Impact And Significance Of Concept-Drift In Twitter Data

Dr.E.Padmalatha S.Sailekya

Abstract : The exponential growth of data in micro blogs with the breadth of the user base requires, drilling of a relevant topics. As there is a growth in data substantial effort is need to filter for the relevance. Detection of relevant, trending information is fundamental building block in drilling of micro blogs. This will help in managing and summarizing the comprehension of people behaviour in emergency situations to take the decisions. A novel approach towards identifying concept drift by initially grouping topics into classes and assigning weights for each class, and finding trends among that classes using sliding window processing model upon Twitter streams. This paper propose an novel approach towards identifying trending topics where the concept-drift occurs, by initially grouping topics into classes and assigning weights for each class, and finding trends among that classes using sliding window processing model upon twitter streams.

index words : concept-drift, sliding window

1 INTRODUCTION

In today's world communication is mainly through social networking sites like, Twitter, Facebook, and Google+. Huge amount of data that is being generated and shared across these micro-blogging sites, serves as a good source of Big Data Streams for analysis. As the topic of discussion changes drastically, the relevance of data is temporal, which leads to concept-drift. Identification and handling of this conceptdrift in such Big Data Streams is present area of interest. The state-of-the-art techniques for identifying trending topics in such data streams mainly concentrate on the frequency of the topic as the key parameter. In the digital world, things that are repeated become reputable or trendy. Different practice of liking, sharing, commenting are judged not on the basis of the content, but on the repetition of the content. This might explain why the more popular a tweet is; the more popular it becomes which increases popularity of the tweet. This depends entirely on the relevance of the topic at that particular instance. The topic keeps changing its position or stick to its previous position as time advances. The topic sometimes slides up or down the rank as it gains public interest or vice-verse. Twitter was therefore selected as a case of study for Concept-Drift Analysis. Concept-Drift Analysis is an integrated study of identifying and handling Concept-Drift in this evolving stream of data.

Related work

In existing system, act miner is used which uses an ensemble classification technique for data problem and solving the other three problem which reduces the cost. Act miner is extended version of mine class.

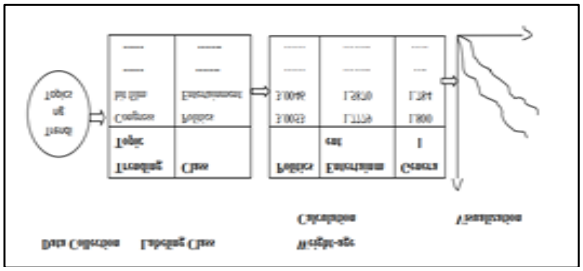
- *Dr.E.Padmalatha is working as Assistant professor in Chaitanya bharathi institute of technology. Email.id :padmalatha@cbit.ac.in*
- *S.Sailekya doing her Masters in block chain technology.*

\

Act miner addresses major problem concept drift. In this method, dynamic feature selection problem and multi class classification in data stream classification based on clustering methods for collecting potential novel instances so memory is required to store. Another disadvantage is that using clustering method first find centroid which also incremental so time overhead occurs. And also not possible classify streamed data continuously. Because of continuous flow of streamed data and classification become continuous task. classify streamed data continuously. Because of continuous flow of streamed data and classification become continuous task.

Methodology

The state-of-the-art techniques for identifying trending topics in such data streams mainly concentrate on the frequency of the topic as the key parameter. Concentrating on such a weak indicator, reduces the precision of mining. Here we propose a novel approach towards identifying concept-drift by initially grouping topics into classes and assigning weight-age for each class, using sliding window processing model upon Twitter streams. Concept-Drift identification System works in four stages. Data Collection is the first step, followed by pre-processing of the gathered topics Second step is to label the topics into four main classes; in the third step class weight-age is calculated for these labelled trending topics by identifying the twelve dynamic class parameters arising from sliding window processing model. In a proposed system twitter data will be collected for trending topics. Collected data is arranged into different classes such as politics, entertainment and general etc by applying data labels.



Proposed Architecture Implementation

Data Collection

Data from tweeter will be collected by following procedure

- Sign into Twitter with twitter account.
- With twitter application management api keys will be obtained.
- then data is pre-processed as follows: (1) removing special symbols, (2) inserting hyphens between words, (3) position was appended to each topic (4) all characters were converted into lower case. After cleaning we get cleaned data.

Labelling

After unique topics were identified it is classified into some categories like Politics, Entertainment, Foreign Affairs, and General Etc based on trending topics. Here manual labelling is done in data editor. The reason for this classification is to reduce the anomalies which arise due to abrupt changes in the trend. So we confined our analysis upon the classified trending topics.

Class Weight-age Calculation

Weight-age for each class was calculated based on the class Parameters identified as shown in Table 3.1. These Class Parameters were identified as Key Parameters for calculating Class weight-age by assigning a weight for each parameter determined empirically and normalized as shown in the Table 3.1. The formula for calculating Class Weight-age is as follows: Class Weight-age:

$$\text{Class Weight-age: } \frac{\sum(parameter\_wt * parameter\_val)}{\sum parameter\_wt}$$

No	Key Parameter	Parameter Weight	Maximum Value (Range)
1	Previous Class Occurrence	50	TRUE 1or 0
2	Class Duration Weight-age	40	1
3	Class Position Weight-age	3	10
4	Updated Class Relevance	2	10
5	Class Relevance	1	10

- Previous Class Occurrence is checked that is whether politics has occurred previously or not and sets the value as follows: -true or 1 if it is occurred -false or NA or 0 (class disappears)

- Class Relevance is calculated by finding class frequency which is nothing but count of occurrence of that class further divided by window size.

$$\text{Class Relevance} = \frac{\text{Class Frequency}}{\text{Window Size}}$$

- Updated Class Relevance is calculated by finding updated class frequency that is count of occurrence of that class which further divide by window size.

$$\text{Updated Class Relevance} = \frac{\text{Updated class frequency}}{\text{window size}}$$

Class Position Weight-age is calculated by finding position weight and position frequency.

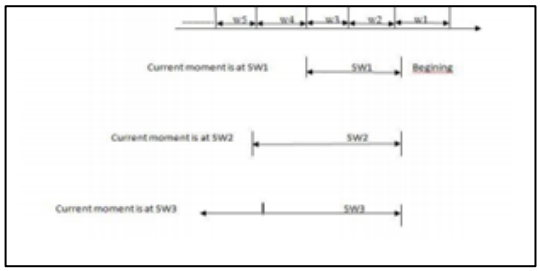
$$\text{Class Position Weight-age}== \frac{\Sigma(pos\_wt * pos\_frequency)}{\Sigma pos\_wt}$$

Class Duration Weight-age i

$$\text{Class Duration Weight-age}= \frac{\text{Current chunk duration+Previous chunk duration}}{\text{Window duration}}$$

s calculated by finding current chunk duration, previous chunk duration and window duration.

After calculating the weights using the above formulas the next step is sliding window process. Sliding window processing model was applied upon modified data-set where the window size was set to 10, 20, 30… 100. The sliding window basic principle is to take a decision on root of current data. Segments of data streams are consider as data strips. The techniques of data stripping called window.



Sliding window structure

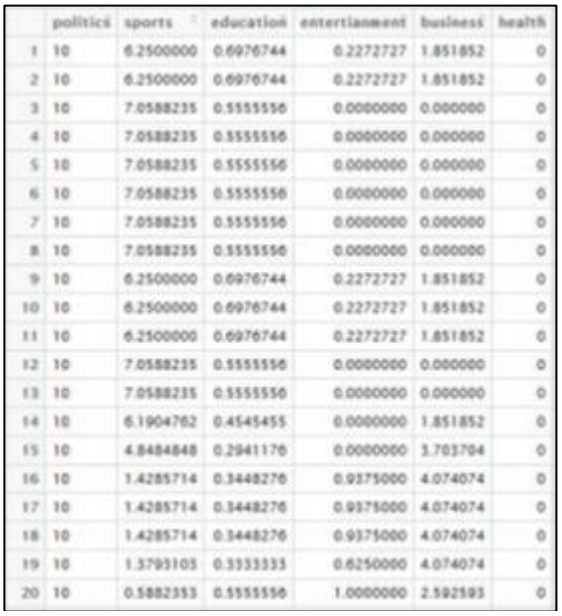
2 RESULTS AND DISCUSSION

The Figure 4.7 is a manual label editor where trend and category column will be appeared. Here in category column manually the category entering will be done depends on trending topic such as politics, sports etc.

	politics	sports	education	entertainment	business	health
1	15	25	3	1	10	NA
2	15	25	3	1	10	NA
3	15	36	3	NA	NA	NA
4	15	36	3	NA	NA	NA
5	15	36	3	NA	NA	NA
6	15	36	3	NA	NA	NA
7	15	36	3	NA	NA	NA
8	15	36	3	NA	NA	NA
9	15	25	3	1	10	NA
10	15	25	3	1	10	NA
11	15	25	3	1	10	NA
12	15	36	3	NA	NA	NA
13	15	36	3	NA	NA	NA
14	16	26	2	NA	10	NA
15	17	16	1	NA	20	NA
16	24	4	1	3	22	NA
17	24	4	1	3	22	NA
18	24	4	1	3	22	NA
19	25	4	1	2	22	NA
20	32	2	2	4	14	NA

Figure 4.*10 Class Relevance (Count)*

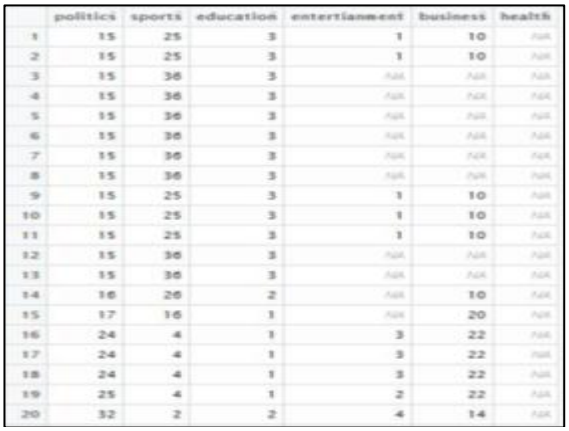
The Figure 4.10 is a class relevance count screen where count of that class depending on occurrence of that class will be displayed. Here class relevance is nothing but frequency count of that class.



	politics	sports	education	entertainment	business	health
1	10	6.2500000	0.6976744	0.2272727	1.851852	0
2	10	6.2500000	0.6976744	0.2272727	1.851852	0
3	10	7.0588235	0.5555556	0.0000000	0.000000	0
4	10	7.0588235	0.5555556	0.0000000	0.000000	0
5	10	7.0588235	0.5555556	0.0000000	0.000000	0
6	10	7.0588235	0.5555556	0.0000000	0.000000	0
7	10	7.0588235	0.5555556	0.0000000	0.000000	0
8	10	7.0588235	0.5555556	0.0000000	0.000000	0
9	10	6.2500000	0.6976744	0.2272727	1.851852	0
10	10	6.2500000	0.6976744	0.2272727	1.851852	0
11	10	6.2500000	0.6976744	0.2272727	1.851852	0
12	10	7.0588235	0.5555556	0.0000000	0.000000	0
13	10	7.0588235	0.5555556	0.0000000	0.000000	0
14	10	6.1904762	0.4545455	0.0000000	1.851852	0
15	10	4.8484848	0.2941176	0.0000000	3.703704	0
16	10	1.4285714	0.3448276	0.9375000	4.074074	0
17	10	1.4285714	0.3448276	0.9375000	4.074074	0
18	10	1.4285714	0.3448276	0.9375000	4.074074	0
19	10	1.3793103	0.3333333	0.6250000	4.074074	0
20	10	0.5882353	0.5555556	1.0000000	2.592593	0

Figure 4.11 *Class Relevance*

The Figure 4.11 is a class relevance screen where it is calculated by a formula. The formula to calculate class relevance is class frequency divide by window size. Here class frequency is nothing but class relevance count which is shown in fig-4.11 and window size means total number of observations i.e. is 18 multiplied by window size that is 3 where the value is 54 which is subtracted by number of empty classes. Again that class relevance is multiplied with 10 to bring it into scale.



	politics	sports	education	entertainment	business	health
1	15	25	3	1	10	NaN
2	15	25	3	1	10	NaN
3	15	36	3	NaN	NaN	NaN
4	15	36	3	NaN	NaN	NaN
5	15	36	3	NaN	NaN	NaN
6	15	36	3	NaN	NaN	NaN
7	15	36	3	NaN	NaN	NaN
8	15	36	3	NaN	NaN	NaN
9	15	25	3	1	10	NaN
10	15	25	3	1	10	NaN
11	15	25	3	1	10	NaN
12	15	36	3	NaN	NaN	NaN
13	15	36	3	NaN	NaN	NaN
14	16	26	2	NaN	10	NaN
15	17	16	1	NaN	20	NaN
16	24	4	1	3	22	NaN
17	24	4	1	3	22	NaN
18	24	4	1	3	22	NaN
19	25	4	1	2	22	NaN
20	32	2	2	4	14	NaN

Figure 4.12 *Updated Class Relevance (Count)*

The Figure 4.12 is an updated class relevance count screen where count of that class depending on occurrence of that class will be displayed after updating. Here updated class relevance is nothing but frequency count of that class which is incremented by 1, when the class appears in the current record. Decremented by 1, when the class disappears till its value is greater than zero and retained as it is when the value becomes zero.

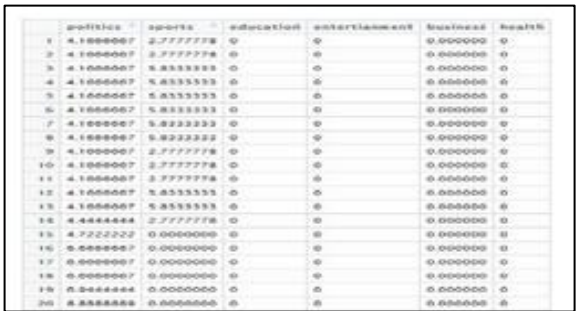
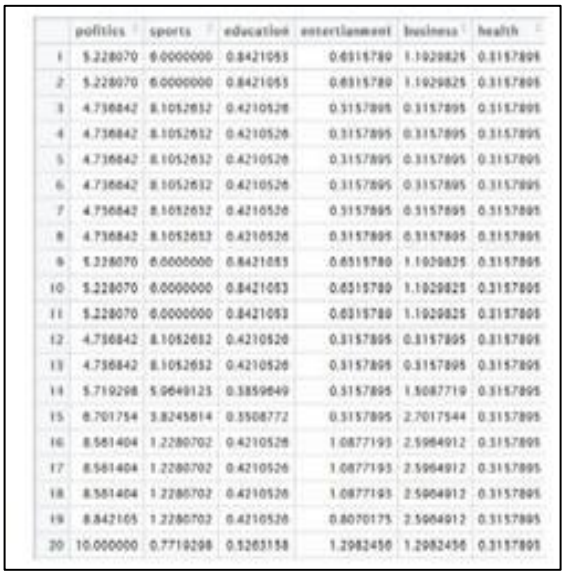


Figure 4.13 *Updated Class Relevance*

The Figure 4.13 is an updated class relevance screen where it is calculated by a formula. The formula to calculate updated class relevance is updated class frequency divide by window size. Here class frequency is nothing but updated class relevance count which is shown in fig-4.13 that is incremented by 1 when the that class is appeared and decremented by 1, when the class disappears till its value is greater than zero and retained as it is when the value becomes zero. Window size here is same as class relevance but along with class appearance count here calculation for class disappearance is also done which give the updated class relevance data. Again that updated class relevance is multiplied with 10 to bring it into scale.



	politics	sports	education	entertainment	business	health
1	5.228070	0.0000000	0.8421053	0.6315789	1.1929825	0.3157895
2	5.228070	0.0000000	0.8421053	0.6315789	1.1929825	0.3157895
3	4.736842	8.1052632	0.4210526	0.3157895	0.3157895	0.3157895
4	4.736842	8.1052632	0.4210526	0.3157895	0.3157895	0.3157895
5	4.736842	8.1052632	0.4210526	0.3157895	0.3157895	0.3157895
6	4.736842	8.1052632	0.4210526	0.3157895	0.3157895	0.3157895
7	4.736842	8.1052632	0.4210526	0.3157895	0.3157895	0.3157895
8	4.736842	8.1052632	0.4210526	0.3157895	0.3157895	0.3157895
9	5.228070	0.0000000	0.8421053	0.6315789	1.1929825	0.3157895
10	5.228070	0.0000000	0.8421053	0.6315789	1.1929825	0.3157895
11	5.228070	0.0000000	0.8421053	0.6315789	1.1929825	0.3157895
12	4.736842	8.1052632	0.4210526	0.3157895	0.3157895	0.3157895
13	4.736842	8.1052632	0.4210526	0.3157895	0.3157895	0.3157895
14	5.719298	5.9649123	0.3859649	0.3157895	1.5087719	0.3157895
15	8.701754	3.8245614	0.3508772	0.3157895	2.7017544	0.3157895
16	8.561404	1.2280702	0.4210526	1.0877193	2.5964912	0.3157895
17	8.561404	1.2280702	0.4210526	1.0877193	2.5964912	0.3157895
18	8.561404	1.2280702	0.4210526	1.0877193	2.5964912	0.3157895
19	8.842105	1.2280702	0.4210526	0.8070175	2.5964912	0.3157895
20	10.000000	0.7719298	0.5263158	1.2982456	1.2982456	0.3157895

Figure 4.14 *Class Position Weight-age*

The Figure 4.14 is a class position weight-age screen. Here class position weight-age is calculated by finding position weight and position frequency where position weight is nothing but weight of that class and class position frequency is nothing but count of that class which is incremented with respective to the position of class in the dataset. Here position weight-age is calculated by taking appearance of that class then 16 subtracted by position of that class. Here 16 is taken because single window size is is 16.Actually it is 18 but after 15th record weights are 0 so 16 is taken here. After this addition of all that vales is done and multiplied by 10 to bring it into range.

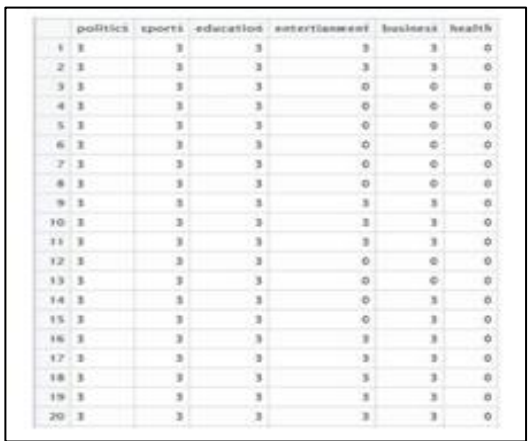
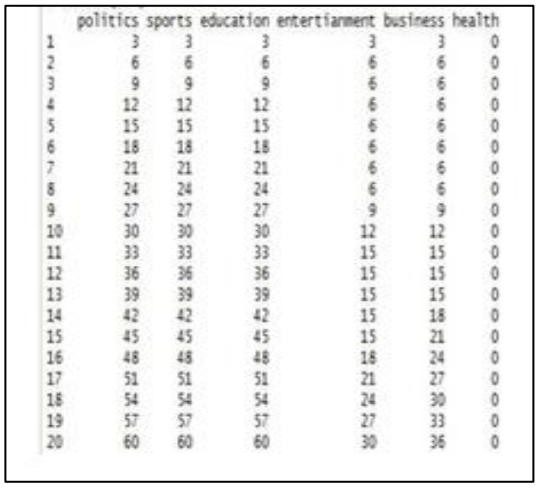


Figure 4.15 *Class Duration*

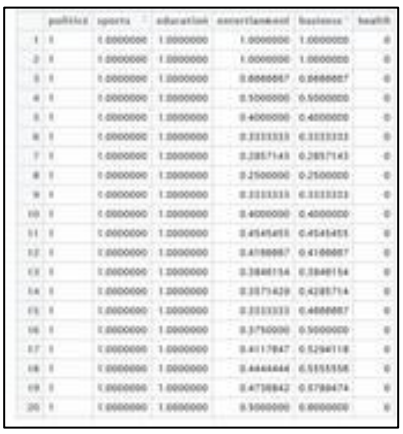
The Figure 4.15 is a class duration screen. Here if that class appears then value 3 is assigned in that place because the window size here is considered as 3 that is only 3 it is not getting incremented .If the class is not appeared then it is assigned as 0.



	politics	sports	education	entertainment	business	health
1	3	3	3	3	3	0
2	6	6	6	6	6	0
3	9	9	9	6	6	0
4	12	12	12	6	6	0
5	15	15	15	6	6	0
6	18	18	18	6	6	0
7	21	21	21	6	6	0
8	24	24	24	6	6	0
9	27	27	27	9	9	0
10	30	30	30	12	12	0
11	33	33	33	15	15	0
12	36	36	36	15	15	0
13	39	39	39	15	15	0
14	42	42	42	15	18	0
15	45	45	45	15	21	0
16	48	48	48	18	24	0
17	51	51	51	21	27	0
18	54	54	54	24	30	0
19	57	57	57	27	33	0
20	60	60	60	30	36	0

Class Duration Weight-age(cumulative sum)

The Figure 4.16 is a class duration weight-age (cumulative sum) screen. Here duration weight-age is calculated by considering current chunk duration, previous chunk duration and window duration. Current chunk duration means current time; previous chunk duration means when the class disappear the previous recorded chunk duration is added. Window duration is calculated as the average window duration for the data considered. Here cumulative sum for class duration is done as where current duration is added with previous duration. i.e... 6+3=9 for 3rd row for politics as shown in Figure 4.16.



Class Duration Weight-age

Figure 4.17 *Class Duration Weight-age*

The Figure 4.17 is a class duration weight-age screen. Here after getting the cumulative sum values as shown in Figure 4.16 that is divided with 3 which again multiplied into record number to bring to normal scale. So in 2 row of politics its weight is 6 that in Figure 4.16 but it is divided by 3 which is again multiplied by record number that is 6 divided by 3*2 where 2 is record number so 6 divide by 6 will be 1 as shown in Figure 4.17.In the same way the class durations are calculated.

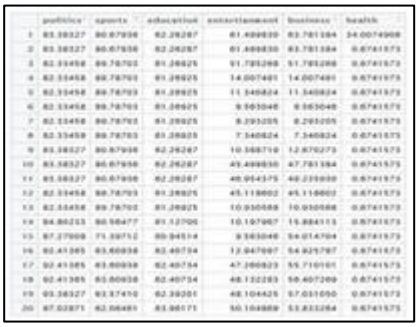


Figure 4.18 *Class Weight-age calculation*

The Figure 4.18 is a class weight-age calculation screen. Weight-age for each class was calculated based on the class Parameters identified as shown in Table 3.1. These Class Parameters were identified as Key Parameters for calculating Class weight-age by assigning a weight for each parameter determined empirically and normalized as shown in the Table 3.1.Here as given in Figure 4.18 weight-ages for politics is 83 which is more when compare to all other classes even though there are changes in weight-ages of politics class where class weight-age is less than other class but by seeing the final scenario it has been concluded that politics has more weight-age than other classes. 8 Class Weight-age calculation

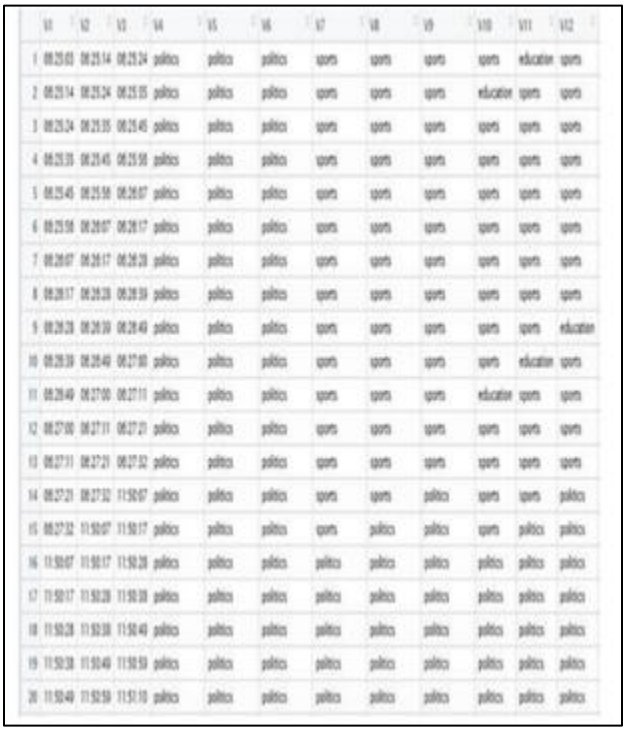


Figure 4.19*(a) Sliding window process*

The Figure 4.19 is a sliding window screen. After the weight age is calculated they are placed in sliding window model to find the topic which has been repeating continuously. Here sliding window model is processed for long period of data that is 12 hours and there are no changes in topic for some time but after sometime there will be a change as shown in Figure 4.19(b). Sliding window process

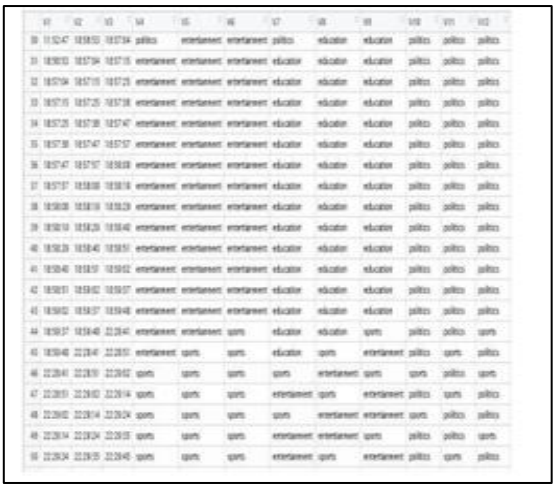


Figure 4.19*(b) Sliding window process*

The Figure 4.19(b) is a sliding window screen. After the weight age is calculated they are placed in sliding window model to find the topic which has been repeating continuously. Here changes are found which means topic get changes according to time. It is also proven that according to time the trending topics get changed where the concept of concept-drift is seen clearly.

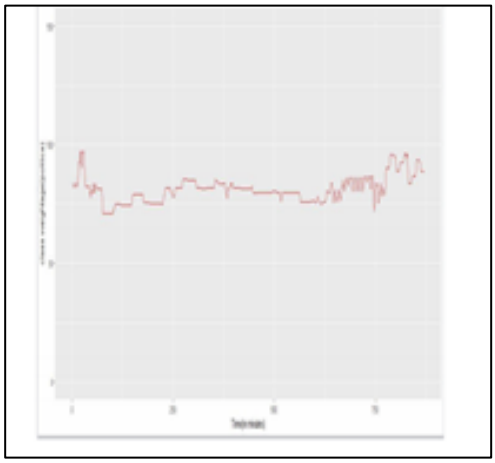


Figure 4.20 *Occurrence of class (politics)*

The Figure 4.20 is an occurrence of class screen. Here graph is plotted using the key parameters on each class to see the occurrence or weight age of that class. The graph is plotted on the politics class which shows the occurrence of the politics class depending on window that is time. Here graph is plotted between class weight-age and time. As the shown in graph the politics class is starting at 83 as the weight-age of class politics is 83 as given in Figure 4.17.Even though the graph gets fluctuated the weight-age of politics class is high when compare to other classes and it is the first trending topic among all the classes.

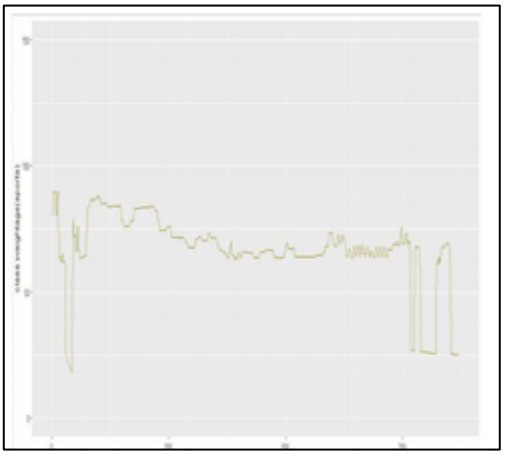


Figure 4.21 *Occurrence of class (Sports)*

The Figure 4.21 is an occurrence of class screen. Here graph is plotted using the key parameters on each class to see the occurrence or weight age of that class. The graph is plotted on the sports class which shows the occurrence of that class depending on window that is time. Here graph is plotted between class weight-age and time. As the shown in graph the sports class is starting at 80 as the weight-age of class politics is 80 as given in Figure 4.17.Even though the graph gets fluctuated that is sometime the weight-age is more than politics but by seeing all the weight-ages it is less than politics but higher than other classes so it is second trending topic among all the classes.

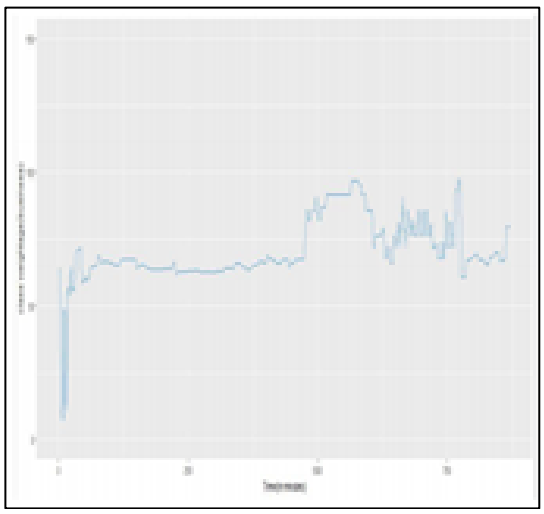


Figure 4.22 *Occurrence of class (business)*

The graph is plotted on the business class which shows the occurrence of that class depending on window that is time. Here graph is plotted between class weight-age and time. As the shown in graph the business class is starting at 63 as it is the weight-age of class as given in Figure 4.17.Even though in some cases the weight-age for the class is getting down by seeing the final results in the Figure 4.17 it has proven that business class has higher weight-age when compare to other three classes that are education, entertainment and health. So even though the graph gets fluctuated the business class is high and it is the third trending topic among all the classes.

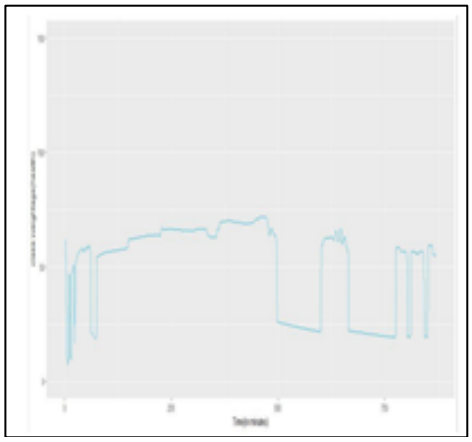


Figure 4.23 *Occurrence of class (entertainment)*

The graph is plotted on the entertainment class which shows the occurrence of that class depending on window that is time. Here graph is plotted between class weight-age and time. As the shown in graph the entertainment class is starting at 61 as it is the weight-age of class as given in Figure 4.17.Here even the class weight-age is starting at less when compare to education which is starting at high weight-age but after seeing all the weigh-ages it has higher when compare to education. So it has proven that entertainment class has higher weight-age when compare to other two classes that are education and health. So even though the graph gets fluctuated the entertainment class is high and it is the fourth trending topic among all the classes.

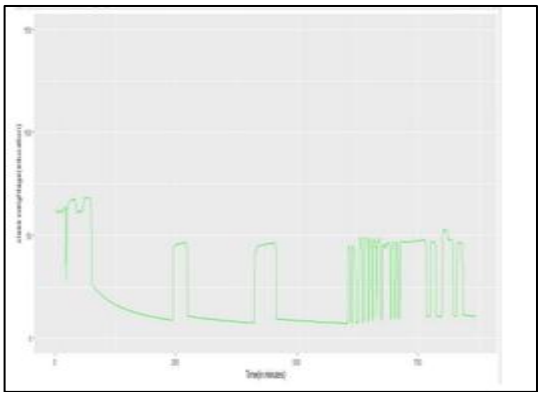


Figure 4.24 *Occurrence of class (education)*

The graph is plotted on the education class which shows the occurrence of that class depending on window that is time. Here graph is plotted between class weight-age and time. As the shown in graph the education class is starting at 62 as it is the weight-age of class as given in Figure 4.17.So it has proven that education is high when comparing health and less than other classes.

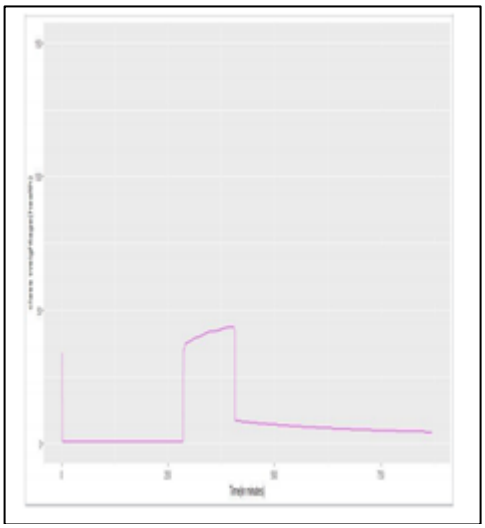


Figure 4.25 *Occurrence of class (health)*

The Figure 4.25 is an occurrence of health class screen. The graph is plotted on the health class which shows the occurrence of that class depending on window that is time. Here graph is plotted between class weight-age and time.So it has proven that health is the sixth and last trending topic among all the classes.

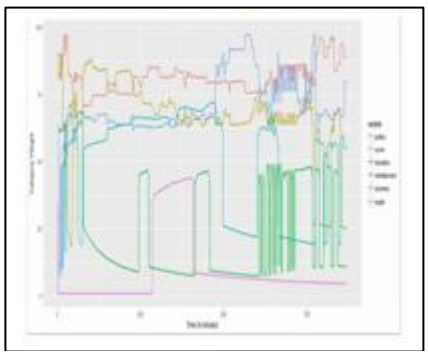


Figure 4.26 *Class Weight-age (long-period)*

Data is collected for 12hour that is from morning 10am to night 10pm and weight-age for each class was calculated based on the class Parameters identified as shown in Table 3.1. These Class Parameters were identified as Key Parameters for calculating Class weight-age by assigning a weight for each parameter determined empirically and normalized as shown in the Table 3.1. in Figure 4.17 it has got to an end that politics is first trending topic, then sports is the second, business third, entertainment fourth, education fifth and finally health is last which has less weight-age when compare to all classes.

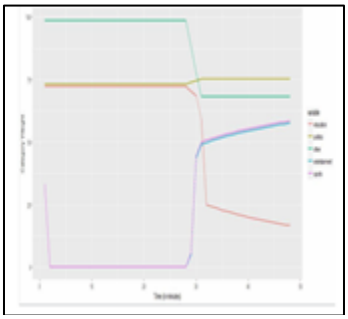


Figure 4.27 *Class Weight-age (short-period)*

Data is collected for 6 minutes and the weight-age for each class was calculated based on the class Parameters identified as shown in Table 3.1. These Class Parameters were identified as Key Parameters for calculating class weight-age by assigning a weight for each parameter determined empirically and normalized as shown in the Table 3.1. Here by seeing this graph it is stated that other class which means not related to any of the classes has more trending as its weight-age is high when compare to other remaining classes. Next the politics class has higher which is second trending, third education, fourth sports and fifth is entertainment is trending. Here in this graph there no much changes in classes because the time taken for collecting the data is less and the records are also less.

3 CONCLUSION AND FUTURE SCOPE

Twitter data is considered as one of the real time data stream. The data collection is done using streaming API, which collects trending topics in real time as they happen and for this data pre-processing is performed. After pre-processing data is manually categorised using data editor. As data is categorised to calculate the class weight-age automatic categorising is performed. The impact of concept-drift is shown in this proposed system with the help of class weight-age calculation. Visualization of the significance of the concept-drift is clearly shown with the help of graph wherever is required. In this proposed system 12 hours data is considered. It can be extended 24 hours data which give the more accurate results and improves the performance.

4 REFERENCES

[1] Lifna C.S, Dr. Vijayalakshmi M.” Identifying Concept-Drift in Twitter Streams.”International Conference on Advanced Computing Technologies and Applications (ICACTA-2015)

[2] Namrata A. Dumasia ,Prof. Ankur Shah “Review Paper on Concept Drift in Process Mining.” International Research Journal of Engineering and Technology (IRJET) Volume: 03 Issue: 04| Apr-2016

[3] Carmela Comito, Clara Pizzuti, Nicola Procopio “Online Clustering for Topic Detection in Social Data Streams” IEEE 28th International Conference on Tools with Artificial Intelligence 2016

[4] Yamini Kadwe1, Vaishali Suryawanshi “A Review on Concept Drift” IOSR

[5] Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661,p-ISSN:2278-8727, Volume 17, Issue 1, Ver. II (Jan – Feb. 2015), PP 20-26.

[6] W. N. Street and Y. Kim, “A streaming ensemble algorithm (SEA) for large-scale classification,” in Proc. 7th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining,pg. 377–382(2010).

[7] D Brzezinski, J Stefanowski,"Reacting to Different Types of Concept Drift:The Accuracy Updated Ensemble Algorithm" IEEE Transactions On Neural Networks And Learning Systems, Vol. 25, pp. 81-9(2014)..

[8] C. C. Aggarwal and K. Subbian, “Event detection in social streams,” in Proc. of the Twelfth SIAM Int. Conference on Data Mining, Anaheim, California, USA, April 26-28, 2012, pp. 624–635.

[9] J. Yin, A. Lampert, M. A. Cameron, B. Robinson, and R. Power, “Using social media to enhance emergency situation awareness,” IEEE Intelligent Systems, vol. 27, no. 6, pp. 52–59, 2012.

[10] [9]T. Sakaki, M. Okazaki, and Y. Matsuo, “Earthquake shakes twitter users: Real-time event detection by social sensors,” in Proceedings of the 19th International Conference on World Wide Web, ser. WWW ’10. ACM, 2010, pp. 851–860.

[11] W. Feng, C. Zhang, W. Zhang, J. Han, J. Wang, C. Aggarwal, and J. Huang,“STREAMCUBE: hierarchical spatiotemporal hashtag clustering for event exploration over the twitter stream,” in 31st IEEE International Conference on Data Engineering, ICDE 2015, 2015, pp. 1561–1572.

[12] [11] C. I. Muntean, G. A. Morar, and D. Moldovan, “Exploring the meaning behind twitter hashtags through clustering,” in Business Information Systems Workshops -BIS 2012 International Workshops and Future Internet Symposium, 2012, pp. 231– 242.

[13] Q. He, K. Chang, E. Lim, and J. Zhang, “Bursty feature representation for clustering text streams,” in Proceedings of the Seventh SIAM International Conference on Data Mining, 2007, pp. 491–496.
[14] Georgiana Ifrim, Bichen Shi, Igor Brigadir. Snow 2014 data challenge: “Assessing the performance of news topic detection methods in social media” In Proceedings of the SNOW 2014 Data Challenge, 2014.
[15] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. Journal of Machine Learning Research, 3:993–1022, 2003.
[16] L. M. Aiello, G. Petkos, C. Martin, D. Corney, S. Papadopoulos, R. Skraba, A. Goker, I. Kompatsiaris, and A. Jaimes. Sensing trending topics in twitter. IEEE Transactions on Multimedia,2013.
[17] S. Petrovi´c, M. Osborne, and V. Lavrenko. Streaming first story detection with application to twitter. In Human Language Technologies: Annual Conference of the North American Chapter of the Association for Computational Linguistics, 2010.
[18] C. Martin, D. Corney, and A. G¨oker. Finding newsworthy topics on twitter. IEEE Computer Society Special Technical Community on Social Networking E-Letter, 2013.
[19] Georgiana Ifrim, Bichen Shi, Igor Brigadir. “Event Detection in Twitter using Aggressive Filtering and Hierarchical Tweet Clustering” In Proceedings of the SNOW 2014 Data Challenge, 2014.
[20] Kathy Lee, Diana Palsetia, Ramanathan Narayanan, Md. Mostofa Ali Patwary,Ankit Agrawal, and Alok Choudhary. “Twitter Trending Topic Classification” In11th IEEE International Conference on Data Mining Workshops, 2011.
[21] Pramod S. & O.P.Vyas. “Data Stream Mining: A Review on Windowing Approach” In Global Journals Inc. (USA), Volume 12 Issue 11 Version 1.0 Year
[22] 2012.
[23] Balwinder Kaur1, Mandeep Kaur2, Pooja Mudgil3, Harjeet Singh.
[24] “IMPORTANCE OF SLIDING WINDOW PROTOCOL” In International Journal of Research in Engineering and Technology eISSN: 2319-1163 | pISSN: 2321-7308, 2013.
[25] Hongmei Wang, Fentian Li, Dongkai Tang, Zeru Wang. “Research on Data stream Mining Algorithm for Frequent Itemsets Based on Sliding Window Model” In IEEE 2nd International Conference on Big Data Analysis, 2017.