outcome(1/0, Yes/No) given a set of independent variables. (Murat Korkmaz et. al. 2012: 1). In our dataset, we represented all the applications which have been patented as 1 and which haven't been patented as 0.

Logistic regression will predict the probability of granting a patent application by fitting the input data to a logit function. Here, we are assuming that there is no linear relationship between the features and the output variable and we are using the logistic regression classifier.

The following features have been used for training the logistic regression classifier:

- Number of inventors for a patent
- Total number of claims
- Number of dependent claims
- The length of first independent claim
- Main classes to which the patent belongs to
- Number of main classes to which the patent belongs to
- Number of sub classes to which the patent belongs to
- Location variables (country, state and city) of assignee
- Location variables (country, state and city) of inventor
- If the applicant is an individual (0), if it is an organization (1)
- Technology relevance of patent applications
- Cosine similarity metric between abstract and claims
- Examiner information
- Art unit information

The feature number of independent claims has not been considered for training the classifier as this information has already been captured by the features 'total number of claims' and 'number of dependent claims', adding this feature would lead to redundancy and correlation between the features. Hence, it has not been added. We have trained the classifier using default parameters of the classifier and we achieved an accuracy of 80.6% on the test dataset. The coefficients of the features, number of dependent claims and total number of claims are negative in the output of the classifier which validates our hypothesis that with increase in total number of claims and dependent claims, the probability of granting a patent will decrease (Alan C. Marco et al. 2016). This is because with the increase in the value of the variable which has negative coefficient, the response variable (here, probability of grant) will decrease. This means that broader applications (in terms of total number of claims) tend to have more probability for not being granted as compared to narrow ones (Alan C. Marco et al. 2016). An increase in the number of dependent claims will decrease the probability of patent grant for particular application. (Alan C. Marco et al. 2016) (Yoshifumi Nakata et al. 2011) (Elizabeth Webster et al. 2014)

The confusion matrix obtained by the classifier on test data set is as follows:

	Predicted_Not Granted	Predicted_Granted
Actual_Not Granted	65326	13290
Actual_Granted	17043	41918

Table 5: Confusion matrix of logistic regression classifier

The accuracy on test dataset is given by $(65326+41918)/(65326+41918+13290+17043) = 77.95\%$

5.2. Decision Tree classifier

Decision tree classifiers are very helpful when there is a non-linear relationship between the features and the response variable. They are the supervised learning models used for classification as well as regression problems. A decision tree classifier splits the population or sample into two or more homogeneous sets based on most significant input variable. The features which have been used for logistic regression classifier are also used as input features for the decision tree classifier. (J. R. Quinlan 1986: 88-90)

The classifier gives an accuracy of 72.34% on the test dataset. The most significant variables along with their percentage of significance have been given as an output by the classifier as follows:

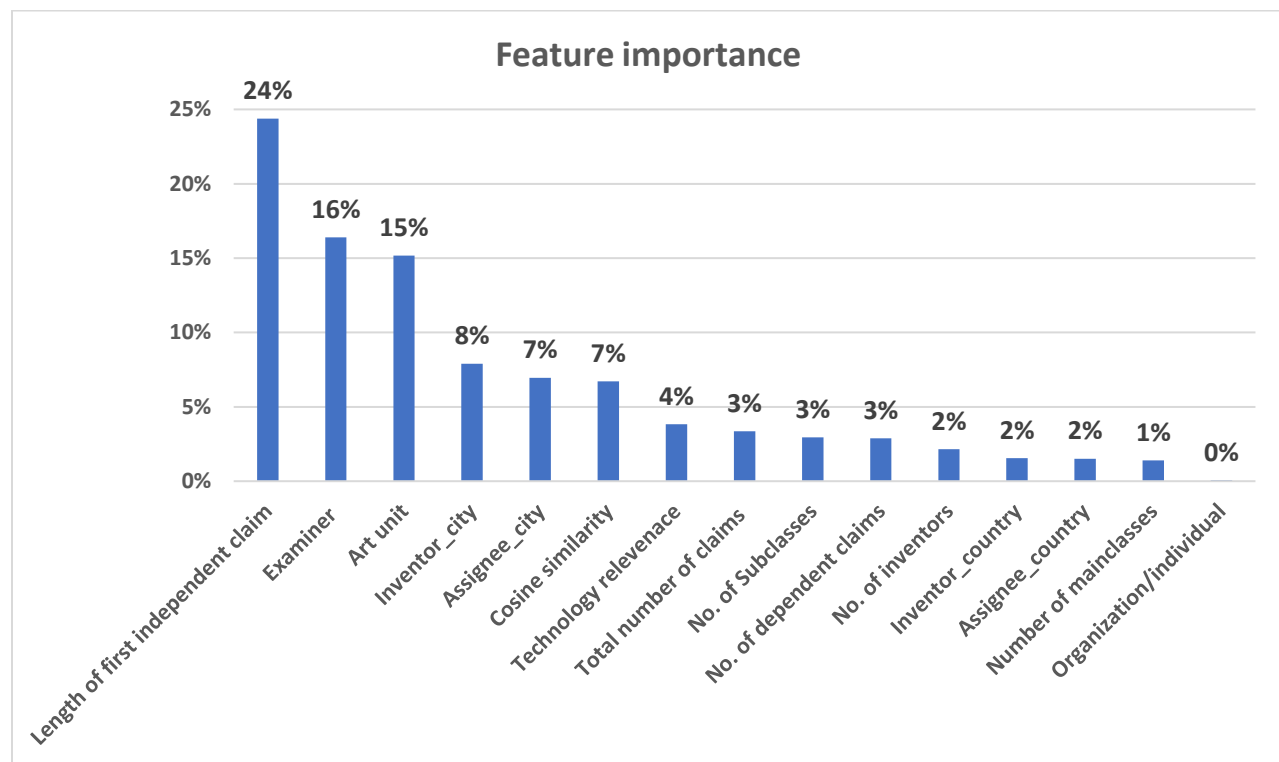


Figure 9: Importance of features

These are the top 15 significant variables in their order of significance as resulted by the decision tree classifier using default parameters. The length of first independent claim is the most significant factor for predicting the probability of grant, followed by examiner information and art unit to which the patent belongs to.

The confusion matrix obtained by the classifier on test data set is as follows:

	Predicted_Not Granted	Predicted_Granted
Actual_Not Granted	59875	18741
Actual_Granted	19303	39658

Table 6: Confusion matrix of decision tree classifier

The accuracy on test data set is given by $(59875+39658)/(59875+39658+18741+19303)= 72.34\%$

Additionally, we have developed random forest classifier, gradient boosting classifier as discussed in Saketa's paper and K-nearest neighbors algorithm as discussed in Yujia's paper. The gradient boosting classifier has resulted in accuracy of 78.05% on test dataset. The accuracy of different classifiers obtained on test data set is shown in figure 10.

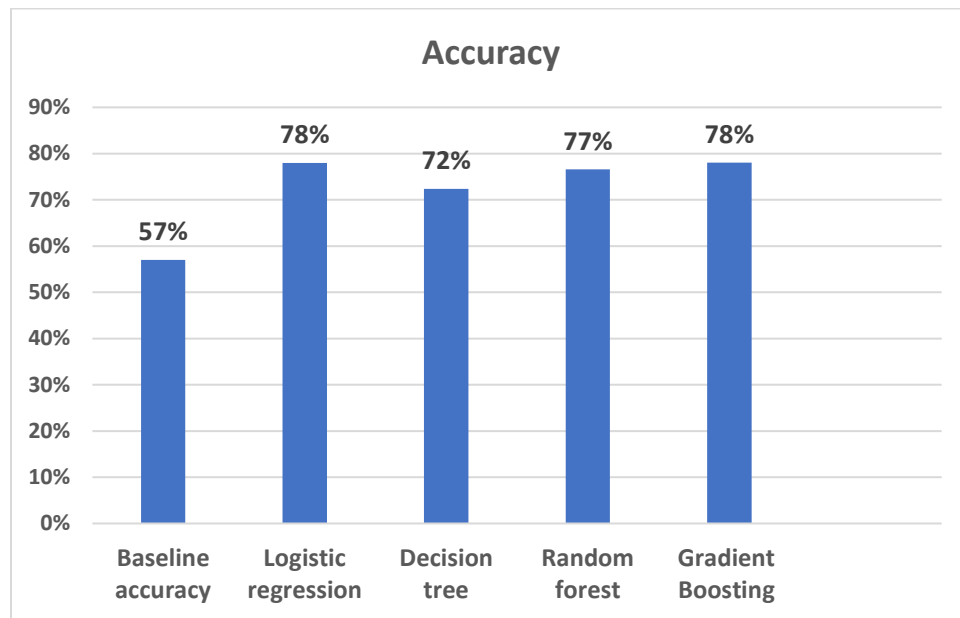


Figure 10: Accuracy of different classifiers

We can see from figure 10 that logistic regression and gradient boosting give the highest accuracy on test dataset. The results of gradient boosting classifier have been mentioned in Saketa's paper and we have also computed the change in accuracy for Gradient boosting classifier with addition of each new feature. This has been described in figure 11.

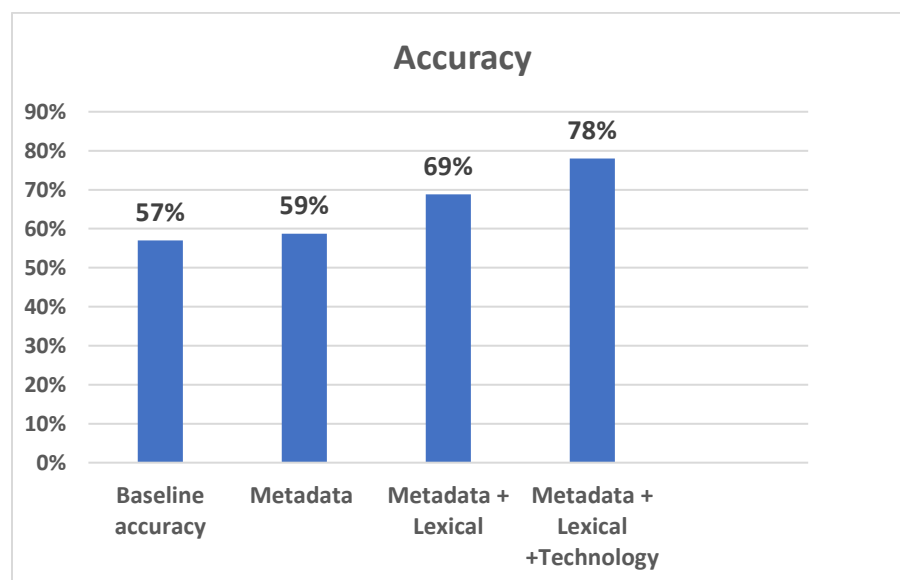


Figure 11: Accuracy improvement with addition of new features

We can see that accuracy has improved with the addition of features like examiner, art unit and similarity metrics.

6. Conclusion

The machine learning classifier has been developed successfully to predict the probability with which a patent application will be granted. The importance of each feature and its effect on the probability of grant of a patent has been described. The fewer the total number of claims and the number of dependent claims, greater will be the probability of a patent to get granted. Also, the length of the first independent claim is the most important feature to be considered while filing a patent which has the highest explanatory power of predicting the outcome. We have achieved 78% accuracy on the test dataset from the year 2011, which means that we can predict with 78% confidence whether a patent will be granted or not using the output of the random forest classifier developed. We are further working in the direction of improving the accuracy of our classifiers by adding more relevant features and by tuning the parameters of the classifiers. Thus, our machine learning classifier will guide the patent applicants involved in the patenting process to reduce their costs of patenting by suggesting the areas of improvement in their applications.

CHAPTER 2

1. Executive Summary

This chapter assesses the commercialization potential of transforming the technique into a product by: 1) conducting a market demand assessment and 2) developing strategies to address the market landscape. The assessment could possibly bring the capstone project (“project”) work into real business world and can be further applied to other patent related projects done in the Berkeley Patent Lab. The Porter’s Five Forces Analysis and SWOT analysis are conducted to assess the market demand. A strategic plan is further developed to address the challenges and threats. The analysis in general validate the market demand and gives a positive attitude on the commercialization of the technique.

2. Introduction

The burgeoning machine learning technology is changing the foundation of the century old patent industry. This trend is backed by one USPTO’s initiatives in its 2014-2018 Strategic Plan(‘USPTO 2014-2018 Strategic plan’), which is clearly stated: “Increase public availability of bulk patent data.” As more patent related data are accessible to the public from an official source, relying on the machine learning to assess a patentability of an application become a valuable application of the technique. Therefore, the commercialization potential of the technique generated out of the project is worth being explored further in detail.

The commercialization assessment of the technique brings the capstone project team (“team”) an idea on how valuable an academic project could be in the business world. As the team has already developed an approach on evaluating the patent grant for a given application with a positive result, validating the commercialization idea could bring direct value of the team and further funding or business development can be expected down the road.

Two major market analysis methods are used to assess the industry and market demand: The Porter’s Five Forces Analysis and the SWOT Analysis. The Porter’s Five Forces Analysis identifies threats from new entrants as machine learning has become a developed industry while the SWOT Analysis validates the opportunities of the product given the big market size.

The overall market demand assessment gives a positive attitude towards launching the product to the market. To address the threats, the team develops strategies such as differentiating the product feature, customizing the product, leveraging Cal affiliated network, reverse ladder pricing, and early establishment of partnership to address those threats. The business model of the product would be providing free trials with limited times of usage while licensing the product to patent service companies charging by the number of usage. The next step of the team is going to be drafting a patent application out of the technique and fit the application into the algorithm developed by the team to get the predicted result.

3. Market Demand Assessment

3.1. Patent Service Industry

The industry being analyzed is the ever growing patent law firm industry which forms the largest market of our product. The market revenue is estimated at 10.55 billion in 2015 with profit estimation at 2.27 billion. The specifically patent related service, exclude the trademark related service, accounts for about 58.3% of the entire industry, which provides a considerable outreach and market value for our product. The annual growth of patent application number has been steadily at about 4.25% over the past 10 years (“U.S. Patent Statistics Chart”, 2016). The corresponding annual revenue growth is estimated to be 2.6% in the past 5 years.

3.2. Major Players

Figure 1 below shows the percentage of patent grant among total patent grant in the calendar year of 2014 among the top 100 patent law firms ranked by IP Today Magazine. Given the complex nature of patent application which requires profound understanding of different industries and fields, there is no single service provider being considered dominant in the market. Considering the 10,004 service providers up to February 2017 suggested by the IBISWorld Industry Report OD4809 , one can infer that more than 99% of the service providers process less than 0.5% of the total application annually.

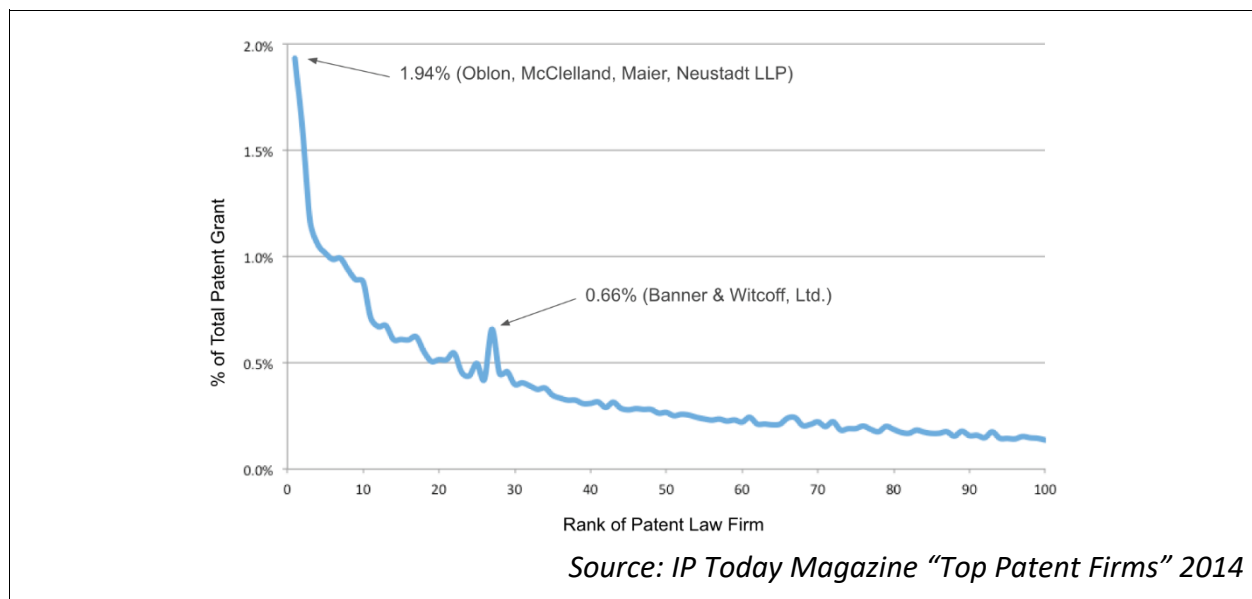


Figure 12: Market Share of Patent Grant for Top 100 Ranked Patent Law Firms

Among current service providers, there is no single one known for providing a similar service to the project’s approach. However, LexisNexis, a business analytic company, has launched a product called: “TotalPatent One” providing patent and non-patent literature

searching. Another product launched by the same company called “PatentOptimizer” utilizes technology to prepare a structured draft for a patent application. Also, by analyzing past public patent data from USPTO including application types, examiners, art units, and assignees, another company named Juristat is able to give future predictions of an application and strategies to optimize it. The introduction of these solution pioneers in the market is, from the team’s perspective, a strong indication of the trend coming soon.

3.3. Porter’s Five Forces Analysis

The overall profitability of the service suggested by the Porter’s Five Forces Analysis is medium as shown in Figure 2 below. The major challenge comes from threat of new entrants as barriers of entry for developing a product with similar approach could be relatively low that any data analyst can propose a similar solution. This challenge also contributes to the increasingly higher bargaining power of buyers that even though the value of the product lies on its accuracy rate of prediction, the market may not be able to distinguish good ones before using it. Note that as there are no comparable products in the market yet, the challenge might occur in the future yet the team will need to be aware of it right now.

Similarly, the competitive rivalry is evaluated to be low at present given only few companies providing a similar service. Although companies in this sphere such as LexisNexis or Juristat are working towards a similar direction, as long as the product itself is able to distinguish from others through a different approach of analysis, the rivalry level is still at a relative low level.

The bargaining power of suppliers are relatively low because of the strong supply of software engineers and data analysts in the market. The maintenance and improvement of the product could be done relatively easy with a few of engineers. Besides, the threat of substitutes is high because for potential customers, they can always have the choice of not using this kind of product but rely on traditional experience-based evaluation on an application. If the product is able to give a highly accurate prediction, our product has the potential to become an indispensable part of patent application process thus making threat of substitutes negligible.

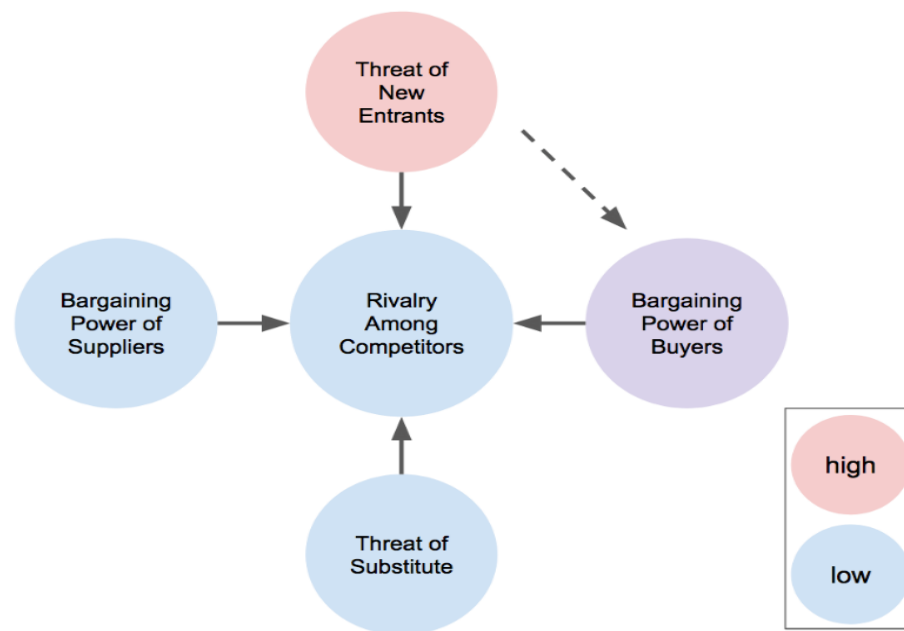


Figure 13: Porter's Five Forces Analysis

3.4. SWOT Analysis

The SWOT Analysis gives a positive standpoint of commercializing the technology into a product. One distinctive strength lies in the affiliation of the product with schools which ensure consistent support from the academy arena and future access to talented new college graduates. Another strength stems from the low or nearly no cost in developing the product, as it is a school project. These two strengths together establish a competitive position of launching the product.

However, the academic based project also has its share of disadvantages. A school project usually ends at the time of graduation, as students are free of obligation to carrying out the project. A very simple algorithm developed could also lead to a less robust product which cannot address the actual business needs of stakeholders.

The opportunity mainly comes from the higher rate of charge of patent agents and the long period of an application. For example, if the prediction accuracy rate of the product is credible, patent applicants would be able to make decisions without consulting patent agents and if the probability of prediction is low they can withdraw the application even before it is rejected. The product itself would hence have a high self-explanatory value in the market. Besides, given the increasing number of patent applications and availability of patent related data released by the USPTO, a good market for the product can be expected.

As the opportunity is considered promising, potential threats may come from a similar product development by current data analytic companies or patent law firms. Another threat arises from other academicians who research on a similar project, given that they may have access to even better resources in future.

Strengths <ul style="list-style-type: none">- Accessibility to support from the academic arena- Low or nearly no cost in product development	Opportunities <ul style="list-style-type: none">- Costly expense on patent agents- Growing annual patent application number- Increasing publication of patent related data from USPTO
Weaknesses <ul style="list-style-type: none">- Potentially end of project after graduation- Potentially oversimplified problem definition	Threats <ul style="list-style-type: none">- Data analytic company launching a similar product- Relatively lower technical threshold in developing a similar product

Figure 14: SWOT Analysis

3.5. Stakeholders

The major stakeholders identified by the team are: patent applicants (entities or individuals), patent service companies (patent search firms or law firm), patent evaluation or valuation firms, patent licensing service providers and the USPTO. Basically all entities involved in the pre or post application prosecution are considered to be stakeholders, for whom knowing whether an application will be granted matters a lot. The USPTO is included as a potential stakeholders because if all patent applications have gone through the product to assess its patentability, application with lower chance to be granted may not be submitted or be improved in its content before submission. In such case, the total number of applications will drop significantly, which reduces the total pendency of processing application accordingly.

Among all stakeholders, patent agents or law firms interestingly play both the roles as customers and competitors. As the process of filing a patent application is complex, which requires certain understanding of the patent law, it explains why majority of the patent applicants request for a patent service company to assist and rely on their advice. As most of the patent applications are filed by agents, the product by being able to assess the patentability of an application would bring the most value to patent agents or law firms' service.

4. Strategic Business Plan

4.1. Technical Strategies

To address the threats of new entrants and weakness of the project as discussed in the previous section, the team has developed the following strategies to address the issues.

- Product Differentiation

As the few products that are available in the market are pretty much focusing on patent search, the product in this project can focus more on modeling lexical features and contents of an application, as well as transaction records received from the USPTO. By taking the analysis on lexical features, the product would be able to evaluate the patentability of an application from the application draft itself. This creates value to customers before even filing an application and can save cost accordingly. On the other hand, the focus on transaction records allows the product to be able to dynamically predict the application's result from transactions that have occurred throughout the lifetime of an application. As the application content and claims usually do not change if not required by the USPTO, this approach provides further reference for customers in assessing the patentability of the application.

- Customized Product Feature

The product also provides customized solutions to various customers. For example, if a customer is specialized in design related patent application, the algorithms can be adjusted to specifically analyze related data in design technology area to give a better predicted result. This strategy is taking advantage of the various classes and industries an application may fall into and to carry out product segmentation within the product itself.

4.2. Marketing Strategies

In addition to strategies involved in the product itself, the team further developed the following marketing strategies in launching the product.

- Leveraging the Cal Affiliated Network

As the product comes out of a school project, the Cal affiliated network is what the team wants to leverage on. The first step would be reaching out to those alumni from Berkeley College of Law who are currently practicing in patent law firms to validate the value in the market and to improve the product in terms of its accuracy rate. The team would also ask for help from those alumni to populating the availability of the product in the industry by offering them a free trial for certain period as an exchange.

- Reverse Ladder Pricing

The pricing would be in a reverse ladder structure that the product would change by the number of usage (i.e. charge by each prediction) while the increased usage number leads to a decreased price per usage. This strategy is to encourage customers to utilize as many times of the product as possible and to make profits by the growing number of usage. This strategy taps on the long-tail distribution of patent service providers and their long-tail number of services provided to their clients.

- Early Establishment of Partnership

As the technical threshold of the product is relatively low, in addition to implementing the technical strategies to address the threats, establishing early partnership with potential customers is also important. Patent agents or law firms is chosen to be our foremost partners at an early stage as they play an indispensable role in the industry value chain with their expertise on preparing, filing, and prosecuting an application. These partnership establishment could be done with Cal affiliated law firms' referral as a start and further expand to new law firms via further referral. The advantage of establishing early partnership is that once similar products become available in the market, the cost of changing suppliers may hinder new entrants to enter the market.

The partnership establishment with patent evaluation or valuation firms follows secondarily important. This is because with the product's capability of assessing a patent application, the patent evaluation or valuation firms would be allowed to estimate a much precise value of a not-yet-filed or ongoing patent application given its chance to be granted. Again the early partnership establishment could be done by the referral from a Cal affiliated patent law firms to enjoy the pioneer advantage.

5. Conclusion

Based on the market demand assessment and the industry analysis, the product could possibly enjoys all the benefits of pioneer advantages while several challenges are to be expected down the road. Benefits include: establish long-term customer loyalty, determine arbitrary product price, set product standards on features, and create an image of pioneer in this type of services. The challenge of threats from new entrants could be mitigated by the strategy of establishing early partnership with patent service firms while the internal challenge resulted by graduation of the team could be solved by further communication. As long as the team is committed to launching the product to the market, the analysis basically supports the commercialization of the product to be a successful business.

Also, as there is no such a comparable product in the market at this moment, the product launch timing and its introduction to a selected group of customers will be the key success of the

business. Therefore, keeping the product in a low profile while introducing the product to Cal affiliated patent law firms secretly should be done in parallel.

After the product has been improved in its performance based on trials on current patent application, a proposed business model providing a free trial of the product online with limited times of usage for the purpose of advertising and licensing the product to customers charging by the number of usage is to be carried out. Further details on terms of licensing like length of license, license fee, amount charged per use, etc. would further be determined based on more market feedback.

The success of the product launch could not only serve as an example of launching a product from a school project but also work as a platform boosting future commercialization of other patent related research projects. The product's service scope would then be further expanded to include patent valuation, patent litigation, patent searching, patent claims scoring, or any other patent research topics in the Berkeley Patent Lab. As such, the project fulfills its full potential of commercialization and creates values in both the academic and business arena.


```
In [ ]: ##### Claim exception data - Final code #####  
import pandas as pd  
import pandas as pd  
import re  
import timeit  
from nltk.corpus import stopwords  
stop = set(stopwords.words('english'))  
start_time = timeit.default_timer()  
  
file = pd.read_csv('C:/Users/saketha lakshmi/Documents/Capstone/Deepika/Data/file_pub1_claims_new_v2.csv')  
print file.head()  
file.columns = [0,1,2,3,4,5,'Title','Claims', 'Abstract', 'total_claims', 'dependent_claims', 'total_length_firstclaim', 'nostop_length_firstclaim']  
#print file.head()  
#to extract the textual information  
def find_between( s, first, last ):  
    try:  
        start = s.index( first ) + len( first )  
        end = s.index( last, start )  
        return s[start:end]  
    except ValueError:  
        return ""  
  
def find_between_idx( s, first, last ):  
    try:  
        start = first  
        end = s.index( last, start )  
        return s[start:end]  
    except ValueError:  
        return ""  
  
for i in range(82641,100000):  
    s = re.findall(r'cancel|Cancel|CANCEL|Cancelled', file.loc[i,'Claims'], re.IGNORECASE)  
    if len(s) > 0:  
## For Claims having 'Cancelled'  
##### Total Claims #####  
        p = re.findall(r'[0-9]+[0-9]*[0-9]*[0-9]*\.[A-Za-z]|[0-9]+[0-9]*[0-9]*[0-9]*\:[A-Za-z]|[0-9]+[0-9]*[0-9]*[0-9]*\-[A-Za-z]|[0-9]+[0-9]*[0-9]*[0-9]*\[A-Za-z]\|[0-9]+[0-9]*[0-9]*[0-9]*\)\-[A-Za-z]|[0-9]+[0-9]*[0-9]*[0-9]*\)[A-Za-z]|[0-9]+[0-9]*[0-9]*[0-9]*\.[A-Za-z]|&|[0-9]+[0-9]*[0-9]*[0-9]*\.\.[A-Za-z]|[0-9]+[0-9]*[0-9]*[0-9]*\[/A-Za-z]|[0-9]+[0-9]*[0-9]*[0-9]*\.\.)&|[0-9]+[0-9]*[0-9]*[0-9]*\-[A-Za-z]&|[0-9]+[0-9]*[0-9]*[0-9]*\.\)&|[0-9]+[0-9]*[0-9]*[0-9]*\.\)[A-Za-z]',file.loc[i,'Claims']) #the remaining formats are not considered in the files except for standard formats because in the chemical compounds they will result in inaccurate outputs  
        p1 = re.findall(r'[0-9]+[0-9]*[0-9]*[0-9]*\.[A-Za-z]', file.loc[i,'Claims'])  
        q = re.findall(r'[0-9]+[0-9]*[0-9]*[0-9]*\.[A-Za-z]|[0-9]+[0-9]*[0-9]*[0-9]*\:[A-Za-z]|[0-9]+[0-9]*[0-9]*[0-9]*\-[A-Za-z]|[0-9]+[0-9]*[0-9]*[0-9]*\[A-Za-z]\|[0-9]+[0-9]*[0-9]*[0-9]*\)\-[A-Za-z]|[0-9]+[0-9]*[0-9]*[0-9]*\)[A-Za-z]|\.[A-Za-z]|[0-9]+[0-9]*[0-9]*[0-9]*\.[A-Za-z]&|[0-9]+[0-9]*[0-9]*[0-9]*\.\.[A-Za-z]|[0-9]+[0-9]*[0-9]*[0-9]*\[/A-Za-z]|[0-9]+[0-9]*[0-9]*[0-9]*\.\.)&|[0-9]+[0-9]*[0-9]*[0-9]*\-[A-Za-z]&|[0-9]+[0-9]*[0-9]*[0-9]*\.\)&|[0-9]+[0-9]*[0-9]*[0-9]*\.\)[A-Za-z]
```



```

file.loc[i,'nostop_length_firstclaim'] = length
else:
    s = find_between( file.loc[i,'Claims'], q[0][:-2], q[1][:-2])
    if len(s) == 0:
        file.loc[i,'total_length_firstclaim'] = 'check'
        file.loc[i,'nostop_length_firstclaim'] = 'check'
    else:
        length = 0
        for j in re.sub("([^a-zA-Z])"," ",s).lower().split():
            if len(j) > 2:
                length = length + 1
        file.loc[i,'total_length_firstclaim'] = length
        length = 0
        for j in re.sub("([^a-zA-Z])"," ",s).lower().split():
            if len(j) > 2 and j not in stop:
                length = length + 1
        file.loc[i,'nostop_length_firstclaim'] = length

#Not Cancelled Patents
else:
    p = re.findall(r'[0-9]+[0-9]*[0-9]*[0-9]*\.[A-Za-z]| [0-9]+[0-9]*[0-9]*[0-9]*\.: [A-Za-z]| [0-9]+[0-9]*[0-9]*[0-9]*\-[A-Za-z]| [0-9]+[0-9]*[0-9]*[0-9]*\[0-9*\)| [A-Za-z]| [0-9]+[0-9]*[0-9]*[0-9]*\)\|[A-Za-z]| [0-9]+[0-9]*[0-9]*[0-9]*\.\&| [0-9]+[0-9]*[0-9]*[0-9]*\.\\-[A-Za-z]| [0-9]+[0-9]*[0-9]*[0-9]*\/[A-Za-z]| [0-9]+[0-9]*[0-9]*[0-9]*\\.\\)&| [0-9]+[0-9]*[0-9]*[0-9]*\.-&| [0-9]+[0-9]*[0-9]*[0-9]*\\.\\)&| [0-9]+[0-9]*[0-9]*[0-9]*\./[A-Za-z]',file.loc[i,'Claims'])

    #q = re.findall(r'[0-9]+[0-9]*[0-9]*[0-9]*\.[A-Za-z]/[0-9]+[0-9]*[0-9]*[0-9]*\.: [A-Za-z]/[0-9]+[0-9]*[0-9]*[0-9]*\-[A-Za-z]/[0-9]+[0-9]*[0-9]*[0-9]*\[0-9*\)| [A-Za-z]/[0-9]+[0-9]*[0-9]*[0-9]*\)\|[A-Za-z]/[0-9]+[0-9]*[0-9]*[0-9]*\.\&/[0-9]+[0-9]*[0-9]*[0-9]*\.\|- [A-Za-z]/[0-9]+[0-9]*[0-9]*[0-9]*/[A-Za-z]',file.loc[i,'Claims'])
    if len(p) != 0:
        file.loc[i,'total_claims'] = len(p)+1
        print(i)

#the Length of list is calculated because, as in below example, there may be many patents which have ' 2008. alphabet ' similar kind of thing in their claims, then if we take the last alphabet 2008 is captured and if we take length 2 is captured, which is like small difference as compared to 2008.

#1. A provisional patent was filed for this invention on Jul. 24, 2008. Application number 610-832-51. The ornamental design for a portable DVD player, as shown and described.
#also many applications dint have something like ' ....claim 2. alphabet' they had more like ' ...claim 2, alphabet...' thing, so here there was no problem and the room for error in this case would be less
##### Dependent Claims #####
    file.loc[i,'dependent_claims'] = len(re.findall(r'(?: [0-9]+[0-9]*[0-9]*\.(?)[Cc]laim(?: [0-9]+[0-9]*[0-9]*\.: (?)[Cc]laim(?: [0-9]+[0-9]*[0-9]*\-(?)[Cc]laim(?: [0-9]+[0-9]*[0-9]*\[0-9*\])(?)[Cc]laim(?: [0-9]+[0-9]*[0-9]*\)(?)[Cc]laim(?: [0-9]+[0-9]*[0-9]*\.)?(?)[Cc]laim(?: [0-9]+[0-9]*[0-9]*\.\&(?)[Cc]laim(?: [0-9]+[0-9]*[0-9]*\.\|(?)[Cc]laim(?: [0-9]+[0-9]*[0-9]*\.\.\\(?)[Cc]laim(?: [0-9]+[0-9]*[0-9]*\./(?)[Cc]laim',file.loc[i, 'Claims']))
    else:
        file.loc[i,'total_claims'] = 1
        file.loc[i,'dependent_claims'] = 0

```

```

##### First Claim Length #####
#file.loc[i,'total_length_firstclaim'] = find_between( file.loc[i,'Claim
s'], p[0][:-2], p[1][:-2])
    if len(p) == 0:
        length = 0
        for j in re.sub("([a-zA-Z])"," ",file.loc[i,'Claims']).lower().sp
lit():
            if len(j) > 2:
                length = length + 1
            file.loc[i,'total_length_firstclaim'] = length
            length = 0
            for j in re.sub("([a-zA-Z])"," ",file.loc[i,'Claims']).lower().sp
lit():
                if len(j) > 2 and j not in stop:
                    length = length + 1
                file.loc[i,'nstop_length_firstclaim'] = length

    else:
        s= find_between_idx( file.loc[i,'Claims'], 0, p[0][:-2])
        if len(s) == 0:
            file.loc[i,'total_length_firstclaim'] = 'check'
            file.loc[i,'nstop_length_firstclaim'] = 'check'
        else:
            length = 0
            for j in re.sub("([a-zA-Z])"," ",s).lower().split():
                if len(j) > 2:
                    length = length + 1
            file.loc[i,'total_length_firstclaim'] = length
            length = 0
            for j in re.sub("([a-zA-Z])"," ",s).lower().split():
                if len(j) > 2 and j not in stop:
                    length = length + 1
            file.loc[i,'nstop_length_firstclaim'] = length

file.to_csv("op10_A11.csv")

op1 = pd.read_csv("E:/machinelearning_capstone/op1.csv")

```



```

In [ ]: import pandas as pd
        from nltk.corpus import stopwords
        from nltk.stem.wordnet import WordNetLemmatizer
        from nltk.stem import PorterStemmer
        import re
        import nltk
        import re, math
        from collections import Counter
        stop = set(stopwords.words('english'))
        file = pd.read_csv('Patent_Part2.csv')
        file.head()
        WORD = re.compile(r'\w+')

#preprocessing the text in claims and abstract using this function
def prepoc_text(x):
    sent_ = x

    #removing the punctuations, numbers and any other symbols other than a
Lphabets
    sent_ = re.sub("([^\a-zA-Z])", " ", sent_).lower()
    sent_ = sent_.split()

    #removing words or letters of length greater than 2
    sent_ = [i for i in sent_ if (len(i) > 2) and (i not in stop)]

    # Lemmatization of words
    lmtzr = WordNetLemmatizer()
    sent_lem = [lmtzr.lemmatize(i) for i in sent_]

    #stemming the words
    ps = PorterStemmer()
    sent_stem = [ps.stem(i) for i in sent_lem]

    #parts-of-speech tagging is performed here
    tagged = nltk.pos_tag(sent_stem)
    #considering only nouns, adjectives, adverbs, etc. and removing all co
njunctions, prepositions and unwanted words
    list_ = ['CC', 'DT', 'EX', 'IN', 'LS', 'MD', 'PDT', 'POS', 'PRP', 'PRP$', 'RP', 'TO', 'SYM', 'WDT', 'WP', 'WP$', 'WRB']
    sent_tagged = [i[0] for i in tagged if i[1] not in list_]
    return sent_tagged

#cosine similarity metric is calculated using this function
def get_cosine(vec1, vec2):
    intersection = set(vec1.keys()) & set(vec2.keys())
    numerator = sum([vec1[x] * vec2[x] for x in intersection])

    sum1 = sum([vec1[x]**2 for x in vec1.keys()])
    sum2 = sum([vec2[x]**2 for x in vec2.keys()])
    denominator = math.sqrt(sum1) * math.sqrt(sum2)

    if not denominator:
        return 0.0
    else:
        return float(numerator) / denominator

```



```
def text_to_vector(text):
    words = WORD.findall(text)
    return Counter(words)

#driver function
for i in range(0, file.shape[0]):
    claims = prepoc_text(file.loc[i,'Claims'])
    abstract = prepoc_text(file.loc[i,'Abstract'])
    claim1 = ' '.join(claims)
    abstract1 = ' '.join(abstract)
    text1 = claim1
    text2 = abstract1

    vector1 = text_to_vector(text1)
    vector2 = text_to_vector(text2)

    cosine = get_cosine(vector1, vector2)

    file.loc[i,'cosine'] = cosine
    print(i)

print(file.head())

#writing the output to file
file.to_csv('Patent_Part2_similarity.csv')
```

```
In [ ]: #merging the final dataset of claims featues
import pandas as pd
file1 = pd.read_csv('Patent_Part1_similarity.csv')
file2 = pd.read_csv('Patent_Part2_similarity.csv')
file3 = pd.read_csv('UnPatent_similarity.csv')
print(file1.head())
print(file1.shape)
file1.drop(['Claims','Abstract','Title'],axis = 1, inplace = True)
file2.drop(['Claims','Abstract','Title'],axis = 1, inplace = True)
file3.drop(['Claims','Abstract','Title'],axis = 1, inplace = True)
file_merged = pd.concat([file1,file2,file3],axis = 0)
file_merged.shape
file_merged.to_csv('file_merged_similarity.csv')
```



```

In [ ]: #Building the classifier on final merged dataset
import pandas as pd
import datetime
from sklearn import tree
from sklearn import linear_model
from sklearn.metrics import confusion_matrix
import matplotlib.pyplot as plt
file = pd.read_csv('Merge1.csv')

#removing duplicates from the dataset
file.drop_duplicates(subset = ['ApplNoSimple'], inplace = True)
file['year'] = file['Appl_Year']
file.shape #(763968, 752)

#splitting the dataset into train and test
train = file[file['year'] <= 2010]
train.shape #(626309, 752)

test = file[file['year'] > 2010]
test.shape #(137659, 752)

#selecting the required features in the dataset for training classifier
train = train.ix[:, [32,33,37,40,43,46,48,49,50,51,55,56,57,731,732,733,750]]+list(range(60,730))
train.shape #(626309, 687)
test = test.ix[:, [32,33,37,40,43,46,48,49,50,51,55,56,57,731,732,733,750]]+list(range(60,730))

#removing missing rows or observations, if any of them are present
train.dropna(axis = 0, how = 'any', inplace = True)
train.shape #(626110, 687)
test.dropna(axis = 0, how = 'any', inplace = True)
test.shape # (137577, 687)

#predictors
x_train2 = train.ix[:,1:]
x_test2 = test.ix[:,1:]

#response variable
y_train = train['Granted_x']
y_test = test['Granted_x']

#building Decision tree classifier
clf2 = tree.DecisionTreeClassifier()
#fitting the classifier
clf2 = clf2.fit(x_train2, y_train)
#predicting the response on test dataset
y_pred2 = clf2.predict(x_test2)
print(confusion_matrix(y_test, y_pred2))

features =
['Inventor_Count', 'AR_ICountry', 'AR_ICity', 'AR_ACountry', 'AR_ACity', 'Org_Ind',
onh', 'SubClass_C', 'MainClass_C', 'dependent_claims', 'nostop_length_firstclaim',
otal_claims', 'art_unit_GRate', 'examiner_GRate', 'Year_Similarity', 'cosine']+'mainclass']*670

```

```

len(features) #686
len(clf2.feature_importances_)
#feature importance given by decision tree classifier
print(sorted(list(zip(clf2.feature_importances_, features))))
#nostop_length_firstclaim(0.23017), examiner_Grate (0.1436),art_unit_grate(0.1
29079),AR_I_city(0.06105), Acity(0.053809), cosine(0.04337732),year_similarity
(0.024598),total_claims(0.024354), dependent claims(0.0193288),month(0.018150
9), subclass_C(0.01530055), inventor count(0.0137569), icountry(0.009).acount
ry(0.010144),mainclass matrix

#building logistic regression classifier
logreg2 = linear_model.LogisticRegression()
#fitting the classifier
logreg2 = logreg2.fit(x_train2,y_train)
#predicting the response on test dataset
y_pred4 = logreg2.predict(x_test2)
print(confusion_matrix(y_test, y_pred4))

#####
#Data visualization
#splitting the dataset into granted and not granted
granted = file[file['Granted_x'] == 1]
ungranted = file[file['Granted_x'] == 0]
#average total number of claims
granted['total_claims'].mean()
ungranted['total_claims'].mean()
#average number of dependent claims
granted['dependent_claims'].mean()
ungranted['dependent_claims'].mean()
#average length of first independent claim
granted['nostop_length_firstclaim'].mean()
ungranted['nostop_length_firstclaim'].mean()

#description statistics of dependent claims, total number of claims and length
of first independent claim
granted['dependent_claims'].describe()
ungranted['dependent_claims'].describe()
granted['total_claims'].describe()
ungranted['total_claims'].describe()
granted['nostop_length_firstclaim'].describe()
ungranted['nostop_length_firstclaim'].describe()

#boxplot of total number of claims for granted and not granted patents
sub = file[['total_claims','Granted']]
sub.boxplot('total_claims', by = 'Granted')
plt.ylim(0,60)
plt.show()

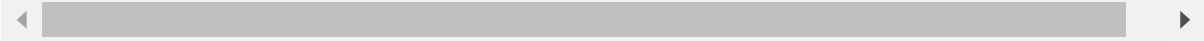
##boxplot of number of dependent claims for granted and not granted patents
sub = file[['dependent_claims','Granted']]
sub.boxplot('dependent_claims', by = 'Granted')
plt.ylim(0,50)
plt.show()

##boxplot of length of first independent claim for granted and not granted pat
ents

```

```
sub = file[['length_of_first_independent_claim','Granted']]
sub.boxplot('length_of_first_independent_claim', by = 'Granted')
plt.ylim(0,300)
plt.show()

#boxplot of cosine similarity for granted and not granted patents
sub = file[['cosine','Granted']]
sub.boxplot('cosine', by = 'Granted')
plt.ylim(0,1.1)
plt.show()
```



BIBLIOGRAPHY

Alan C. Marco, Joshua D. Sarnoff, Charles A. deGrazia

2016 Patent Claims and Patent Scope, August 2016. accessed November 13,2016.

Anna Huang

2018, Similarity Measures for Text Document Clustering, April 2008. accessed March 6, 2017.

Brown & Michaels,(2017, March 11)

2017, How do I read a patent? – the Claims, retrieved from

<http://www.bpmlegal.com/howtopat5.html> accessed on March 11, 2017.

Cristian Moral, Angélica de Antonio, Ricardo Imbert and Jaime Ramírez

2014, A survey of stemming algorithms in information retrieval, Vol. 19 No. 1, March 2014
accessed March 11, 2017.

Elizabeth Webster, Paul H, Jensen and Alfons Palangkaraya

2014, Patent examination outcomes and the national treatment principle, published on
May 7, 2014

Gang Luo (corresponding author)

2016, A Review of Automatic Selection Methods for Machine Learning Algorithms and
Hyperparameter Values, December 2016 retrieved from

http://pages.cs.wisc.edu/~gangluo/automatic_selection_review.pdf accessed March 9,
2017.

IBISWorld Industry Report OD4809, February 2017

2017, Trademark & Patent Lawyers & Attorneys Market Research Report, NAICS
OD4809, February 2017.

J.R. Quinlan

1986, Induction of Decision Trees, Machine Learning 1: 81-106, 1986
<http://hunch.net/~coms-4771/quinlan.pdf> accessed March 11, 2017.

Joan Farre-Mensa, Deepak Hedge, Alexander Ljungqvist

2015 The Bright Side of Patents, USPTO Economic Working Paper No. 2015-5, December 2015. <https://www.uspto.gov/sites/default/files/documents/Patents%20030216%20USPTO%20Cover.pdf> accessed November 13, 2016.

Kristen Osenga

2012 The Shape of Things to Come: What We Can Learn From Patent Claim Length, Santa Clara High Technology Law Journal Volume 28, Issue 3, March 2012 accessed March 12, 2017.

M. Ravichandran, G. Kulanthaivel, T. Chellatamilan

2015 Intelligent Topical Sentiment Analysis for the Classification of E-Learners and Their Topics of Interest, February 2015 accessed March 11, 2017.

Murat Korkmaz¹, Selami Güney, Şule Yüksel Yiğîter

2012, The importance of logistic regression implementations in the Turkish livestock sector and logistic regression implementations/fields, J.Agric. Fac. HR.U., 2012, 16(2): 25-36, accessed March 9, 2017.

Rachel Tsz-Wai Lo, Ben He, Iadh Ounis

Automatically Building a Stopword List for an Information Retrieval System, http://terrierteam.dcs.gla.ac.uk/publications/rtlo_DIRpaper.pdf accessed March 10, 2017.

Riddhi Dave, Prem Balani

2015, Survey paper of Different Lemmatization Approaches, International Journal of Research in Advent Technology (E-ISSN: 2321-9637) Special Issue 1st International Conference on Advent Trends in Engineering, Science and Technology "ICATEST 2015", March 2015 accessed on March 1, 2017.

Robert D. Fish Esq.

2014 Strategic Patenting, <http://www.fishiplaw.com/preface> accessed March 11, 2017.

Saif Hassan, Fern'andez Miriam, He Yulan, Alani Harith

2014 On stopwords, filtering and data sparsity for sentiment analysis of Twitter, 2014 accessed on March 11, 2017.

Stephen Yelderman

2014 Improving Patent Quality with Applicant Incentives, Harvard Journal of Law & Technology, Fall 2014 accessed March 10, 2017.

Snowball (The Porter stemming algorithm)

Retrieved from <http://snowball.tartarus.org/algorithms/porter/stemmer.html> accessed March 13, 2017.

Thomas Ewing, Carlos Olarte, Kanika Radhakrishnan, Markus Engelhard et. al .

WIPO Patent Drafting Manual: IP Assets Management Series,
http://www.wipo.int/edocs/pubdocs/en/patents/867/wipo_pub_867.pdf accessed March 10, 2017.

Top Patent Firms

Intellectual Property Today, Feb 2015 Retrieved from <http://mqrlaw.com/wp-content/uploads/2017/02/2015-top-patent-firms.pdf> accessed on March, 2017.

U.S. Department of Commerce, Patent and Trademark Office, (Washington, DC 20231)

USPTO: A Guide to Filing A Design Patent Application. Retrieved from https://www.uspto.gov/web/offices/com/iip/pdf/brochure_05.pdf accessed March 12, 2017.

U.S. Patent Statistics Chart, (2016, June 15)

Retrieved from https://www.uspto.gov/web/offices/ac/ido/oeip/taf/us_stat.htm accessed November 13, 2016.

United States Patent and Trademark Office, Patent Examiner Positions (2017, March 11)

Retrieved from <https://www.uspto.gov/web/offices/pac/exam.htm> accessed on March 11, 2017.

U.S. PATENT AND TRADEMARK OFFICE, Patent Technology Monitoring Team (PTMT)(2017, March 11)

Retrieved from https://www.uspto.gov/web/offices/ac/ido/oeip/taf/tecstca/explan_clstc_gd.htm accessed on March 11, 2017.

USPTO 2014 - 2018 Strategic plan

Retrieved from <http://www.uspto.gov/strategicplan> accessed on February 5, 2017

USPTO.GOV, The United States Patent and Trademark Office an agency of the Department of Commerce (2017, March 11)

Retrieved from <https://www.uspto.gov/web/offices/pac/mpep/s301.html> , 301 Ownership/Assignability of Patents and Applications [R-07.2015], accessed on March 11, 2017.

USPTO.GOV, The United States Patent and Trademark Office an agency of the Department of Commerce (2017, March 11)

Retrieved from <https://www.uspto.gov/web/offices/pac/mpep/s1826.html> , 1826 The Abstract [R-07.2015], accessed on March 11, 2017.

Walter G. Park

2010 THE WIPO JOURNAL:Analysis of Intellectual Property Issues, 2010 Volume 2 Issue 1, Articles : On Patenting Costs,

http://fs2.american.edu/wgp/www/2010_2_WIPO_Issue_1_Park.pdf accessed November 13,2016.

Yan Liu, Pei-yun Hseuh, Rick Lawrence, Steve Meliksetian, Claudia Perlich, Alejandro Veen

2011 Latent Graphical Models for Quantifying and Predicting Patent Quality accessed November 13,2016.

Yoshifumi Nakata, Xingyuan Zhang

2011 A survival analysis of patent examination requests by Japanese electrical and electronic manufacturers accessed January 11, 2017.