Automated News Classification using N-gram Model and Key Features of Nepali Language

Dinesh Dangol[*], Rupesh Dahi Shrestha[+], Arun Timalsina[#]

[*] Department of Computer Science & Engineering, dines.dangol@gmail.com
[+] Department of Electronics & Communication Engineering, rupeshd.shrestha@gmail.com
[#] Department of Electronics & Computer Engineering, t.arun@ioe.edu.np

Dinesh Dangol *is currently working as Assistant Professor and Head of the Department of Computer Science and Engineering (in Nepal Engineering College). He has more than 10 years of teaching and practical experience in networking and software development and has few research publications in the field of machine learning and natural language processing.*

Rupesh Dahi Shrestha *received the BE degree in Electronics and Communication from Nepal Engineering College in 2005, and MSc degree in Information and Communication Engineering from Pulchowk Campus in 2013. He has experience of 12 years in teaching and is currently working as Assistant Professor and Head of the Department of Electronics and Communication Engineering, Nepal Engineering College. His research interests include signal processing, robotics and embedded system.*

Arun Timalsina *is Assistant Professor in Department of Electronics and Computer Engineering and Deputy Director of Center for Applied Research and Development in Institute of Engineering, Pulchowk Campus, Tribhuvan University. He completed his Ph.D. from Wayne State University and Masters from AIT.*

Abstract:

With an increasing trend of publishing news online on website, automatic text processing becomes more and more important. Automatic text classification has been a focus of many researchers in different languages for decades. There is huge amount of research repository on features of English language and their uses on automated text processing. This research implements Nepali language key features for automatic text classification of Nepali news. In particular, the study on impact of Nepali language based features, which are extremely different than English language, is a more challenging because of the higher level of complexity to be resolved. The research experiment using vector space model, n-gram model and key feature based processing specific to Nepali language shows promising result compared to bag-of-words model for the task of automated Nepali news classification.

Keywords: *Document Similarity, Nepali Text Classification, Morphological analysis, Vector Space Model, Bag-of-words Model, N-gram, Bi-gram, Nepali News Classification*

I. Introduction

The availability of news in electronic form is ever increasing mainly because of the rapid growth of the World Wide Web. The task of automatic classification of such colossal news is emerging as the key challenge for organizing that huge information and also for any processing with an aim of knowledge discovery in this information. Proper classification of online news needs text level analysis using various natural language processing and machine learning techniques to get meaningful knowledge. Text classification is becoming interesting to many researchers, which consists of several challenges like appropriate document representation, dimensionality reduction to handle algorithmic issues and an appropriate classifier function to obtain good generalization. The language level specifics, which might be entirely different from one language to another language both at syntax and semantics adds other challenging issues on text classification problem.

Vector Space Model (VSM) represents each document as a feature vector of the terms in the document. Each feature vector contains term weights of the terms in the document. First step in classification system based on Vector Space Model (VSM) is the generation of words. Words are generated using analyzer. Different analyzers available for English text cannot be used for Nepali text due to the differences in the language. Preprocessing is done by analyzer before generating final set of words for VSM.

An analyzer tokenizes text by performing number of operations on it, which include extracting words, discarding punctuation, removing accents from characters (for certain languages), lowercasing, removing common-words which is also termed as stopwords, reducing words to a root form, etc. These operations are language dependent and exact implementations vary from language to language.

For an example, in English, words such as the, an, a, etc., are very common and are used as stopwords. In Nepali text, these stopwords do not exist and new stopwords are needed to be identified. It is also common to process verbs to reduce them to their roots for English text. In Nepali language, root verbs are combined in different ways to form new verbs. Similarly, different suffices are added to nouns and pronouns to form different words. There are various forms of word reduction based on suffices as per the nouns and pronouns used in a particular sentence and context. Such feature is not seen in English text. Lowercasing is common in English but is not applicable in Nepali text. Similarly, accent is used in Nepali text. Another feature common in Nepali language is different forms of words with same meaning. These words are not the synonym words, rather similar words but written in different forms with minor variations on characters being used in writing. Examples are अवरुद्ध - अबरुद्ध, उपलव्धि - उपलब्धि, कमिटि - कमिटी, इकाइ - इकाई, आइतवार - आईतवार, अवकास - अवकाश, etc. It is required to substitute such different forms with single form. N-gram is a continuous sequence of n-words from the given text. While N-gram can be used in character level, this paper uses N-gram in word level. A sentence म भात खान्छु, results in म-भात and भात-खान्छु using N-gram when N = 2. Although experiments were done using N>2 also, the paper shows results with N = 2. N-gram with N = 2 is often referred to as bi-gram model.

In addition to these all syntax level processing, this paper also investigates the effects of combining semantics based processing methods and preprocessing based on Nepali language features. Although the principle focus of this research work is in news classification, this research contribution is equally applicable in building news clustering models for news search engine optimization similar as Zhang and Dimitroff models [13] and Kresteland Mehta [8] for the purpose of designing efficient news recommendation systems.

The rest of this paper is organized as follows. Section II discusses previous works related to this paper. Section III introduces baseline approach and some other methods. It then discusses the proposed approach using preprocessing based on Nepali language features. Section IV then discusses the experimental setup for the methods. Based on the experiments, section V shows the evaluation of these methods and discusses the performances of the classifier. Finally, section V summarizes and concludes the paper.

II. Related Works

VSM represents each document as a feature vector of the terms including both words and phrases [12]. Each feature vector contains term weights of the terms in the document. In the simplest form, it does not consider the dependency between the terms and ignores the sequence and structure of the term in the document. A document is represented by n-dimensional term vectors in VSM. A term vector is a collection of term-weight pairs. The weight of a term depends on the frequency of occurrence of the term called Term Frequency (TF). Term Frequency - Inverse Document Frequency (TF-IDF) weight outperforms TF and is commonly used weights [5,6]. A wide variety of distance functions and similarity measures have been used, such as squared Euclidean distance, cosine similarity, and relative entropy. Cosine similarity is a commonly used similarity measure [3,4].

Khan et al. proposed the semantics based feature vector using Parts-of-Speech (POS) model for text classification [5]. The feature space was reduced remarkably considering POS. Masuyama and Nakagawa presented the roles of the different POS in feature selection and showed that nouns best describe a category's contents [10]. Krestel and Mehta used Latent Dirichlet Allocation (LDA) to find the hidden factors of important news stories [8]. Authors focused on verb, adjective and noun for part-of-speech (POS) information. For Named-Entity information, the most common named-entities like person, locations, organization and job-titles were used for efficient results. Lo, He and Ounis generated stopwords automatically using frequency of occurence of words for information retrieval system [9]. Basnet and Pandey have shown different aspects of morphological analysis of verbs in Nepali [1]. The study focused on inflection aspect of verb morphology and the use of compounding of verbs for Nepali language.

Latent Semantic Indexing (LSI) remained one of the predominantly used techniques in information retrieval systems because of the method featuring in dimensionality reduction and overall quality improvement in results [2,3]. LSI uses Singular Value Decomposition (SVD) to find low rank approximation of the original term space. Users in different contexts, or with different needs, knowledge, or linguistic habits will describe the same information using different terms. In contrary, same word may have more than one distinct meaning. LSI captures the essential meaningful semantic associations while reducing redundant and noisy semantic information [2]. Paulsen and Ramampiaro combined LSI with a new clustering algorithm to retrieve and cluster biomedical information using Lucene and JAMA API in Java [11]. They obtained best result when rank was reduced to few hundreds. Gleich and Zhukov used SVD to recommend related terms at varying levels of generality by varying the rank of SVD space [3]. Kafle et. al. compares classification of Nepali documents using cosine similarity and SVM along with word2vec and TF-IDF [14].

III. Methodology

The proposed method uses various Nepali languages features to reduce the dimension of feature space. These feature based improvements are cascaded in similar style as Masuyama and Nakagawa [10] nevertheless the authors used different method than the proposed method. Finally the results are verified using cross-fold validation techniques detailed out as in [7].

A. Algorithm of Baseline Method

- Split the whole documents into training and testing datasets. The splitting ratio will be different in each iterations of validation cycle following k-fold cross-validation.
- Analyze the news document using StandardAnalyzer in Lucene[1].
- Calculate the term weight using the standard TF-IDF calculation scheme.
- Using the training documents, compute the term-weight vectors, which is also known as document vectors.
- Use term-weight vectors associated with training documents of each class to compute corresponding centroid vector.
- Compute term-weight vector of test documents, exactly repeating similar steps of training documents.
- Compute similarity measure between the test document and news classes. Similarity measure between a document and a class is calculated by computing the cosine similarity between the test vector of the former and centroid vector of the latter.
- Classify the test document to a class which gives minimum similarity measure.
- Repeat whole processes to evaluate using k-fold cross-validation.

B. Algorithm of Proposed Method

- Generate stopwords using frequency thresholding where frequency is the total number of documents in which a term appears. Remove stopwords from the complete set consisting total terms.
- Replace ब with व, इ with ई, ु with ू ि withी, ढ with ध, ड with द, ठ with थ, ट with त and श with स.
- Remove word suffices such as बाट, द्‌वारा, मार्फत, देखि, सँग, तर्फ and सम्म.
- Generate terms using n-grams from each document.

[1] Lucene 3.6.1 and JDK 1.7.0 were used for processing and baseline evaluation purpose. Lucene can be downloaded from http://lucene.apache.org/

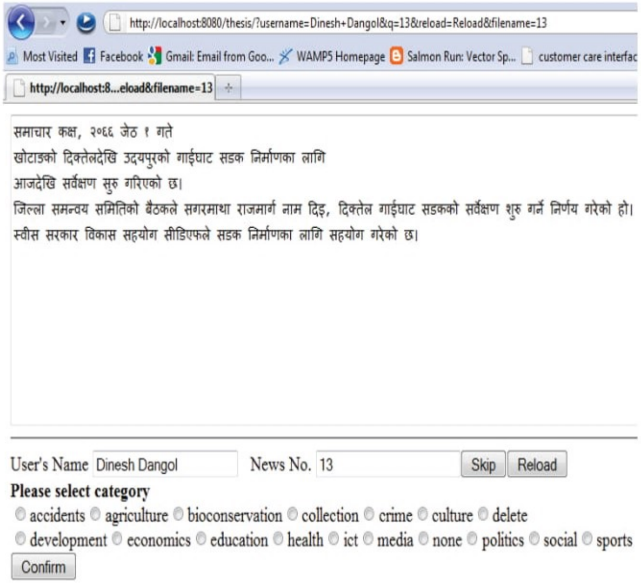
- Calculate weights of the reduced term space.
- Calculate term-document matrix from the reduced term space.
- Generate term-weight vector from the new term-document matrix.
- Evaluate the classifier using k-fold cross-validation.

IV. Experiment and Evaluation

The data required for this work was prepared from the raw dataset used for morphological analysis of Nepali text by Basnet and Pandey [1]. The original dataset consists of 71981 news documents, occupying 453 MB disk space, collected from various sources of Nepali daily newspapers such as Gorkhapatra, Kantipur, Samacharpatra, Mahanagar, Nagariknews and Annapurna Post published during January 2009 to November 2009. The dataset was used for morphological analysis of Nepali text by Basnet and Pandey [1]. Data cleaning was required before the dataset could be used. Duplicate news documents were replaced with unique news documents. The dataset containing non-Nepali encodings were also removed. Many news documents contained more than one news items in the same document, which was useless for classification, as each news document must belong to a single category.

To test the performance of the classifier, actual category of news document must be compared with the category predicted by the classifier. A web-based application was developed for this purpose. A web-interface displays news from the dataset and human user can select respective category. After selection, the provided category was stored in a database.The web-interface was developed using HTML, PHP and mysql. The database consists of two tables: "**ncategory**" and "**news**". The table "**ncategory**" stores the news categories which are loaded as options for classification. A user is provided the content of the news to be categorized and available options. The information submitted by the user is stored in the table "**news**". The information includes the news, category selected and user's name. Fig. 1 shows the screenshot of the web-interface developed for manual categorization of news documents.

After going through such rigorous process of cleaning and manual categorization, dataset of 700 news documents is prepared. Initially, 50% were used for training and 50% were used for testing. Later the document set splitting was done following k-fold cross-validation. Different categories chosen for the model experimentation are shown in Table I.



http://localhost:8080/thesis/?username=Dinesh+Dangol&q=13&reload=Reload&filename=13

http://localhost:8...eload&filename=13

समाचार कक्ष, २०६६ जेठ १ गते
खोटाङको दिक्तेलदेखि उदयपुरको गाईघाट सडक निर्माणका लागि
आजदेखि सर्वेक्षण सुरु गरिएको छ।
जिल्ला समन्वय समितिको बैठकले सगरमाथा राजमार्ग नाम दिइ, दिक्तेल गाईघाट सडकको सर्वेक्षण शुरु गर्ने निर्णय गरेको हो।
स्वीस सरकार विकास सहयोग सीडिएफले सडक निर्माणका लागि सहयोग गरेको छ।

User's Name Dinesh Dangol News No. 13

Please select category

○ accidents ○ agriculture ○ bioconservation ○ collection ○ crime ○ culture ○ delete
○ development ○ economics ○ education ○ health ○ ict ○ media ○ none ○ politics ○ social ○ sports

Fig. 1 Screenshot of web interface used for manual classification

TABLE I

Categories of the News Documents

No.	Category	No. of documents
1	Accidents (A)	100
2	Politics (P)	100
3	Sports (S)	100
4	Development (D)	100
5	Health (H)	100
6	Education (E)	100
7	Crime (C)	100
	Total	**700**

Confusion matrix was created from the result of classification. Precision, recall and accuracy were used for evaluation of the classifier. Precision of a class measures the ratio of correctly classified documents to the total documents classified in the class. Recall of a class measures the ratio of correctly classified documents to the total documents in a class. Table II shows the confusion matrix of one of the experiments using baseline method. Table III shows the calculation of precision, recall and accuracy using True Positives (TP), False Negatives (FN), False Positives (FP) and True Negatives (TN) derived from confusion matrix of Table II.

TABLE II
Confusion Matrix of an Experiment on Baseline Method

P		**Predicted class**						
		A	**D**	**C**	**S**	**H**	**E**	
Actual class	**P**	93	2	2	0	0	2	1
	A	0	83	5	4	0	0	8
	D	0	1	70	16	8	1	4
	C	0	0	13	70	4	1	12
	S	0	0	9	10	71	6	4
	H	0	0	8	2	3	83	4
	E	0	2	6	22	3	0	67

TABLE III
Performance Measures of an Experiment on Baseline Method

Class	**TP**	**FN**	**FP**	**TN**	**Pr**	**Re**	**Acc**
P	93	7	0	600	1	0.93	0.99
A	83	17	5	595	0.94	0.83	0.96
D	70	30	43	557	0.61	0.7	0.89
C	70	30	54	546	0.56	0.7	0.88
S	71	29	18	582	0.79	0.71	0.93
H	83	17	10	590	0.89	0.83	0.96
E	67	33	33	567	0.67	0.67	0.90
Avg.	**76.71**	**23.28**	**23.28**	**576.71**	**0.78**	**0.76**	**0.93**

Average precision for 7 classes were calculated for each experiment using 2-fold cross-validation. Experiments were repeated 20 times and average values were recorded. Baseline method uses standard analyzer to generate terms. First step was to filter commonly occurring words. Stopwords were generated using frequency thresholding. Top two most frequent terms 5 and / were selected as stopwords. Second step was to replace various word forms. Different forms of replacements and corresponding number of terms reduced are shown in Table IV. After this, suffices that occur in nouns and pronouns were removed to reduce the dimension further. Suffices that gave best results are बाट, द्वारा, मार्फत, देखि, सँग, तर्फ and सम्म.

Table V shows the effect of addition of new methods applied to the baseline method on performance measures These three methods remarkably reduced the dimension of the feature to be used for classification and also increased average precision and recall compared to the baseline method as shown in Fig. 2.

TABLE IV
Effect of Word Replacements on Dimension Reduction

Word replacement forms	**No. of terms reduced**
बव	233
इई	51
	228
ि ा	134
ढध	3
डद	7
ठथ	4
टत	11
शस	82
All combined	**746**

Fig. 3 shows the comparison of average accuracy percentage among different methods.

TABLE V
Effect of Preprocessing on Classifier Performance

Code	**Methods**	**Average precision**	**Average recall**	**Average accurary**	**No. of terms**
B	Baseline	0.7863	0. 7699	93.42	28903
BS	Baseline with stopword filtering	0.7878	0.7773	93.63	28901
BSR	Baseline with stopword filtering and word replacement	0.7905	0.7800	93.71	28157
BSRS	Baseline with stopwords filtering, word replacement and suffices removal	0.7902	0.7803	93.72	27292
BSRNN	Baseline with stopwords filtering, word replacement, suffices removal and bi-gram	0.8176	0.8005	95.81	27292

V. Conclusion and Future Works

The language features are always important to be considered in text analytics. This paper presented the Nepali language specific features which are different than English language features. Theresults with feature based processing are quite promising in comparison to baseline methods. Both stopword filtering and word replacement are key processing to be included for the increment in precision and recall. The reduction in number of terms was also observed. Morphological analyzer that removed specific suffices increased both the average precision and average recall compared to the baseline method. The results were further improved by using n-gram based word selection.

Only suffices that occur in nouns and pronouns were only considered in this paper. Stemming algorithms in Nepali may be used to reduce different forms of verbs. Only common forms of word replacements were considered in this paper. Other forms of replacements are still possible. Examples of replaceable word pairs are ओखलढुंगा - ओखलढुङ्गा, एवं - एवम्, काठमाडौँ - काठमाडौं, इन्ष्टिच्युट - इन्स्टिच्युत, etc. More replacements may improve dimension reduction and classifier performance further.

References

[1] S. B. Basnet and S. B. Pandey, “Morphological Analysis of Verbs in Nepali”, *Nepalese Linguistics 24*, 2009, pp. 21-30.

[2] S. Deerwester, S.T. Dumais, G. W. Furnas, T. K. Landauer and R.Harshman, “Indexing by Latent Semantic Indexing”, *Journal of the American Society of Information Science*, 1990.

[3] D. Gleich and L. Zhukov, “SVD Subspace Projections for Term Suggestion Ranking and Clustering”, *SIGIR-2004*, 2004.

[4] A. Huang, “Similarity Measures for Text Document Clustering”, *New Zealand Computer Science Student Conference 2008*, 2008, pp. 49-56.

[5] A. Khan, B. Baharudin and K. Khan, “Semantic Based Features Selection and Weighting Method for Text Classification”, *IEEE, ITSIM*, 2010, pp. 850-855.

[6] A. Khan, B. Baharudin and K. Khan, “Efficient Feature Selection and Domain Relevance Term Weighting Method for Document Classification”, *ICCEA-2010*, 2010.

[7] R. Kohavi, “A Study of Cross-validation and Bootstrap for Accuracy Estimation and Model Selection”, *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, 1995.

[8] R. Kresteland B. Mehta, “Learning the Importance of Latent Topics to Discover Highly Influential News Items”, *33rd Annual German Conference on AI* 2010, pp. 221-218.

[9] R. T. Lo, B. He and L. Ounis, “Automatically Building a Stopword List for an Information Retrieval System”, *5th Dutch-Belgium Information Retrieval Workshop*, 2005.

[10] T. Masuyama and H. Nakagawa, "Cascaded Feature Selection in SVMs Text Categorization", *CICLing 2003*, 2003, pp. 588-591.

[11] J. R. Paulsen and H. Ramampiaro, “Combining Latent Semantic Indexing and Clustering to Retrieve and Cluster Biomedical Information: A 2-step Approach”, *NIK-2009 Conference*, 2009.

[12] P. D. Turneyand P. Pantel, “From Frequency to Meaning: Vector Space Models of Semantics”, *Journal of Artificial Intelligence Research 37*, 2010.

[13] J. Zhang and A. Dimitroff, "The Impact of Webpage Content Characteristics on Webpage Visibility in Search Engine Results", *Information Processing & Management*, Vol. 41, Issue 3 2005, pp. 665-690.

[14] K. Kafle, D. Sharma, A. Subedi and A. K. Timalsina Improving Nepali Document Classification by Neural Network

* * *