International Journal of Applied Information Systems (IJAIS) ISSN : 2249-0868
Foundation of Computer Science FCS, New York, USA
Volume 12 - No. 25, November 2019 - www.ijais.org

A Proposal of Weight based Similarities Hybrid Algorithm on Social Media Posts through Crowdsourcing to Achieve High Performance Recommendation

Fayza Amreen, Md. Golam Muktadir, Tonmoy Hossain and Nazmus Sakib
Department of Computer Science & Engineering
Ahsanullah University of Science and Technology
Dhaka, Bangladesh

ABSTRACT

Recommendation based system on social media posts through crowd-sourcing is an ambitious task. This paper has formulated a hybrid algorithm acquaint with a new approach which is based on weight-based similarity to classify the social media posts as a positive or negative directional. In this paper, it has been proposed a scheme to use social media platforms taking crowd-source reactions and gathering information from the comments and posts by a user. The initial base post is generated by using the Raindrop algorithm where the credibility of the user account is factored as weight. The reaction from this base post can be used to determine whether the community is accepting the information of the base post or rejecting it with a negative impression. To find positively relevant comments and posts regarding the base post, the MinHash algorithm is used. Firstly, using substantial steps of Natural Language Processing (NLP) for pre-processing the data. Then the generalized MinHash algorithm is used to extract the relevant data from all the comments and the posts with the signature. Finally, Longest Common Subsequence (LCS) Algorithm is implemented to categorize the supporting most similar data, thus the post that triggered by the user will get the directional data from the relatively matched comments from the shingles.

Keywords

Crowdsourcing, Data Mining, Natural Language Processing (NLP) , MinHash, Confusion Matrix, Raindrop, Longest Common Subsequence (LCS) Algorithm

1. INTRODUCTION

Adam and Eve are the first human beings, since then socialization has begun. Humans are social beings in nature. In this process from ancient society, people are living and sharing not only for their safety but also enrich them also. In this phase of the civilization era, people are more likely to tend towards socialization. They want to share their beliefs, expressions, excitements, values with the existing world. For this, crowds are trying to develop a convivial or friendly relationship with others. Despite that, almost each and every class of society, communities are getting more dependent on this kind of platform where they can express every bit of imagination related to their daily lives. This paper is articulating about a platform where people are spending a meticulous amount of time Social Media. Nowadays social network has given the scope to make the socialization process on a broader scale. Today, people are being connected not for getting news and views along with making choices and options through voting and sourcing. Crowdsourcing is day by day being a buzz word in the internet market, perhaps within it, no giants are carried forward to future technology. Social media is the most embraced and significant platform where people can accomplish, connect, share and do lots of things.

Living in the most technologically advanced generation, every aspect of life is affected by digital media. As the availability of the Internet extends its reach to every corner of the world, people are getting used to relying on the Internet for almost everything. It has become a source of entertainment in a way to meet the necessity. Therefore, to get any kind of information, people usually take help from the Internet asserting that social media has become a prominent platform to get information. In every second, on average, around 6,000 tweets are tweeted on Twitter, which means approximately 500 million tweets are posted per day [1]. Therefore, social media is the perfect platform for crowdsourcing data for information. It is cost-effective and enough participants are willing to contribute as a crowd. People need to know how news or information impacts society and may predict the impression the news causes on society but there is zero credibility of these predictions. So, automating this system would be more conceivable to accept the aftermath.

Crowdsourcing is a model from which we can glean information and services related to our studies. It is mostly used to expand the capabilities of the model, find a solution to unaccountable technical challenges. For our proposed model, crowdsourcing is the perfect mechanism by which we can accumulate data to train our model. To assemble information related to a piece of news or incident, a social media platform can be used very effectively [2]. Currently, social media is one of the primary means of expressing ones opinion to share with everyone around the world who has access to the Internet. Nowadays, people using social media are more willing and comfortable to give an opinion or information rather than in the physical world. For any kind of information gathering through social media, the first important thing is to generate a post on social media, which will act as the base post for target information. The base post must be generated according to proper calculation. Raindrop algorithm is a method of solution for the non-linear equation. Factoring the verification of the account, duration of the account and, friends or followers of the account as weight, the base post can be generated. Crowdsourcing informative data from the posts or reactions of social media users on the base post can be used to generate an efficient dataset to extract information relevant to an event or statement.

Prediction of the acceptance of the society for any news can be assured using the Confusion Matrix. Implementing Confusion Matrix on the dataset of reactions gathered from crowdsourcing can give the result whether the post is accepted by the society or rejected with a negative impact. From the generated dataset, the positively relevant data to the base information can be extracted, and discarding negatively related posts, using many methods. But before this extraction, each datum from the dataset has to be pre-processed by several steps of natural language processing. For extracting the appropriate data, the data needs to be pre-processed for employing methods of NLP and its sub-fields. Each datum has to be broken down into tokens by the NLP process and only then the relevancy can be analyzed. MinHash is a probabilistic hashing algorithm that can be used to determine the similarity between two sets. In the generalized MinHash algorithm, the Jaccard Index is used to represent how relevant two sets are. This can also be used to extract which data from the dataset are relevant to the event or statement. Finally, the Longest Common Subsequence algorithm can extract the most similar posts from the dataset using threshold.

In the following sections, literature review, a brief proposed methodology and conclusion with the future plans will be discussed.

2. LITERATURE REVIEW

A few existing articles along with some basic knowledge reviewed to build a competent model for the recommendation. In this section, we will present a brief description of the basics behind the model.

2.1 Raindrop Algorithm

This Algorithm is an iterative process to find the global optimal solution of a non-linear program, which is inspired by the nature of the raindrop. The optimal solution is determined by using Random Walk model in this algorithm. If *N* raindrops falls on the ground *S* and the location of i^{th} raindrop is denoted by $x_i \epsilon S$, we can assume that after the raindrops fall, they will move at each time interval to reach the local optimal point [18]. Considering that the raindrops have restriction in movement, they can only move to one of *2n* directions where *n* is the number of dimensions [3]:

$$d_{i,k} = \begin{cases} v_i e_k & k \leq n \\ -v_i e_{k-n} & n < k \leq 2n \end{cases} \tag{1}$$

where $[e_1, e_2, ..., e_n]$ = In with I_n an $n \times n$ identity matrix. Here, v_i is the velocity of raindrop *i*. Each time interval the direction is changed based *n* the optimal location. If the raindrop is moving toward the optimal location, the direction does not change, but the velocity of the raindrop decreases by half. If the direction is not right then the direction is of the raindrop changes to the following equation-

$$d_{i,j+1} = arg_{d_{i,k}} minf(x_{i,j} + v_i d_{i,k}) \tag{2}$$

where *j* is the time interval. Finally, when the raindrop reaches the local optimal location, it gains the velocity of zero and stops [3]. Global optimal solution is one of the local optimal solution found by the raindrops.

2.2 Crowdsourcing

A thousand minds are better than one. Crowdsourcing is a phenomenal method of using many minds to find a solution to a problem. Crowdsourcing is not a new approach but it has been reinvented by Jeff Howe and Mark Robinson in 2005 [4]. Crowdsourcing is an approach to outsourcing. It outsourcing tasks to a large and undefined crowd through an open-call. By crowdsourcing, useful information can be extracted efficiently from data acquired by the crowd participants. Crowdsourcing has changed the paradigm of a business approach to software development. It is one of the significantly low costing approaches to complete a task as basic as collecting data to solve crucial problems like image processing.

The Internet is the main enabling factor of crowdsourcing every internet user may become a potential contributor. Crowdsourcing is evolving as a distributed problem-solving model [5]. It is mainly depended on the intelligence and contribution of a large number of people. Crowdsourcing can be a mean of validating, modifying and improving a hypothesis [6]. There are three main factors of crowdsourcing:

—**Requester:** Requester is the person who is publishing the task or the problem that needs a solution.

—**Crowd:** Crowd of people is the responders of the problem published by the requester.

—**Platform:** And finally, a platform where these interactions between requester and crowd take place.

There are some issues in crowdsourcing that needs to be taken under consideration:

—**Quality:** Not everyone in a crowd has the same level of expertise, therefore it tends to create a quality management issue. There may be people intentionally gives the wrong answer or not have enough training. To maintain the quality of the data, a standard quality answer is inferred from the noisy dataset.

—**Cost:** Crowd is not necessarily free. Which means crowdsourcing can be costly. When using crowdsourcing for a task, the cost control factor needs to be taken into consideration.

—**Latency:** Crowds involvement can cause excessive latency. The crowd may be distracted or unavailable for the task or the task may not be appealing to the crowd. If the requester has a time constraint, then latency in crowdsourcing is an important issue.

2.3 Data Mining

Every feature of the current world is factored by data. But the main concern is whether useful and meaningful information is extracted from data. The term “Data Mining” develops from this query. Data mining is defined as a process extracting hidden information from the data set which was previously unknown and is potentially useful [7]. It has many advantages, such as it provides useful information that queries and reports which unable to provide us efficiently. This technique can also be explained as finding the correlations in a large relational database based on the different depth of angles.

However, Data mining is composed of several phases. The first phases are used for data pre-processing where data is prepared in a format for further use and the rest are used to work on the data where hidden information is retrieved. In the modern world, the massive use of the internet and computer has provided an endless stream of data for any given requirement and the term Big Data has come as an outcome from the massive use. Information on Big data is collected from any complex social network. However, the study of data mining from social networks is of great value [8]. Crowdsourcing data can reflect various patterns of human thought processes, experience, and activities due to having social, economic and cultural information in it. The process of data analysis is very important for the focused mining data for specific information from data produced by social networks.

2.4 Natural Language Processing (NLP)

NLP is a tract of Artificial Intelligence and Linguistics, which analyzes human language computationally. It is a major factor associated with the branch of science, which focuses on the development and improvement in the process of learning the human language as a machine [9]. Natural language processing is a branch of computer science and artificial intelligence which is concerned with the interaction between computers and human languages. Natural language processing is the study of mathematical and computational modeling of various aspects of language and the development of a wide range of systems. These include the spoken language systems that integrate speech and natural language.

Natural language processing has a role in computer science because many aspects of the field deal with linguistic features of computation [10]. Natural language processing is an area of research and application that explores how computers can be used to understand and manipulates natural language text or speech to do useful things. The applications of Natural language processing include fields of study, such as machine translation, natural language text processing, and summarization, user interfaces, multilingual and cross-language information retrieval (CLIR), speech recognition, artificial intelligence (AI) and expert systems processing language [11]. Understanding a language is hard, so it is done by breaking a sentence in a token step by step and understanding each token [12]. The steps of NLP that we adopted can be pipelined as:

—Sentence Segmentation
—Tokenization
—Lemmatization
—Dependency Parsing
—Named Entity Recognition (NER)
—Coreference Resolution

2.5 MinHash Algorithm

Processing large-scale data, finding similarity, distance and data computation is an elementary research area of this modern phenomenon. In recent times, the hashing techniques have been certified to efficiently conduct similarity estimation in terms of both theory and practice. MinHash (mine-wise independent permutations) technique is usually used to estimate similarities between the two sets and weights [13]. MinHash is generalized to estimate the generalized Jaccard similarity of weighted sets. Let *U* be a set and *A* and *B* be subsets of *U*, then the Jaccard index is defined to be the ratio of the number of elements of their intersection and the number of elements of their union [14]:

$$J(A, B) = |AB|/|AUB| \tag{3}$$

This value is 0 when the two sets are disjoint, 1 when they are equal, and strictly between 0 and 1 otherwise. Two sets are more similar when their Jaccard index is closer to 1. In this work, the concentration is on finding the percentage of similarity between two sets that one is considered as hashtags and another one is the user feedback against the task message. MinHash generates a score against each of the feedback and a threshold value is fixed to filter feedback initially based on their similarity score and these strainer outputs are used in raindrop algorithms to retrieve more relevant feedback precisely [15].

2.6 Confusion Matrix

In the field of Text data mining, Text classification has become one of the most important techniques. The task is to automatically classify data into predefined classes based on their content. In classification problems, higher accuracy in classification is the primary concern. However, the identification of the features having the largest separation power is also important [16]. Even more, this is mainly because the larger the number of attributes, the sparser the data become and thus many more training data are necessary to accurately sample such a large domain.

Text classification has become one of the most important techniques in text data mining [17]. Data is classified into predefined classes based on their content automatically by this technique. Higher accuracy in classification always remains as a primary concern in classification problems. But the most important thing is the identification of the features having the largest separation power [18]. There is a proportionate relation between the attributes and data. The larger the number of attributes, the sparser the data becomes. As a result, for very large data sets (such as Twitter response), the classification is highly dependent on feature selection. However, many more training data are necessary to accurately sample such a large domain. A confusion matrix of size $n \times n$ associated with a classifier shows the predicted and actual classification, where n represents the number of different classes. The Confusion matrix assumes that the occurrence of each word in a document is conditionally independent of all other words in that data given its class.

2.7 Longest Common Subsequence (LCS)

LCS is a technique to find the longest common sequence in the given information or data. To find the approximate identical information, we have operated the LCS algorithm [19]. If *A* is a sequence of length *p* and *B* is another sequence of length *q* then we have to look at each and every possible subsequence from *A* whether it is a subsequence of *B* [20]. We employed the dynamic approach of this algorithm to reduce the time complexity of our proposed model. The time complexity of the algorithm is *O (pq)* where *p*, *q* are the length of the strings or data [20]. Thus, we can find the most common patterns or indistinguishable information by which we can perform classification.

3. PROPOSED METHODOLOGY

The proposed system designs an efficient Data Mining technique that can support analyzing a post on social media in a manner of society accepting the event of the post in a positive demeanor or in rejecting it with a negative attitude. Afterward, special methods are constructed to identify the comments which are positively relevant to the event of the base post and distinguish which ones are negatively relevant.

3.1 System Architecture

Using the social media platform, a user posts a statement or a piece of eventful news including distinctive keywords. As social media is a place of virtual socialization, there are enough public network users to form an appropriate crowd. The reaction from a user is used to determine the factor of how the community is accepting the post. So, reaction is used as a feature by which we can find the discrepancies. The comments using the particular keywords are collected to form an effective dataset to use for assembling relevant comments segregated by whether it is positive or negative. Figure-1 depicts the proposed methodology hybrid algorithm. Following are the steps by which we can generate comments and classify them-

3.1.1 Data Collection:. Day by day data is promulgating exponentially. Because Social Media consists of a huge amount and various types of data, we glean data from this platform.

Firstly, the post generation, we enumerate three factors of the Raindrop Algorithm weight, direction and average velocity. Weight can be calculated considering three properties of a user the creation of the social media account, user's social media rating and verification of the account. For the direction of the post, we can segregate the post in a positive or negative post. And, the velocity of the post can be measured by considering the number of comments and how fast a post reaches to the people's. Finally, we will create a post based on these factors. Here, three major factor has been focused. The nature of the user as the ratting of the user, follower of that particular user or year of use

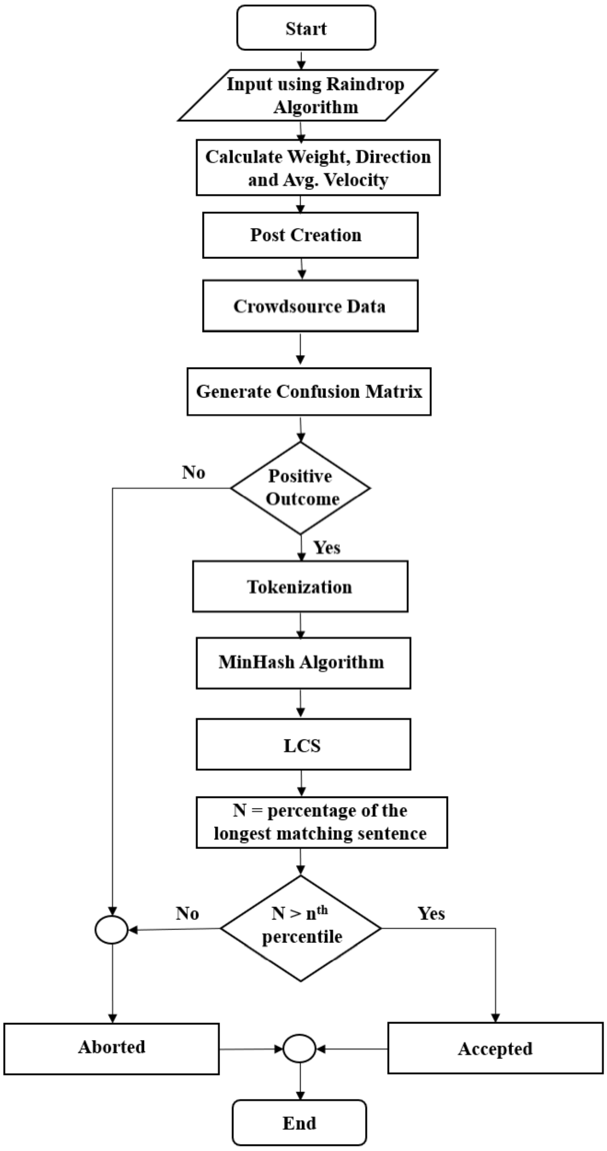
Start
Input using Raindrop Algorithm
Calculate Weight, Direction and Avg. Velocity
Post Creation
Crowdsource Data
Generate Confusion Matrix
No
Positive Outcome
Yes
Tokenization
MinHash Algorithm
LCS
N = percentage of the longest matching sentence
No
N > nth percentile
Yes
Aborted
Accepted
End

Fig. 1. Proposed Methodology of the weight based similarities hybrid algorithm

in the social media. When the post is been made, the keywords also to be thought of shingles of the future NLP works. User also focus on the zone of the comfort about the comments as the user is expecting the vibration about the nature of information. From there the post will be ready using the raindrop algorithm.

3.1.2 Confusion Matrix Generation:. After creating a post based on the three factors of the Raindrop Algorithm, a confusion matrix will be generated. From the confusion matrix, we get the percentage of True Positive, True Negative, False Positive, False Negative, the Precision, and the Sensitivity. From there, it is determined whether the society accepts the base post positively or have negative impression over it. We will accept the data if the outcome is positive otherwise reject the data for further processing.

3.1.3 Pre-processing and Classification:. At first, pre-processing is done on the positive outcome data. For the pre-processing we adopt tokenization. Tokenization is a process of replacing data with identical identification symbols that maintain the structure of the data. After that, the MinHash algorithm will be applied to the data. MinHash technique is used to find out the similarity between the user response and the post. Jaccard similarity is calculated to sort the feedback comments and posts in order of relevancy. Jaccard Index is calculated using the tokens form each feedbacks and distinctive hashtags from the base post. The comments which has a higher Jaccard Index or score are more relevant. Taking threshold value, theta (θ)= 0.75, the feedbacks which has a higher or equal score than the threshold are filtered to the relevant data. Then, we find the longest common subsequence of the user response and the generated pre-processed post.

```
Algorithm 1: Weight based Similarities Hybrid Algorithm
   Input: InputText, Comments, HashTag
   Output: Relevant Comments
   Data: Crowdsource from Social Media
 1 Fetch data from the Crowd or Social Network
 2 Enumerate the Confusion Matrix
 3 Check if True positive is greater than False positive, if YES than goto
     step-4 otherwise goto step-10
 4 Find the Tokens by applying Tokenization
 5 Operate MinHash Algorithm to generate Signatures
 6 Find the individual weight of the data
 7 Apply Least Common Significant (LCS) Algorithm to find the relevant
     comments and generate the performance
 8 Check if Performance if greater than nth percentile, if YES then goto-9
     otherwise goto step-10
 9 Accept data
10 Abort data
```

Fig. 2. Algorithm of the proposed model

Employing the LCS algorithm, we will find the most common substring between the user post and generated post as the inessential words should be negligible and eliminating them before extracting the efficiency can increase the performance of the comparison In the flowchart, *N* is the percentage of the longest matching sentence. Hence, we compare the percentage with a percentile which defines the approval performance of the post. If *N* is higher than the percentile, then we approve the post otherwise abort the post. Figure-2 represents a brief working flow of the complete methodology.

4. CONCLUSION

In this work, a confusion matrix defines the post-acceptance from the reductions given by the users. We apply natural language processing to extract words and phrases from the text which is used in the MinHash algorithm to determine the similarities between the shingles and creating a meaningful signature. Finally, LCS algorithm is applied to the outcome dataset from MinHash to find the optimal result. We have already implemented NLP on the text and MinHash algorithm. The similarities between the shingles vary by the number of keywords inserted for searching. The original raindrop algorithm implemented to find out the weigh vector of the post but the percentile formulation with that weight vector could not plotted yet on this work due to lack of time and we will work on it in the future.

5. REFERENCES

[1] https://www.internetlivestats.com/twitter-statistics/. Accessed on 8^{th} October, 2019

[2] Li, Guoliang & Wang, Jianan & Zheng, Yudian & Franklin, Michael. (2016). “Crowdsourced Data Management: A Survey”. *IEEE Transactions on Knowledge and Data Engineering. 28. 1-1.* 10.1109/TKDE.2016.2535242.

[3] Z. Wei, “A Raindrop Algorithm for Searching the Global Optimal Solution in Non-linear Programming”, *Cornell University, 2013.*

[4] J. Howe, “The Rise of Crowdsourcing,” *Wired Magazine, vol. 14, n4. 6, pp. 14, 2006.*

[5] Brabham, Daren (2008), “Crowdsourcing as a Model for Problem S5lving: An Introduction and Cases”, *Convergence: The International Journal of Research into New Media Technologies, 14 (1): 7590*

[6] Antonio Ghezzi, Donata Gabelloni, Antonella Martini, Angelo N6talicchio, “Crowdsourcing: A Review and Suggestions for Future Research”, *International Journal of Management Reviews, Vol. 00, 121 (2017).*

[7] Kantardzic, Mehmed (2003). “Data Mining: Concepts, Models, M7thods, and Algorithms”. *John Wiley & Sons. ISBN 978-0-471-22852-3.*

[8] Zafarani, Reza; Abbasi, Mohammad Ali; Liu, Huan (2014). “Social Media Mining: An Introduction”.
[9] A Gelbukh, “Natural Language Processing and its Applications”, *Research in Computing Science, 2010*

[10] D. Hindle, M. Rooth, “Structural Ambiguity and Lexical Relations”, *Computational Linguistics, 1993.*

[11] Chowdhury, Gobinda G. “Natural language processing.” *Annual review of information science and technology 37.1 (2003): 51-89.*

[12] Tomas Mikolov, “Distributed Representations of Words and Phrases and their Compositionality”, *NIPS 2013.*

[13] J. Howe, “The Rise of Crowdsourcing,” Wired Magazine, vol. 14, no. 6, pp. 14, 2006.

[14] Kosub, Sven; “A note on the triangle inequality for the Jaccard distance”

[15] Ioffe, Sergey. “Improved consistent sampling, weighted minhash and l1 sketching.” *2010 IEEE International Conference on Data Mining. IEEE, 2010.*

[16] Sofia Visa, Brian Ramsay, Anca Ralescu, Esther van der Knaap, “Confusion Matrix-based Feature Selection”, *22nd Midwest Artificial Intelligence and Cognitive Science Conference, Ohio, USA, 2011*

[17] Ronnie Merin George and Dr. Jose Alex Mathew, “Emotion Classification Using Machine Learning and Data Preprocessing Approach on Tulu Speech Data”, *IJCSMC, Vol. 5, Issue. 6, June 2016*

[18] B. P. Salmon, W. Kleynhans, C. P. Schwegmann and J. C. Olivier, “Proper comparison among methods using a confusion matrix”, *Geoscience and Remote Sensing (IGARSS), IEEE International Symposium, 2015*

[19] L. Bergroth and H. Hakonen and T. Raita (2000). “A Survey of Longest Common Subsequence Algorithms”. *SPIRE. IEEE Computer Society. 00: 3948.*

[20] Wagner, Robert; Fischer, Michael (January 1974). “The string-to-string correction problem”. *Journal of the ACM. 21 (1)*