IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 17, Issue 2, Ver. IV (Mar – Apr. 2015), PP 42-50
www.iosrjournals.org

Feature Selection And Vectorization In Legal Case Documents Using Chi-Square Statistical Analysis And Naïve Bayes Approaches

Obasi, Chinedu Kingsley[1], Ugwu, Chidiebere[2]
[1, 2]*(Department of Computer Science, University of Port Harcourt, Choba, Rivers State, Nigeria)*

Abstract : *Most machine learning techniques employed in the area of text classification require the features of the documents to be effectively selected owing to the large chunk of data encountered in the classification process and term weights built from document vectors for proper infusing into the respective classifier algorithms. Effective selection of the most important features from the raw documents is achieved by implementing more extensive pre-processing techniques and the features obtained were ranked using the chi-square statistical approach for the elimination of irrelevant features and proper selection of more relevant features in the entire corpus. The most relevant ranked features obtained are converted to word vectors which is based on the number of occurrences of words in the documents or categories concerned, using the probabilistic characteristics of Naïve Bayes as a vectorizer for machine learning classifiers. This hybrid vector space model was experimented on legal text categories and the study revealed better discovered features using the pre-processing and ranking technique, while better term weights from the documents was successfully built for machine learning classifiers used in the text classification process.*

Keywords – *Chi-square statistics, Feature selection, Feature ranking, Pre-processing, Stemming, Vectorization*

I. Introduction

Text classification has been introduced in diverse domains for the correct classification of texts into specific categories. A typical architecture of a text classification system includes a text extraction, feature selection and feature weighting stage in its pre-processing phase according to [1]. The most common issue faced in the classification of documents is the high dimensionality encountered in the feature space. The introduction of these methods in the text classification system is to reduce the high dimensionality of these features, and to achieve such results, effective feature selection techniques needs to be implemented.

Feature selection is removing redundant and irrelevant features from the feature space, and the selected feature set should contain sufficient and reliable information about the original feature set [2]. In text classification studies, there are different selection schemes that have been used in literature, they include term strength [3], document frequency [4], odds ratio [5], mutual information [6], information gain [7], chi square [8], and so many others.

Documents are represented using a “bag-of-words” approach [9] where the exact ordering of words or terms, in the documents is ignored but the number of occurrences of each term is considered. Features or terms in the documents are usually assigned weights to depict the importance of such terms to the documents. This is referred to as vectorization. Several weighting schemes such as Term Frequency (TF) and Term Frequency-Inverse Document Frequency (TF-IDF) [10] have been used in the vector space model i.e. where documents are represented as features or identifier vectors.

This paper demonstrates the hybrid approach, chi square statistics and Naïve Bayes algorithm for successful discovery of relevant features and their respective term weighting in the text classification process. The remaining parts of this paper discusses the hybrid scheme in details, experiments and discussion of results, conclusion and references.

II. Chi-Square-Naïve Bayes Hybrid Scheme

The chi square-Naïve Bayes scheme is a hybrid approach demonstrated in this paper for proper selection of features using the ranking method and the vectorization of the features. This section describes the experimental procedures used in preparing the raw legal case text documents by extensive pre-processing, and how the feature ranking procedure, a feature selection process and vectorization of the features is implemented thereafter to obtain the term weights from the raw documents provided.

A. Experimental Study

This experiment is performed with a set of 57 legal case text documents from six (6) different categories: land, finance, politics_and_government, housing and criminal obtained from [11] and [12] and converted to text format for the purpose of this experiment. In Table 1, we can see the total number of features calculated from the documents in each category before the pre-processing process starts. The results of the pre-processing stages is shown in Fig 1.

Table 1 – Number of Features on Legal Case Documents

Categories	Total no. of features before pre-processing	Total no. of features after stop word removal and stemming	Total no. of features after filtering out tokens <= 3	Prior Probability of the categories
Civil	60491	22439	20899	0.077422342
Land	168486	57541	53569	0.198451479
Finance	92653	38467	35937	0.880873594
Politics_&_Government	260229	110542	103079	0.381866005
Housing	41227	16769	15654	0.057991739
Criminal	121538	44603	40797	0.151136385
Total	744624	290361	269935	1.747741543

Error! Not a valid link.

To show how we implemented the extensive pre-processing stages, we will be using an extract of one of the land cases, S.O. Adole vs. Boniface B. Gwar, converted to land_case_1.txt from the Land category. The document extract is as below:

"Yet it went on to hold that there was no evidence before the trial court proving the location of the land in issue in order to arrive at a decision that falls within the area designated as urban under the 1984 Order and therefore failed to prove his case."

B. Pre-processing

The pre-processing stage involved sub-stages like tokenization, stop words and non-alphanumeric words, stemming, token filtering by length, and case transformation. The tokenization of the words or features in the documents entails breaking the contents of the documents into smaller tokens which can be words, symbols or phrases. The tokenization algorithm implemented is described as follows:

```
1)  Read in the text documents as strings from each document.
2)  Initialize a variable for handling the tokens.
3)  For each document;
3a) Read each character till a space is encountered.
3b) Count the characters gotten as a token, move to the next character.

4)  Store the tokens gotten as input to the next stage.
```

After tokenization the document extract was broken down into tokens ignoring the commas used in separating the tokens thus:

"", Yet, it, went, on, to, hold, that, there, was, no, evidence, before, the, trial, court, proving, the, location, of, the, land, in, issue, in, order, to, arrive, at, a, decision, that, falls, within, the, area, designated, as, urban, under, the, 1984, Order, and, therefore, failed, to, prove, his, case, . ,"

Table 1 shows that the civil category has 60491 features, land has 168486 features, finance has 92653 features, politics_and_government has 260229 features, housing has 41227 features, and criminal has 121538 features after the documents have been tokenized. The obtained features when passed through the stop word removal algorithm allows the stop words and alphanumeric words to be eliminated from the features in all the documents in different categories. The stop word algorithm used is stated below:

```
1)  Read in the tokenized text documents.
2)  For all the tokens formed from all the documents;
2a)     Pass them iteratively through the stop word file.
2b) Also pass them through a declared set of non-alphanumeric words.
2c) Remove all matches of stop word characters and alpha-numeric words in the documents.
3)  Store the remaining tokens in the text document as input for the next stage.
```

The stop words list, 571 in number were gotten from [13], and was used in removing them from the text documents. The non-alpha numeric words are ! @ # & ( ) – [{ }] : ; ', ? / *. The algorithm above shows that each token is allowed to pass through a file englishST.txt which contains words seen as noise and less relevant, and also passed through the set of non-alphanumeric characters. The tokens from our example document extracted after going through this stage returns the following tokens:

hold evidence trial court proving location land issue order arrive decision falls within area designated urban 1984 order failed prove case

All the stop words were clearly removed, leaving us with the remaining features or tokens.

The stemming process selects the root form of these words from many occurrences of any word or feature, by removing the suffixes or prefixes of the words. The root words helps us to eliminate related words with similarities. In this work, the Porters' Stemming algorithm [14] was used to achieve this process. In words with the same inflectional forms in documents, such as connect and connection, begin and beginning, end and ending, etc., the root word for instance connect-, begin-, and end- are selected in replacement of all such instances, this enhances the a better selection of features and removes unnecessary ones. Considering the features from our example document extract, it can be seen that the feature "proving" and "prove" can have one root word "prov" representing them. The effect of the elimination of noise (unwanted features) from the stop word removal and the stemming process can be clearly observed from Table 1 as all the document categories showed a clear reduction in the initial number of features recorded. For instance the land category had 168486 features, but after these stages, the land category now has 57541 features, showing that noisy features of 110945 were eliminated, and so on. Our example document extract returns the following tokens after this stage:

hold evidence trial court prov location land issue order arrive decision falls within area designated urban 1984 order failed case

These tokens from different categories were filtered based on their length (i.e. the number of characters they contain). This was achieved by passing the tokens through a loop, we set the minimum number of characters in a word to be 3. This helped in filtering out more noise from the corpus. The following algorithm was implemented in the filtering stage to eliminate words seen as noise like ab, aaa, bbb, etc.

```
1)  Read in the text documents from the stemming stage.
2)  Initialize a counter to count each character in a token and store as token size.
3)  For all tokens in each document,
3a) Check for tokens with token size less than 3.
3b) Remove all such tokens from the corpus (filtering).
4)  Store the remaining tokens in the text document as input for the next stage.
```

Our tokens are returned exactly as before in this case as there are no tokens to filter as listed:

hold evidence trial court prov location land issue order arrive decision falls within area designated urban 1984 order failed case

After filtering the features, the legal case documents in different categories shows more reduction as can be observed in Table 1. For instance the civil case now has 20899 features after filtering as against 22439 features after the stemming process. The result of the pre-processing stages on the raw case documents is shown in Fig. 1. The case transformation stage converts all tokens repeated in both lower and upper cases to all lower cases. This was necessary to eliminate same tokens in different cases. In the document extract we used, all our tokens are already converted to lower cases. The feature "order" appears twice in the land_case_1.txt extract used, while others appear only once. Table 2 shows some of the selected features and their number of occurrences in some selected cases of different categories that we obtained from the text-converted legal case repositories' website after the case transformation stage. This forms our basis for ranking the features in the document so as to select better features. It can be observed that the term "court" appeared 71 times in the "housing_case_1" document of the housing category, while a term like "land" appeared 331 times in the "land_case_1" document of the land category. The description shows how all the occurrences of the features are distributed in the documents of different categories.

Table 2 - Sample Features of The Legal Cases And Occurrences (Law Reports Of Courts Of Nigeria And Law Aspire, 2001- 2011)

Categories	Sample legal cases	court	appel	state	appeal	plaintiff	defend	case	issu	evid	respond	law	trial	learn	section	judgment	act	land
housing	housing_case_1.txt	71	29	26	50	44	40	17	21	16	29	4	22	21	0	37	0	53
housing	housing_case_2.txt	55	76	15	35	14	14	14	9	10	54	12	10	19	4	26	5	28
housing	housing_case_3.txt	91	27	11	74	10	15	21	27	0	17	7	15	11	1	22	7	19
housing	housing_case_4.txt	33	23	12	27	36	33	10	3	16	8	5	8	6	0	13	4	13
housing	housing_case_5.txt	56	75	24	48	14	31	13	30	13	68	13	14	30	8	21	11	0
housing	housing_case_6.txt	94	26	32	56	51	49	47	29	86	23	40	44	58	3	34	7	2
housing	housing_case_7.txt	76	33	2	95	6	5	5	2	3	8	8	7	6	0	20	0	0
housing	housing_case_8.txt	33	23	12	27	36	33	10	3	16	8	5	8	6	0	13	4	13
criminal	criminal_case_1.txt	97	74	41	51	2	11	83	19	33	7	37	132	47	38	25	13	0
criminal	criminal_case_2.txt	75	102	47	29	0	0	45	32	54	9	15	36	43	3	27	21	7
criminal	criminal_case_3.txt	67	82	49	32	0	0	49	15	75	6	12	51	26	13	12	9	0
criminal	criminal_case_4.txt	164	54	68	34	0	0	54	47	112	66	24	117	59	16	22	18	0
criminal	criminal_case_5.txt	142	183	87	83	0	1	129	27	140	6	33	70	68	69	31	226	1
criminal	criminal_case_6.txt	189	342	59	73	0	24	98	40	265	12	22	156	88	11	35	30	0
criminal	criminal_case_7.txt	223	356	100	127	0	0	85	86	196	18	8	96	92	8	24	14	0
criminal	criminal_case_8.txt	106	156	67	83	0	0	65	67	127	10	20	83	73	13	21	21	1
criminal	criminal_case_9.txt	64	21	33	23	0	0	30	26	35	17	4	26	17	3	13	5	0
land	land_case_1.txt	152	121	65	84	28	34	92	111	60	132	38	27	24	159	57	192	331
land	land_case_10.txt	134	43	13	91	146	143	52	42	38	30	27	46	50	4	43	14	221
land	land_case_11.txt	92	27	11	74	10	15	22	27	0	17	7	15	11	1	22	7	19
land	land_case_12.txt	55	48	2	29	80	87	19	20	36	66	21	15	10	1	32	11	65
land	land_case_13.txt	98	67	7	85	63	48	52	69	27	28	9	17	12	5	35	4	72
land	land_case_14.txt	27	6	2	36	34	40	19	30	4	7	5	15	5	0	18	2	51
land	land_case_15.txt	54	23	7	29	58	57	39	21	31	18	22	30	30	9	25	12	113
land	land_case_16.txt	110	87	14	59	101	54	32	33	43	39	19	53	29	0	23	15	163
land	land_case_17.txt	153	54	7	97	69	82	51	21	22	46	5	54	42	4	32	10	57
land	land_case_2.txt	58	7	1	32	80	73	28	15	29	11	9	18	18	3	31	5	45
land	land_case_3.txt	101	81	9	128	11	11	21	4	7	65	69	14	19	16	20	1	0
land	land_case_4.txt	76	65	3	41	58	52	32	14	61	39	8	44	26	0	22	13	67
land	land_case_5.txt	124	63	30	102	153	109	27	28	18	80	53	38	16	3	38	2	38
land	land_case_6.txt	153	69	20	88	43	35	39	30	4	53	22	35	31	2	120	14	38
land	land_case_7.txt	61	4	1	60	20	32	11	7	0	11	8	11	12	20	14	0	12
land	land_case_8.txt	31	11	2	35	58	55	9	6	17	9	5	14	16	0	21	2	109
land	land_case_9.txt	258	70	27	114	197	240	151	181	80	25	10	81	67	1	91	12	278
civil	civil_case_1.txt	87	57	33	44	20	10	39	35	16	37	66	39	39	6	28	8	1
civil	civil_case_2.txt	88	133	56	40	76	22	41	42	12	30	15	20	12	2	30	8	0
civil	civil_case_3.txt	81	132	28	58	22	12	40	44	8	43	9	19	16	5	21	4	0
civil	civil_case_4.txt	167	134	70	74	1	1	29	47	20	49	36	69	82	20	21	20	2
civil	civil_case_5.txt	379	80	28	181	73	57	104	29	129	19	18	72	32	24	139	17	35
civil	civil_case_6.txt	168	53	10	129	37	15	34	77	3	40	18	47	25	0	47	4	5
finance	finance_case_1.txt	159	21	29	90	323	259	48	100	64	24	13	33	15	3	48	16	0

C. Feature Ranking-A Feature Selection Process

Creating a subset of quality features is of great importance in text classification as this enhances the performance of any classifier used, reduces training time, hence the various feature selection stages. The feature selection process selects a subset from the original feature set according to some criteria of importance of features [15]. In essence, feature selection should both reduce the high dimensionality of the feature space, and also provide a better understanding of the features, in order to improve classification result [16]. One of the variants of feature selection is feature ranking, which includes different methods like information gain (IG), chi-square (χ^2) statistics, reliefF, etc. Feature ranking is a kind of feature selection process which ranks the features based on their relevancies and importance with respect to the problem [17].

The feature ranking method used in our hybrid approach is the Chi-squared (χ^2) statistical analysis as seen in [18] which is one of the most popular and effective approaches of selecting features of more relevance in text documents. The Chi-square statistics tells us how relevant a word/term is to each class/category, and we will remove from the features, the words that are not relevant for that class/category. At the end, terms of high relevance are chosen. The formula for the Chi-square statistics is as follows in equation (1) below:

$$\chi^2(t,c) = \frac{N \times (AD - CB)^2}{(A+C)\times(B+D)\times(A+B)\times(C+D)} \qquad (1)$$

This is the value for a term (t) and a category(c) where:

A is the number of documents of category c containing the term t; B is the number of documents of other category (not c) containing t; C is the number of documents of category c not containing the term t; D is the number of documents of other category not containing t; N is the total number of documents.

Applying the chi-square to a feature and a category of a legal case document, we will consider a term like “plaintiff” in a category “housing” from the sample features we have in Table 2, the chi-square value can be calculated as thus:

N = number of sample documents = 57; A = 8, B = 41, C = 0, D = 8

$$\chi^2(plaintiff, housing) = \frac{57 * ((8*8) - (0*41))^2}{(8+0)*(41+8)*(8+41)*(0+8)} = 1.519366$$

Using the same method, we calculated for each feature in their respective categories or classes to obtain their chi-square values. Terms that have higher values are of higher importance than terms of lower values. We removed the less important features being marked by the chi-square method, by choosing a threshold value of 1.5, which means removing any words that has a chi-square value of less than "1.5", we can either increase or decrease our benchmark by changing the value of chi-square, to obtain important and more effective features. The features selected can now be vectorized for text classification.

The vectorization process i.e. term weighting was done using the Naïve Bayes algorithm which is a known classifier in the machine learning field. It is a very simple and effective method used in text classification [19], and it is based on the Bayes Rule which is based on probability theory. Using the probabilistic characteristics of the algorithm, the posterior probability distribution over the categories of text documents for each feature is calculated from the prior probability, the likelihood and evidence of the features and the values form a multi-dimensional vector for the classifier [20]. The list of features or words generated from the ranking method can be depicted as word1, word2, word3,…wordn, where n is for the total number of words in the document.

Assuming we have different categories of documents represented by Cat_a, Cat_b, Cat_c,… and the words in the documents represented as $word_{i,}$ where i = 1, 2, 3, ……

a. The prior probability of a particular category, $Pr(Cat_a)$ or $Pr(Cat_b)$ or $Pr(Cat_c)$ denoted as Pr(Cat) can be computed as:

$$\Pr(\text{Cat}) = \frac{total\ number\ of\ words\ in\ the\ category}{total\ number\ of\ words\ in\ the\ training\ set} \qquad (2)$$

The categories in our legal case documents are civil, land, finance, politics_and_government, housing and criminal. To calculate the prior probabilities of these categories, we considered the number of features remaining in the documents after the pre-processing the stage. For the civil category, the total number of features after stop word removal and stemming is 2648 + 4352 + 2491 + 4454 + 5462 + 3032 = 22439. Filtering out tokens that have less than or equal to three characters, we have 2411 + 4133 + 2363 + 4214 + 4963 + 2815 = 20899 tokens for the vectorization stage. The total number of tokens remaining in all categories from Table 1 is 269935 tokens. The prior probability of the civil cases is 20899/269935 = 0.07742234 as can be seen in Table 1.

Comparing the result with the individual prior probabilities in the civil cases, we got the same value we had 0.008931779 + 0.015311093 + 0.008753959 + 0.015611165 + 0.018385907 + 0.010428436 = 0.07742234. For the Land category, the prior probability can be calculated in the same way and we have 53569/269935 = 0.19845147. For the finance category, the prior probability can be calculated in the same way and we have 35937/269935 = 0.13313205. The prior probability of the remaining categories are calculated likewise.

Table 3 shows some sample features from all the legal case documents used in this experiment, their number of occurrences in each category, and the likelihood and evidence of the features.

b. The likelihood of a particular document category, say, $Cat_{a,}$ with respect to a particular word, $word_{i,}$ denoted as Pr(Word|Cat) can be computed as:

$$\Pr(\text{Word}\,|\text{Cat}) = \frac{number\ of\ occurence\ of\ the\ word\ in\ the\ category}{total\ number\ of\ occurence\ of\ all\ words\ in\ the\ category} \qquad (3)$$

To calculate the likelihood of a particular document category with respect to a word in the document, say land, case, etc., we need to calculate the total number of occurrence of all words in that category.

From Table 3, the total number of occurrence of all words in the civil category is 970+589+225+526+229+117+… = 23133, the word court appeared 970 times in the civil category, so the likelihood of the word "court" in the civil category is 970/23133 = 0.04193144. The word "appel", a stemmed word of its instances appeared 589 times in the civil category, resulting in a likelihood of 589/23133 = 0.0254614. This method applies for each word in any category.

Table 3 - Sample Features, Occurrences, and their Likelihood

Word/Attribute	Total number of occurences in every category	Occurrence in the civil category	Occurrence in the criminal category	Occurrence in the finance category	Occurrence in the housing category	Occurrence in the land category	Occurrence in the politics&government category	Likelihood = Probability (Word|Civil Category)	Likelihood = Probability (Word|Criminal Category)	Likelihood = Probability (Word|Finance Category)	Likelihood = Probability (Word|Housing Category)	Likelihood = Probability (Word|Land Category)	Likelihood = Probability (Word|Pol. & Govt. Category)	Evidence = Probability(Word)
court	7598	970	1127	1018	509	1737	2237	0.0419314	0.0244241	0.0256591	0.029228	0.02926065	0.0196777	0.0253766
appel	4449	589	1370	441	312	846	891	0.0254615	0.0296903	0.0111156	0.017916	0.0142513	0.00783765	0.0148592
state	4139	225	551	148	134	221	2860	0.0097264	0.0119411	0.0037304	0.007695	0.00372286	0.0251579	0.0138239
appeal	3952	526	535	572	412	1184	723	0.0227381	0.0115944	0.0144175	0.023658	0.01994508	0.00635985	0.0131993
plaintiff	3272	229	2	858	211	1209	763	0.0098993	4.334E-05	0.0216263	0.012116	0.02036622	0.0067117	0.0109282
defend	2986	117	36	722	220	1167	724	0.0050577	0.0007802	0.0181983	0.012633	0.01965871	0.00636864	0.0099729
case	2907	287	638	290	137	696	859	0.0124065	0.0138266	0.0073096	0.007867	0.01172447	0.00755617	0.0097091
issu	2558	274	359	341	124	659	801	0.0118446	0.0077802	0.008595	0.00712	0.01110119	0.00704597	0.0085435
evid	2456	188	1037	220	160	477	374	0.0081269	0.0224736	0.0055452	0.009187	0.00803531	0.00328988	0.0082028
respond	2390	218	151	421	215	676	709	0.0094238	0.0032724	0.0106115	0.012346	0.01138756	0.0062367	0.0079824
law	2233	162	175	119	94	337	1346	0.007003	0.0037926	0.0029994	0.005398	0.00567694	0.01184005	0.007458
trial	2230	266	767	228	128	527	314	0.0114987	0.0166222	0.0057468	0.00735	0.00887758	0.00276209	0.007448
learn	2097	206	513	198	157	418	605	0.008905	0.0111176	0.0049907	0.009015	0.00704142	0.00532186	0.0070038
section	1998	57	174	138	16	228	1385	0.002464	0.0037709	0.0034783	0.000919	0.00384078	0.01218311	0.0066731
judgment	1992	286	210	303	186	644	363	0.0123633	0.0045511	0.0076372	0.01068	0.01084851	0.00319312	0.0066531
act	1937	61	357	86	38	316	1079	0.0026369	0.0077368	0.0021677	0.002182	0.00532318	0.00949139	0.0064694
land	1913	43	9	9	128	1679	45	0.0018588	0.000195	0.0002268	0.00735	0.02828361	0.00039584	0.0063892
constitut	1725	79	44	70	8	74	1450	0.003415	0.0009536	0.0017644	0.000459	0.00124657	0.01275488	0.0057613
claim	1653	116	18	218	98	633	570	0.0050145	0.0003901	0.0054948	0.005627	0.01066321	0.00501399	0.0055209
said	1596	96	247	203	107	464	479	0.0041499	0.0053529	0.0051167	0.006144	0.00781632	0.00421351	0.0053305
feder	1510	38	13	83	82	25	1269	0.0016427	0.0002817	0.0020921	0.004709	0.00042114	0.01116272	0.0050433
order	1429	121	82	283	78	357	508	0.0052306	0.0017771	0.0071331	0.004479	0.00601385	0.00446861	0.0047727
counsel	1338	124	280	121	59	215	539	0.0053603	0.0060681	0.0030499	0.003388	0.00362178	0.0047413	0.0044688
provis	1291	57	40	59	19	82	1034	0.002464	0.0008669	0.0014871	0.001091	0.00138133	0.00909555	0.0043118
parti	1279	66	26	227	74	322	564	0.0028531	0.0005635	0.0057216	0.004249	0.00542425	0.00496121	0.0042717
power	1254	21	14	36	10	65	1108	0.0009078	0.0003034	0.0009074	0.000574	0.00109496	0.00974649	0.0041882
fact	1250	106	329	122	67	288	338	0.0045822	0.00713	0.0030751	0.003847	0.00485151	0.00297321	0.0041749
exhibit	1243	83	214	324	180	195	247	0.0035879	0.0046378	0.0081666	0.010336	0.00328487	0.00217273	0.0041515
nwlr	1218	194	173	200	66	227	358	0.0083863	0.0037492	0.0050411	0.00379	0.00382393	0.00314914	0.004068
made	1215	75	200	186	80	252	422	0.0032421	0.0043344	0.0046882	0.004594	0.00424507	0.00371211	0.004058
person	1181	39	330	98	35	138	541	0.0016859	0.0071517	0.0024701	0.00201	0.00232468	0.00475889	0.0039444
respect	1110	63	72	124	97	228	526	0.0027234	0.0015604	0.0031255	0.00557	0.00384078	0.00462694	0.0037073
ground	1105	141	173	134	77	385	195	0.0060952	0.0037492	0.0033775	0.004421	0.00648552	0.00171531	0.0036906
see	1093	105	212	159	64	230	323	0.004539	0.0045944	0.0040077	0.003675	0.00387447	0.00284126	0.0036505
gener	1025	64	34	77	23	64	763	0.0027666	0.0007368	0.0019408	0.001321	0.00107811	0.0067117	0.0034234
judg	1022	115	342	102	68	253	142	0.0049713	0.0074117	0.002571	0.003905	0.00426191	0.0012491	0.0034134
disput	1021	30	12	32	89	603	255	0.0012968	0.0002601	0.0008066	0.005111	0.01015784	0.0022431	0.00341
nigeria	966	47	61	98	39	84	637	0.0020317	0.001322	0.0024701	0.002239	0.00141502	0.00560335	0.0032263
matter	953	62	53	92	46	126	574	0.0026802	0.0011486	0.0023189	0.002641	0.00212253	0.00504917	0.0031829

c. The evidence of a particular word, say, word$_i$, denoted as Pr(Word) can be computed using the equation below:

$$\mathrm{Pr}(\mathrm{Word}) = \frac{total\ number\ of\ occurence\ of\ the\ word\ in\ every\ category}{total\ number\ of\ occurence\ of\ all\ words\ in\ every\ category} \qquad (4)$$

The total number of occurrence of all words in every category is 299410, so to calculate the probability of each word/feature in any category, we will need the total number of its occurrence in all the categories.

The word “court” for instance has a total of 7598 occurrences in all the categories as can be seen in Table 3, so the evidence is 7598/299410 = 0.025376. The evidence of the word “land” in all the categories = 1913/299410 = 0.0063892. The evidences of all the words was calculated in the same manner.

d. The posterior probability of the word in a document for a particular category, denoted as Pr(Cat|Word), can be computed by combining the first three equations in a calculative manner, as follows:

$$\mathrm{Pr}(\mathrm{Cat}|\mathrm{Word}) = \frac{\mathrm{Pr}(Word\,|\,Cat).\mathrm{Pr}(Cat)}{\mathrm{Pr}(Word)} \qquad (5)$$

After the values of equation 5 has been computed, we used the values to create a table of every word that occurred in the documents and their respective posterior probabilities. For example, Cat$_a$|Word$_1$, Cat$_a$|Word$_2$, Cat$_a$|Word$_3$…, Cat$_b$|Word$_1$, Cat$_b$|Word$_2$, …and so on.

The posterior probability of the word “land” in the Category “Land” can be calculated as thus: First we need to know the likelihood of the feature “land” with respect to the category “Land”, Pr(land|Land) = 0.028283611, this is shown in Table 3, we can also deduce from Table 1, that the prior probability of the Land category, Pr(Land) = 0.198451479 and the evidence of the word “land”, Pr (land) = 0.006389232 as can be seen from table 3, so the probability of Land|land can be calculated this way:

$$\frac{0.028283611 * 0.198451479}{0.006389232} = 0.878497514$$

The posterior probabilities of the words arising from the different categories forms the multi-dimensional vectors which are the term weights used in the text classification process. Table 4 shows some of the features and their posterior probabilities as calculated for different categories of the legal case documents.

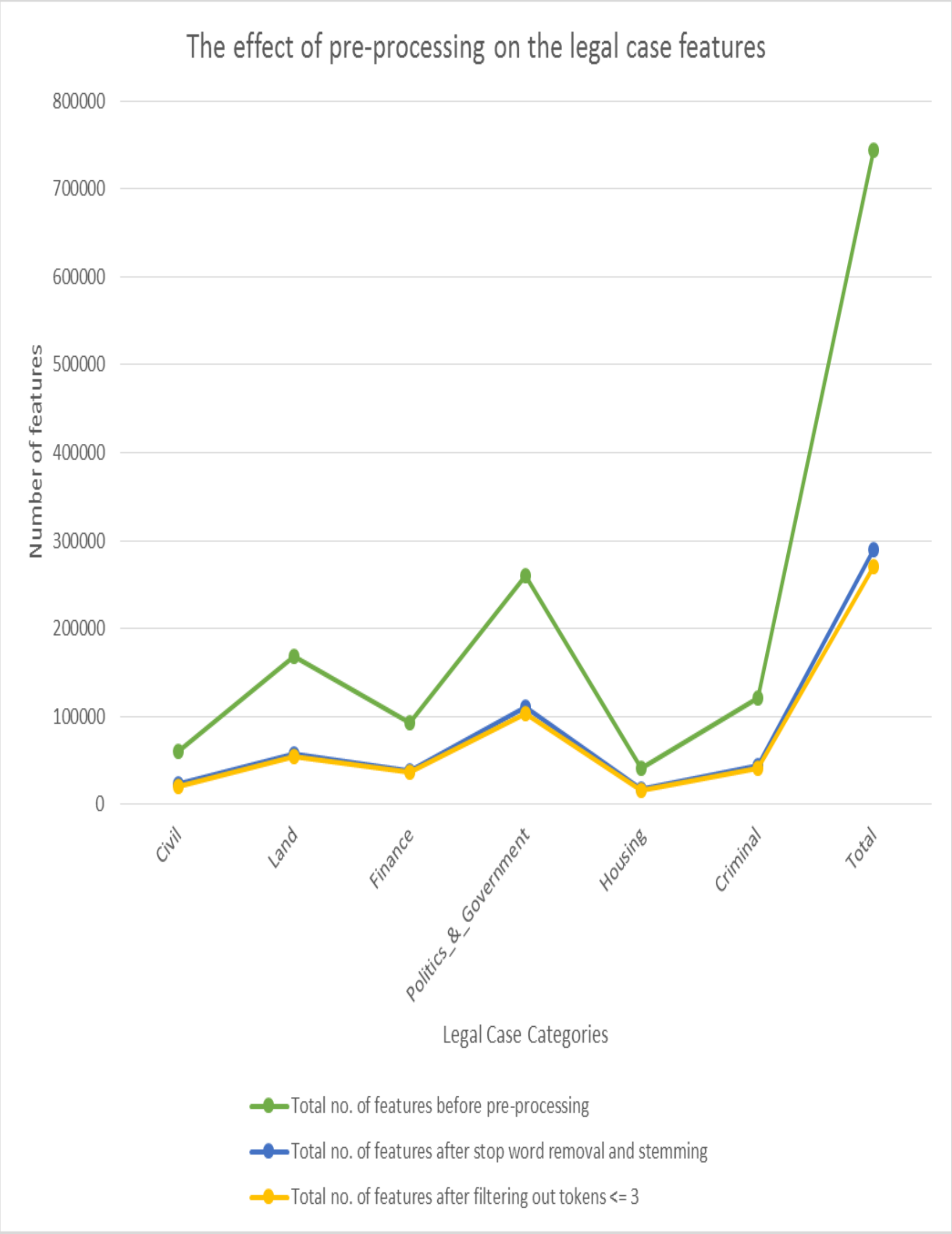

Figure 1 – The result of pre-processing on the legal case document features.

Table 4 - Some Features and Their Posterior Probabilities in Different Categories

Word/Attribute	Posterior Probability = Prob (Civil Category|Word)	Posterior Probability = Prob (Criminal Category|Word)	Posterior Probability = Prob (Finance Category|Word)	Posterior Probability = Prob (Housing Category|Word)	Posterior Probability = Prob (Land Category|Word)	Posterior Probability = Prob (Pol. & Govt. Category|Word)
court	0.127930204	0.145463533	0.890681419	0.066792461	0.228825979	0.296109464
appel	0.132664137	0.301986595	0.658946396	0.069919955	0.190332415	0.201419212
state	0.054473808	0.130552646	0.237706027	0.03227887	0.053444331	0.694954226
appeal	0.133373472	0.132759727	0.962172625	0.103941561	0.299874521	0.183995395
plaintiff	0.070133068	0.000599441	1.743202726	0.064295135	0.369843343	0.234529154
defend	0.039264196	0.011823395	1.607390132	0.073458462	0.391188297	0.243856511
case	0.098932165	0.215231163	0.663173051	0.046987733	0.239645351	0.297189681
issu	0.107337301	0.137633265	0.886191831	0.048331481	0.257863382	0.314932564
evid	0.076706141	0.414075847	0.595481429	0.064953203	0.194399354	0.153154173
respond	0.091402745	0.061959592	1.171003245	0.089691134	0.283108948	0.298355417
law	0.072698749	0.076856186	0.354268176	0.041970881	0.151058782	0.60623629
trial	0.119530138	0.337302846	0.679679056	0.057228724	0.236543246	0.141615366
learn	0.098439509	0.239910049	0.627683429	0.074646625	0.199518261	0.290163337
section	0.028587753	0.085404994	0.459153084	0.007984238	0.114220528	0.697171914
judgment	0.143872353	0.10338546	1.011177036	0.093096331	0.323594651	0.183274855
act	0.031557374	0.180745752	0.295149964	0.019559733	0.163291005	0.560244262
land	0.022524447	0.004613782	0.031275297	0.066711999	0.878497494	0.023658278
constitut	0.045892175	0.025014572	0.269763287	0.004623915	0.042938555	0.845404385
claim	0.070321121	0.010678965	0.876713196	0.059110166	0.383297209	0.346806793
said	0.060275246	0.151772669	0.84554564	0.066843602	0.290997926	0.301847942
feder	0.025217806	0.008442983	0.365405474	0.054143442	0.016571731	0.84522098
order	0.08485038	0.056274433	1.316521968	0.054421603	0.25005802	0.357533777
counsel	0.092868027	0.205225547	0.601178112	0.043964775	0.160837438	0.405152241
provis	0.044243478	0.030385281	0.303808327	0.014673588	0.063575884	0.805526607
parti	0.051709941	0.019935738	1.179856573	0.057685961	0.251993958	0.44350054
power	0.016781177	0.010948636	0.190844157	0.007950811	0.051882467	0.888644009
fact	0.084976042	0.258116275	0.648819242	0.053440898	0.230614853	0.271952015
exhibit	0.066912555	0.168838758	1.73279739	0.144381095	0.157024812	0.199853343
nwlr	0.159608159	0.139292688	1.091582617	0.054026347	0.186544895	0.295611505
made	0.061856541	0.161429619	1.017678431	0.065648176	0.207600819	0.349318645
person	0.033091416	0.27402712	0.5516328	0.029547933	0.116959092	0.460715646
respect	0.056874582	0.063611996	0.742630206	0.087127993	0.205596951	0.476593768
ground	0.127866708	0.153537099	0.80615105	0.069476415	0.348741196	0.177483476
see	0.096265304	0.190215194	0.967054312	0.058380629	0.21062624	0.297213104
gener	0.062568646	0.032530037	0.49939106	0.022372418	0.06249725	0.74866282
judg	0.112758059	0.328174411	0.663472886	0.066338702	0.247784666	0.139740741
disput	0.029443956	0.01152617	0.208352224	0.086910694	0.591148199	0.251188662
nigeria	0.048755249	0.06192731	0.674408217	0.040252831	0.087037611	0.663205164
matter	0.06519277	0.054539666	0.641754364	0.048125348	0.132337353	0.605765569

III. Result Discussion

The effect of the pre-processing stages on the legal case documents can be seen in table 1, and it can be observed that there is a great effect of stop word removal and stemming in the documents than the filtering stage, for instance, the total number of features of all the categories is 744624 while the number of features remaining after stop word removal and stemming is 290361, the number of features after filtering is 269935. It shows that the number of features considered as noise before filtering is 454263, while the number of noisy features after filtering is only 20426, but the total effect of all the stages cannot be over-emphasized when compared to the features before the pre-processing stages. Figure 1 shows the graph depicting the effect of the pre-processing stage on the number of features against each of the legal case categories. The undulating green line in the graph shows the different number of features obtained from each legal case category before pre-processing. Each point denotes the exact number of features present at each stage. The blue line shows the number of features remaining after the features have been passed through the stop word removal and stemming stage, while the yellow line shows the number of features obtained after filtering out the unwanted features or tokens. The demarcation between the lines in the graph shows a clear effect of the pre-processing stage on the number of features as the pre-processing stages are implemented.

The feature ranking method was used to select features of more importance and thereby shredding less important features from the available selected features from the pre-processing stage. Table 3 shows the top-ranked features in the documents of the various categories, for instance the feature "court" has the highest number of occurrences, 7598 times in all categories, and so on, followed by the stemmed word appel- which appeared 4449 times. This method has shown its effectivity in the feature selection process as more important features to the least important features are listed in descending order as shown in Table 3.

Table 4 depicted that the term weights of the relevant features were successfully calculated using the Naïve Bayes algorithm, these term weights can be used by any classifier of choice to classifier the documents to their respective categories. The posterior probabilities of the features are the term weights and can be used to form a multi-dimensional vector for a classifier like the Support Vector Machine. The new hybrid approach has introduced a new method of creating word vectors from raw text documents.

IV. Conclusion

In this paper, we developed a hybrid approach for feature selection and vectorization using a scheme referred to as Chi Square-Naïve Bayes hybrid scheme. Extensive pre-processing techniques like tokenization, stop words and non-alphanumeric character elimination, stemming, token filtering and case transformation was implemented for the extraction of the features from the raw corpus, more relevant features were selected using the chi-square (χ^2) statistical analysis, while the term weights were successfully formed using the Naïve Bayes probabilistic method for text classification. This developed vector space model successfully selected features and produced term weights from the provided corpus which are best used in the text classification process and the results obtained shows that this hybrid approach better than the existing weighting schemes.

Future study can incorporate other feature selection and vectorization (term weighting) methods to yield more interesting research results.

References

[1]. M. Duoqian, D. Qiguo, Z. Hongyun and, J. Na, Rough Set Based Hybrid Algorithm for Text Classification, Proceedings from Journal of Expert Systems with Applications, 36(5), 2009, 9168 – 9174.

[2]. G. Forman, An Extensive Empirical Study of Feature Selection Metrics for Text Classification, The Journal of Machine Learning. Res. 3., 2003, 1289 – 1305.

[3]. Y. Yang, Noise Reduction in a Statistical Approach to Text Categorization, Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 1995, 256 – 263.

[4]. Y. Yang and J.O Pedersen, A Comparative Study of Feature Selection in Text Categorization, Proceedings of the 14th International Conference on Machine Learning, 1997, 412 – 420.

[5]. D. Mladeni and M. Grobelnik, Feature Selection on Hierarchy of Web Documents, Journal of Decision Support Systems, 35, 2003, 45 -87.

[6]. H. Liu, J. Sun, L. Liu and H. Zhang, Feature Selection with Dynamic Mutual Information, Journal of Pattern Recognition, 42, 2009, 1330 – 1339.

[7]. C. Lee and G.G. Lee, Information Gain and Divergence-based Feature Selection for Machine Learning Based Text Categorization, Journal of Information Processing and Management, 42, 2006, 155 – 165.

[8]. Y.T. Chen and M.T. Chen, Using Chi Square Statistics to Measure Similarities for Text Categorization, Journal of Expert Systems with Applications, 38, 2011, 3085 – 3090.

[9]. T. Joachims, A Probabilistic analysis of the Rocchio Algorithm with TF-IDF for Text Categorization, Proceedings of the 14th International Conference on Machine Learning, 1997, 143 – 151.

[10]. C.D. Manning, P. Raghavan and, H. Schutze, Introduction to Information Retrieval (Cambridge, Cambridge University Press, 2008).

[11]. http://www.lawreportsofcourtsofnigeria.org

[12]. http://www.lawaspire.com.ng

[13]. http://members.unine.ch/jacques.savoy/clef/englishST.txt

[14]. M.F. Porter, An Algorithm for Suffix Stripping, Program (Auto. Lib. Info. Syst.), 14(3), 1980, 130-137.

[15]. L. Liu, J. Kang, J. Yu and Z. Wang, A Comparative Study on Unsupervised Feature Selection Methods for Text Clustering, Proceedings of NLP-KE'05, 2005, 597-601.

[16]. H. Liu, J. Sun, L. Liu and H. Zhang, Feature Selection with Dynamic Mutual Information, Journal of Pattern Recognition, 42, 2009, 1330 – 1339.

[17]. Y. Hong, S. Kwong, Y. Chang and Q. Ren, Consensus Unsupervised Feature Ranking from Multiple Views, Patt. Rec. Letters, 29, 2008, 595-602.

[18]. Z. Zheng, R. Srihari and S. Srihari, A Feature Selection Framework for Text Filtering, Proceedings of the Third IEEE International Conference on Data Mining, 2003, 705-708.

[19]. F. Sebastiani, Machine Learning in Automated Text Categorization, ACM Computing Surveys, 34(1), 2002, 1-47.

[20]. D. Isa, L.H Lee, V.P Kallimani and, R. RajKumar, Text Documents Preprocessing with the Bayes Formula for Classification Using the Support Vector Machine, IEEE, Traction of Knowledge and Data Engineering, 20(9), 2008, 1264-1272.