A Scalable Feature Extraction Technique for Social Media Analysis

Tauqeer Ahmad, Muhammad Rizwan Rashid Rana, Muhammad Aun Akbar, Asif Nawaz
University Institute of Information Technology
PMAS-Arid Agriculture University, Rawalpindi, Pakistan
Corresponding Author: rizwanrana315@gmail.com

Abstract—Social media can be considered as a tool for expressing opinions and allow people to comment on a different topic. People especially youth are interested to use the different type of social media sites, messenger, blogs, microblogs etc. People comment positive and negative on published tweets both locally and globally. In social media it was very difficult to find out about specific event accrued in any part of the world. Feature extraction is now becoming the very active area of research. The data comes from various types of systems in the enterprise. Feature extraction technique used to reduce noisy data and increase the accuracy of the system. In the Past, many researchers work on feature extraction to improve semantic similarity between words using feature extraction techniques. These techniques which researchers used in the past not enough for better result and for improvement of accuracy of the system. With improving the old techniques, results should be improved. Semantic similarity between words can be identified through different feature extraction techniques by using social media sites. Results are taken on the basis of semantical way .so, The Proposed technique will be accomplish using Candidate Terms (Natural Language Tool Kit) and the Refining these Terms through (Pointwise mutual information) for Most suitable features (ACO) at the end Final Features created and system accuracy increased 82% using these techniques.

Keywords - component; Data Mining; Features; Pointwise Mutual Information; Ant Colong Optimization.

I. INTRODUCTION

Data is collected and can be accessed from every part of the world. Managing and balancing of created dataset contribute the effective approaches from given framework which has delivered by a dataset. Data mining is the very huge area which can manage the transferring of big data into very important information [1]. Data mining always contain set of many algorithm and process to extract the valuable dataset from given big dataset. Data mining contains many methods like web data mining, regression, association rules clustering, regression, supervised learning and unsupervised learning. Our most significant strategy of data mining is feature extraction process. Feature Extraction intends to extract most valuable data from given dataset [2]. The main feature space of feature extraction issue is finding the most valuable features from a dataset. Without strong knowledge, it is difficult to recognize that which are most valuable features.

Microblogging like Twitter is one of the fast developing tools for conveying opinion, which allows people to publish short tweets on different topics. People remark on real-world occasions both local and globally. Social media contained different types of tweets on different events which are positive and negative in the big database. It was very difficult to find out the specific event especially disruptive events and non-disruptive events. Some experiments were done on a different type of features that distinguish disruptive events from other events. But the length constraint of a tweet limits the amount of sentiment that can be expressed [3]

People are globally like a joint family through social media websites and they share their feelings, emotion and other information with each other through text.at present days, many scientists are working in the field of feeling extraction from text. The most difficult task for internet is to force people shows their personal feelings with a word-based standard in the period of virtual interaction. Direct observation of emotion from the text was very tough to show internal feelings of a user.

Feature extraction is now becoming the very active area of research. The data comes from various types of systems in the enterprise such as Wikipedia, Twitter and from all social media sites. The dataset carried from the big database of different sites and social media. This type of data consists of the bag of a word and not inconsistent form. Data is not in unique form, it contains raw data and information. Feature extraction process extracts the valuable information from a database and collects related dataset which is required [4]. Feature extraction technique used to reduce noisy data and increase the accuracy of the system.

II. RELATED WORK

Dimensionality reduction techniques play an important role in different fields like research, weather information and extraction of features in many fields. Recently, huge amount of data carried to identify different valuable features through domain. This type data set consists of the useful features and this type of features help to create accurate result and also find out the greater accuracy. Generated Features are acts like given input variables and also attributes of the dataset. Feature extraction methods are used in the detection and highlighted the valuable, useful and very important features from the database which consist of bag of words and also raw materials [5]. The feature extraction task is to identify useful data which help and very benefit, and in some time it is important to applying dimensionality reduction techniques.

Feature selection (fs) is generally used as part of machine learning, in particular with the quantity of data is massive [6]. Characteristic selection or extraction is carried out through removing redundant and beside the point features from the selected dataset [7]. The huge quantity of data is a superb undertaking to classification project. Due to the fact if the dataset has large amount of attributes it can preference a significant quantity of parameters throughout the category manner. Preferably, every attribute utilized as a part of the classification manner have to included ballance collection of statistics. But, mostly attributes are strongly correlated, it increases stage of redundancy inside the dataset which can negatively have an effect on classification accuracy.

Similarly, the multivariate technique may cope with redundant and inappropriate attributes, it complements the precision of the classification in contrast to univariate based totally feature selection techniques. Minimum-redundancy–maximal-relevance (MRMR) is a popular multivariate approach it turned into presented by [8]. Another paper proposed any other method for measuring the correlation between a discrete and non-stop function [9]. They offered a filter method for feature selection. the proposed technique chooses a function subset by casting off inappropriate factors as indicated by way of the correlation most of the characteristic and target characteristic, and getting rid of redundant features most of the applicable attributes. The benefits of filter strategies are that it is computationally mild and is proven to paintings well for certain datasets [10].

III. PROPOSED METHODOLGY

Our proposed model is showing in Figure 1. It includes Dataset, Pre-processing, Semantic similarity, Mutual information, Ant colony algorithm and at the end Optimal features.

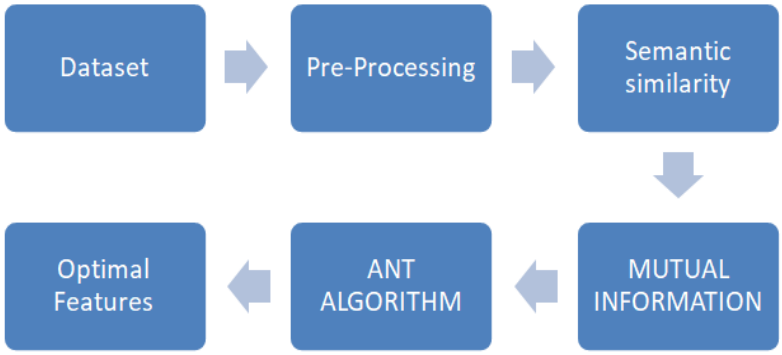

Figure 1. Proposed model.

A. *Dataset*

Datasets are very important in any experiments. We are using two different types of datasets for feature extraction. They are tweets from twitter and product reviews datasets. We extracted the 10,000 tweets from twitter using Twitter API. Product review datasets are the benchmark datasets very much used by researchers [11]. Product review dataset consists of Computer reviews, wireless router reviews and speaker reviews.

B. *Pre-Processing*

When data set is preprocessed then it contain raw data .In this dataset raw materials, and bag of word. Mostly, deferent dialects utilize specific preprocessing procedures on account of syntactic and morphological reasons. The objective of this stage is to diminish intentional types of words to a typical base frame. In this segment the essential preprocessing methods are examined.

Tokenization is the procedure of gap the records into tokens and expels the undesirable crude characters. A similar procedure must be connected to content informational index and inquiry to guarantee that an arrangement of characters in content will coordinate a similar succession wrote which is specifying in the question [12].

Some words are present in document's which creates small impact when features are extracted from text documents. These word extracted from text dataset and we called these stop words. To extricates the rundown of stop words (stop list) the terms in the content record gathering are arranged by accumulation recurrence (number of events of terms in archive gathering), and the most regular terms with pretty much nothing or none semantic esteem in respect to the area of the content reports are then disposed of. Semantic comparability of content reports must be considered while choosing the stop words [13].

C. Semantic Similarity

Lexical semantics is the region of NLP in which we ponders the significance of the words. As we realize that, there are a ton of words that have more than one importance. Consider for instance the accompanying sentences:

- I went by Transport from Rawalpindi to Lahore.
- Ahmad and Ali prepare each day at the exercise center. The word prepare has deferent significance in the above sentences.
- A progression of associated Transports carriages or wagons moved by a train or by vital engines.
- A man fit by appropriate exercise, eat less carbs, hone, and so on, concerning an athletic execution

Along these lines, we can state that there are many words that have spelling similarly however contain deferent faculties. For the most part, a sense is one of the conceivable implications of a given word. In the event that two deferent kind of faculties word are not semantically related between them we are discussing a homonymy connection, as the say case with transport we just observed. Something else, if two faculties of a word are semantically related we are discussing polysemy. Consider the case of the word Creature:

- A classification of a creature.
- A geological range with numerous creatures.

Finally, it can state that ideas which we talk about identified with some importance are semantically comparable. For instance, steed and jackass are more semantically related than cycle and transport. In any case, in the event that we contrast stallion and jackass and cycle and transport then comparability between them can't be clear: steed and jackass are creatures, and both cycle and transport are on wheels methods for transportation, with the goal that the two sets of words are greatly related between them.

D. Mutual information in collection extraction

Pointwise Mutual Information is additionally one of the correct affiliation measures in feature extraction. Pointwise Mutual Information was brought into etymology [14]. Therefore, in the computational semantic writing, PMI is regularly alluded to as just MI, while in the data theoretic writing, MI alludes to the arrived at the midpoint of measure. In this case in Table 1, the bigram, Mr. President gets total score of I (Lmr=yes, Rpresident=yes) = 4.972. In the Europarl test of 20k sorts, Mr. President listed on 1573th as far as Pointwise Mutual Information. Although MI and PMI are related, their behavior as association measures is not fundamentally the same as. A perception frequently made about PMI is that low recurrence occasions get generally high scores. For example, occasional word sets have a tendency to rule the highest point of bigram records which is positioned after Pointwise Mutual Information. One this method its conduct can be comprehended is by taking a gander at the Pointwise Mutual Information estimation of extraordinary cases. At the point when both sections of a bigramme just happen with each other, here it contain $p(x,y) = p(x) = p(y)$. In this circumstance, Pointwise Mutual Information contain an estimation of $-\ln p(x,y)$. On the off chance that we take a gander at where the MWE 2008 shared assignment comes about, here it carry the reason that Pointwise Mutual Information performs moderately good as an affiliation measure in those situations where exposed event recurrence does not. Therefore, there are some feature extraction assign where similar features of a relationship with event recurrence is engaged their property. Mutual Information cannot su□er from an affectability at very low recurrence information, as it is a normal of Point Wise Mutual Information weighted by$p(x,y)$ – as $p(x,y)$

The effect of the expanding Point Wise Mutual Information on the normal turns out to be less. Indeed, in this sort of information we have in feature extraction and can be expected the upper bound of Mutual Information to be emphatically related with recurrence. Mutual Information rises to the decline of the two marker factors when they are consummately related. It's most extreme along with these lines maximum for all the more equitably circulated factors. In Table 1, by a wide margin more likelihood mass is in the base right (Lw =no,Rv =no). It takes after that entropy, and hence maximal Mutual Information, is (marginally) higher for mixes that happen all the more regularly. Similarly as with Point Wise Mutual Information, in any case, the absence of a fixed upper headed for MI means that it is simpler to translate it as a measure of separation to 0 than the measure of correlation.

E. Ant Colony Optimization

Gotten ideas taken from ant rummaging ACO calculation utilized as a part of the subterranean insect include determination calculation [15]. Like the first ACO calculation, various manufactured ants utilized to constant develop arrangements in the given calculation. Be that as it May, rather than collecting pheromones, as the first ACO calculation does, the proposed calculation evaluates the pheromone powers at every emphasis.This will support investigation and diminish the likelihood of being caught in nearby minima. Furthermore, not at all like the first ACO that assembles consecutive arrangements at every cycle, the proposed calculation just changes few highlights in subsets that are chosen by the accurate ants. This will diminish the computational intricacy as the measure of the chose highlight set gets bigger. A half breed assessment measure is utilized to gauges the general execution of subgroup and in addition the nearby significance of highlights. An order calculation is utilized to gauge the execution of subsets (i.e., wrapper assessment work). Then again, the nearby significance of a given component is computed utilizing by the given three channel measures depicted in the past segment (FC, MIFS or MIEF).There are some parameters are used in this algorithm:

- n: shows the number of features that can constitute from the original given set, F = {f1, …, fn}.
- na: it shows the amount of artificial ants to search from the feature$\alpha = 0.3$, $\beta = 1.65$ and $\gamma = 3$, these all are found to be most valuable choice from this and other classification tasks space.
- Ti: intensity of pheromone trail linked with feature fi.
- Sj= {s1, sm}: a list which carry the selected feature subset for ant j.• PL: previously tested subsets list
- k, best k subsets (k <na) used to influence the feature subsets of the next iteration.
- BL: list of the best k subsets. In the starting iteration, every ant will randomly choose a subset of m features. In the next and following iterations, every single ant will start with m – p features that are randomly selected from the previously selected k-best subsets, where p is an integer that contain the ranges between 1 and m – 1. In this method, the features that constitute the best k subsets will be a more chances to be present in the subsets of the next iteration. Anyhow, it still be possible for each ant to consider other features also. For instance, ant j will consider those features that contain the best compromise between previous knowledge, for example, pheromone trails, and local importance.

IV. RESULTS

In this chapter, we report our experiment and obtain results. The experiment was performed on different data set with whole attributes and after performing extraction by using our proposed technique. Experiment results are calculated on the basis of accuracy. Experiment result is shown in table 1.

TABLE I. REULTS

#	*Dataset Type*	*Dataset*	*Reviews/Tweets*	*Accuracy (%)*
1	Product	Computer reviews	531	87.40
2	Product	Wireless router reviews	879	85.90
3	Product	Speaker reviews	689	86.20
4	Twitter	Tweets	10,000	84.60

Our proposed technique for feature extraction and selection shows very promising results. This feature extraction is very important for some data mining tasks such as sentiment analysis, as it improved feature extraction also improves the results of many data mining tasks. Proposed technique also shows the stability as it also works wells on large datasets but achieving high accuracies.

V. CONCULSION

The main objective of this research is to create Scalable feature extraction algorithm based on Aunt Colony Optimization Technique. Our result shows that the suggested technique produces more significant and valuable features from large datasets. These high quality features improves are results other many tasks of data mining. In future our proposed approach may be improved by improving some pre-processing techniques and also by using the improved Ant colony optimization.

REFERENCES

[1] K. Thearling. "An introduction to data mining.", 2017.
[2] A. Bandhakavi., N. Wiratunga, D. Padmanabhan and S. Massie, “Lexicon based feature extraction for emotion text classification”. Pattern recognition letters, pp.133-142, 2017.
[3] A. Chauhan and A. L. Hughes, “ Overview of Event Based Resources on Facebook and Twitter”: Fort McMurray Wildfire, May 2016.
[4] T. Yamashita, N. Nakashima and S. Hirokawa, “Classification and Feature Extraction for Text-based Drug Incident Report”, In Proceedings of the 2018 6th International Conference on Bioinformatics and Computational Biology , pp. 145-149, 2017.
[5] B. Ramesh, C. Xiang and T. H. Lee, “Shape classification using invariant features and contextual information in the bag-of-words model”, Pattern Recognition, pp.894-906., 2015.
[6] S. Kashef and H. Nezamabadi-pour.,“An advanced ACO algorithm for feature subset selection”, Neurocomputing, pp.271-279, 2015.

[7] A. T. D. Souza., V. J. Vieira, M. A. D. Souza., S. E. Correia., S. C. Costa, and W. C. D. A. Costa., "Feature selection based on binary particle swarm optimisation and neural networks for pathological voice detection". International Journal of Bio-Inspired Computation, pp.91-101, 2015.
[8] H. Peng, F. Long and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy", IEEE Transactions on pattern analysis and machine intelligence, pp.1226-1238, 2005.
[9] J. Novakovic. "Toward optimal feature selection using ranking methods and classification algorithms". Yugoslav Journal of Operations Research, 2016.
[10] J. R. Vergara and P. A. Estévez, "A review of feature selection methods based on mutual information", Neural computing and applications, pp.175-186, 2014.
[11] C. Manning, M. Surdeanu, J. Bauer, J. Finkel., S. Bethard and D. McClosk, "The Stanford CoreNLP natural language processing toolkit", In Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations, pp. 55-60, 2014.
[12] C. Manning, M. Surdeanu, J. Bauer, J. Finkel., S. Bethard and D. McClosk, "The Stanford CoreNLP natural language processing toolkit", In Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations, pp. 55-60, 2014.
[13] M. Chaovalit., P. Raghavan and H. Sch¨utze, "Introduction to InformationRetrieval.Cambridge University Press, New York, NY, USA, 2008.
[14] P. Isola, D. Zoran, D. Krishnan and E. L. Adelson, "Crisp boundary detection using pointwise mutual information", In European Conference on Computer Vision , pp. 799-814, 2014.
[15] M. A. Mellal and E. J. Williams, "A survey on ant colony optimization, particle swarm optimization, and cuckoo algorithms", In Handbook of research on emergent applications of optimization algorithms, pp. 37-51, 2018.