Active Online Learning for Social Media Analysis to Support Crisis Management

Pramod Gadekar[1], Pallavi Wadekar[2], Rohit Divare[3], Akshay Kale[4], and Kshitij Thosar[5]

[1] *Lecturer, Computer Technology, P.Dr.V.V.P.Instt of Tech & Engg.(Polytechnic),Loni ,Maharastra ,India*

[2,3,4,5] *Student, Computer Technology, P.Dr.V.V.P.Instt of Tech & Engg.(Polytechnic),Loni ,Maharastra ,India*

ABSTRACT

People use social media (SM) to describe and discuss different situations they are involved in, like crises. It is therefore worthwhile to exploit SM contents to support crisis management, in particular by revealing useful and unknown information about the crises in real-time. Hence, we propose a novel active online multiple-prototype classifier, called AOMPC. It identifies relevant data related to a crisis. AOMPC is an online learning algorithm that operates on data streams and which is equipped with active learning mechanisms to actively query the label of ambiguous unlabeled data. The number of queries is controlled by a fixed budget strategy. Typically, AOMPC accommodates partly labeled data streams. AOMPC was evaluated using two types of data: (1) synthetic data and (2) SM data from Twitter related to two crises, Colorado Floods and Australia Bushfires. To provide a thorough evaluation, a whole set of known metrics was used to study the quality of the results. Moreover, a sensitivity analysis was conducted to show the effect of AOMPC's parameters on the accuracy of the results. A comparative study of AOMPC against other available online learning algorithms was performed. The experiments showed very good behavior of AOMPC for dealing with evolving, partly-labeled data streams.

Keyword: - *Online Learning, Multiple Prototype Classification, Active Learning, Social Media, Crisis Management*

1. INTRODUCTION

The primary task of crisis management is to identify specific actions that need to be carried out before, during, and after a crisis. In recent years, research studies have investigated the use of social media (SM) as a source of information for efficient crisis management. Our previous work on SM in emergency response focused on offline and online clustering of SM messages. We propose a learning algorithm, AOMPC that relies on active learning to accommodate the user's feedback upon querying the item being processed. The primary goal in using user-generated contents of SM is to discriminate valuable information from irrelevant one.

The classifier plays the role of filtering machinery, recognizing the important SM items (e.g., tweets) that are related to the event of interest. The selected items are used as cues to identify sub-events. An original online learning algorithm, AOMPC, is proposed to handle data streams in an efficient way. It uses unlabeled and labelled data which are tagged through active learning. The number of queries is controlled by a budget and the requested items help to direct the AOMPC classifier to a better discriminatory capability. AOMPC is evaluated on synthetic datasets and real-world datasets related to two crises.

2. SYSTEM ARCHITECTURE

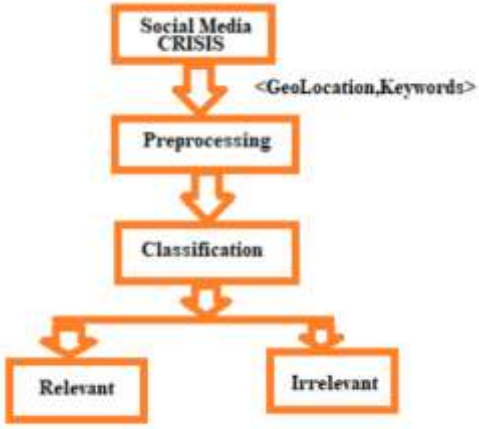

Figure 1:System Architecture

2.1 FLOWCHART

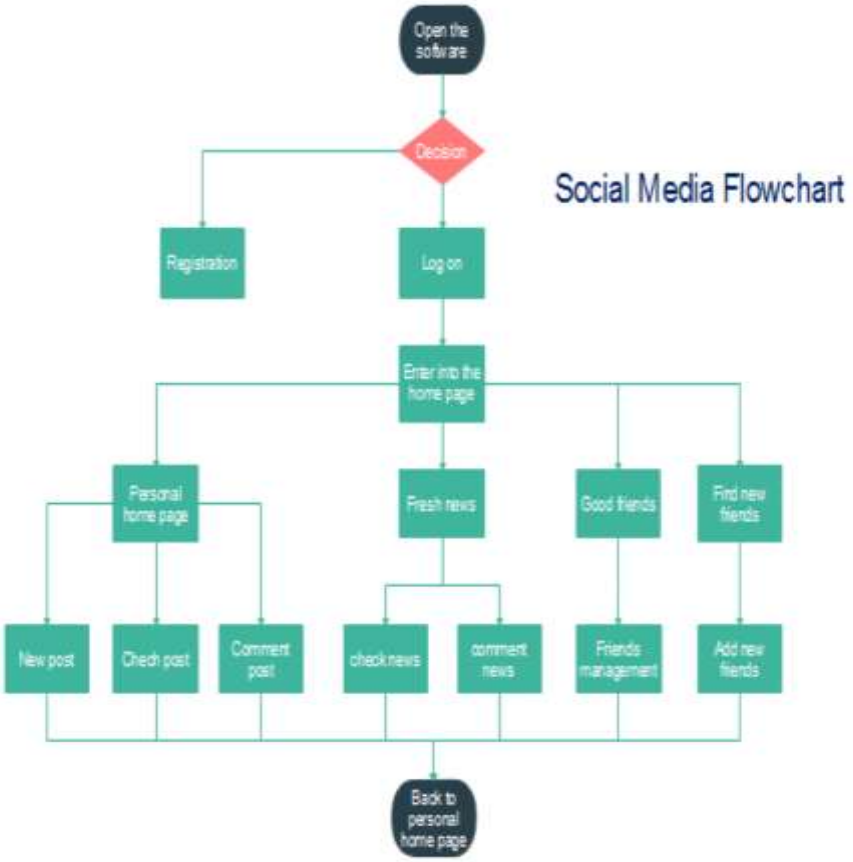

Fig -2: Flowchart

2.1 Table

TABLE 1
List of symbols used

Variable	Description
$\mathbf{x}$	Input (one item) received by the data stream X with bt_{CT} batches
V	Set of currently known prototypes
α	A parameter used in Alg. 1 to compute the staleness of a prototype. It is given as: $\alpha = e^{\frac{-log2}{\beta}}$, where β is the half-life span, denoted hereafter as $(1/2)$-life-span, described in [31] that refers to the amount of time required for a quantity to fall to half its value as measured at the beginning of the time period.
I	Set of indices i indicating the prototypes $\mathbf{v}_i$
$dist$	Appropriate distance measure; see Algorithm 2
UT	Threshold used to identify uncertainty
CT	Current time
LTU	Last time the prototype was updated (i.e., the winner)
S	List of nearest prototypes in ascending order to the current input $\mathbf{x}$
$label$	Labels are: $relevant$, $irrelevant$, and $unknown$

3. RELATABLE WORKS

The problem addressed in this paper is related to several topics: multiple prototype and Learning Vector Quantization (LVQ) classification, online learning for classification, active learning with budget planning, and social media analysis (i.e., natural language processing). A short overview of these topics is presented in the following.

3.1 MATERIALS AND METHODS

A prototype-based classification approach operates on data items mapped to a vector representation (e.g., vector space model for text data). Data points are classified via prototypes, which are adapted based on items related/similar to them. Self organizing maps (SOM) are an unsupervised version of prototype-based classification, also known as LVQ. LVQ has been applied to several areas, e.g., robotics, pattern recognition, image processing, text classification etc. There are several offline multiple prototype classifiers, such as LVQ, fuzzy LVQ, and the deterministic Dog-Rabbit (DR) model. Bouchachia [8] proposes an incremental supervised LVQ-like competitive algorithm that operates online. The time-based learning rate of our algorithm considers concept drift (i.e., changes of the incoming data) directly during the update of the prototypes.

3.2 DISCUSSION AND FUTURE WORK

The advantage of AOMPC compared to the other algorithms is the continuous processing of data streams and incremental update of knowledge, where the existing prototypes act as memory for the future. Here forgetting of outdated knowledge is controlled by α, which also depends on the budget. Learning serves to adapt and/or create clusters in a continuous way. The algorithm queries labels on-the-fly for continuously updating the classification model. In summary, it can be said that budget B and threshold UT are related to each other. Increasing their values increases the quality of the algorithm. B has also an influence on the number of clusters that are created (i.e., the more often the user is asked, the more hints for new clusters are given). The advantage of our algorithm compared to the others is the transferred knowledge from one batch to the next creating a continuous view on the arriving data. The already known prototypes act as memory (i.e., forgetting is based on α and learning is based on the new creation of clusters, see Algorithm 1). In terms of performance, Tab. 5 shows the best results of AOMPC for different budget values using the CQM measure. For GD, the variable learning rate α and the fixed α rate in the case of SSMD show good performance. For CF, the variable learning rate seems to be more suitable considering the number of queries. AOMPC produces good results on AB using a fixed learning rate. The reason is that the data items are very similar and that changes within the textual data happen slowly and near the boundary. Finally, comparing the active learning strategies ("DCN" options), we can notice that very good performance is achieved especially for SSMD and CF. The quality of clustering increases even for low values of B. Overall, AOMPC shows a quite good performance (see Tables 4, 3 and 5), despite the fact that it operates online and handles labelling just-in-time. Moreover, AOMPC was run on batches just for the sake of feature selection (see Sec. 3.3). AOMPC can run in purely point-based online mode (i.e., item-by-item) as well. In the future, we plan to extend this algorithm by deleting clusters when they lose their importance. This could also be done for features in order to obtain an evolving feature space. We also plan to implement a variable budget strategy so that, for instance, the number of queries (i.e., budget) is bigger for cold start and gets reduced afterward, depending on the uncertainty and the performance of the algorithm. Finally, it would be interesting to identify drift, without defining a threshold, but by considering the general case, where classes are non-contiguous.

4. CONCLUSIONS

This paper presents a streaming analysis framework for distinguishing between relevant and irrelevant data items. It integrates the user into the learning process by considering the active learning mechanism. We evaluated the framework for different datasets, with different parameters and active learning strategies. We considered synthetic datasets to understand the behavior of the algorithm and real-world social media datasets related to crises. We compared the proposed algorithm, AOMPC, against many existing algorithms to illustrate the good performance under different parameter settings. As explained in Sec. 4.6, the algorithm can be extended to overcome many issues, for instance by considering: dynamic budget, dynamic deletion of stale clusters, and generalization to handle non-contiguous class distribution.

5. ACKNOWLEDGEMENT

We would like to express our sincere gratitude to everyone who contributed to the successful completion of this project on Active Online Learning for Social Media Analysis to Support Crisis Management.

First and foremost, we would like to thank our supervisor [Supervisor's Name], for providing us with the guidance, support, and expertise necessary to complete this project. His/her invaluable feedback, insights, and suggestions helped us to refine our ideas, analyze our data, and present our findings effectively.

We are also grateful to [Name of the Institution or Organization] for providing us with the necessary resources, including access to software, data, and other facilities, which enabled us to carry out our research effectively.

We would also like to extend our appreciation to the participants who generously gave their time to participate in our study. Without their willingness to share their experiences and perspectives, our research would not have been possible.

Finally, we would like to thank our families, friends, and colleagues who provided us with the support, encouragement, and motivation necessary to complete this project. Their unwavering belief in us and our abilities kept us going during the most challenging times.

Thank you all for your support, encouragement, and contributions to our project.

6. REFERENCES

[1] F. Abel, C. Hauff, G.-J. Houben, R. Stronkman, and K. Tao, "Semantics + Filtering + Search = Twitcident. Exploring Information in Social Web Streams," in Proc. of the 23rd ACM Conf. on Hypertext and Social Media. ACM, 2012, pp. 285–294.

[2] U. Ahmad, A. Zahid, M. Shoaib, and A. AlAmri, "Harvis: An integrated social media content analysis framework for youtube platform," Information Systems, vol. 69, pp. 25 – 39, 2017.

[3] G. Backfried, J. Gollner, G. Qirchmayr, K. Rainer, G. Kienast, G. Thallinger, C. Schmidt, and A. Peer, "Integration of Media Sources for Situation Analysis in the Different Phases of Disaster Management: The QuOIMA Project," in Eur. Intel. and Security Informatics Conf., Aug 2013, pp. 143–146.

[4] BBC News Europe. (2012, Aug.) England Riots: Maps and Timeline. [Online]. Available: http://www.bbc.co.uk/news/ uk-14436499

[5] H. Becker, M. Naaman, and L. Gravano, "Learning Similarity Metrics for Event Identification in Social Media," in Proc. of the Third ACM Int'l Conf. on Web Search and Data [48], [51]. We propose a Learning Vector Quantization (LVQ)- like approach based on multiple prototype classification. The classifierMining, ser. WSDM '10. NY, USA: ACM, 2010, pp. 291–300.

[6] J. Bezdek, T. Reichherzer, G. Lim, and Y. Attikiouzel, "MultiplePrototype Classifier Design," IEEE Trans. on Systems, Man, and Cybernetics, Part C: Applications and Reviews, vol. 28, no. 1, pp. 67– 79, Feb 1998.

[7] M. Biehl, B. Hammer, and T. Villmann