Indian Journal of Science and Technology, Vol 8(32), DOI: 10.17485/ijst/2015/v8i32/92127, November 2015
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

Hybridization of Bag-of-Words and Forum Metadata for Web Forum Question Post Detection

Adekunle Isiaka Obasa[1,2*], Naomie Salim[1] and Atif Khan[1]

[1]Faculty of Computing, Universiti Teknologi Malaysia, 81310, Skudai, Johor, Malaysia; iaobasa@yahoo.com, naomie@utm.my, atifkhan@icp.edu.pk
[2]Department of Computer Science, College of Science and Technology, Kaduna Polytechnic, P.M.B 2021, Kaduna, Nigeria

Abstract

Background/Objective:A web forum is a problem-solving online community.Web forum research activitieshave been focused on answer mining with the assumption that the starting post is a question post. This paper proposes methods for mining standard web forum questions. **Methods/Statistical Analysis**:Popular methods for web forum question post detection are question mark, question words, higher n-grams and sequential pattern mining. These methods have problem of low detection rate and implementation complexity. Implemented in this paper is hybridization of simple bag-of-words model with web forum metadata, simple rule of question mark and question words. Dimensional reduction was performed using chi-square and wrapper techniques. **Findings**:The quality of web forum question posts varies from excellent to mediocre or even spam. Detecting good question posts is non-trivial. It requires utilization of salient features. Combination of simple rule of question mark and question words with forum metadata performed better than each of the two.Integration of bag-of-words model with simple rule of question marks, question words and forum metadata enhances question post detection. Dimensionality reduction using chi-square were found to perform better than other popular filters like info gain, gain ratio and symmetric uncertain. **Applications/Improvements**: Three publicly available datasets of varying technical degrees were used for the experiments.The experimental results revealed that an enhanced bag-of-words model can perform better than complex techniques that implement higher N-gram with part-of-speech tagging.

Keywords: Bag-of-words, Forum Metadata, Web Forum,Question Detection, Dimensionality Reduction, Web Forum Question

1. Introduction

A web forum is a typical social medium that has grown in popularity. A forum is a typical Web 2.0 site as it allows users to interact and collaborate with each other. In web forum, both technical and less technical issues are discussed. The Forum brings together experts from all walks of life. The scenario is that users with specific problems post questions to the forum and rely on other members to provide good answers. The answers provided by users are referred to as reply posts while the posted question known as the initial post. The combination of the initial post and the reply posts is known as a thread. People mostly use the discussion boards (i.e. web forum) as problem-solving platforms. A number of commercial organizations such as Microsoft, Dell and IBM directly use online forums as problem-solving domain for answering questions and discussing needs raised by customers. [1]found that 90% of 40 discussion boards they studied contain question-answering knowledge. By using speech acts investigation on several sampled forums, [2,3] discovered that question answering content is usually the largest type of content on forums.

The collaborative activities within the forum offer a lot of benefits. In technical forums such as hardware or software forum, a lot of issues such as installing software or hardware, troubleshooting codes, fixing bugs, implementing tools, etc. are being discussed on a daily basis. For non-technical forum like travel, members share their travel

*Author for correspondence

experience with others. Good opinions are generated by members for the benefit of other members. It will be highly desirable to mine human knowledge being generated in the forum for the benefit of mankind. The aim of this paper therefore, is to mine standard initial posts as web forum question posts using an enhanced bag-of-words. That is, bag-of-words combined with forum metadata and simple rule of question mark and question words.

The rest of the paper is organized as follows. Section 2 gives description of the problem. Section 3 discusses related work while Section 4 presents the proposed approach. Experimental design is done in Section 5. Section 6 concludes the paper and discusses the future work.

2. Problem Formulation

We consider web forum question post detection as classification problem. The problem is about getting salient features that can effectively classify web forum initial post as a question or not question. The initial post of a thread is considered as a whole as a question post if it contains a specific problem that needs to be solved otherwise it is non-question post. This problem definition is similar to that of [4]and[5]. For example, the following statements constitute a question post from Photography on the net, a digital camera forum.

"I have found that when I take pictures of scenery or landscape with no particular focus, that the camera has a difficult time focusing. I have tried the landscape mode but that does not work very well. I am mainly trying to do manual focus, however it is so difficult to tell by just looking through the viewfinder. Are there any techniques or tips that anyone would recommend?"

The last sentence in the post is a question sentence; it gives little information about the real problem. The problem is the entire scenario that the author described using several sentences as a whole. It is therefore practical to treat the whole post as a question post.

Web forum initial post is often being considered as question post when mining answers from web forum [6-8] without due consideration for what the post is all about. Initial post can be an announcement, a report or an acknowledgement which does not require any answer from the members. Furthermore, some initial posts are trivial questions that cannot be mined for any knowledge discovery. An example of such that can be found in forum due to its less restrictiveness is a question post like "*Hi guys, check out my pictures on facebook. Can anybody say that I'm not handsome?"* In view of all these issues, it is desirable to first identify the web forum question post before looking for its answer.

3. Related Work

Mining of web forum questions have been tackled using approaches that ranges from simple rules to complex techniques. The simple rules are the question marks, the 5W1H question words (who, what, where, when, why or how) and modal words (can, will, would etc.). The simple rule approaches are popular methods but have been found to be inadequate for mining web forum questions [1,5]. In a forum, many questions don't end with question mark and most of the statements that contain question words are not questions e.g. *"Everybody knows how he behaves"*. It is also difficult for simple rule to detect imperative questions such as *"I wonder if anybody could direct me to her office"*

The inadequacies of simple rules called for the implementation of some complex techniques such as sequential pattern mining (SPM) [1,4], n-grams [4,5], and regular expressions [9]. All these approaches could enhance the performance of the simple rules but have their own shortcomings. The SPM requires POS tagger. Accuracy of the method depends on POS tagging. The casual language of forum may affect tagging. Also, the computational effort of the approach may be impractical for large dataset. Higher n-grams are computationally expensive to generate. For question mining, differences between question and non-question n-gram must be well established [5]. [10]used regular expression to achieve F1 Score of 96% in E-mail domain for interrogative questions. This approach can be built around question words to mine interrogative questions. Itcannot handle the complex questions that dominate web forum questions. Notable research activities that involve the use of bag-of-words combined with some other approaches based on news articles, community-based question answering (CQA) and web forum corpus are shown in Table 1.

In Table-1, [11] used news article of Wall Street Journal corpus to determine opinion questions using BoW combined with n-gram. Their BoW was simply collection of opinion words which are positive and negative adjectives, nouns, verbs and adverbs. This in a way is a form of filtering out some word identities on a larger scale compared to the works of [5,12] . This influenced the performance of the

Table 1. Review of bag-of-words combination with other approaches

Reference	What is combined with BoW	Feature Selection used for BoW	Learning Method	Motivation	Result Accuracy (F-1 measure) (%)
[11]	BoW + 2-gram + 3-gram	Considered mainly opinion words	Naïve Bayes	Answering opinion questions by separating opinions from facts	87
[13]	BoW + POS + 2-gram + 3-gram	None	LibSVM	Determining whether CQA question has Objective or subjective orientation	72
[15]	BoW + 2-gram	Chi-square	LibSVM	Community QA question classification	75.3
[14]	BoW + 2-gram + 3-gram	None	Multinomial Naïve Bayes	Evaluation of subjectivity analysis in web forum.	72.3
[18]	BoW only	None	Multinomial Naïve Bayes	Classification of web forum posts	57.7

BoW and n-gram's higher result compared to others. [13] and [14] used BoW with 2 and 3-grams without feature selection to achieve similar results using different classifiers. [15]combined BoW with 2-gram and applied chi-square as feature selection to obtain a slightly higher result.A very low result was realized by [16] that used only BoW. This confirms that BoW needs to be combined with some other approaches to enhance its performance. It is also worth noting that feature selection enhances BoW performance. A question one could ask here is what dimensionality reduction will be most suitable for enhancing BoW performance? This question, to a large extent is addressed in this study.

4. Proposed Approach

One of the research questions of this paper is to confirm the effectiveness of using simple rules and forum metadata in mining web forum questions. By simple rule, we mean question marks (?) and question words (i.e. Wh-word types). The forum metadata are the forum structural features such as no. of words in a post, position of a post in the thread, etc. These features, their descriptions and types are shown in Table 2. Some researchers [1,5] have expressed the view that using simple rule for web forum question detection may be inadequate. Also, some researchers [4,12] have indicated that forum metadata are helpful in detecting forum questions. It is therefore the concern of this paper to confirm whether the combination of the two can enhance performance. It will also be worthwhile to examine progressively the performance of each feature so as to check feature redundancy by using add-one-in approach.

Another research question of this paper is to investigate the effectiveness of dimensionality reduction on bag-of-words approach for web forum question post detection. The implementation of bag-of-words can be summarized as follows:

i) Detect and extract keywords,
ii) Build a keyword dictionary,
iii) Use keyword dictionary to build term-document matrix
iv) Use machine learning to train a classifier for the classification.

The above procedure will generate a set of keywords known as bag-of-words. These keywords are the features that will be used to mine the question post. The keywords are represented in term-document matrix using weighting schemes. We experimented with three term weighting schemes: Binary, TF, and TF*IDF. The binary represents text as binary vector of terms. Each unit of the vector represents a term and its value is '1' if the term appears in the document, '0' otherwise. The TF and TF*IDF are term frequency and product of term frequency with the inverse document frequency. The binary exhibited better performance in our implemented experiments, so we use this weighting scheme for all the experiments in this paper.

Most of the values of the term-document matrix will be zeros since for a given document, a small fraction of it will be found in keyword dictionary. In view of this, bag-of-words are said to be typically high-dimensional sparse datasets that require a lot of memory. In addition, some of the non-zero features could be redundant or less effective

Table 2. List of features, descriptions and types

Feature	Code	Description	Type
Question Mark (?)	QM	No. of question marks in the post. The higher the number, the more likely the post is a question.	Simple rule
Wh-word type	WH	No. of question words in the post. The higher the number, the more likely the post is a question.	Simple rule
No. of words in post	NW	Question post are expected to be precise, therefore the lesser the number of words the more likely the post is a question.	Forum metadata
No. of posts in thread	NP	A thread with too many posts is subjective; hence its question may not be factual. Threads with more than 20 posts are considered as non-questions	Forum metadata
No. of threads created by the user	NT	The higher the number the more likely the post is a question.	Forum metadata
No. of replies created by the user	NR	The higher the number the more likely the post is not a question.	Forum metadata

Table 3. Question detection dataset summary

Instances	CAM	Ubuntu	NYC
Total No. of Positive Instances (i.e. Questions)	204	223	225
Total No. of Negative Instances (i.e. Non-Questions)	204	223	225
Total No. of Initial Posts	408	445	450

for the task of question detection. In order to overcome these problems, we experiment with both filter and wrapper feature selections to obtain the most salient features. The dimensionally reduced feature sets were combined with the simple rules and forum metadata to form feature vectors used for training and testing.

5. Experimental Design

In this section, we show how our proposed approach is actualized.

5.1 Dataset and Dataset Annotation

Three different datasets were used for the experiments conducted in this paper. We collected 16,853 threads of Photography On The Net[1], a digital camera forum (CAM dataset) and 41,078 threads of Ubuntu Forum[2], an Ubuntu Linux community forum (Ubuntu dataset). In addition, we also collected 31,998 threads of Trip Advisor-New York[3] that contains travel related discussions on New York City (NYC dataset). All the datasets are made available publicly by [4,14,17]. These three forums are considered so as to evaluate the implemented methods on different domains of online forums. The Ubuntu dataset that contains a lot of configuration parameters and codes represents highly technical domain, CAM dataset that contains more of technical terms and some settings but no codes represents less technical domain while NYC dataset that does not contain codes, configuration settings and more of technical terms represents non-technical domain.

5.2 Experimental Settings

We used different supervised learning algorithms for our classification task. These algorithms include Multinomial Naïve Bayes (MNB), Support Vector Machines (LibSVM), Decision tree (J48), Sequential Minimal Optimization (SMO) and Multilayer Perceptron (MP). In order to aid the experimentation carried out in this research, a freely available machine learning toolkit called weka is used. Weka is a pool of machine learning algorithm for data mining activities. The version of weka implemented in this study is weka 3.7.12.

Classification results are obtained using 10-fold cross-validation and 80% split (i.e. 80% training, 20% testing). The performances of our classifiers were evaluated using

[1]http://photography-on-the.net/
[2]http://ubuntuforums.org/
[3]http://www.tripadvisor.com/ShowForum-g60763-i5-New_York_City_New_York.html.

precision (P), recall (R) and F-1 measure (F) metrics. Basic pre-processing such as removal of HTML tags and lower casing all words were performed on the corpus of initial posts used for the experiments. Bag-of-words dimensionality reduction is performed using both filters and wrapper. The filters considered in the study are: Chi-square, Information gains (Info. Gain), Gain ratio, Symmetrical uncertainty (Sym. Uncert.). These filters are experimented using three thresholds of 0, 5 and 10. The wrapper method is based on SMO classifier. SMO was determined empirically for the wrapper.

5.3 Experimental Results and Discussions of Simple Rule and Forum Metadata

As stated above, five machine learning algorithms were used for the experiments. The Multinomial Naïve Bayes gave best result for CAM dataset (a less technical dataset). The tree classifier (J48) gave best results for Ubuntu (a highly technical dataset) and NYC (a non-technical dataset). Classification results for the two classifiers are shown in Tables 4 and 5. The comparative analyses of the two classifiers on the three datasets using F-scores of the 10-folds cross validation are depicted in Figures 1 and 2 for the individual and combined features respectively. The performance observations of the individual features, the simple rules, the forum metadata, and other various combinations as well as the scalability of the two classifiers are as follows:

i) The J48 performs better than MNB classifier using individual features for classification. The MNB gave the same results for almost all the features irrespective of whether the validation is 10-fold cross validation or

Table 4. Empirical results of simple rule and forum metadata ofNYC dataset using MNB and J48

Dataset	Feature	Val.	MNB			J48		
			P	R	F	P	R	F
NYC	QM	Cross	40.5	63.6	49.5	85.3	83.8	84
		80%	30.1	54.8	38.8	81.3	81.5	81.4
	WH	Cross	40.5	63.6	49.5	49.4	62.3	50
		80%	30.1	54.8	38.8	29.2	51.6	37.3
	NW	Cross	40.5	63.6	49.5	65.9	65.6	56.5
		80%	30.1	54.8	38.8	60.8	56.5	44.7
	NP	Cross	40.5	63.6	49.5	67.5	66.6	58.6
		80%	30.1	54.8	38.8	73	64.5	58.7
	NT	Cross	40.5	63.6	49.5	40.5	63.6	49.5
		80%	30.1	54.8	38.8	30.1	54.8	38.8
	NR	Cross	40.5	63.6	49.5	71.5	69.5	63.9
		80%	30.1	54.8	38.8	77.9	62.9	54.7
	QM+WH	Cross	69.1	70.1	68.4	85.3	83.8	84
		80%	66.2	66.1	65.3	81.3	81.5	81.4
	NW+NP+NT+NR	Cross	70.2	68.8	63.1	69.9	69.8	65.8
		80%	79.1	66.1	60	73	64.5	58.7
	QM+WH+NW	Cross	85.6	84.7	84.9	84.9	83.4	83.7
		80%	85.5	85.5	85.5	81.3	81.5	81.4
	QM+WH+NW+NP	Cross	81.3	81.5	81.4	84.7	83.4	83.7
		80%	78.8	75.8	74.5	91.1	90.3	90.3
	QM+WH+NW+NP+NT	Cross	77.6	77.9	77.3	82.6	81.5	81.7
		80%	77.7	74.2	72.6	85.5	85.5	85.5
	QM+WH+NW+NP+NT+NR	Cross	72.4	69.8	64.1	83.6	82.8	83
		80%	79.1	66.1	60	81.3	81.5	81.4

Table 5. Results of simple rule and forum metadata of CAM and Ubuntu using MNBand J48

Dataset	Feature	Val.	MNB			J48		
			P	R	F	P	R	F
CAM	QM	Cross	49	49	48.5	62.8	58.8	55.3
		80%	23.8	48.8	32	56.3	54.9	50.6
	WH	Cross	49	49	48.5	49	49	48.5
		80%	23.8	48.8	32	23.8	48.8	32
	NW	Cross	49	49	48.5	62	57.8	53.9
		80%	23.8	48.8	32	55.5	53.7	47
	NP	Cross	49	49	48.5	68.3	55.1	45.3
		80%	23.8	48.8	32	63.5	53.7	41.7
	NT	Cross	49	49	48.5	48.7	48.8	48
		80%	23.8	48.8	32	23.8	48.8	32
	NR	Cross	49	49	48.5	59.4	55.6	50.7
		80%	23.8	48.8	32	50.7	51.2	44.2
	QM+WH	Cross	53.7	53.7	53.7	63.2	60.8	58.9
		80%	49	48.8	48.4	56.3	54.9	50.6
	NW+NP+NT+NR	Cross	54	52.5	47.3	62.8	58.8	55.3
		80%	50.7	51.2	40.2	50.9	51.2	49
	QM+WH+NW	Cross	66.3	66.2	66.1	63.3	63	62.7
		80%	63.6	63.4	63.4	56.3	56.1	55
	QM+WH+NW+NP	Cross	60.3	60.3	60.3	64.3	63.5	62.9
		80%	61.2	61	60.9	58.8	58.5	57.8
	QM+WH+NW+NP+NT	Cross	54.9	54.7	54	64.3	63.5	62.9
		80%	50.4	50	48.9	58.8	58.5	57.8
	QM+WH+NW+NP+NT+NR	Cross	54.5	52.7	47.4	66.5	65.7	65.2
		80%	50.7	51.2	40.2	58.8	58.5	57.8
Ubuntu	QM	Cross	25.2	50.2	33.6	78.3	78.2	78.1
		80%	24.4	49.4	32.7	71.1	70.8	70.6
	WH	Cross	25.2	50.2	33.6	25.2	50.2	33.6
		80%	24.4	49.4	32.7	24.4	49.4	32.7
	NW	Cross	25.2	50.2	33.6	68	58.1	51.6
		80%	24.4	49.4	32.7	66	53.9	42.6
	NP	Cross	25.2	50.2	33.6	60.5	57.9	54.9
		80%	24.4	49.4	32.7	79.2	64	58.9
	NT	Cross	25.2	50.2	33.6	54.1	52.3	46.7
		80%	24.4	49.4	32.7	54.5	52.8	46.5
	NR	Cross	25.2	50.2	33.6	55.3	54.7	53.4
		80%	24.4	49.4	32.7	55.8	55.1	54
	QM+WH	Cross	62.2	61.7	61.2	78	77.9	77.9
		80%	57.6	57.3	57	71.1	70.8	70.6
	NW+NP+NT+NR	Cross	61.5	61	60.6	59.3	59.2	59.2
		80%	53.1	52.8	52.3	73.5	64	60.3
	QM+WH+NW	Cross	77	76.8	76.8	78.5	78.2	78.1
		80%	76.7	76.4	76.3	70.9	69.7	69.1
	QM+WH+NW+NP	Cross	67.6	66.7	66.2	80.9	80.2	80.1
		80%	67.2	65.2	64.2	75.5	73	72.3
	QM+WH+NW+NP+NT	Cross	66.1	64.6	63.8	81	80.4	80.3
		80%	61.6	59.6	57.9	75.5	73	72.3
	QM+WH+NW+NP+NT+NR	Cross	67.5	66.2	65.6	80.8	80.6	80.6
		80%	61.8	60.7	59.9	76.1	75.3	75.1

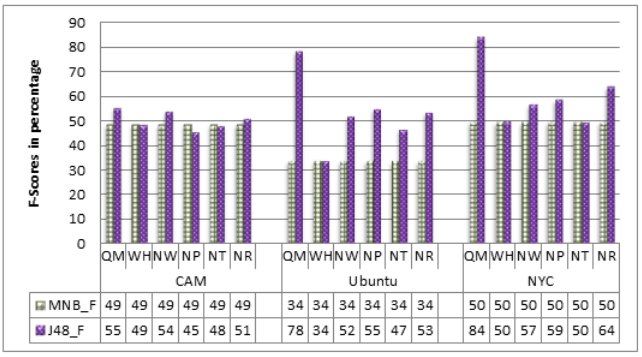

Figure 1. Comparative analysis of individual features for the two classifiers on the three datasets

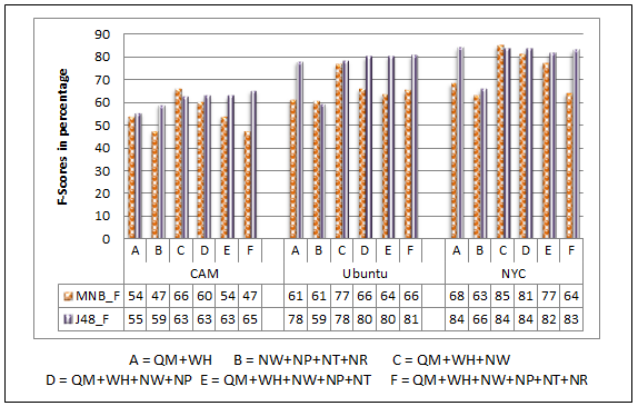

Figure 2. Comparative analysis of combined features for the two classifiers on the three datasets

80% split. This happens for all the three datasets. The results are much lower than that of J48.

ii) The QM feature outperforms all other individual features. It gave up to 84% F-measure for NYC dataset. This shows that the question mark is still a good feature to use in mining web forum question post. But this performance does not cut across the three datasets hence the need for more scalable features.

iii) The simple rule (QM+WH) outperforms the forum metadata (NW+NP+NT+NR) in all the 3 datasets. The simple rule is highlighted in blue while forum metadata is highlighted in pink in Tables 4 and 5. The combination of the forum metadata features is better than their respective individual features. Whereas the simple rule combination is not generally better than the QM alone. This shows that the question word scales down the performance of the question mark. This may be due to the fact that both are often used in a question but question mark may have more occurrences because of questions that do not contain question words. For example, *"Is it right for a man to slap a lady?"*

iv) With other combinations, the MNB classifier gives the best results using the QM+WH+NW features with all the 3 datasets. On the other hand, the J48 uses all the 6 features to give better results than the MNB.

v) Generally, these set of features favour cross validation than 80% split.

vi) Mostly, the combination of the simple rule with the forum metadata in some forms outperforms either of the two being used separately and this confirms that their combination can enhance question post detection.

5.4 Experimental Results and Discussions of Bag-of-words

In Table 6, screening of the three datasets (CAM, Ubuntu and NYC) using different reduction methods confirm chi-square, information gain, gain ratio and symmetrical uncertainty exhibiting the same feature reduction with only chi-square giving discriminative features for thresholds of 5 and 10. In the table, the 1775 features of CAM dataset were reduced to 253 features for all the four filters using threshold of 0. Chi-square gave 93 and 15 features for thresholds of 5 and 10 respectively. Classification results of the four filters for threshold of 0 are the same. In view of this, our empirical analyses are based on chi-square, wrapper and non-filtering.

The results of MNB and SMO are the best of the 5 classifiers. SMO gave best result for CAM dataset (a less technical dataset) while MNB gave best results for Ubuntu (a highly technical dataset) and NYC (a non-technical dataset). A comparative analysis of the MNB and SMO is shown in Figure 3. SMO works better with the wrapper method while MNB favours chi-square with lower threshold. Cross validation favours CAM dataset (a less technical dataset) and 80% split favours both Ubuntu and NYC. The MP classifier takes much longer time to generate results. Its computation for thousands of features was ignored in this study since such results cannot be better than the filters method.

As expected, the BoW without dimensionality reduction performed poorly with all the classifiers. The use of chi-square with different thresholds gives some improvements. An amazing observation with the use of chi-square thresholds is that higher thresholds with fewer feature space does not guarantee better performance. This reveals

Table 6. Dataset feature reduction analyses

Dataset	Filter /Wrapper	Thresholds		
		0	5	10
CAM	Chi-square	253	93	15
	Info. Gain	253	0	0
	Gain Ratio	253	0	0
	Sym. Uncertain	253	0	0
	Wrapper(SMO)	63		
	No. Filter	1775		
Ubuntu	Chi-square	139	74	10
	Info. Gain	139	0	0
	Gain Ratio	139	0	0
	Sym. Uncertain	139	0	0
	Wrapper(SMO)	44		
	No. Filter	1626		
NYC	Chi-square	99	98	33
	Info. Gain	99	0	0
	Gain Ratio	99	0	0
	Sym. Uncertain	99	0	0
	Wrapper(SMO)	22		
	No. Filter	124		

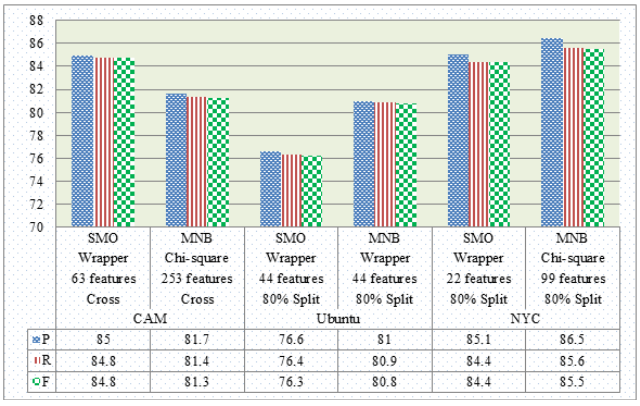

Figure 3. Comparative analysis of bag-of-words' results for the two best classifiers (SMO and MNB)

that higher threshold of chi-square does not optimize feature selection.The wrapper method with higher number of features often performs better than the higher threshold of chi-square with lesser number of features. Out of the five classifiers, multinomial Naïve Bayes works much better with chi-square using lower threshold especially on NYC dataset (less technical dataset). The general results of the two best classifiers for the three datasets are shown in Table 7.

Comparative analysis of the bag-of-words with simple rule and forum metadata discussed in the previous section reveals that the bag-of-words is generally better than the simple rule and forum metadata. The only exception is the NYC dataset in which simple rule is better. The results of BoW are much more generalized than the simple rule. The J48 classifier favours the simple rule while SMO and MNB favour BoW. Apart from the fact that BoW performs better, its feature generation is also simpler than the simple rule approach.Figure 4 depicts comparative analyses of the best results of the two approaches on the three datasets. In the next section, we will confirm whether the integration of the simple rule and forum metadata with bag-of-words will be better than either of the two.

5.5 Experimentation, Results and Discussions of Hybrid Approach

We also propose the same five machine learning algorithms used in the previous sections of the paper for exploration on the newly formulated feature sets. The three datasets (CAM, Ubuntu and NYC) were also being used for the experimentation. The implemented features are: QM, WH, NW, NP, NT, NR, BoW+QM+WH+NW and BoW+ALL_SR. The BoW is the bag-of-words features generated using SMO-based wrapper method while ALL_SR is the combination of all simple rule and forum metadata individual features, that is, QM+WH+NW+NP+NT+NR. All these features have been described in previous sections.

Out of the 5 classifiers used for the experiments, multilayer perceptron (MP) gave the best results for the three datasets (CAM, Ubuntu and NYC). Best results were recorded for the features BoW+QM and BoW+QM+WH+NW on CAM dataset, BoW+QM on Ubuntu dataset and BoW+QM+WH+NW on NYC dataset. This shows that QM discriminates well with other features in enhancing performance. The combination QM+WH+NW is also a good feature to be combined with other features to improve performance. The multilayer perceptron results for the 3 datasets are shown in Table 8.

The hybrid approach outperformed the simple rule and forum metadata (SR&FM) and the bag-of-words (BoW) approaches. The approach scales well with multilayer perceptron classifier on the 3 datasets. Comparative analyses of the three approaches using their precision

Table 7. Empirical results of bag-of-words for the two best classifiers using different dimensionality reductions

Dataset	Feature Selection Method	No. of Feature	Validation Method	SMO			MNB		
				P	R	F	P	R	F
CAM	No Filter	1775	Cross	62.7	62.7	62.7	73	73	73
			80% Split	65	64.6	64.6	66.4	65.9	65.8
	Chi-square	253	Cross	81.7	79.9	80.2	81.7	81.4	81.3
			80% Split	74.8	72	71.5	79.4	78	77.9
		93	Cross	74	72.8	72.5	69.6	68.9	68.6
			80% Split	66.5	63.4	62.3	61.5	59.8	59
		15	Cross	72.2	71.1	70.7	56.9	56.9	56.9
			80% Split	74.8	72	71.5	58.6	56.1	54.4
	Wrapper	63	Cross	85	84.8	84.8	65.2	64.7	64.4
			80% Split	73	69.5	68.8	75.7	73.2	72.8
Ubuntu	No Filter	1626	Cross	59.3	59.3	59.3	64.1	63.8	63.6
			80% Split	66.3	66.3	66.1	65.7	64	63.7
	Chi-square	139	Cross	74.6	73.3	72.9	74.7	73.9	73.7
			80% Split	75.2	71.9	71.4	80.6	75.3	74.5
		74	Cross	69.1	68.1	67.6	70.3	69	68.5
			80% Split	71	67.4	66.6	76.6	70.8	69.6
		10	Cross	66.1	66.1	66.1	62.9	62.9	62.9
			80% Split	63.6	62.9	62.9	70.8	64	61.9
	Wrapper	44	Cross	75.5	75.2	75.2	77.9	76.9	76.6
			80% Split	76.6	76.4	76.3	81	80.9	80.8
NYC	No Filter	1224	Cross	70.5	70.5	70.5	70.3	70.1	70
			80% Split	69.1	68.9	68.9	76.5	75.6	75.4
	Chi-square	99	Cross	76.3	76.3	76.3	82.5	81.8	81.7
			80% Split	73.8	73.3	73.3	86.5	85.6	85.5
		98	Cross	76.5	76.5	76.5	82.7	82	82
			80% Split	73.8	73.3	73.3	86.5	85.6	85.5
		33	Cross	79.7	79.4	79.3	82.3	81.8	81.8
			80% Split	83.3	82.2	82.1	84	82.2	82
	Wrapper	22	Cross	83.2	82.9	82.9	84.6	84.3	84.2
			80% Split	85.1	84.4	84.4	84	81.1	80.8

(P), recall (R) and F-score (F) are shown in Figure 5. In Figure 5, simple rule and forum metadata, bag-of-words and integration of the two are labelled as SR&FM, BoW and SR&FM+BoW respectively.

5.6 Comparison with Baselines

In order to establish where this study stands in the research area, we consider the five works that are closely related to the study as baselines. These works are presented in Table 1. The comparative analysis of the baselines and our proposed integration approach is shown in Figure 6. The F-score metric is used for the comparison. The proposed approach outperforms the five baselines. It then means that the use of filter and wrapper method for feature selection in this domain is better than using part-of-speech tagging. The reason for this may be attributed to the noisy nature of web forum which may not allow part-of-speech tagging to work effectively.

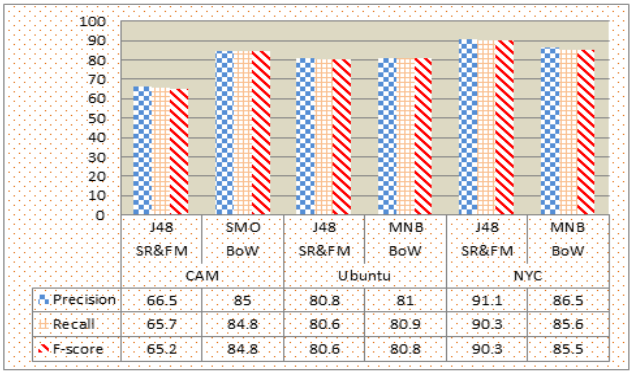

Figure 4. Comparative analyses of bag-of-words with simple rule and forum metadata

Table 8. Empirical results of integrating BoW with SR&FM using MP

Dataset	Feature	Val.	MP		
			P	R	F
CAM	BoW+QM	Cross	80.9	80.9	80.9
		80% Split	87.8	87.8	87.8
	BoW+WH	Cross	80.9	80.9	80.9
		80% Split	85.4	85.4	85.4
	BoW+NW	Cross	81.1	81.1	81.1
		80% Split	85.4	85.4	85.4
	BoW+NP	Cross	82	81.9	81.8
		80% Split	84.2	84.1	84.1
	BoW+NT	Cross	81.4	81.4	81.4
		80% Split	86.6	86.6	86.6
	BoW+NR	Cross	81.4	81.4	81.4
		80% Split	86.6	86.6	86.6
	BoW+QM+ WH+NW	Cross	81.1	81.1	81.1
		80% Split	87.8	87.8	87.8
	BoW+ALL_SR	Cross	80.9	80.9	80.9
		80% Split	83	82.9	82.9
Ubuntu	BoW+QM	Cross	81.7	81.6	81.6
		80% Split	86.6	86.6	86.6
	BoW+WH	Cross	80.2	80.1	80.1
		80% Split	85.4	85.4	85.4
	BoW+NW	Cross	80.2	80.1	80.1
		80% Split	85.4	85.4	85.4
	BoW+NP	Cross	80.2	80.1	80.1
		80% Split	80.7	80.5	80.4
	BoW+NT	Cross	79.9	79.9	79.9
		80% Split	85.4	85.4	85.4
	BoW+NR	Cross	79.4	79.4	79.4
		80% Split	84.2	84.1	84.1
	BoW+QM+ WH+NW	Cross	80.4	80.4	80.4
		80% Split	84.2	84.1	84.1
	BoW+ALL_SR	Cross	80.2	80.1	80.1
		80% Split	80.7	80.5	80.4
NYC	BoW+QM	Cross	86.1	86	86.1
		80% Split	83.9	83.9	83.8
	BoW+WH	Cross	77.6	77.9	77.2
		80% Split	78.8	75.8	74.5
	BoW+NW	Cross	76.3	76.6	75.7
		80% Split	78.8	75.8	74.5
	BoW+NP	Cross	74.5	75	73.9
		80% Split	79.6	74.2	72.1
	BoW+NT	Cross	75.3	75.6	74.6
		80% Split	76.3	74.2	73
	BoW+NR	Cross	76.6	76.9	76.2
		80% Split	78.8	75.8	74.5
	BoW+QM+ WH+NW	Cross	88.7	88.6	88.7
		80% Split	86.2	85.5	85.3
	BoW+ALL_SR	Cross	86	86	86
		80% Split	86.2	85.5	85.3

(*Continued*)

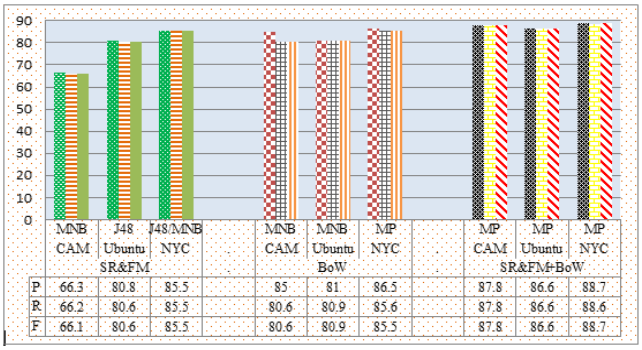

Figure 5. Comparative Analyses of the three Approaches – SR&FM, BoW and SR&FM+BoW on the three datasets

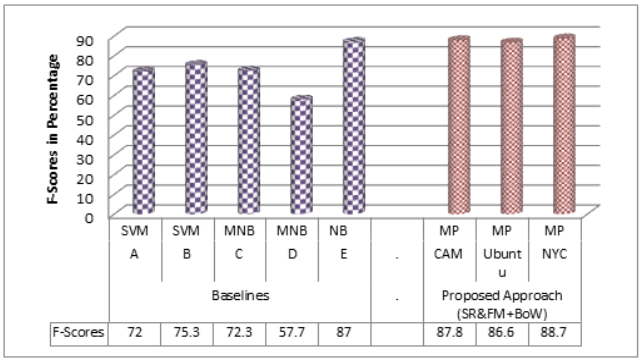

Figure 6. Comparing proposed approach with the baselines

6. Conclusion

In this paper, we investigate the performance of simple rules of question mark and question words, web forum metadata and bag-of-words for web forum question post detection. In the paper, we confirm that:

i. The combination of simple rule with forum metadata performs better than each of the two implemented separately.
ii. Dimensionality reduction using both filters and wrapper enhances the performance of bag-of-words in detecting web forum questions.
iii. The use of higher thresholds for chi-square can reduce feature dimension without necessarily optimizing it.
iv. The performance of classifier for this task depends on the technicality of the dataset. That is, different datasets will require some specific classifier for optimal performance.
v. The performance of bag-of-words can be enhanced by simple rule and forum metadata for web forum question post detection.

Our future work shall address the following problems:

1. Evaluating the performance of bag-of-ngrams using different filters and wrapper to determine the best N-gram for the task.
2. Investigating the performance of feature selection using evolutionary algorithms for web forum question post detection.

7. Acknowledgements

This work was supported by the Ministry of Education Malaysia, Kaduna Polytechnic, Kaduna, Nigeria and Soft Computing Research Group (SCRG) of Universiti Teknologi Malaysia (UTM). The work was also supported in part by grant from Vote R.J130000-7828.4F719.

8. References

1. Cong G, Wang L, Lin C-Y, Song Y-I, Sun Y. Finding question-answer pairs from online forums. In: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval: ACM; 2008, pp. 467–74.
2. Kim J, Chern G, Feng D, Shaw E, Hovy E. Mining and assessing discussions on the web through speech act analysis. In: Proceedings of the Workshop on Web Content Mining with Human Language Technologies at the 5th International Semantic Web Conference; 2006.
3. Kim J, Shaw E, Feng D, Beal C, Hovy E. Modeling and assessing student activities in on-line discussions. In: Proc. of the AAAI Workshop on Educational Data Mining; 2006.
4. Hong L, Davison BD. A classification-based approach to question answering in discussion boards. In: Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval: ACM; 2009, pp. 171–8.
5. Sun L, Liu B, Wang B, Zhang D, Wang X. A study of features on Primary Question detection in Chinese online forums. In: Fuzzy Systems and Knowledge Discovery (FSKD), Seventh International Conference on: IEEE; 2010, pp. 2422–7.
6. Catherine R, Singh A, Gangadharaiah R, Raghu D, Visweswariah K. Does Similarity Matter? The Case of Answer Extraction from Technical Discussion Forums. In: COLING (Posters); 2012, pp. 175–84.
7. Deepak P, Visweswariah K. Unsupervised Solution Post Identification from Discussion Forums. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics; Baltimore, Maryland, USA: Association for Computational Linguistics; 2014 June 23-25, pp. 155–64.
8. Kumar N, Srinathan K, Varma V. Unsupervised Deep Semantic and Logical Analysis for Identification of Solution Posts from Community Answer. 2015.
9. Atkinson J, Figueroa A, Andrade C. Evolutionary optimization for ranking how-to questions based on user-generated contents. Expert Systems with Applications 2013, 40(17), pp. 7060–8.
10. Kwong H, Yorke-Smith N. Detection of imperative and declarative question–answer pairs in email conversations. AI Communications 2009, 25(4), pp. 271–83.
11. Yu H, Hatzivassiloglou V. Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In: Proceedings of the 2003 conference on Empirical methods in natural language processing: Association for Computational Linguistics; 2003, pp. 129–36.
12. Wang B, Liu B, Sun C, Wang X, Sun L. Extracting Chinese question-answer pairs from online forums. In: Systems, Man and Cybernetics, 2009. SMC. IEEE International Conference on: IEEE; 2009, pp. 1159–64.
13. Li B, Liu Y, Agichtein E. Cocqa: co-training over questions and answers with an application to predicting question subjectivity orientation. In: Proceedings of the conference on empirical methods in natural language processing: Association for Computational Linguistics; 2008, pp. 937–46.
14. Biyani P, Bhatia S, Caragea C, Mitra P. Using non-lexical features for identifying factual and opinionative threads in

online forums. Knowledge-Based Systems 2014, 69(October 2014), pp. 170–8.
15. Aikawa N, Sakai T, Yamana H. Community QA Question Classification: Is the Asker Looking for Subjective Answers or Not? IPSJ Online Transactions 2011, 4, pp. 160–8.
16. Bhatia S, Biyani P, Mitra P. Identifying the role of individual user messages in an online discussion and its use in thread retrieval. Journal of the Association for Information Science and Technology 2015, pp. 1–13.
17. Sumit B, Prakhar B, Prasenjit M. Classifying User Messages For Managing Web Forum Data. Proceedings of the Fifteenth International Workshop on the Web and Databases (WebDB 2012), Scottsdale, AZ, USA. 2012 May 20, pp. 1–6.