Effect of machine learning algorithms for information extraction from microblogs for emergency relief and preparedness by comparative analysis study.

Harshadkumar Prajapati[1], Hitesh Raval[2]

1 Research Scholar, Faculty of Computer Science, Sankalchand Patel college of Engineering, Sankalchand Patel University, SK Campus, Visnagar, Gujarat.

2 C. J. Patel college of computer studies, Sankalchand Patel University, SK Campus, Visnagar, Gujarat.

Abstract:

The frequency and severity of natural disasters around the world have increased due to excessive human interference in the environment. Particularly during major emergencies, social media has significantly changed information-sharing and information-gathering methods. A significant possibility to automatically extract information from a large amount of digital data has also been made possible by sophisticated information extraction models built on machine learning methods. In this paper, different machine learning algorithms are compared based on number of parameters such as precision, recall, F1 score and accuracy. In this paper, we conclude that the accuracy achieved using Support Vector Machine (SVM)linear model is better than all other algorithms.

Key Words: Microblogs, Precision, Recall, Information Extraction, Machine Learning, SVM

Introduction

Several natural disasters happened around the world were very extreme[1,2,3,4] Government and non-government organizations that respond to disasters form the foundation of a nation's infrastructure for handling major incidents. These organizations work at various levels (local, regional, federal, and international) and are in charge of various stages. Emergency responders heavily rely on numerous sorts of information, which are crucial for communication both within and across organizations and for mobilizing diverse resources.

Since the advent of web 2.0, the Internet's purpose has shifted from providing information to supporting community development and communication[5]. With the development of technology and the widespread affordability of the internet, there are now more ways than ever before to produce, access, and share information via different social networking sites like Facebook, YouTube, LinkedIn, and Twitter[6]. One of the biggest social media sites in the world is Twitter, where users may interact with one another by exchanging brief tweets about current events, entertainment, updates, or just odd ideas. Twitter provides a search tool where users may enter search terms and receive a list of results. Users can choose to view the most popular tweets, the most recent tweets posted, or users who most closely relate to the search term. Twitter play a vital role for information extraction in various domains. Twitter plays an important role in emergency situation. In proposed research, Twitter data is used for information retrieval in emergency relief and preparedness situation.

The next section of paper will discuss about information retrieval and role of machine learning algorithms. The next section will deal with methodologies and data Set used for the experimentation followed by results and discussions. Finally, the paper will end with conclusion section.

Role of Machine Learning in Information Retrieval from Microblogs

Software that organizes, stores, retrieves, and evaluates information from document repositories, particularly textual information, is known as information retrieval (IR). Microblogs like Twitter became an efficient and powerful source of valid textual information.

Definition: Information Retrieval from Microblog: Organization, Storing, Retrieving and Evaluating useful and valid information from Microblog like Twitter using machine learning based software.

Retrieving information is a crucial component of machine learning[7]. It is the procedure for locating and obtaining data from a database. Algorithms that search through the database or user input can be used for this.

Finding data patterns is made possible by information retrieval, which is crucial for machine learning. Finding data patterns through supervised or unsupervised learning[8] is the foundation of machine learning. One option for doing this is by employing information retrieval techniques to discover pertinent facts.

When solving supervised learning problems, the algorithm searches the database for pertinent data using keywords or other criteria. In unsupervised learning problems, the algorithm searches for relevant data by looking for patterns in different data sources rather than using predefined keywords or features.

Machine learning applications include question-answering systems, web crawling systems, and many others require information retrieval. It may be applied to practically any field where people make queries. It can be used to find information in books, databases, or any other kind of source.

The area of information retrieval is wide and has numerous subfields. Each area focuses on a distinct component of information retrieval and offers fresh, innovative approaches to the process.

Although information retrieval has been automated for many years, machine learning has just recently been used in this field. Indexing, query formulation, comparison, and feedback are the four separate stages of information retrieval, all of which offer chances for learning.

Adaboost is well known ensemble classifier that is used in various domains. Information retrieval is the main domain where this algorithm can be used extensively[9]. By combining a small number of independent learners and training the same data set, ensemble learning overcomes the constraints of a single classifier and dramatically enhances an algorithm's capacity to generalize. AdaBoost is an iterative technique that constantly trains the incorrectly classified data, particularly continuously, and then serializes these weak classifiers to build a strong classifier. The self-adaptation involves decreasing the individual weights of incorrectly classified samples while raising the weight of correctly classified samples. After uploading, all sample weights are sent to the lower layers for training until the predetermined minimum error rate is reached or the predetermined maximum number of iterations is reached.

Finally, a strong classifier is created by serially combining the base classifiers that were obtained from each training.

The non-parametric supervised learning approach used for classification and regression applications is the decision tree. It is organized hierarchically and has a root node, branches, internal nodes, and leaf nodes. It is also used extensively for information retrieval[10] .

The Nave Bayes algorithm[11] is a supervised learning method for classification issues that is based on the Bayes theorem. It is mostly employed in information retrival with a large training set. The Naive Bayes Classifier is one of the most straightforward and efficient classification algorithms available today. It aids in the development of quick machine learning models capable of making accurate predictions.

SVM[12] is a classifier that can handle large amounts of data without suffering significantly from performance issues. SVM is increasingly more commonly employed. SVM has also been frequently utilized in information retrieval, particularly in the classification procedure. The text data, which has high-dimension data, is a great candidate for use of SVM's capacity to analyses high-dimension data.

One of the strongest algorithms for classification jobs is Random Forest (RF)[13]. The fundamental tenet of RF is that a collection of poor learners can create a strong learner. Large datasets can be accurately and precisely classified using RF. Every tree in RF serves as a classifier and is reliant on a random vector value. When training, RF creates several decision trees, each of which predicts the outcome of the modality using bootstrap samples of the training data and a random selection of attribute values. By integrating decision trees with a majority vote or by averaging them all, predictions are created.

Data Set and Methodologies

In proposed research, numerous tweet-ids via FIRE Microblog Track are utilized to access the tweets. From these tweets, which included tweet content, tweet id, and other metadata, research kept the pertinent information. From the complete list of tweets, several tweets are filtered out.

Python programming tool is used to to perform text-based analysis on tweets. Samples are taken to understand the text attributes from tweets. Several pre-processing techniques are applied to remove unusual text from the data.

Results and Discussions

Six main machine learning algorithms are compared in terms of various parameters such as: Precision, Recall, F1 Score and accuracy.

The percentage of accurate predictions to all other guesses is known as accuracy. It is one of the simplest model measurements there is.

Accuracy = (Correct Prediction)/ (Correct Prediction + Incorrect Prediction) -----------(1)

In other way ,

Accuracy = (True Positive + True Negative) / (True Positive + False Positive + True Negative + False Negative) -------------(2)

Algorithm	**Accuracy**
Adaboost	0.42
Decision Tree	0.50
Naïve Bayes with Gaussian Filter	0.29
SVM (Linear)	0.51
Random Forest	0.39

Table-1 describes the accuracy of all algorithms based on sample data that was provided for training purpose.

(Table -1 Accuracy of Machine Learning Algorithms for Twitter Data)

From, results it is derived that the accuracy of SVM model is good in comparison of all other models. Decision Tree algorithm also provides accuracy near to SVM but implementation of decision tree in context to information retrieval is quite difficult. The accuracy of all remaining algorithm is not up to the mark.

Precision, Recall and F1 Score measures are also vital for the quality of the model. The proportion of correctly made positive forecasts to all positive predictions is known as precision.

Precision = (True Positive) / (True Positive + False Positive) -----------(3)

Recall determines the proportion of actual positives to all positive labels.

Recall = (True Positive) / (True Positive + False Negative) ------------- (4)

The F1 score, which is the harmonic mean of the two values, depends on both recall and precision.

F1 Score= 2 * ((Recall * Precision)/ (Recall + Precision)) ---------------(5)

Table-2 describes the Precision, Recall and F1 Score of Twitter data.

Algorithm	**Precision**		**Recall**			**F1 Score**	
	Macro	**Micro**	**Macro**	**Micro**		**Macro**	**Micro**
Adaboost	0.68	0.72	0.57	0.61		0.62	0.66
Decision Tree	0.63	0.66	0.59	0.62		0.61	0.64
Naïve Bayes with Gaussian Filter	0.49	0.52	0.48	0.55		0.48	0.53
SVM (Linear)	0.75	0.77	0.62	0.68		0.67	0.72
Random Forest	0.78	0.80	0.39	0.44		0.49	0.57

(Table -2 Precision, Recall and F1 Score of Machine Learning Algorithms for Twitter Data)

The results indicate that the SVM algorithm is best suited with all parameters. Random Forest is good as far as Precision is concern but not performs well for Recall and F1 Score. The result of Adaboost algorithm is also good in all parameters but not as good as SVM. The implementation is also difficult for Adaboost Algorithm.

Conclusion:

The invention of social media sites is boon for the human being if it is used in positive manner. Twitter is one of the most popular social media platforms utilized in the world in which users across world may communicate with one another by sharing quick tweets on news, entertainment, updates, or just strange ideas. The information available in numerous tweets are beneficial for many domains. Information extraction from microblogs such as Twitter for emergency relief and preparedness play a vital role for quick and useful insights to human being. Information Extraction from Tweets can be carried out using machine learning algorithms. The insights of our study is, six prominent algorithms are compared with various parameters such as accuracy, precision, recall, F1 Score. SVM linear algorithm gave better results in comparison with all other algorithms. However, still these results are based on various data which may further optimized to improvise the accuracy of the model which can enhance the ability of machine learning algorithm.

References

1.Duni, L., & Theodoulidis, N. (2019). *earthquake: Strong ground motion with emphasis indurres.* Albania: EMSC on Line Report.

2.Cerrai, D., Yang, Q., Shen, X., Koukoula, M., & Anagnostou, E. N. (2020). Hurricane dorian: automated near-real-time mapping of the “un-precedented” flooding in the bahamas using synthetic aperture radar. *Natural Hazards and Earth System Sciences*, 1463–1468.

3.Gupta, K. (2020). Challenges in developing urban flood resilience in india. *Philosophical Transactions of the Royal Society*.

4.Ward, M. e. (2020). mega-fires on australian fauna habitat. *Nature Ecology & Evolution*, 1-6.

5.Tuten, T. L. (2008). Advertising 2.0: social media marketing in a web 2.0 world. *ABC-CLIO*.

6.Zeng, D., Chen, H., Lusch, R., & Li, S.-H. (2010). Social media analytics and intelligence . *IEEE Intelligent Systems*, 13-16.

7.Jin, L. S. (2011). Machine learning for information retrieval. *SIGIR '11: Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval* (pp. 1293–1294). New York,USA: ACM.

8.Reddy, Y. C., & Varma, N. K. (2011). Review on Supervised Learning Techniques. *Emerging Research in Data Engineering Systems and Computer Communications* (pp. 577–587). Springer.

9.Chen, Y. L. (2020). Research on Stock Selection Strategy Based on AdaBoost Algorithm. *Proceedings of the 4th International Conference on Computer Science and Application Engineering*, (pp. 1-5).

10.Bindhia, K. F., Vijayalakshmi, Y., Manimegalai, P., & Babu, S. S. (2017). Classification Using Decision Tree Approach towards Information Retrieval Keywords Techniques and a Data Mining Implementation Using WEKA Data Set. *International Journal of Pure and Applied Mathematics*, 19-29.

11.Yang, F. J. (2018). An implementation of naive bayes classifier. *n 2018 International conference on computational science and computational intelligence (CSCI)* (pp. 301-306). IEEE.

12.Drucker, H., Shahrary, B., & & Gibbon, D. C. (2002). Support vector machines: relevance feedback and information retrieval. *Information processing & management*, 305-323.

13.Svetnik, V., Liaw, A., Tong, C., Culberson, J. C., Sheridan, R. P., & Feuston, B. P. (2003). Random forest: a classification and regression tool for compound classification and QSAR modeling. *Journal of chemical information and computer sciences*, 1947-1958.