International Journal For Advanced Research In Science & Technology

A peer reviewed international journal www.ijarst.in

ISSN: 2457-0362

SMS Classification Method for Disaster Response using Naïve Bayes Algorithm

[1]Dr.R.SATHEESKUMAR, [2]MONDEDDU NAVYA REDDY, [3]SHAIK MOHAMMADAFRID, [4]SEELAM JDEEPTHI, [5]KAKANI DHATA SAI SANDEEP

[1,2,3,4,5]Assistant professors, Department of CSE in Narasaraopet Institute Of Technology

ABSTRACT

The bulk of SMS that the Quick Response Team and Rescue Agencies received during disasters made it hard for them to categorize responses based on priorities. This paper provides a method that classifies SMS received by the agency as Spam, Invalid, Alert 1 Alert 2, and Alert 3. This method allows proper response to be extended to those asking for it based on prevailing needs. This also provides a chance to ignore insignificant messages and save precious time that may be incurred by merely dealing with unimportant messages. The implementation of Naïve Bayes Algorithm, a self-learning algorithm, and together with Natural Language Processing was utilized in this research. Extension of the method is however devised in order to cover the irregularity of the data to process. Test results of the classification method showed success in its implementation and since it is a self-learning process, the method gets better and became more accurate through time.

1. Introduction

Mining information from natural languages sent through text message using mobile devices could be of significance during a disaster and crisis management. With the creation of the different government agencies [3] to respond and mitigate crises and disasters brought by climate change, risk reduction and response became accessible and readily available. However, these agencies rely heavily on only information provided and also through requests for their intervention during disasters. It would be of great help to the agency if they can initiate the intervention from their position. With the introduction of wireless communication, where almost everybody from all walks of life is hooked to this technology, the posts, messages, and conversations that they contribute to this technology are filled with unharnessed information. If proper technology is utilized, the unstructured information in the form of SMS could be of something significant. However, in-depth processes using technology must be devised in order to turn this information into something useful. Data Mining Technology provides a platform for information extraction from unstructured information to solicit useful data. The significance of this technology including the performance of popular algorithms associated with it was scrutinized in the study of Patil [1]. The same research also covers the capability of the Naïve Bayes Algorithm which is part of this study. This study is focused on classifying SMS information sent through mobile devices that are coursed through authorized agencies to determine the authenticity of the message and its prevailing needs. The primary intention is to provide relevant and correct action towards a valid request and to set aside spam and insignificant messages as well. This was made possible through the use of basic NLP processes and Naïve Bayes Algorithm. The expected system has the capability to classify SMS messages as Spam, Invalid, Alert 1 (does not need immediate attention), Alert 2 (requires attention within the day) and Alert 3 (requires immediate attention ASAP).

2. LITERATURE SURVEY:

Toman, et.al., "Influence of Word Normalization on Text Classification", University of West Bohemia, Faculty of Applied Sciences, Plzen, Czech Republic.

In this paper we focus our attention on the comparison of various lemmatization and stemming algorithms, which are often used in nature language processing (NLP). We describe the algorithm in detail and compare it with other widely used algorithms for word normalization on two different corpora. We present promising results obtained by our EWN-based lemmatization approach in comparison to other techniques. We also discuss the influence of the word normalization on classification task in general.Office of the Civil Defense Website, Mandate, Mission, Vision, and Objective m by providing leadership in the continuous development of strategic and systematic approaches as well as measures to reduce the vulnerabilities and risks to hazards and manage the consequences of disasters.

Stuart E. Middleton, Lee Middleton, and Stefano Modafferi, "RealTime Crisis Mapping of Natural Disasters Using Social Media", 2014, University of Southampton IT Innovation Centre.

The proposed social media crisis mapping platform for natural disaster s http://ocd.gov.ph/index.php/aboutocd/mandate-missionand-vision.

The Office of Civil Defense (OCD), as the implementing arm of the National Disaster Risk Reduction and Management Council, shall have the primary mission of administering a comprehensive national civil defense and disaster risk reduction and management progress uses locations from gazetteer, street map, and volunteered geographic information (VG Michal I) sources for areas at risk of disaster and matches them to geoparser real-time tweet data streams. The authors use statistical analysis to generate real-time crisis maps. Geoparsing results are benchmarked against existing published work and evaluated across multilingual datasets. Two case studies compare five-day tweet crisis maps to official post-event impact assessment from the US National Geospatial Agency (NGA), compiled from verified satellite and aerial imagery sources.

3. SYSTEM ANALYSIS

EXISTING SYSTEM:

The study of Patil which is focused on the performance analysis of Naïve Bayes and J48 classification algorithm showed an accuracy of both in classifying specimens. Naïve Bayes being probabilistic in nature is more appropriate to use in this study. The paper which is focused on normalization of text provided a pre-processing technique that this study considered employing especially in data cleansing and preparation prior to the text classification process. Stop words are first removed prior to normalization which converts words into its basic forms. Mobile SMS which serves as the input of the operation goes through a process that sanitizes the input from unnecessary words and ambiguous terms. The data undergo stop-words removal and normalization. The sanitized data will serve as the input to the next process which is the actual classification process. The result is the classified text message

Disadvantages:

Unable to predict the disaster

no accuracy in time

PROPOSED SYSTEM:

The Bayesian classifier in this research is used to represent the probability distribution of identified text messages of SMS. This method was formulated to solve the classification problem associated with this research. Considering the independence of the tokens among each other and also with respect to the output, the Naïve Bayesian Classifier is the most appropriate method that can efficiently do the job. If the input is considered to be a set of attributes, like words, then using a Bayesian network, we can calculate the probability of whether a message belongs to a specific class or not The process is composed of two phases: the training phase and the classification phase. In the training phase, the filter is trained using a known collection of words for classified messages. A database of tokens appearing in each corpus and their total occurrences are maintained in a database. Based on their occurrences in each set of classified messages, each token is assigned a probability for its capacity to determine a message of its classification. Then, using this collection of tokens, the filter classifies every new incoming message in the classification phase. Once the status of a new message is confirmed, all its tokens are also recorded, thus updating the database. This self-learning function of the filter makes it unique among the other available message filters.

Text Classification Process includes Natural Language Processing technique which eliminates linguistic mistakes (language flaws), noise (unnecessary words), stop words (words with little significance to the semantics) and repeated words and letters in the message. The remaining items are considered as tokens that are part of the Bag of Words (Corpus) y (message) = c. The bag of words (c) is a subset of the message clean from linguistic mistakes, noise, stop words, and repetitions. Raw messages need to be subjected to NLP to eliminate unnecessary words. The generated bag of words will then be used to create training set y which is a learned classifier needed for the actual classification process. The learned classifier will be the basis for determining the classification of the text messages using the Naïve Bayes Classifier. Since Naïve Bayes is a machine learning algorithm, the authenticity and veracity of the training set would be essential for the success of its succeeding operations. The method has two phases. The first phase is the training of the learned classifier. Pre-processed classified data are first fed in the system allow the learning process of the method to take it phase. The higher the number of pre-processed classified SMS processed by the system during the learning process would provide a more accurate implementation of the method. The second phase is the actual classification method of the raw SMS received.

Advantages:

predict the disaster

accuracy in time

4. IMPLEMENTATION

The process starts with the training of the Learned Classifier by feeding the process with messages previously classified (Spam, Invalid, Alert1, Alert2, Alert3). The more messages involved, the better the machine will become. More accurate results will be derived from a well-trained system.

The training process or the creation of the learned classifier is illustrated in Fig. 1. Classified messages are subjected to Pre-Processing for cleansing of the unnecessary words like "Stop Words", Noise, and repetitions. The words are then included in the database of the learned classifier which will eventually become part of the classifier's knowledge.

The processed messages presented are eventually added to the Learned Classifier and which will be used in the further classification process. The success of the Naïve Bayes Technique is dependent on the training method conducted in the system. After the process, messages m1 to mn will become classes of c1 to cn of the learned classifier. There are five (5) probabilities P(c|m) that needs to be determined, the probability that the message is Spam, Invalid, Alert1, Alert2, or Alert3. Equation is used to determine the probability of each classification. The classification of the message is determined using the decision criteria.

INPUT DESIGN

The input design is the link between the information system and the user. It comprises the developing specification and procedures for data preparation and those steps are necessary to put transaction data in to a usable form for processing can be achieved by inspecting the computer to read data from a written or printed document or it can occur by having people keying the data directly into the system. The design of input focuses on controlling the amount of input required, controlling the errors, avoiding delay, avoiding extra steps and keeping the process simple. The input is designed in such a way so that it provides security and ease of use with retaining the privacy. Input Design considered the following things:

➢ What data should be given as input?

➢ How the data should be arranged or coded?

➢ The dialog to guide the operating personnel in providing input.

➢ Methods for preparing input validations and steps to follow when error occur.

OBJECTIVES

1. Input Design is the process of converting a user-oriented description of the input into a computer-based system. This design is important to avoid errors in the data input process and show the correct direction to the management for getting correct information from the computerized system.

2. It is achieved by creating user-friendly screens for the data entry to handle large volume of data. The goal of designing input is to make data entry easier and to be free from errors. The data entry screen is designed in such a way that all the data manipulates can be performed. It also provides record viewing facilities.

3. When the data is entered it will check for its validity. Data can be entered with the help of screens. Appropriate messages are provided as when needed so that the user will not be in maize of instant. Thus the objective of input design is to create an input layout that is easy to follow

OUTPUT DESIGN

A quality output is one, which meets the requirements of the end user and presents the information clearly. In any system results of processing are communicated to the users and to other system through outputs. In output design it is determined how the information is to be displaced for immediate need and also the hard copy output. It is the most important and direct source information to the user. Efficient and intelligent output design improves the system's relationship to help user decision-making.

1. Designing computer output should proceed in an organized, well thought out manner; the right output must be developed while ensuring that each output element is designed so that people will find the system can use easily and effectively. When analysis design computer output, they should Identify the specific output that is needed to meet the requirements.

2. Select methods for presenting information.

3. Create document, report, or other formats that contain information produced by the system.

The output form of an information system should accomplish one or more of the following objectives.

❖ Convey information about past activities, current status or projections of the

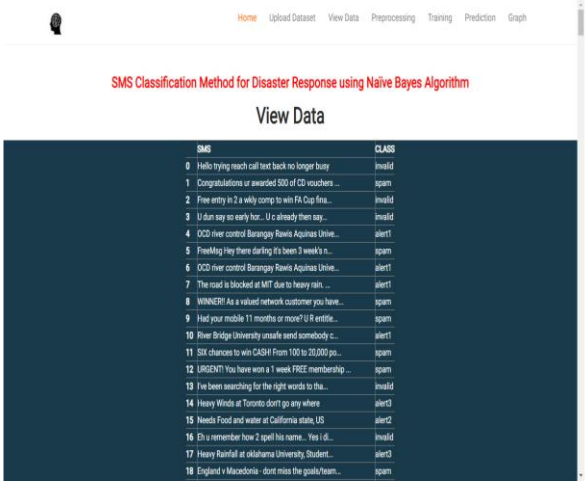
❖ Future.

❖ Signal important events, opportunities, problems, or warnings.

❖ Trigger an action.

❖ Confirm an action.

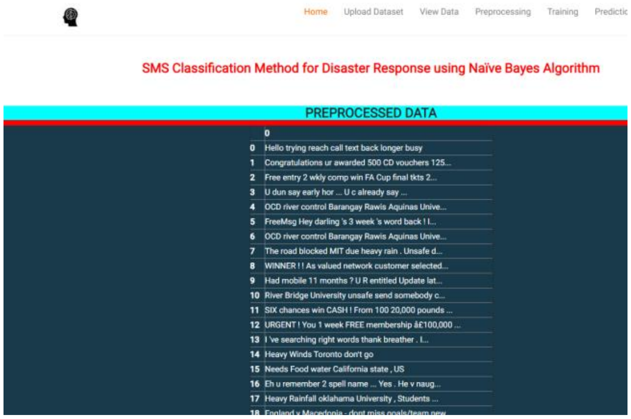
5. Results:



SMS Classification Method for Disaster Response using Naïve Bayes Algorithm

View Data

	SMS	CLASS
0	Hello trying reach call text back no longer busy	invalid
1	Congratulations ur awarded 500 of CD vouchers ...	spam
2	Free entry in 2 a wkly comp to win FA Cup fina...	invalid
3	U dun say so early hor... U c already then say...	invalid
4	OCD river control Barangay Rawis Aquinas Unive...	alert1
5	FreeMsg Hey there darling it's been 3 week's n...	spam
6	OCD river control Barangay Rawis Aquinas Unive...	alert1
7	The road is blocked at MIT due to heavy rain. ...	alert1
8	WINNER!! As a valued network customer you have...	spam
9	Had your mobile 11 months or more? U R entitle...	spam
10	River Bridge University unsafe send somebody c...	alert1
11	SIX chances to win CASH! From 100 to 20,000 po...	spam
12	URGENT! You have won a 1 week FREE membership ...	spam
13	I've been searching for the right words to tha...	invalid
14	Heavy Winds at Toronto don't go any where	alert3
15	Needs Food and water at California state, US	alert2
16	Eh u remember how 2 spell his name... Yes i di...	invalid
17	Heavy Rainfall at oklahama University, Student...	alert3
18	England v Macedonia - dont miss the goals/team...	spam



SMS Classification Method for Disaster Response using Naïve Bayes Algorithm

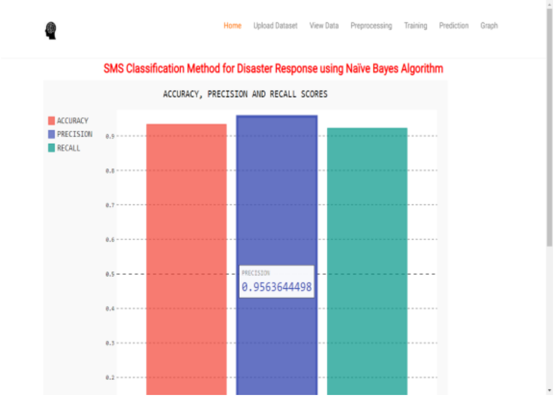
PREPROCESSED DATA

	0
0	Hello trying reach call text back longer busy
1	Congratulations ur awarded 500 CD vouchers 125...
2	Free entry 2 wkly comp win FA Cup final tkts 2...
3	U dun say early hor ... U c already say ...
4	OCD river control Barangay Rawis Aquinas Unive...
5	FreeMsg Hey darling 's 3 week 's word back ! I...
6	OCD river control Barangay Rawis Aquinas Unive...
7	The road blocked MIT due heavy rain . Unsafe d...
8	WINNER ! ! As valued network customer selected...
9	Had mobile 11 months ? U R entitled Update lat...
10	River Bridge University unsafe send somebody c...
11	SIX chances win CASH ! From 100 20,000 pounds ...
12	URGENT ! You 1 week FREE membership â£100,000 ...
13	I 've searching right words thank breather . I...
14	Heavy Winds Toronto don't go
15	Needs Food water California state , US
16	Eh u remember 2 spell name ... Yes . He v naug...
17	Heavy Rainfall oklahama University , Students ...
18	England v Macedonia - dont miss goals/team new...

Model Training:

Performance Metrics:

Graph:



6. CONCLUSION:

This study proved the capability of the extended Naïve Bayes Formula to classify SMS messages according to five different classifications based on the collection of preclassified information used as the learned classifier. The study generally correctly classified test data according to specific classifications up to 89% accuracy. The 11% that falls within the False-Negative result is attributed to the number of entries in the dataset that served as the learned classifier of the Naïve Bayes Algorithm. Clearly, there is a necessity to increase the number of pre-classified entries in the learned classifier to further improve the capability of the method to correctly classify SMS inputs. However, since the method is self-learning, it recommended that the system be used in actual operation to meet the required high accuracy classification capability of the process.

REFERENCES :

[1] Tina R. Patil, et.al. “Performance Analysis of Naive Bayes and J48Classification Algorithm for Data Classification” International Journal Of Computer Science And Applications Vol. 6, No.2, Apr 2013 ISSN: 0974-1011 (Open Access)

[2] Michal Toman, et.al., “Influence of Word Normalization on Text Classification”, University of West Bohemia, Faculty of Applied Sciences, Plzen, Czech Republic

[3] Office of the Civil Defense Website, Mandate, Mission, Vision, and Objectives http://ocd.gov.ph/index.php/about-ocd/mandate-missionand-vision

[4] Stuart E. Middleton, Lee Middleton, and Stefano Modafferi, “RealTime Crisis Mapping of Natural Disasters Using Social Media”, 2014, University of Southampton IT Innovation Centre

[5] Nikhil Dhavase, “Location Identification for Crime & Disaster Events by Geoparsing Twitter”, 2014, (M.E. Student)Dept. of Information TechnologyPune Institute of Computer

Technology Pune, India, International Conference for Convergence of Technology – 2014 978-1-4799-3759-2/14/$31.00©2014 IEEE1

[6] Anubrata Dasכ,Neeratyoy Mallikכ ,Somprakash Bandyopadhyay†, Sipra Das Bitכ ,Jayanta Basa, “Interactive Information Crowdsourcing for Disaster Management Using SMS and Twitter: A ResearchPrototype”

[7] Eirini Takoulidou, Vilelmini Sosoni, Katia Kermanidis, 2016, Department of Informatics, “Social Media and NLP tasks: Challenges in Crowdsourcing Linguistic Information”

[8] Tanzim Mahmud, K. M. Azharul Hasan, Mahtab Ahmed, Thwoi Hla Ching Chak, "A Rule-Based Approach for NLP Based Query Processing"Proceedings of International Conference on Electrical Information and Communication Technology (EICT 2015)

[9] Monisha Kanakarajand Ram Mohana Reddy Guddeti† Dept. of Information Technology, National Institute of Technology Karnataka, Surathkal, Mangalore 575025 India “NLP Based Sentiment Analysis on Twitter Data Using Ensemble Classifiers”, 2015 3rd International Conference on Signal Processing, Communication and Networking (ICSCN)