Detection of Natural Calamities from Assamese Posts in Social Media

S Kalita¹, R R Deka², M P Bhuyan³, K Kashyap⁴, S K Sarma⁵

Department of Information Technology, Gauhati University, Assam, India.

¹itsme.simantakalita@gmail.com, ²rrd0794@gmail.com, ³mpratim250@gmail.com, ⁴kb.guwahati@gmail.com, ⁵sks001@gmail.com

Abstract. Social media users and online news portals are rising exponentially in the North-eastern state of India, where every small instance of daily life are posted on social media platforms such as Facebook, Twitter by the users in their native language. Every social media user post about their daily experiences on such kind of social media platform which gives explicit information happening in a particular place. From those posts, extracting of information has been tried regarding natural calamities, a detection system which detects real world happening of natural calamities in a particular place. The objective of this paper is to detect such events from Assamese text posted on social media. Suppose an earthquake has happened, the online news portals, social media users started reporting about the happening in various platforms, just by observing the post an earthquake can be detected easily. A powerful statistical model Conditional Random Field (CRF) is used to detect those natural calamities as events from the posts which are being posted using the Assamese language. This model has the objective to capture the real-world happenings while a goal have been achieved by introducing an event extraction rule. The CRF model is trained with a large dataset NCED20 which is develop manually. The model is trained on a set of features and selecting those features is a significant step in the learning process. In this paper, an algorithm to capture events from the social media post has been proposed.

Keywords: Natural calamities detection, Event detection, Assamese, Social media, Conditional Random Field, CRF.

1. Introduction

Microblogging and social media services are considered as the views of new generation, in such a blog it allows users to give their views in terms of the text on the Internet. The importance of such services is that they are real-time [5] in nature, where updates of particular events posted very frequently. The advantage of using these blogs or posts is that those are real-time and numerous posts of a similar topic are generated at the same time by the social media users. To detect events from Facebook posts we have basically used Assamese news portals reports, posted on Facebook to find useful text. As
many as tens of Assamese news pages are available on Facebook that generates thousands of news-post on real-time basis. There are a huge number of Assamese news channel pages as well as independent news portals available. Because the Assamese is widely used language in the Northeastern part of India. The neighboring states of Assam such as Arunachal, Nagaland, Meghalaya uses Assamese [31], and that is why Assamese language is widely understandable in many parts of Northeast India which makes it easier for the researchers to use Assamese text and implement the model of event detection which can impact a large population. For example "ভূমিকম্প″ ("bhumikompo") ("Earthquake"), "বানপানী" ("baanpani") ("flood"), "ভূমিস্খলন″ ("bhumiskholon") (landslide) and "ধূমুহা″ ("dhumuha") ("storm") and its synonyms are the words in Assamese which means different natural calamities that occurs in this part of region. As flood, landslide and, it is earthquake prone region, such news are reported regularly in local pages of social media. For the requirement of our model the texts are filtered, that are collected from the such social media posts which is done by pre-processing texts by removing special characters such as ‘#’, '@', ',', and '|' (full stop in Assamese) using a python program before feeding the post or the input text to the model for classification of events.

In this research, a natural calamities detection system is developed that depends upon Conditional Random Field (CRF) [8]. The model is designed to seize such events which identify the “prompt” as the event from the input texts [14] e.g. “ভূমিকম্পই জোকাবি গল মণিপুরক বাতি ১ বজাত” (Bhumikompoi jukari gol Manipurok rati 1 bojat)(“Earthquake shook Manipur at 1pm midnight”)

Where the “ভূমিকম্পই” (Bhumikompo) is the event and “মণিপুরক” (Manipur) is the location. This proposed model detects natural calamities such as Earthquake, Flood, Storm, and landslide along with its location. The problem of natural calamity detection is a classical identification problem [13]. A CRF probabilistic method has been used to detect the events from the Assamese post. The model is being trained on a large dataset for the classification of the events. The CRF is trained over a sequence of features and choosing those features is also an important step in the training process. It is a smart algorithm for observing the sequence of text and it has shown a great amount of success which includes feature detection [21], [22], [23], parts of speech tagging [9]-[11], [24] and many more. Rest of the research is managed in the following ways:-

Section 2 states the related works in the field of event detection. Section 3 describes the proposed method and methodology. In Section 4 details analysis of the dataset is discussed, Section 5 is for the explanation and discussion of the experimental results and finally, Section 6 concludes the research in this paper and proposed future work.

2. RELATED WORKS

Perol in their research paper [1] have proposed a technique to study the induced seismicity in Oklahoma, USA. With the recent development of induced seismicity in the central part of the United States, it sets up for extensive catalogues for the improvement of seismic hazard evaluation. From the last few decades, a great amount of seismic data has maximized exponentially, which had created an urgent need for efficient algorithms to unfailingly detect and track down earthquakes. The authors have made use of the current advances in artificial intelligence and impart ConvNetQuake, an exceedingly dependable convolutional neural network [25] for detecting and tracking down earthquake from a particular waveform. They have detected more than 17 times more earthquakes than previously noted by the Oklahoma Geological Survey. The algorithm is many times faster than the well-known models. Whereas, Beyreuther in their research have developed an original technique to resolve the detection and categorization difficulty of an earthquake in a single stage by using Hidden Markov Model (HMM) [18]. They have applied the Hidden Markov Modelling classifier to a much normal problem of detection and distance-based classification of minor to majorly intensified earthquakes. The authors have found overall good results with the HMM in the previously generated classification tasks and considerable results in the constant cases. They have put special importance upon the important problem of selecting the properties, which superlatively depicts the characteristics
of the different signals that are to be categorized [2]. Differently, Tonon in their research have developed a model named ArmaTweet [3], an augmentation of the Social Media Analysis (SMA) system in an association between armasuisse and Fribourg and Oxford Universities that support semantical event detection. The Social Media Analysis (SMA) developed by the R&D agency for the Swiss Armed Forces called the Armasuisse Science and Technology, helps in detecting natural disaster events, etc. by evaluating Twitter posts. The system is workable solely on a keyword that is not effective for complex events. The system devised by the authors pulls out a structured and detailed representation from the ‘tweets’ text using Natural Language Processing technology, which it puts together with DBpedia and Wordnet in an RDF knowledge graph. Thus the security analysts could illustrate precisely the events that are of interest, using SPARQL inquiries over the graph. Through experiments, the authors have found that ArmaTweet can detect many complicated events that are not detectable solely by keywords. While Feng have developed a language-independent neural network to hold both the sequence and chunk data from particular contexts and make use of them for training an event detector for many languages without encoding features manually [4]. All the previous methods were heavily reliable on knowledge-centred around language and Natural language Processing toolkits. With the new method, the experiments have shown that the method can get rough, structured and precise results for many languages. In the task of event detection of ACE 2005 English, with the method, the authors have achieved a 73.4% F-score with a mean of 3% total improvement in contrast to the state of the art model. Besides, their experimental results are competitive with Spanish and Chinese languages. In their research paper, Atefeh have provided a survey of methods and techniques for event detection from streams of Twitter. The methods try to find real-world happenings that unwrap over space and time [5]. To general media, the detection of events from Twitter streams gives new challenges and problems. Twitter contains a huge amount of incoherent and incomprehensible messages and hogwash content, which badly affect the performance of event detection. The event detection methods are given in literature label this problem by adjusting with techniques from different fields to the unlikeness of Twitter. This article categorizes these methods and techniques following the type of events, the task of detection, and detection method and examines mostly used features. Tweets are also used to detect some topics which are not events such as ‘ipads’, ‘iphone’, which were trending on Twitter. Such topic events were tried to extract from tweet using Hierarchical Dirichlet Process (HDP) [15] which is based on Term Frequency-Inverse Document Frequency TF-IDF [16]. In this research paper of event detection from Twitter [17] where the researchers have claimed to detect events along with location using CRF [8]. Another article to detect real-time events from Twitter [20] is based on incremental tf-idf [19] where term vector is generated to extract the minor events from the twitter messages.

Here Bernhard in their article have put forward a method to filter the real-time media flow by examining the gravity and seriousness of posts, drawing out facts by the use of Natural Language Processing and group posts using an unconventional event detection system. The authors have found that depending upon a social media feed corpus that was tagged, fake or failed alarms are restricted to posts that have highly vague information with small values for the rescue elements [6]. And finally, Chen and their co-researchers have presented a model named LIPED (Life Profile based Event Detection) [7] which is based on the abstraction of life profiles to predict the activeness or dynamism of events for detection of events efficiently. A cluster of events with comparable activeness layouts shares a life profile that is patterned by a Hidden Markov Model (HMM) [14]. Reviewing the burst and diverse characteristics of events, LIPED recognizes the activeness of state of events, for which, LIPED stabilizes the grouping accuracy and recollect to get better F1 scores than the rest of the most known methods processed on the official TDT1 corpus.

Conditional random field is a probabilistic approach which is so versatile such as, it has been implemented for text segmentation [8], POS tagging [10], extraction of features [21], and information extraction from research papers [22]. CRFs are just not limited to natural language processing, it is implemented also in other engineering topics such as detection of sign in images [26], as CRF learns from its previous and later input in the training phase that helps the model in predicting the output
accurately. Because of these advantages, it has been implemented in biomedical relation extraction [27], text summarization in a document [29], finding protein and gene from text [28]. CRF is also being used in the detection of multilane of a road which is implemented in cars for road assist [30].

As this research paper is focused on natural calamity detection based on Assamese text, Assamese researchers are putting efforts to process the Assamese language, and few current researches are Bhuyan et al. [33] have clustered Assamese words based on their context using the n-gram models. They have designed 733 clusters with an accuracy of around 60%. Research such as to check the correctness of sentences based on stochastic model [34]. Assamese language processing is in its initial phase there is no model to predict the next in a sentence, in this article Bhuyan et al. have developed a word prediction model using N-gram [32], and with it the prediction accuracy is calculated of different input length of automatic prediction list [35] is another development. Along with it, to check and correct words by a statistical model [36] which will help users to correctly type Assamese text if all these tools can be incorporated in a single model.

3. METHODOLOGY

This application is to predict an output vector N={N₁, N₂, N₃,……Nₚ} of the sequence of random variables after observing M=(M₁, M₂, M₃,…….Nₚ). Each of the word variable Nᵢ will be labelled at position “P”. The Nᵢ holds various information about the word in that particular position.

3.1. Conditional Random Field

Conditional Random Field (CRF) is a model of undirected graph, which calculates the conditional probability of a target output values over the target input values. The input sequence such as N=(N₁, N₂, N₃,……Nₚ) noticing a sequence M=(M₁, M₂, M₃,……Nₚ) [8], [9] and [11] is given by :

\[ P(N|M) = \frac{1}{Z} \sum_{g=1}^{G} \lambda_R f_R(N_{g-1}, N_g, M, G) \]  

Where, \( f_R(N_{g-1}, N_g, M, G) \) is a function of feature and \( \lambda_R \) is its weight which is calculated during the phase of training while the conditional probability no sum upto 1. So, a factor is calculated which normalization factor which is given as

\[ Z = \sum \exp(\lambda_R f_R(N_{g-1}, N_g, M, G)) \]  

Consideration of parameters of CRF is given by the feature \( f_R \) to calculate each weight \( \lambda_R \) that increases the probability of the training dataset NCED20 = \{N(x) | M(x)\}_{x=1, 2, 3,……X} \)

To train the CRF a log sequence is required which is given by a sequence observing another sequence

\[ L_n = \sum_{t=1}^{X} \log( P(N^{(x)}|M^{(x)}) ) - \sum_{t=1} \frac{\lambda_R^2}{2\alpha^2} \]

Where \{N(x) | M(x)\} is the value of labelled data which is used to trained the model. And \( \frac{\lambda_R^2}{2\alpha^2} \) where \( \alpha^2 \) is a factor which avoid an overfit that is why \( \frac{\lambda_R^2}{2\alpha^2} \) is calculated.

3.2. Applying the CRF for Natural Calamities detection

The CRF is used to solve the event detection problem, where a sequence of text or a post is taken into observation. An important Factor about CRF is the selection of features for event detection which plays a very crucial role in detecting the events. The combination of different labelled sets and words will have different probabilities. These features are considered as prime features for detecting the events using this CRF model. The feature list also includes prefix and postfix for each word. The
previous word and the next word are also considered as a feature. The prefix words have various information to identify the event as well as the postfix words are also important to detect the event successfully. The model calculates the conditional probability considering the feature such as weight, prefix word, and postfix words which determines the events.

\[ P(N | M) = \frac{1}{Z} \sum_{g=1}^{G} \sum_{R} \lambda_{R} f_{R}(N_{g-1}, N_{g}, M, G) \]

(4)

P(N | M) gives the conditional probabilities considering the input N sequence, observing the sequence M. The value of P(N | M) will determine whether the word is an event or not.

4. DATASET

A training dataset NCED20 is prepared from Facebook posts of news portal which were reported and similar text sentences were created for each events such as flood, earthquake, landslide and storm to extend the dataset to a sizeable number by which the CRF model can be trained smoothly for better results.

The Dataset is manually developed where it is labelled sequentially and sentences wise. The words are manually labelled against each of the words which are events such as “Earthquakex”, “floodx”, “Landslidex”, “Stormx” and “Placex” and the irrelevant word are labelled as “Otherx”. NCED20 dataset, that is labelled sentence wise by the above tags where the Earthquakex implies to “bhumigompo” earthquake, floodx implies to “banpani” flood, Landslidex implies to “bhumiskholon” landslide and stormx implies to “dhumuhua” storm and the Placex implies to a location where the natural calamity has happened.

5. RESULTS AND DISCUSSIONS

Using this CRF model for natural calamity detection using a dataset named NCED20. It is a large and versatile dataset that contains multiple events of different calamities. It has shown an average of 72.94% of accuracy. The experiment is done by creating different posts and collecting posts from news portals and social media. Where a mix posts has been created of different calamities and single calamities also while clubbing together many sentences to test the model. The sentences vary from 20 to 75.

Table 1. Experiment results of CRF

| Exp No. | No. of sentences | Total No. of events | No. of events detected | Accuracy % |
|---------|------------------|---------------------|------------------------|------------|
| 1.      | 20               | 3                   | 2                      | 66.66%     |
| 2.      | 43               | 10                  | 7                      | 70.00%     |
| 3.      | 49               | 25                  | 19                     | 76.00%     |
| 4.      | 73               | 21                  | 17                     | 80.00%     |
| 5.      | 75               | 43                  | 31                     | 72.09%     |

In the first experiment, unknown data has been used where an accuracy of 66.66% has been encountered and in the later experiments mix sentences has been used, few are known to the model and few are unknown sentences to conduct the experiment as shown in Table 1.
Figure 1. Show Accuracy against Input size graph

The graph is shown in Figure 1, shows how the model performs when the versatile input is given to conduct the experiment. The graph is stable between 70% to 72% and shows a sudden spike in 76% and 80% as mix data is given as input.

Figure 2. Show natural calamity detection of earthquake

Figure 2, shows the detection of natural calamity from the input sentence. The figure shows it has detected two Earthquakes as events and place Manipur.
Figure 3. Show natural calamity detection of flood

Figure 3, shows the detection of natural calamity from two sentences club together to detect the event flood.

Figure 4. Show natural calamity detection of multiple events

In Figure 4, multiple sentences has been taken and clubbed together to gather a versatile output which includes multiple events that are detected.

6. CONCLUSION AND FUTURE WORK

In this research work, a method has been proposed using a probabilistic model known as Conditional Random field to detect natural calamities. It is an event detection approach where every natural calamity is considered as an event. As shown in multiple experiments in this research paper to detect different calamities with an average accuracy of 72.94%. The model is tested with numerous input from different sources. This model will help in generating rapid response when current incidence will be provided from online portals to detect such events. The future scope is to define which relevant sentences to detect the events and neglect the irrelevant sentences automatically.
References

[1] Perol T, Gharbi M and Denolle M 2018 Convolutional neural network for earthquake detection and location. *Science Advances* 4(2) p.e1700578.

[2] Beyreuther M and Wassermann J 2008 Continuous earthquake detection and classification using discrete Hidden Markov Models. *Geophysical Journal International* 175(3) pp.1055-1066.

[3] Tonon A, Cudré-Mauroux P, Blaré A, Lenders V and Motik B 2017 May ArmaTweet: detecting events by semantic tweet analysis. *In European Semantic Web Conference* (pp. 138-153). Springer, Cham.

[4] Feng X, Qin, B. and Liu, T 2018 A language-independent neural network for event detection. *Science China Information Sciences* 61(9) p.092106.

[5] Atefeh F. and Khreich W 2015 A survey of techniques for event detection in twitter. *Computational Intelligence* 31(1) pp.132-164.

[6] Klein B, Castanedo F, Elejalde I, Lopez-de-Ipina D and Nespral A P 2013 Emergency event detection in twitter streams based on natural language processing. *In Ubiquitous Computing and Ambient Intelligence. Context-Awareness and Context-Driven Interaction* (pp. 239-246). Springer Cham.

[7] Chen C C, Chen M C and Chen M S 2005 August. LIPED: HMM-based life profiles for adaptive event detection. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining* (pp. 556-561).

[8] Lafferty J, McCallum A and Pereira F C 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. *University of Pennsylvania Scholarly Commons Departmental Papers (CIS) Department of Computer & Information Science*.

[9] Ekbal A, Haque R and Bandyopadhyay S 2007 December. Bengali part of speech tagging using conditional random field. In *Proceedings of seventh international symposium on natural language processing (SNLP2007)* (pp. 131-136).

[10] Silfverberg M, Ruokolainen T, Linden K and Kurimo M 2014. Part-of-speech tagging using conditional random fields: Exploiting sub-label dependencies for improved accuracy. *In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*.

[11] Khan W, Daud A, Nasir, J A, Amjad T, Arafat S, Aljohani N and Alotaibi F S 2019. Urdu part of speech tagging using conditional random fields. *Language Resources and Evaluation* 53(3) pp.331-362.

[12] Zhang C 2008. Automatic keyword extraction from documents using conditional random fields. *Journal of Computational Information Systems* 4(3) pp.1169-1180.

[13] Kalita S, Sarma S K, Bhuyan M P, and Deka V 2020. Event Detection in Assamese Text using Conditional Random Field. *In Journal of Advanced Research in Dynamic and Control System* 0, pp.1370-1375

[14] Blunsom P 2004. Hidden markov models. *Lecture notes August 15*(18-19) p.48.

[15] Teh Y W, Jordan M I, Beal M J and Blei D M 2005. Sharing clusters among related groups: Hierarchical Dirichlet processes. In *Advances in neural information processing systems* (pp. 1385-1392).

[16] Ramos J 2003 December. Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning* (Vol. 242, pp. 133-142).

[17] Wang X, Zhu F, Jiang J and Li S 2013 June. Real time event detection in twitter. In *International Conference on Web-Age Information Management* (pp. 502-513). Springer Berlin Heidelberg.

[18] Kupiec J 1992. Robust part-of-speech tagging using a hidden Markov model. *Computer speech & language*, 6(3), pp.225-242.

[19] Brants T, Chen F and Farahat A 2003 July. A system for new event detection. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 330-337).
[20] Hasan M, Orgun M A and Schwitter R 2019. Real-time event detection from the Twitter data stream using the TwitterNews+ Framework. *Information Processing & Management* 56(3) pp.1146-1165.

[21] Wang T, Li J, Diao Q, Hu W, Zhang Y and Dulong C 2006 June. Semantic event detection using conditional random fields. In *2006 Conference on Computer Vision and Pattern Recognition Workshop (CVPRW’06)* (pp. 109-109). IEEE.

[22] Peng F and McCallum A 2006. Information extraction from research papers using conditional random fields. *Information processing & management* 42(4) pp.963-979.

[23] Peng X, Cao H, Prasad R and Natarajan P 2011 September. Text extraction from video using conditional random fields. In *2011 International Conference on Document Analysis and Recognition* (pp. 1029-1033). IEEE.

[24] Patel C and Gali K 2008. Part-of-speech tagging for Gujarati using conditional random fields. In *Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages*.

[25] Hansen L K and Salamon P 1990. Neural network ensembles. *IEEE transactions on pattern analysis and machine intelligence* 12(10) pp.993-1001.

[26] Weinman J, Hanson A and McCallum A 2004 September. Sign detection in natural images with conditional random fields. In *Proceedings of the 2004 14th IEEE Signal Processing Society Workshop Machine Learning for Signal Processing* (pp. 549-558). IEEE.

[27] Bundschus M, Dejori M, Stetter M, Tresp V and Kriegel H P 2008. Extraction of semantic biomedical relations from text using conditional random fields. *BMC bioinformatics* 9(1) p.207.

[28] McDonald R. and Pereira F 2005. Identifying gene and protein mentions in text using conditional random fields. *BMC bioinformatics* 6(S1) p.S6.

[29] Shen D, Sun J T, Li H, Yang Q and Chen Z 2007 January. Document summarization using conditional random fields. In *IJCAI* (Vol. 7 pp. 2862-2867).

[30] Hur J, Kang S N and Seo S W 2013 June. Multi-lane detection in urban driving environments using conditional random fields. In *2013 IEEE Intelligent Vehicles Symposium (IV)* (pp. 1297-1302). IEEE.

[31] Deka R R, Kalita S, Bhuyan M P and Sarma S K 2019 September. A Study of Various Natural Language Processing Works for Assamese Language. In *International Conference on Innovation in Modern Science and Technology* (pp. 128-136). Springer Cham.

[32] Bhuyan M P and Sarma S K 2019. An N-gram based model for predicting of word-formation in Assamese language. *Journal of Information and Optimization Sciences* 40(2) pp.427-440.

[33] Bhuyan M P, Sarma S K, and Sarma P 2020 Context-based Clustering of Assamese words using n-gram model Second International Conference on Advances in Electrical and Computer Technologies 2020 (ICAECT 2020) Springer (in press).

[34] Bhuyan M P, Sarma S K and Rahman M 2020 June. Natural Language Processing based Stochastic Model for the Correctness of Assamese Sentences. In *2020 5th International Conference on Communication and Electronics Systems (ICCES)* (pp. 1179-1182). IEEE.

[35] Bhuyan M P and Sarma S K 2019 February. Effects of Prediction-Length on Accuracy in Automatic Assamese word prediction. In *2019 IEEE International Conference on Electrical Computer and Communication Technologies (ICECCT)* (pp. 1-4). IEEE.

[36] Bhuyan M. P. and Sarma S. K. 2019 A statistical model for automatic Error Detection and Correction of Assamese Words International Journal of Recent Technology and Engineering (IJRTE) vol.8:2 pp. 6111-6116 DOI:10.35940/ijrte.B3859.078219.