RESEARCH ARTICLE

SAMRIDDHI Volume 13, Special Issue 1, 2021 Print ISSN : 2229-7111 Online ISSN : 2454-5767

Usage of "Techniment" Technical Sentiment Value to Determine Whichlogs to Send for Log Analysis Tools to Save Not Only Bandwidth But Also Compute

Ankur Verma*[1], Yogesh Kumar Sharma[2]

[1] JJTU, India; e-mail : anshverma@hotmail.com

[2] Department of Computer Science, Shri J.J.T. University, Churela, Jhunjhunu, Rajashthan, India; e-mail : dr.sharmayogeshkumar@gmail.com

ABSTRACT

In most log analysis tool its designed to feed in all the logs generated. Using "Techniment" which means Technical Sentiment, is an Acronym wrapped consisting of the word technical and Sentiment. This is basically a combination of Sentiment analysis and Classification machine learning algorithm usage. This helps determine a value which can help lower not only network bandwidth but also compute.

Key Words: Techniment, opinion, minion, log, analysis, product, program, tools and text.

SAMRIDDHI : A Journal of Physical Sciences, Engineering and Technology, (2021); DOI : 10.18090/samriddhi.v13iS1.10

INTRODUCTION

Every Entity needs to connect to another entity to perform a task as desired. The likelihood of that persona to be informed of the state of the entity is crucial and hence we will use this technique to determine that state and based on classification solely identify the techniment of that entity.

Because Opinion matters and mining of it is essential for the existence of any entity in the computing world. This is synonymous to Sentiment that is more preferably and commonly used in NLP, analysis of text & computational linguistics. This is done to not only identify but also to extract subjective information in source materials such as other techniment entity in this case. In common analysis based on Sentiment was widely applied to reviews & social media for a variety of applications with an aim of to determine the attitude of a speaker or a writer with respect to some topic, however with techniment we are trying to determine the way a system would respond based on the state it is currently running at the moment.

Corresponding Author : Ankur Verma, JJTU, India; e-mail : anshverma@hotmail.com

How to cite this article : Verma, A., Sharma, Y.K. (2021). Usage of "Techniment" Technical Sentiment value to determine whichlogs to send for Log Analysis Tools to save not only bandwidth but also Compute.

SAMRIDDHI : A Journal of Physical Sciences, Engineering and Technology, Volume 13, Special Issue (1), 42-48.

Source of support : Nil

Conflict of interest : None

"Techniment" is where we dig deeper in determining how a product functions and classify those into known features. We used stochastic dual coordinate ascent .NET library that will help to extract log statements based on features. We used this algorithm as it Doesn't look at the model error function directly, it looks at a dual function and maximizes it (ascent). And the technique doesn't look at all weights at once, it examines one weight at a time (holding the others constant), i.e.,

coordinate. And because the technique samples one or a few training items at a time.

Concepts such as {BYOP} i.e. Bring Your Own Product will help anyone to bring in their own products and plug the data logs into this analytics tool and help provide direct feedback to the end system which will not only improvise the learning but also track the number of times a steps is used in the system so that the new person that will be utilizing this approach/tool will know not only the techniment of the log but also the number of times it has been used i.e. the success factor that will help determine how we can use this to the benefit of the entire ecosystem to the products in the market place or any company that will be able to take advantage of this system.

WHAT IS TECHNIMENT?

Definition of Techniment

Techniment can be defined as a terminology that is used to predict an outcome of a log statement that is been looked at or analyzed. It works on a granular level in trying to not only classify based on the parts of log but also predict the toxic level of what the log line would mean when its analyzed. In Sentiment analysis that data sets play a critical part in the issue in this field. [8] The product reviews form the main sources of data. Business holders find such review important as these reviews are important as they can take business decisions based on the analysis results of user's opinions about their products.

This is a fervent way to find out how a component feels in a system from a technical aspect and the best way to do that is to put those into different clusters or boxes of similar types. Because the statement in a log could be of various type and technical modules mentioned in it, its most likely that two dissimilar logs having a common class to be worded differently that confuses the log reader. Because when support engineer or services personnel aiming to troubleshoot or fix a problem tries to determine the outcome, he/she would most likely struggle and would have to search multiple knowledge base system to help determine what the logs mean to the real world.

Why do we need Techniment terminology?

Windows Events, event tracing, debug logs, performance counters, Product logs and ELK(Elasticsearch, Logstash, Kibana) are already there and tools such as event view already preview visualization of what's wrong with a system with details then why would we need something called as techniment to show product sentiment and what value does it offer. Below is one example of Windows event log that we have considered for this discussion. None of the tool provide a technical sentiment analysis i.e., technical opinion of a particular log statement with respect to the value techniment gives to not only put a similar log in a cluster group name and then further process it provides toxic level of a log statements. These were done on human emitted words such as "I don't like this product" or "I find it hard to use to product" and then machine learning was used to determine the sentiment of a particular user but we have used similar concept in a different way customized for technical products treating them as a individual identity by giving them feature names and then determining the positive and negative output it gives on a per log basis that can further help determine the power of techniment analysis to benefit the log analysis world. This is similar to finding a rusty needle in a soiled haystack.

Every product and its component produce logs that gets logged in plain text file or binary in the windows file system that stories this either indefinitely or in a particular time frame or frequency that it gets overwritten. These provide details envisioned at that particular time. So, if a product component PC1 is unable to connect to product component PC2 using a channel CH1 then we might see something like PC1 unable to connect to PC2 using CH1. Basically, this is either determined by seeing at a log or via the system behavior impacted via this log. This is a very time-consuming process that takes into account.

ROLE OF LOGS

When any software works and is fully operational it's really not required for it to generated an awfully lots of logs as generated logs is expensive operation which means that it takes IO cycles to generated log

that to a system is an expensive operation if done synchronously and usually it is done asynchronously. There are many libraries like log4j, log4net etc which also do the job asynchronously depending on the implementation. However, the key point here is that there is no real need of performing this expensive operation if everything is healthy and working optimally. It's like our health report or diagnostic we would not visit a doctor, or they would not require a report unless something was not quite right about us.

Similarly, if everything is healthy there is no requirement, but if something is not working or working a bit slow then there comes a need for logs. Many real time systems do not really generate logs unless we have a switch or a way to turn them on or off. Because every instance of generating a log and to a more granular level would have performance impact on the working of the system in such a way that it could deviate away from the real purpose of its primary goal whatever it might be.

Figure 1: Flow showing Log generation operation and should be avoided

So even if the logs are time consuming and reduce performance of a running system, we still leave them running but not always in debug mode so that more logs are generated but instead have short information logs enough to understand if something goes wrong and logs events related to high level operation.

Bottom line is that logs should not be in place unless there is a real need for it but unless we enable them, we won't know if we need it or not because when a problem occurs it's the log that help us dig deeper because not always it's possible to have the source code of a running software in debug mode to determine where the issue is, it's only the error code in the log or some unique statements that help to nail down to the exact issue where the problem is i.e., the root cause.

How Logs Are generated, Analyzed and outcome

There are many ways logs can be generated, usually in a piece of code we write logger async classes that output error code along with class name and other useful information from the stack trace that help in overall problem rectification process.

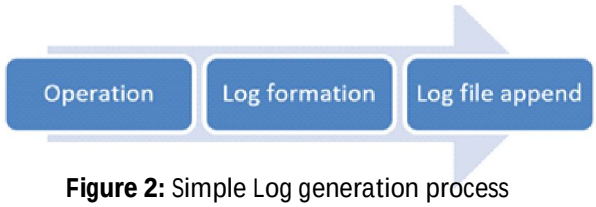

Figure 2: Simple Log generation process

COMBINING SENTIMENT ANALYSIS WITH CLASSIFICATION

Due to the sheer volume of information that is generated in log file its required that logs need to be analyzed and categorized so that the person using the information find it useful. Sentiment analysis combined with classification gives use a quick way to see the problem much more efficiently and this overall process is what we are calling it as techniment.

Table-1: Sentiment Toxic Value

Sentiment	Sentiment Text
1	There was no endpoint listening at http://There was no endpoint listening at http://
1	Error: cannot connect
0	INFO: able to connect
0	INFO: <COMPONENT><Module>is invalid
1	ERROR: network error
1	ERROR: SQL error
1	{Product}. <COMPONENT>`1.Sender Could not connect to <COMPONENT>service 10

A value of 1 indicate negative sentiment and a value of 0 indicate positive sentiment.

The above data is used to train the machine learning module, which helps in prediction if a log statement that belongs to a component or cluster/ feature is positive or negative. The training data could be loaded from a flat file, tab separated or comm separated i.e., tsv or csv. ML.net even allows to load the data from the database. The important point to note is that ML.net allows to load the training data at the time of starting the program so this is once in program lifetime item.

Once the Machine learning algorithm is trained a model is created, once the module is formed, we Evaluate the model and show accuracy stats. Then we Load evaluation/test data. The machine learning pipeline is what creates a model file that is for post consumption by some other program and we do no need to depend on the same program to create the model as the model can be created by another program or even another programming language. Because the format for a particular model of a algorithm is same as per the definition any machine learning program can not only create a model but anyone can consume it.

A lot of time gets spend in analyzing data and moreover enormous amount of system resources is spent to analyze all logs that are categorized to look at only unhappy techniment data or sad techniment i.e., log items gets sent only when relevant to be analyzed and cross referenced.

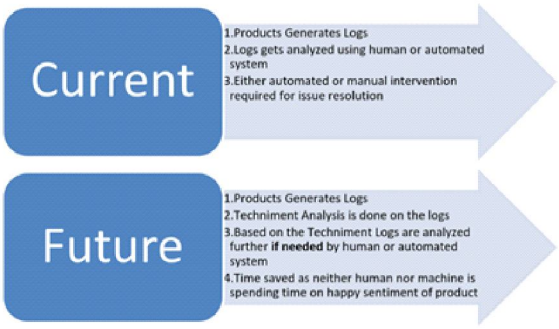

Figure 2: Flow showing Log generation operation and should be avoided

The researcher found that many software products that were implemented in the field provided immense number of logs and it started becoming difficult at the production area where this software were deployed to be able to identify the problem when it occurs. Most of the time anyone with specialized skill resolves the problem on certain system but that goes undocumented. The need for such a research happens because most of the discovery techniques today relies solely on the completely event log being available to carry out machine learning or any mining activity but if something is missing then the activity does not provide the accurate result hence it is important to repair the missing activities in the event log and provide a better data so that analysis could be done more effectively and the result of process mining classification happens much more effectively will stop [3].

The approach we have found out is that we would be able to do use two major machine learning algorithms in combination, one is to do classification and other to do sentiment analysis. However, based on the dataset and the ability to do techniment analysis we would be able to tentatively predict the outcome of possible issues on the system that would need looking into. Currently problem resolution system does not work based on techniment analysis which is a new terminology that we are trying to introduce by this research. The main controversy finding all the missing activities to be able to successfully between them to achieve that real world that distance measure and also the number of events in the cluster overall [4].

The researcher found that many software products that were implemented in the field provided immense number of logs and it started becoming difficult at the production area where this software were deployed to be able to identify the problem when it occurs. Most of the time anyone with special skill resolves the problem on certain system but that goes undocumented.[3] Both synthetic a real-world datasets showed a number of improvement and better result when sorting based algorithm was used for doing the comparison and other scenarios related to it such as 1 activation Of RELU function in the deep neural network and hinge loss that had empirical risk minimization problem. They have consistently improved the top K accuracy, but we all know that the main ingredient in this optimization scheme is the stochastic dual coordinate ascent which relies on the sorting method. The SDCA machine learning algorithm was also part of machinelearning. net provided by Microsoft and is most extensively used algorithm that the.net community is offering with the ML dot NET Framework.

Table-2: Size of Techniment

Log text line size in characters	Techniment size always in bit	Percentage of Techniment	Percentage reduction
84	1	1.19%	98.81%
21	1	4.76%	95.24%
21	0	0.00%	100.00%
24	0	0.00%	100.00%
20	1	5.00%	95.00%
16	1	6.25%	93.75%
110	1	0.91%	99.09%
161	1	0.62%	99.38%
102	1	0.98%	99.02%
47	1	2.13%	97.87%
65	1	1.54%	98.46%
47	1	2.13%	97.87%
11	1	9.09%	90.91%
729	**11**	**1.51%**	**98.49%**

We need to divide the set of log statement from a log file in different groups based on the features of the log statement. Those features are the predetermined and belongs to different modules of a product that we are analyzing. When the log file first comes into the engine, we don't know what it is and hence its feature is unknown. We want to learn the structure of a data set from the features and predict how a data instance fits this structure.

Log size is of a big concern and shipping huge logs across network is very time consuming, imagine all the efforts to be put in that to isolate a issue at hand only to find out that the logs are not relevant which happens a lot in today's scenario. Hence, we need to have a way to not only compress logs but compress meaning from the log i.e. the sentiment aka the techniment of the log statement so that we get a binary bit value of a big log statement.

In today's environment if something doesn't work e.g. We have a problem_statement_1, that we need to fix using solution_1, then we need a combination of resource_1 and tool_1 to look at data_1

Now when data_1 is too big i.e., of size_1 then resource_1 would take time_1 to analyze_1 and find solution_1

Similarly, tool_1 would take time_2 to find solution_1 or solution_2

Now time_2 could be less than time_1 but resource_1 could take more time_1

Thus, to save both time_1 and time_2 we need techniment_1 to help determine the concise data_1 i.e. data_1(techniment) and only pick data_1 if techniment is true or positive indicating the product component in question is exhibiting a negative sentiment.

The time and resource problem at hand is too huge when we have many systems running on premise and also in cloud and to be able to look at that data without a techniment could waste enormous amount of time and energy that could be utilized for other aspects of the product customer front end experience.

Once we build the techniment factor in the product itself, image the product won't be generating logs if the techniment is happy or positive so to save the expensive file system operation.

CONCLUSION

Sample Techniment data set that will be used to determine the state of system can be categorized in multiple ways. This multiple way determines the toxic level of data such that it is for process preventive action on the system that it is trying to process. When the information is processed it will capture the good state helping in supervised learning and a human would accept the feedback marking it done that adds the human factor so that the next person that will be using the system will know it.

This data sent of Techniment and Classification together will jointly be used to determine and predict the output based on each product log statement. System logs of a product will be regularly scanned for any new data and once a data is found it gets broken down into json data set using random GUID file names to ensure its unique a lot of files will be created. These files will be then be displayed on a GUI which will further add human factor to the already classified data massaged with sentimental data and create a hit counter that will help determine how much prediction was not only liked by a human for prediction and executed but also how many ran successful.

Once data is analyzed as much as possible, we extract the known feature that would be basis of our research going forward and will be listed below in a table. This initially would be a static set and later added as per need. however, we wanted to cover on a broad level what kindly of features can exist in a particular product that is installed on premise, i.e. it would have a backend database that it would need to communicate to persist data and it would need other component that it would interact with.[6] concluded that a system can treat boat type of data, whether it is text or number based and using experimental result they were able to figure out that concrete ideas could be achieved in the upcoming framework that involves both techniques of mining data and text. Once a rule is created it became important to identify which part of the text matches to which part of the rule so that the analysis process can split the data into different parts and send it to the respective mining system to avoid any issues altogether in computation.

In simple terms we take the data in the following format in csv format based on the table shared above and assign toxic flag to determine which type of error is toxic i.e., negative to a product health

1,0,0,0,1, SQL-Error

0,1,0,0,1, Cached-Settings

0,0,1,0,1, Some Other

0,0,0,1,1, License

This will help not only categorize the data into various classed but also provide the technical sentiment per sample record. “Classification paper” [5] discusses the many phases in opinion and sentiment analysis experimented by using some new approaches. Those are left for upcoming researchers. Due to the interdisciplinary nature, a successful opinion mining requires expertness in NLP techniques, traditional data mining techniques and domain knowledge of opinions.

We parse the data using “Stochastic Dual Coordinate Ascent” Multiclass classification algorithm where each feature is passed to categorize each type of error that the system generates. “Clustering event logs using iterative partitioning” [1] explains how iterative partitioning can be helpful in clustering the event logs. When an even occurs which is actually a push event rather than a pull event i.e., when an event occurs the subscribers of the event are notified via a remote delivery service and the subscribers are not supposed to pull events on a regular basis in hoping that a problem has occurred. And carefully clustering the event helps overall as this information is passed to the prediction engine.

Table-3: Classification of System Log and Feature Extraction

Features for Data Input Sample				SDCA Prediction
SQL	LICENCE	ERROR	CACHE	
1	0	1	0	Database Error
1	0	1	0	Database Error
1	0	1	0	Database Error
0	1	1	0	{COMPONENT 1 } Error
0	0	1	1	{COMPONENT 2 } Setting Error
2	0	1	0	SQL

In any given log statement of a product, we see a lot of unwanted messages or basically happy or neutral techniment as mentioned in the below table that still gets analyzed or sent to tools for analysis merely wasting a lot of resources over the network and storage that is unnecessary and be avoided by sending logs with a high techniment score as elucidated in the table below. Evaluated the interdisciplinary field which is able to extract information Using various information retrieval techniques, such as mining of data,[2] learning using machine or computer software and linguistic which is based on computational methods. The ultimate goal was to extract high-quality data an information from the information provided so that it is to structure data and analyze it along with adding

some linguistic features to it. This was done using the approach of mining text, for example if you have a PDF file which is having unstructured data it can be converted to an XML, email message, and other forms. Text mining of the data that was preprocessed associated words with their corresponding structure and categorization to get text classification and word Association. If a data stem contains a double suffix, then it is connected with a single suffix and so on. This particular paper focused mainly on K-means an hierarchical agglomerative clustering algorithm methods which was used to form the cluster. As we know document can be clustered into a hierarchical form which then become suitable for browsing. However, this form of algorithm has an efficiency problem. If multiple terms mainly to Co-occur within the same paragraph, they constitute an Association.

Table-4: Sample log Folder Analyzed

Techniment	Name	Count of lines
0	Positive	3501
2	Neutral	1683
4	Negative	7

As we can see that every log file statement that is analyzed forms a techniment score which clearly indicate that only. 10 % is a negative techniment data and that constituent to more than 99.86 % of the file size that we end up analyzing.[7] a good survey was done of most of the text mining techniques and application used in today's world. This helps very much of been able to know how useful this process is and moreover we are able to compare the data obtained from different text mining applications. Without been able to mine and extract useful information we are unable to use that information and make decisions based on those.

ACKNOWLEDGMENT

Would like to take an opportunity to be able to express my profound gratitude to my family, my mother Meena Verma, her sisters and my sister Disha Verma., who has been regular inspiration for all the time.

Words fail me to express my appreciation and absolute indebtedness to my wife Mrs. Soumya Verma for her continuous patience, tolerance and tremendous help. My heartiest thanks to my child AnshVerma for his unconditional love and patience during the research.

REFERENCES

[1] A. Makanju, A. Zincir-Heywood, and E. Milios (2009). Clustering event logs using iterative partitioning. In KDD'09: Proc. of Conference on Knowledge Discovery and Data Mining. Publication: KDD '09: Proceedings of the 15th ACM SIGKDD internatGARSON, G. David. *Guide to writing empirical papers, theses, and dissertations.* New York: Dekker, c2002. 350 s. ISBN 0-8247-0605-6.

[2] Deepak Agnihotri ,Kesari Verma and Priyanka Tripathi (2014). Pattern and Text Data.Fourth on Communication Systems and Network Technologies. DOI: 10.1109/CSNT.2014.92

[3] DejunChu et al.(2018). Optimizing Top- k Multiclass SVM via Semismooth Newton Algorithm. DOI: 10.1109/TNNLS.2018.2826039

[4] Jiuyun Xu and Jie Liu (2019). A Profile Clustering Based Event Logs Repairing Approach for Process Mining. DOI: 10.1109/ACCESS.2019.2894905

[5] Hidayath Ali Baig , Dr. Yogesh Kumar Sharma and Syed Zakir Ali (2020) . Privacy-Preserving in Big Data Analytics: State of the Art

[6] Poongodi S and Radha N (2013). Classification Techniques. Advanced Research in Computer Science and Software Engineering, 5: 9025-9034.

[7] TomoyaMatsumoto ,WataruSunayama , Yuji Hatanaka and Kazunori Ogohara (2017). Data Analysis Support by Combining Data Mining and Text Mining. 6th IIAI on Advanced Applied Informatics (IIAI-AAI). DOI: 10.1109/IIAI-AAI.2017.165

[8] Vishal Gupta and Gurpreet S. Lehal (2009). A Survey of Text Mining Techniques and Applications. JOURNAL OF EMERGING TECHNOLOGIES IN WEB INTELLIGENCE, 1:60-76.

[9] Walaa Medhat, Ahmed Hassan and HodaKorashy (2014). Sentiment analysis algorithms and applications: A survey. Elsevier B.V. on behalf of Ain Shams University. DOI: 10.1016/j.asej. 2014.04.011

[10] Dr. Yogesh Kumar Sharma and Ghouse Mohiyaddin Sharif G.M(2018, Framework for privacy Preserving Classification in Data Mining. [27] 5(9):178-183.