Classification of Human Loss News

Ansar Abbas[1], Uzma Farooq[1], Adnan Abid[1]
{15026050006;uzma.farooq; adnan.abid}@umt.edu.pk

***Abstract*—Text classification is an important research area, and researchers have worked in many different aspects of this research area. In this work we aim to classify news articles related to human loss into different categories. The principal contribution of this work is the provision of different variants of Naïve Bayes classification algorithm that help categorizing the news articles into different defined categories. We have also presented a data set for the evaluation of this work which has been generated by a web crawler which extracts the news from the leading daily newspapers of Pakistan. The results show a promising accuracy of 89% correct classification.**

***Keywords*— Text Classification; Naïve Bayes; News Classification; Crime; Accident; Disaster; Terrorism; Stemming; Human Loss News;**

I. Introduction

News is not only about gathering and reporting facts and figures; it is in fact information that impacts our society. News affects subjects of a society in a way that how they perform their duties and make their decisions. Information gathered from news broadcasts also affects their choices.

Among all the news, crime and order news are of great importance for a civilized society, no matter how big or small these are. Avoiding human loss due to crime or disaster is a big challenge for any society and it cannot be tolerated at any level. Timely reporting and analysis for different patterns of such events and then forward out this information to take decisions is very crucial to the concerned field formations so that they can take appropriate measures to avoid such incidents in future. This information must be updated and readily available on daily basis so that ill-fated decisions and measures could be avoided.

Organization and management of vast volumes of electronic text information is a great challenge. Text classification could be used as an essential technique to handle this issue. Text classification is assigning predefined categories to textual data. There are lot of applications of text classification in natural language processing like prediction of user preferences, news filtering and email filtering and many more. A number of machine learning techniques have been used to classify texts like rule induction, Naïve Bayes, decision tree induction, KNN, Rocchio and SVM.

Focus of this paper is to define and implement the ways to collect crime and order related news from different news websites and classify them in different crime and order related categories so that different kinds of relevant public and private officials can make use of these reports for decision making and to improve effectiveness of their work.

To achieve above described goal we developed a model which consists of three phases. First phase was formation of data set, second phase was training of the classifier and third and final phase was classification of test data.

The rest of the articles have been presented in the following manner: next section presented the related work. Whereas, Section III presents the process of generating the data set, and discusses the generated data set as well. The proposed approach and its variants have been discussed in Section IV. We discuss the experimental evaluation of our work in Section V. Lastly, the article is concluded and some future directions have been discussed in Section VI.

II. Related Work

Text classification research is very mature and presents many classification models. Many classification standard data sets and evaluation techniques of related standards are well established [4][5][11].

Nahm [2], very few documents were manually indexed and this index was used for a large body of text for the construction of database for data mining.

Macskassy [3], key approach is use of user's specification to categorize history documents. This data set is then used to train text classifier.

III. Data Set Generation

A. News Crawler

First of all we developed a crawler in PHP that collects crime, order and disaster related news from news websites once a day and saves them in the form of text files, one file for one story. The technique used in extracting news from a web news page is as discussed in ECON [1]. Every web news page is represented as a DOM tree, then by using the features of a DOM tree, algorithm counts number of punctuations of each text node, returns big-node (parent of text-nodes) having maximum number of punctuations. This technique is very simple, robust and accurate. Its accuracy is above 93%.

B. News Source

Our crawler crawls through 10 top English news websites of Pakistan on daily basis and collects human loss related news. Names of these sources and corresponding websites are given in Table 1.

Table 1: News Sources

Source	Website
Dawn	http://www.dawn.com/
The Nation	http://nation.com.pk/
The News	http://www.thenews.com.pk/
Daily Times	http://www.dailytimes.com.pk/
The Express Tribune	http://tribune.com.pk/
Pakistan Observer	http://pakobserver.net/
Business Recorder	http://www.brecorder.com/
The Frontier Post	http://www.thefrontierpost.com/
The Friday Times	http://www.thefridaytimes.com/tft/
Pakistan Today	http://www.pakistantoday.com.pk/

C. Resultant Data Set

Our data set consisted of 62 news in English language. All news were related to Pakistan. All news were labeled manually into 5 categories. Categories and number of news falling against each category are shown in Table 2. Size of every news is between 128b and 2.5Kb.

Table 2: Categories and count of news

Category	No of News
Accident	10
Crime	30
Disaster	9
Operation	4
Terrorism	9

Simple News: Simple news is such news in which only one story is reported as shown in Fig 1.

> QUETTA: A man was shot dead by unidentified gunmen in Qambrani area of the provincial capital. As per reports, the victim, identified as Mehrab Khan, a resident of Dera Murad Jamali, was returning home when armed assailants opened fire on him and fled. As a result, he died on the spot.

Fig 1: Simple news

Complex News: During preprocessing of data set we noticed that all news are not simple rather some news are multi-story i.e. aggregation of more than one story belonging to different categories as shown in Fig 2.

> Karachi: An unidentified alleged criminal was killed in a police encounter in Lyari, area of the metropolis on Saturday. The unidentified alleged criminal, was killed in a shootout with a police party in Lyari. The body was shifted to Civil Hospital Karachi for medico-legal formalities. **Meanwhile**, some unidentified armed men shot and injured one man in Garden area of the metropolis on Monday. A man, Umair, 29, resident of Garden, was shot and injured by unidentified men in Garden. He was shifted to Civil Hospital Karachi for treatment. **Moreover**, two women, two men and a minor girl were injured after a speeding Rickshaw overturned in Sakki Hassan area here on Monday. Fareeda, 40, Shugufta, 30, Ammad, 20, Sohail, 25, and a minor girl Aisha, 3, were injured in the road accident when the Rickshaw overturned in Sakki Hassani. They were shifted to Abbasi Shaheed Hospital.

Fig 2: Complex news

Since our model strongly assumes independence of categories therefore complex news was broken up and each story was placed in a different text file.

IV. Proposed Approach

A. NAIVE BAYES CLASSIFIER

The model used for classification of news was Naïve Bayes. This is an independent feature and probabilistic model based on Bayes' theorem assuming strong independence. Naïve Bayes classifier calculates the probability of news belonging to a certain category for all categories and assigns it to that category having highest probability. There are different models on different assumptions for Naïve Bayes. Two most common models are Multivariate Bernoulli which uses binary word occurrences as Boolean weights and Multinomial which uses word occurrence frequencies. We considered Multinomial Naïve Bayes model because it is generally superior and yields more accurate results than Multivariate Bernoulli model [7][15][16][17][18][19].

Feature Selection: Since Naïve Bayes classifier uses set of feature words to calculate probabilities therefore we implemented an algorithm that extracted candidate feature set from the benchmark data set. Our algorithm removed stop words and words that occurred in all the categories. Then this set was further fine-tuned [8][9][10][20].

Stemming: We performed our experiments both ways without stemming and with stemming. We implemented Potter stemmer as our stemming algorithm because of its better performance and simplicity [12][13][14].

B. SYSTEM MODEL

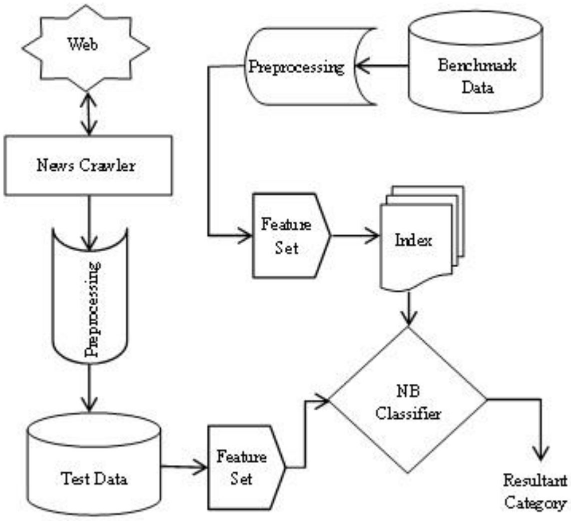

Fig 3: News text classification system model

Our news web text classification system model considers collection of news data set to be classified through web via news crawler. After collection, news text is preprocessed and placed in a repository in the form of text files and then it is filtered through feature text and passed on to Multinomial Naïve Bayes classifier, which classifies the input news by using inverted index of benchmark data. Our proposed model is shown in Fig 3.

C. Algorithm

We implemented Multinomial Naïve Bayes Classifier in Java. Algorithms for training and test are shown in Fig 4 & Fig 5.

```
TrainNB(C, D, F)
1   I ← BuildIndex (C, D, F)
2   N ← CountDocs(D)
3   for each c ∈ C
4   do N_c ← CountDocsInClass(I, c)
5      catprob[c] ← N_c / N
9      for each t ∈ I
10        do termprob[t][c] ← ComputeTermProb(t, c, I)
11  return I, catprob, termprob
```

Fig 4: Multinomial Naïve Bayes (Training)

```
TestNB(C, I, classprob, termprob, d)
W ← ExtractTokensFromDoc(I, d)
for each c ∈ C
do score[c] ← log catprob[c]
   for each t ∈ W
   do score[c] *= log termprob[t][c]
return max_{c∈C} score[c]
```

Fig 5: Multinomial Naïve Bayse (Test)

First of all it builds an inverted index to represent vector space representation [6] of training data according to feature set then it computes category probabilities. In next step it computes each term's probability within each category using the function *ComputeTermProb* at line 10 of Fig 4.

In testing phase, Fig 5, firstly given news is tokenized according to feature set then logs of pre computed probabilities are multiplied for terms of given news for each class. Category with maximum score is returned as resultant category of given news.

V. EVALUATON AND RESULTS

A. Experimental Setup

The proposed algorithm has been evaluated on the aforementioned data set. Naïve Bayes algorithm was implemented using two variants, where one variant involves *stemming*, whereas the other does not incorporate *stemming*. We discuss them with the notions of *Naïve Bayes without Stemming* and *Naïve Bayes with Stemming*.

Furthermore, we tested both variants in 3 different rounds. In each round 1/3 news of each class were tested against 2/3 trained news of that class. In every round 'tested news' of previous round were swapped with 'trained news' of previous round keeping in consideration that overall ratio of tested vs. trained remains unchanged for each category and whole data set. Thus each news was tested for its classification.

B. Evaluation Metrics:

Accuracy of classification was considered as a primary measure to evaluate the proposed approach and compare the effectiveness of its variants. Apart from this, it is evident that the problem at hand is a multinomial classification, therefore, we have incorporated confusion matrix as another evaluation measure to visualize the correct and wrongly classified news. It is pertinent to mention here that we managed to refine our feature set with the help of this confusion matrix to improve the accuracy of the proposed approaches.

C. Results

In most of the cases Naïve Bayes classifier assigns correct category as proved from confusion matrix. Sometimes incorrect category is assigned due to feature words that occur in more than one category. This situation is unavoidable as naturally categories in our experiment are very close to each other especially crime verses operation. To find completely disjoint sets of feature words for every category in our problem is almost impossible.

Naïve Bayes without Stemming: It is clear from the accuracy graph presented in Fig. 6 that overall accuracy of this variant is almost 66%, while the confusion matrix presented in Table 3 reveals that many of the news have been classified into irrelevant class heads.

Table 3: Confusion Matrix (without stemming)

	Accident	Crime	Disaster	Operation	Terrorism
Accident	5	1	1	1	2
Crime	1	20		9	
Disaster		2	7		
Operation		1		3	
Terrorism	1	1		1	6

Naïve Bayes with Stemming: Interesting, the results after incorporating stemming in the preprocessing were much better than those of without stemming. The results showed in Figure 6 show that overall accuracy of the system increase from 66% to 89%. While at the same time a comparison of the confusion matrices presented in Table 3 and 4 reveals that wrong classified news has been reduced and stemming has resulted in fixing the wrong reported results.

Table 4: Confusion Matrix (with stemming)

	Accident	Crime	Disaster	Operation	Terrorism
Accident	10				
Crime		28		2	
Disaster			9		
Operation		2		2	
Terrorism	2			1	6

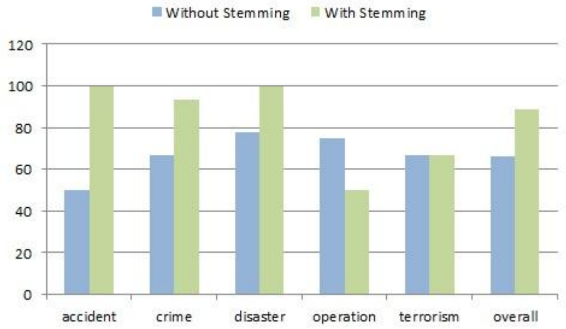

Fig 6: Comparison of accuracy

VI. Conclusion and Future Directions

In this work we have presented an approach to classify news related to human loss into different classes. We have used Naïve Bayes classification algorithm for this purpose, and implemented its two different variants. The results reveal that the variant which involves stemming of terms exhibits much better results as compared to the one without stemming. Another contribution of this work involves generation of a data set of such news, to this end, we have written a crawler to extract news from leading Pakistani daily newspapers.

In future, we intend to improve this work in both dimensions, i.e. in terms of generating a larger data set, and in terms of improving the classification accuracy. We also intend to increase the scope of our research by involving other Meta information e.g. the geographical locations of the news, which will help mitigating problems in these different domains in a pro-active manner.

References

[1] Guo, Yan, et al. "ECON: an approach to extract content from web news page." *Web Conference (APWEB), 2010 12th International Asia-Pacific*. IEEE, 2010.

[2] Nahm, Un Yong, and Raymond J. Mooney. "Text mining with information extraction." *AAAI 2002 Spring Symposium on Mining Answers from Texts and Knowledge Bases*. Vol. 1. 2002.

[3] Macskassy, Sofus A., and Foster Provost. "Intelligent information triage."Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2001.

[4] Yang, Yiming, and Xin Liu. "A re-examination of text categorization methods." Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 1999.

[5] Lewis, David D., et al. "Training algorithms for linear text classifiers."Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 1996.

[6] Salton, Gerard, Anita Wong, and Chung-Shu Yang. "A vector space model for automatic indexing." *Communications of the ACM* 18.11 (1975): 613-620.

[7] Rish, Irina. "An empirical study of the naive Bayes classifier." *IJCAI 2001 workshop on empirical methods in artificial intelligence*. Vol. 3. No. 22. IBM New York, 2001.

[8] Yang, Yiming, and Jan O. Pedersen. "A comparative study on feature selection in text categorization." *ICML*. Vol. 97. 1997.

[9] Koller, Daphne, and Mehran Sahami. "Toward optimal feature selection." (1996).

[10] How, Bong Chih, and Kulathuramaiyer Narayanan. "An empirical study of feature selection for text categorization based on term weightage."*Proceedings of the 2004 IEEE/WIC/ACM International Conference on Web Intelligence*. IEEE Computer Society, 2004.

[11] Pope, Mark W. "Automatic classification of online news headlines." *A Master's paper, University of North Carolina at Chapel Hill* (2007).

[12] Ramasubramanian, C., and R. Ramya. "Effective pre-processing activities in text mining using improved porter's stemming algorithm." *International Journal of Advanced Research in Computer and Communication Engineering*2.12 (2013): 2278-1021.

[13] Moral, Cristian, et al. "A survey of stemming algorithms in information retrieval." *Information Research: An International Electronic Journal* 19.1 (2014): n1.

[14] Smirnov, Ilia. "Overview of stemming algorithms." *Mechanical Translation* 52 (2008).

[15] Kim, Sang-Bum, et al. "Some effective techniques for naive bayes text classification." *IEEE transactions on knowledge and data engineering* 18.11 (2006): 1457-1466.

[16] Ting, S. L., W. H. Ip, and Albert HC Tsang. "Is Naive Bayes a good classifier for document classification?." *International Journal of Software Engineering and Its Applications* 5.3 (2011): 37-46.

[17] Metsis, Vangelis, Ion Androutsopoulos, and Georgios Paliouras. "Spam filtering with naive bayes-which naive bayes?." *CEAS*. 2006.

[18] Rennie, Jason D., et al. "Tackling the poor assumptions of naive bayes text classifiers." *ICML*. Vol. 3. 2003.

[19] Rennie, Jason DM. *Improving multi-class text classification with naive Bayes*. Diss. Massachusetts Institute of Technology, 2001.

[20] Zheng, Zhaohui, Xiaoyun Wu, and Rohini Srihari. "Feature selection for text categorization on imbalanced data." *ACM Sigkdd Explorations Newsletter*6.1 (2004): 80-89.