Journal of
Data Acquisition and Processing

ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.7766233

A REVIEW ON MULTIDIMENSIONAL CLASSIFICATION TECHNIQUES OF NATURAL LANGUAGE SITUATIONAL INFORMATION BASED ON NN AND BERT MODEL

Mr. Dattatray S. Shingate[1] Dr. Shyamrao V. Gumaste[2]
[1]PhD Scholar, Department of Computer Engineering, Mumbai Education Trust's Institute of Engineering, Nashik
[2]Professor, Department of Information Technology, Mumbai Education Trust's Institute of Engineering, Nashik
Savitribai Phule Pune University

Abstract

Nowadays, the use of social networking sites is becoming more popular especially in the disaster like situations. Social media has changed the way of communication. People now get news through it. Apart from all these positive features, social media is also slowly turning out to be a live-saving tool. Various social media platforms, including Twitter, blogs, News aggregators, etc., include heterogeneous material in a variety of forms. An oversized quantity of helpful data is shared on social networking throughout associate emergency, together with the users' sympathies and opinions. So, normally people uses their natural language in which they speak for communication which computer doesn't understand. Therefore, there should have a reliable methodology which required for extracting helpful information (Context) from the natural language that people normally used for the communication and sharing their available information regarding current scenario happening nearby them during the disaster. Further down this information can categorized into situational and non-situational information. Situational information indicates a current consequences happening due to disaster and where immediate action is required which may reduce further losses that may likely to be happened. Social media today are crucial for disseminating information from the real world, experiences from daily life, and ideas through online groups and networks. Real-time events like disasters, power outages, traffic, etc. can leverage this information. Due of the noisy data, unrelated data, and data in many formats present on social media, analysing and comprehending such information can be difficult. Because of this, this study examines and classifies numerous event detection techniques in various forms of social media.

Keyword : Disasters, RoBERTa model, situational information, Event detection

1. Introduction

Social media are hugely influential in today's world. Social media can be used to connect people all over the world through exchanging information, ideas, and common interests. Many people used social media to send and receive messages in order to communicate. There are various forms of social media, including Twitter, news archives, multimedia, blog posts, web pages, Facebook, etc. Social media produces a tonne of information that the public will find really helpful. The information can be characterised as an occurrence that took place outside. Social media offers both information about events and other topics. Social media can be used to post enormous amounts of information during a crisis. Finding event-related material on social media is a difficult undertaking because of this. Several social media platforms have different traits.

During a disaster, situational awareness information enables high-level concerned authorities to have a better understanding of the situation. Given the urgency of the situation during a disaster, particularly in the early stages, humanitarian groups must plan their relief operations accordingly. It provides a comprehensive picture of the disaster, indicating where resources and medical facilities are required and, consequently, the magnitude of the disaster.

By separating it from non-situationally shared information, the situationally shared information has been identified.[4] The shared situational information comprises victims, deceased individuals, and resource availability and demand. Moreover, disseminated information that offers accurate, useful information promotes situational awareness. Content that demonstrates understanding of the seriousness of the problem and specifics of the situation are included in this shared information. The authors employed a feature-based technique to find shared information in a crisis situation.

The situational shared information in a communication language like English was identified using a feature-based technique and an SVM classifier.[6] Yet, to find situational shared knowledge during a crisis, feature-based approaches have been utilised in every existing study.

2. Literature Survey

In [4], focusses mainly on automatically identifying of tweets that contributed to situational information's using filtered approach and explain why it is beneficial for those seeking information during mass emergency. They have utilized a combination of hand annotated and automatically-extracted linguistic features. Authors used search and filter information mechanism. They opted mechanism of Maximum entropy and Naïve-bayes approach for feature extraction in the year of 2011.

In [5], authors classified information that people post during disasters into a set of user-defined categories of keywords from information (e.g., "needs", "damage", etc.). The system continuously imports data from Twitter for this purpose, analyses it using machine learning classification techniques, and uses real-time crowdsourcing to benefit from human input. For the very first time they used word based context finding of information based on N-gram feature extraction approach. Relies on vocabulary that is present in training phase of model so, produces bad results on new vocabulary present at testing phase and provides good result if training and testing performed on similar dataset.

A novel framework is proposed in [6] which organizes tweets into categories to extract situational information before summarizing the material. The suggested paradigm takes into account the common characteristics of disaster situations, where a single tweet frequently combines situational and non-situational information, as well as some numerical information, such the number of casualties. The objective was to develop a model which is independent of vocabulary present at training and testing phase of model. They noted that in order to aid decision-making processes when there is a pressing need for time, situational information can be extracted from the vast amounts of feeling and opinion and summarized. The methodology opted was Used TF/IDF vector generation approach for feature extraction (quite Superior to N-gram approach). Model doesn't relies on vocabulary that is present in training phase of model. So, produces good results on new vocabulary present during testing phase

A method was developed for extracting and summarizing Situational Information from the social media during emergency situation, they prominently highlighted that it is necessary to not only extract the situational information from the large amounts of sentiment and opinion, but also to summarize the large amounts of situational information posted in real-time [7]. They created a new framework for categorization and summarizing that can handle tweets in both Hindi and English. They first extract tweets that contain situational information, and then they describe it. Their suggested methodology was created on an understanding of how various concepts change on Twitter in times of crisis. With the aid of this knowledge, we are able to perform better on English-language tweets than cutting-edge tweet classifiers and summarization techniques. Also, this is the first effort that we are aware of to extract situational information from tweets that are not in English. Authors used feature based approach with SVM (rbf) is used and compared with SVM(linear), Logistic regression, Naïve-Bayes Classification algorithm.

In [8], Author first time introduced the concept of convolution neural network for the classification and well aware about the fact that in order to attain optimal results, the most recent state-of-the-art classification systems need a large amount of labelled data that is specific to a certain event for training. In this study, they offered neural network-based classification techniques for the problem of categorizing information into binary and multiclass categories. They demonstrated that neural network-based models outperform state-of-the-art techniques and don't need any feature engineering. Their suggested strategy makes the best use of the out-of-event data in the early stages of a disaster when no labelled data is available and produces positive results.

For locating valuable tweets in a crisis situation, authors of [9] presented neural network-based classification algorithms. Their suggested strategy makes the best use of the data from previous events (out-of-event data) at the beginning of a disaster when no labelled data is available and produces positive results. They used CNN model for classifying tweets into useful vs. not useful for a crisis event. They used pre-trained word embedding i.e. Google embedding and crisis embedding to better initialising the models, and they fine-tune them before delivering it to CNN for improved performance.

Disastro, a real-time disaster response system based on artificial intelligence proposed by the authors of [10], helps the volunteers by locating significant tweets in real-time Twitter data and categorising them as "rescue" and "donation"-related. They have used SVM with count vectorizer representation for the implementation and claimed an accuracy of above 82%. The tweets posted during the Chennai rains and Kerala floods were used to empirically assess disaster using several machine learning algorithms. Disastro is adaptable to a variety of disasters and has enhanced categorization accuracy, making it robust and able to manage any location-based crisis.

They proposed [11] to use inductive semi-supervised technique to utilize unlabeled data along with fewer labeled data. They specifically use a deep learning framework based on graphs to construct an inductive semi-supervised model. When compared to solely using labelled data, their findings indicate a considerable improvement using unlabeled data as well. To construct a graph, they choose k-nearest neighbor-based approach for finding nearest neighbours of instances. The nearest neighbour graph has n vertices and an edge set made up of a subset of n occurrences for each vertex. The distance between tweets ti and tj (d(i, j)) determines the edge, and the value of d indicates how similar the two tweets are.

Authors [12] suggested an ensemble strategy for tweet-level spam detection. They created five CNNs, and an ensemble is made up of one feature-based model and five CNNs. Different word embeddings (Glove, Word2vec) are used by each CNN to train the model. The user-based, content-based, and n-gram features used in the feature-based model. Their method uses a multilayer neural network that serves as a meta-classifier to integrate deep learning and conventional feature-based models. They tested their approach using both balanced and unbalanced datasets.

A novel strategy based on the skipgram model was put forth by authors [14], in which each word is represented as a collection of character n-grams. Each character in an n-gram has a corresponding vector representation, with words being represented as the total of these representations. Their approach made it possible to swiftly train models on enormous corpora and calculate word representations for terms that weren't in the training set.

BERT (Bidirectional Encoder Representations by Transformers), which will be a useful concept in the field of sentiment analysis, was introduced by the authors of [13]. Google created the profound learning model known as BERT. Since Google opened it, a number of researchers and businesses have embraced it and used it for a variety of text classification jobs. As a result, in this research, they used BERT to a dataset of tweets about disasters. This study will assist rescue and emergency responders in developing efficient knowledge management strategies. BERT makes use of a transformer, an attentional process that learns the relationship between words in a text. The converter consists of two different mechanisms, the encoder receives text input, the decoder creates a job forecast in its purest form. BERT just needs an encoder mechanism because it is designed to create language models. Pretraining and fine-tuning are the two phases employed in BERT. The model is pre-trained using several pre-training activities that condition it on unlabeled data. All parameters are refined by the named downstream data once the relevant parameters launch BERT models to be precisely targeted.

A novel framework is proposed in [1] which broadly classifies tweets into 2 categories: Situational and Non-Situation Tweets based on generating word based word2vec matrix for each word for finding out the context. For feature extraction SVM (rbf) is used in their implementation and also used Bidirectional Encoder Representations from Transformers BERT which is an open source machine learning framework for natural language processing (NLP). BERT uses the surrounding text to provide context in order to help computers understand the meaning of ambiguous words in text. Using Wikipedia text for pre-training, the BERT framework can be fine-tuned utilizing question and answer datasets. Their study bounded to binary classification of information which creates the scope for multidimensionality classification.

From the thorough study of all the available related papers we came up with the following summarized table:

SN	Author , Title , Publisher / Year of Publication	Objectives	Tehniques / Methods / Technology Applied or explained	Finding / Results / Conclusion / Research Gap identified
1	**Author :** Sudha Verma , Sarah Vieweg , William J. Corvey , Leysia Palen1 , James H. Martin1 , Martha Palmer , Aaron Schram1 & Kenneth M. Anderson1 **Title:** Natural Language Processing to the Rescue? Extracting "Situational Awareness" Tweets During Mass Emergency **Publisher :**Association for the Advancement of Artificial Intelligence (www.aaai.org). **Year : 2011**	1]Automatically identifying of tweets that contributed to situational informations 2]To provide relative information to intended user who is seeking for help during emergency	1]Used Maximum entropy and Naïve-bayes approach for feature extraction.	1] Used Search and filter based mechanism 2]Accuracy 80% Average
2	**Author :** Muhammad Imran, Carlos Castillo , Ji Lucas, Patrick Meier, Sarah Vieweg Qatar Computing Research Institute Doha, Qatar svieweg@qf.org.qa **Title:** AIDR: Artificial Intelligence for Disaster Response **Publisher :** ACM Proc. 23rd Int. Conf. World Wide Web, **Year : 04/2014**	1]Identifying User defined categories of Tweets 2]Used word based context finding.	1]Used N-gram feature extraction approach	1]Relies on vocabulary that is present in training phase of model. 2] Model produces bad results on new vocabulary present at testing phase which wasn't part during training a model. In short testing was solely dependent on the way model training has done 3]Gives good result if training and testing performed on similar dataset
3	**Author :** Koustav Rudra, Department of CSE, IIT Kharagpur, India Niloy Ganguly, Department of CSE, IIT Kharagpur, India Pawan Goyal, Department of CSE, IIT	1]To develop a model which is **independent of vocabulary** present at training and testing phase of model	1]Used TF/IDF vector generation approach for feature extraction (quite Superior to N-gram approach)	1]Does Not Relies on vocabulary that is present in training phase of model. 2] It produces good results on new vocabulary present during testing phase of

	Kharagpur, India Saptarshi Ghosh, Department of CST, IIEST Shibpur, India **Title:** Extracting Situational Information from Microblogs during Disaster Events: a Classification-Summarization Approach **Publisher :** Proc. 24th ACM Int. Conf. Inf. Knowl. Manage. **Year : 10/2015**			model as model training was not dependent on vocabulary words present during training phase.
4	**Author :** Koustav Rudra, Department of CSE, IIT Kharagpur, India Niloy Ganguly, Department of CSE, IIT Kharagpur, India Pawan Goyal, Department of CSE, IIT Kharagpur, India Saptarshi Ghosh, Department of CST, IIEST Shibpur, India **Title:** Extracting and Summarizing Situational Information from the Twitter Social Media during Disasters **Publisher** :ACM Transactions on the Web · **Year : 01/2019**	1]To develop a Novel classification-summarize framework to classify Situational and Non-Situation Tweets 2]To develop a hybrid framework which can classify tweets in both English and Hindi languages	1]Used feature based approach with SVM (rbf) classifier is used.	1]Used feature based approach with SVM (rbf) is used and compared with SVM(linear), Logistic regression, Naïve-Bayes Classification Algo. 2] Accuracy 84% Avge
5	**Author :** Sreenivasulu Madichetty and Sridevi M **Title:** A Neural-Based Approach for Detecting the Situational Information From Twitter During Disaster	1]Classification of Tweets into 2 categories: Situational and Non-Situation Tweets 2]Generating word based word2vec	1]For feature extraction SVM (rbf) is used 2]Used BERT- Base and BERT- Large NLP Framework 3]Used CNN Deep learning model for	1]Binary classification of Tweets- generated the scope for multidimensionality classification 2]It does not work well if tweets

	Publisher :IEEE TRANSACTIONS ON COMPUTATIONAL SOCIAL SYSTEMS **Year** : **06/2021**	matrix for each word for finding out the context 3]Improving performance over LSTM	binary classification	contains insufficient number of features

So, based on the findings from various related papers There have been attempts by several authors [4], [6], [7], and [1] to categorise situational and non-situation information during a crisis but there isn't yet a reliable model that can do multidimensional classification of situation information.Their focus was only on binary categorization. A feature-based technique based on the BOW model was utilised by Verma et al. [4] to detect the situational tweets during a crisis. The n-grams feature-based method was then utilised by Imran et al. [5] to detect the user-defined categories of tweets sent during an emergency. The new vocabulary that is only present in the testing tweets, however, performs poorly since it depends on the vocabulary that is present in the training tweets. When training and testing are done solely using the same disaster data set, it works well for the same disaster case. Later, Rudra et al. [7] created a model that can recognise situational tweets during a crisis regardless of the vocabulary used in the training tweets.

We proposed a system that is a neural-based approach with the mix of the RoBERTa model (RoBERTa is a transformer-based language model that uses self-attention to process input sequences and generate contextualized representations of words in a sentence) and feature-based technique for extracting the information can does a better for categorizing it into situational and non-situational information. Further down proposed method will also outperforms among all available strategies on totally different disaster information datasets as RoBERTa is a pre-processed model by the power of Google.

3. Conclusion

Social media can be used to spread the vast volume of user-generated information. Users can benefit greatly from context analysis and understanding of shared information during emergencies, power outages, traffic jams, etc. One of the key duties and goals in recognising real-world occurrences is event detection. On social media data, event detection and its multidimensional classification must be effective and precise. This study describes a survey of numerous techniques based on event detection and their classification approaches using a variety of social media platforms. Future work will provide a new method for categorising information, particularly situational information in social media, in order to increase accuracy and speed.

References

[1] M. Sreenivasulu and M. Sridevi "A Neural-Based Approach for Detecting the Situational Information From Twitter During Disaster" IEEE TRANSACTIONS ON COMPUTATIONAL SOCIAL SYSTEMS 2021, pp 1-11.

[2] M. Sreenivasulu and M. Sridevi, “Mining informative words from the tweets for detecting the resources during disaster,” in Proc. Int. Conf. Mining Intell. Knowl. Explor. Hyderabad, India: Springer, 2017, pp. 348–358.

[3] M. Sreenivasulu and M. Sridevi, “A survey on event detection methods on various social media,” in Recent Findings in Intelligent Computing Techniques. Singapore: Springer, 2018, pp. 87–93.

[4] S. Verma et al., “Natural language processing to the rescue? extracting ‘situational awareness’ tweets during mass emergency,” in Proc. 5th Int. AAAI Conf. Weblogs Social Media, 2011, pp. 385–392.

[5] M. Imran, C. Castillo, J. Lucas, P. Meier, and S. Vieweg, “Aidr: Artificial intelligence for disaster response,” in Proc. 23rd Int. Conf. World Wide Web, 2014, pp. 159–162

[6] K. Rudra, P. Goyal, N. Ganguly, S. Ghosh, and M. Imran, “Extracting and summarizing situational information from the Twitter social media during disasters,” ACM Trans. Web, vol. 12, no. 3, p. 17, Jul. 2018.

[7] K. Rudra, S. Ghosh, N. Ganguly, P. Goyal, and S. Ghosh, “Extracting situational information from microblogs during disaster events: A classification-summarization approach,” in Proc. 24th ACM Int. Conf. Inf. Knowl. Manage., Oct. 2015, pp. 583–592

[8] Dat Tien Nguyen, Kamela Ali Al Mannai, Shafiq Joty, Hassan Sajjad, Muhammad Imran, Prasenjit Mitra “Rapid Classification of Crisis-Related Data on Social Networks using Convolutional Neural Networks”in arXiv:1608.03902v1 [cs.CL] 12 Aug 2016.

[9] Dat Tien Nguyen, Kamela Ali Al Mannai, Shafiq Joty, Hassan Sajjad, Muhammad Imran, Prasenjit Mitra,“Robust Classification of Crisis-Related Data on Social Networks Using Convolutional Neural Networks” Proceedings of the Eleventh International AAAI Conference on Web and Social Media (ICWSM 2017)

[10] Krishna Kanth A, Abirami S, Chitra P, Gayathri Sowmya G, “Real time Twitter based disaster response system for indian scenarios,” in Proc. 26th Int. Conf. High Perform. Comput., Data Analytics Workshop (HiPCW), Dec. 2019, pp. 82–86.

[11] Firoj Alam,1 Shafiq Joty,2 Muhammad Imran1,“Graph Based Semi-Supervised Learning with Convolution Neural Networks to Classify Crisis Related Tweets” Proceedings of the Twelfth International AAAI Conference on Web and Social Media (ICWSM 2018) pp 556-559.

[12] Sreekanth Madisetty and Maunendra Sankar Desarkar, “A Neural Network-Based Ensemble Approach for Spam Detection in Twitter” IEEE TRANSACTIONS ON COMPUTATIONAL SOCIAL SYSTEMS, VOL. 5, NO. 4, DECEMBER 2018 pp 973-984

[13] A K Ningsih□ and A I Hadiana ,“Disaster Tweets Classification in Disaster Response using Bidirectional Encoder Representations from Transformer (BERT)” IOP Conf. Series: Materials Science and Engineering 1115 (2021) 012032 IOP Publishing doi:10.1088/1757-899X/1115/1/012032.

[14] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, “Enriching word vectors with subword information,” 2016, arXiv:1607.04606. [Online]. Available: http://arxiv.org/abs/1607.04606

[15] F. Chollet. (2015). Keras. [Online]. Available: https://github.com/ fchollet/keras

[16] Josh Xin Jie Lee, “Multi-Label Text Classification with XLNet”, https://towardsdatascience.com/multi-label-text-classification-with-xlnet

b5f5755302df#:~:text=Achieve%20state%2Dof%2Dthe%2D,class%20text%20classification%20with%20XLNet&text=At%20the%20time%20of%20its,sentiment%20analysis%2C%20and%20document%20ranking.

[17] Prashant Yawalkar, Dr. M. U. Kharat, “Automatic Handwritten Character Recognition of Devnagri Language: a hybrid training algorithm for Neural Network”, Evolutionary Intelligence, Springer Publication, April 2021

[18] Prashant Yawalkar, Dr. M. U. Kharat, “A Fuzzy Neural Hybrid Approach for Recognition Of Hand Written Devnagri Characters”, Asian Journal of computer science and technology, April –June 2019.

[19] Prashant Yawalkar, Dr. M. U. Kharat, “Recognition Of Hand Written Devnagri Characters Using Effective Thinning And Fuzzy Logic”, Journal of Advanced Research in Dynamical and Control Systems, Special Issue, December 2018, pp. 553–566.

[20] Prashant Yawalkar, Dr. M. U. Kharat, “Effective Thinning Algorithm for Recognition of Hand Written Devnagri Compound Characters Using Neural Network”, International Journal of Applied Engineering Research, Volume 13, Number 12 (2018) pp. 10539–10550.

[21] Prashant Yawalkar, Dr. M. U. Kharat, Dr. S. V. Gumaste, “Segmentation of Multiple Touching Hand Written Devnagari Compound Characters:Image Segmentation for Feature Extraction”, Feature Dimension Reduction for Content Based Image Identification, IGI Global, A volume in the Advances in Multimedia and Interactive Technologies (AMIT) Book Series, pp. 140–163.

[22] Pallavi Patil, Prof. P. M. Yawalkar, “Recommendation of Conversation Documents With Keyword Based Clustering”, International Journal of Innovative Research in Computer and Communication Engineering, June 2016.

[23] Prof. D. S. Shingate, Harshada Godse, Snehal Shinde, Ankita Kshirsagar, and Vidya Kale, "Product reviews and sentiment analysis using machine learning," International Journal Of Advance Research And Innovative Ideas In Education, vol. 8, no. 3, pp. 2116-2120, May-Jun 2022. [Online]. Available: http://ijariie.com/AdminUploadPdf/Product_reviews_and_sentiment_analysis_using_machine_learning_ijariie16952.pdf [Accessed :]

[24] Vaishali A. Mahale, Prashant M. Yawalkar, “Advanced Mechanism in Learning of OPS from Weakly Labeled Street View Images”,International Journal For Scientific Research And Development, July 2016

[25] Pallavi Patil, Prof. P. M. Yawalkar "Review on Extraction of Keywords and Recommendation of Documents in Conversation", International Journal of Advanced Research in Computer and Communication Engineering, Vol. 4, Issue 12, Dec 2015

[26] Vaishali A. Mahale, Prashant M. Yawalkar, "A Review on Advanced Mechanism in Learning and Recognition of Ops from Weakly Labeled Street View Images", International Journal of Advanced Research in Computer and Communication Engineering Vol. 5, Issue 1, Jan 2016