Statistical Detection of Data Drift in Real-time Social Network Conversations

Chetana Pujari
Dept. of I&CT
Manipal Institute of Technology,
Manipal Academy of Higher Education
Manipal, India
chetana.pujari@manipal.edu

Sumith N.
Dept. of I&CT
Manipal Institute of Technology,
Manipal Academy of Higher Education
Manipal, India
sumith.n@manipal.edu

Pooja S
Dept. of I& CT
Manipal Institute of Technology,
Manipal Academy of Higher Education
Manipal, India
pooja.s@manipal.edu

Chandrakala C B
Dept. of I&CT
Manipal Institute of Technology,
Manipal Academy of Higher Education
Manipal, India
chandrakala.cb@manipal.edu

Vibha Prabhu
Dept. of I&CT
Manipal Institute of Technology,
Manipal Academy of Higher Education
Manipal, India
vibha.prabhu@manipal.edu

2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT) | 979-8-3503-7024-9/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICCCNT61001.2024.10724556

***Abstract*—The increasing reliance on conversational datasets for natural language processing (NLP) applications necessitates a comprehensive understanding of potential data drift phenomena. This paper investigates the phenomenon of data drift within conversational datasets over time, aiming to develop effective methods for detection and mitigation.**

Our approach involves the analysis of temporal changes in the distribution of conversation data, focusing on linguistic patterns, user preferences, and contextual nuances. A novel framework leveraging advanced statistical methods and machine learning techniques to quantify and detect data drift within the dataset is proposed here. The methodology is designed to adapt to the evolving nature of language use, capturing subtle shifts in conversational dynamics that may impact model performance.

Furthermore, experimental results on a diverse set of conversational datasets, demonstrating the efficacy of our approach in identifying and characterizing data drift is presented here. The findings highlight the importance of continuous monitoring and adaptation to evolving linguistic patterns, ensuring the robustness and generalization capability of NLP models over time.

This research contributes to the broader understanding of data drift in conversational datasets and provides a foundation for the development of adaptive NLP models capable of maintaining high performance in dynamic linguistic environments. The proposed framework not only enhances the reliability of existing models but also lays the groundwork for future research in addressing the evolving challenges posed by data drift in natural language conversations.

***Index Terms*—NLP, data drift, temporal, statistical model**

I. Introduction

In the realm of natural language processing (NLP), the burgeoning use of conversational datasets has become integral to the development and refinement of language models. As these models are deployed across various applications such as chatbots, virtual assistants, and sentiment analysis, the dynamic nature of language use over time poses a significant challenge. One critical aspect that demands attention is the concept of data drift within conversational datasets which is the gradual and often subtle changes in linguistic patterns, user preferences, and contextual nuances that occur over time.

Understanding and managing data drift is imperative for maintaining the effectiveness and reliability of NLP models. Conversational datasets are inherently susceptible to temporal variations, reflective of the evolving nature of language itself. As the dynamics of conversations shift, existing models may encounter performance degradation, necessitating adaptive strategies to ensure continued accuracy and relevance. The concept underlying this research revolves around comprehensively identifying the phenomenon of data drift in conversational datasets. Data drift, in the context of our study, refers to the temporal changes in the distribution of linguistic elements within the dataset, encompassing shifts in vocabulary usage, syntactic structures, and conversational context. Recognizing these changes is crucial, as they directly impact the performance and generalization capability of NLP models trained on such datasets.

Our conceptual framework involves an analysis of conversational data spread over time. Linguistic patterns, user preferences, and contextual nuances to unveil subtle shifts that may otherwise go unnoticed are investigated here. The proposed concept not only acknowledges the dynamic nature of language but also emphasizes the need for adaptive models that can autonomously adjust to these evolving linguistic landscapes. By conceptualizing data drift as a continuous and dynamic process, our research aims to show the existence of such phenomenon which is a motivation for better NLP models.

II. Background work

The study by Ackerman et al. [1] demonstrates that the absence of statistical similarity between confidence values of production and baseline classes indicates data drift. Data drift refers to a significant deviation of production data from the training data, leading to potentially inaccurate classification results. Sahiner et al. [2] corroborate these findings, highlighting data drift as a major factor contributing to the deterioration of performance in medical machine learning systems.

Concept drift, a phenomenon gaining traction across various domains like sensors, robotics, and anomaly detection, is discussed in [3]. This shift in the behavior of generative processes is increasingly evident in research involving data streams with structural breaks or regime changes, as indicated by works such as Masegosa et al. [4]and [5].

TThe significance of online incremental ML algorithms, particularly their scalability in continuous learning scenarios and adaptability to non-stationarities, shifts, and drifts in data is discussed in [6]. However, the persistent challenge of handling recurring concepts in stationary scenarios remains a subject of ongoing study [7], [8].

ADaptive WINdowing 2 (ADWIN2), a drift detection method proposed by Bifet and Gavaldà [9]. This method utilizes a sliding window that dynamically adjusts its size based on the mean difference between old and new data subwindows. ADWIN2 is employed to detect concept drift using online classification error rates, with a separate instance for issuing warnings.

Drift detection methods including Drift Detection Method (DDM) by Gama et al. [10], Reactive Drift Detection Method (RDDM) proposed by Barros [11] as an improvement of DDM, Drift Detection Method based on Hoeffding's Inequality (HDDM) introduced by Frias [12], and Early Drift Detection Method (EDDM) as a variant of DDM by Baena-García et al. [13] emphasize the impact of data drift on various learning models. These methods employ various statistical and algorithmic approaches to monitor and detect concept drift in online learning scenarios. Probabilistic Real-Drift Detection (PRDD) proposed by [14], is designed to track and respond to concept drift based on its probabilistic definitions. PRDD utilizes the classifier's prediction errors and confidence levels to detect specifically the Real conflict drift.

III. Problem description

When data exhibits drift and is used for Natural Language Processing (NLP) applications, several issues can arise, impacting the performance, accuracy, and reliability of NLP models. Drift in the data distribution may lead to a mismatch between the training and deployment environments, causing a decrease in the model's overall performance. NLP models trained on outdated or insufficiently diverse data may fail to accurately understand or generate language in real-world scenarios. Drift can exacerbate existing biases present in the training data, leading to biased predictions and reinforcing stereotypes over time. NLP models may misinterpret or fail to capture the evolving semantics of language, affecting their ability to understand user intent accurately.Further, drift can result in the introduction of out-of-distribution data that the model has not encountered during training. Continuous monitoring and adaptation strategies are required to address data drift effectively. Without adaptive mechanisms, models may become outdated quickly, hindering their ability to generalize to new linguistic patterns and user behaviors.

In summary, identifying data drift is imperative for maintaining the relevance, fairness, and effectiveness of NLP models in the face of evolving language use. It allows for proactive adjustments, preventing performance degradation and ensuring that models remain robust and reliable over time.

IV. Methodology

Identifying data drift is a crucial aspect of maintaining the performance and reliability of machine learning models, including those used in Natural Language Processing (NLP). Several approaches can be employed to identify data drift. For our study statistical approach is employed.

Statistical Measures: Statistical measures involve comparing statistical properties of the source and target datasets to detect differences. This can include measures such as mean, variance, covariance, and distributional comparisons. Statistical tests, like the Kolmogorov-Smirnov test or the Wasserstein distance, can be used to quantify the dissimilarity between the two datasets.

The Kolmogorov-Smirnov (KS) test is a statistical method used for comparing two probability distributions to assess if they are significantly different from each other. In the context of finding data drift, the KS test can be applied to compare the distributions of features or characteristics in two datasets, such as a source dataset (training) and a target dataset (deployment), to identify potential changes in data distribution over time. Select a Significance Level typically of 0.05 is chosen. This represents the threshold below which the null hypothesis is rejected. Next the KS Statistic based on the maximum vertical distance (supremum) between the cumulative distribution functions (CDFs) of the two group of text spread in a fixed time period is calculated. - The formula for the KS statistic is given by:

$$D = \max |F_1(x) - F_2(x)|$$

- $F_1(x)$ and $F_2(x)$ are the empirical cumulative distribution functions of the two datasets.

4. Determine the critical value from the KS distribution table based on the significance level and the size of the datasets. If the calculated KS statistic is greater than the critical value, reject the null hypothesis. suggesting that there is a significant difference between the two distributions. This indicates the presence of data drift.

The insights received from the Kolmogorov-Smirnov (KS) test for data drift can provide valuable information about the differences between two datasets. Specifically, the test helps in determining whether there is a statistically significant shift in the distribution of a feature or set of features over time.

V. Experiments

In our experiment, the phenomenon of data drift in online conversations is studied in two conversation datasets namely, IT start up discussion [15] and Topical chat [16].

The first step of our experiment involved preprocessing the conversation data to extract relevant information, namely timestamps and the actual text messages exchanged during the discussions. This preprocessing step ensured that noise is removed. Once the data preprocessing was complete, the next step is to analyze the occurrence of data drift within these conversations.

IT start up discussion consisted of a total of 256 conversations, collected over the span of one week from various online platforms or social networks frequented by IT startup communities. These conversations encompassed a wide range of topics, including technology trends, business strategies, and software development practices.

Initially, the overall data drift represented by p-value which trends over a period of one day, as depicted in Fig. 1 is observed. This allowed us to visualize how the characteristics of the conversations evolved over time within the larger timescale of a day. Average Magnitude of Shifts of 0.149, Average KS Statistic of 0.149 and Frequency of Shift of 1 was observed.

Furthermore, a more granular analysis by investigating data drift at an hourly level was conducted. By breaking down the dataset into smaller time intervals, subtle shifts or variations in the conversation dynamics with higher temporal resolution was captured. Fig. 2 presents the p-value results of this hourly analysis, highlighting the fluctuations in data drift within shorter time spans.Also, in this case, Average Magnitude of Shifts was 0.249,Average KS Statistic was 0.24 and Frequency of Shifts was 3. Interestingly, our observations revealed that the occurrence of data drift was more pronounced when analyzed at an hourly basis compared to the broader daily perspective. This suggests that data drift can be more effectively detected and understood when examined at finer time intervals, particularly when working with relatively smaller datasets such as those in our study.

Overall, our experiment sheds light on the temporal dynamics of data drift in online conversations within the IT startup domain, emphasizing the importance of considering time granularity in monitoring and analyzing such phenomena. A parallel experiment using a larger dataset, denoted as the Topical dataset [16]. This dataset comprised a significantly larger volume of conversations, totaling 18,878 discussions covering eight distinct topics of interest. Due to the size of the Topical dataset,data drift over an extended time period could be studied in detail. This allowed insights into how conversation dynamics evolve across longer duration within specific topical contexts.

Upon preprocessing the Topical dataset to extract timestamps and conversation text, similar to the previous experiment, the occurrence of data drift was observed. Fig. 3 presents the results of this analysis, showcasing the observed data drift within the conversations. It is observed that Average Magnitude of Shifts was 0.163, Average KS Statistic was 0.1637 and Frequency of Shifts was 1051.

The analysis of Fig. 3 demonstrates that data drift is evident within well-defined topic-based conversations. This conclusion is supported by the identification of noticeable shifts in conversation characteristics over time, despite the larger dataset size and longer observation period. The evaluation metrics shown here collectively provide robust quantitative evidence of data drift within the dataset, reinforcing the qualitative observations from the visualization.

This finding underscores the significance of monitoring data drift even in contexts where conversations are centered around specific topics. By doing so, there can be a better understanding on how discourse within distinct topical domains evolves and adapts over time, thereby enabling more informed decision-making and strategy development for various applications.

In summary, experiment on the Topical dataset reaffirms the importance of detecting and understanding data drift in online conversations, particularly in scenarios involving large and diverse collections of discussions across different topical domains.

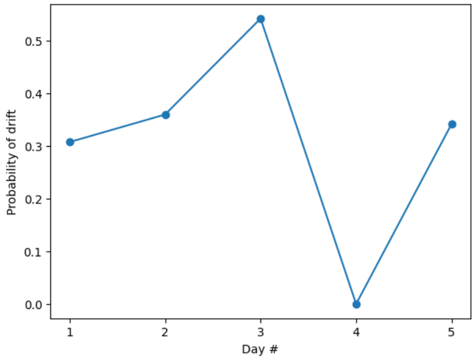

Fig. 1. Data drift seen in Startup discussion dataset

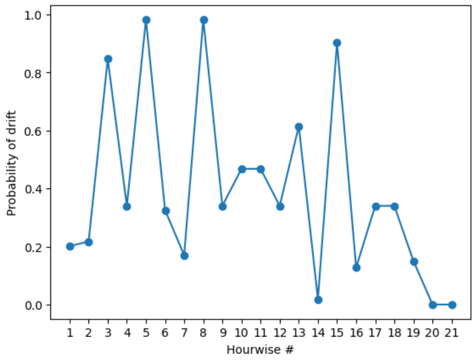

Fig. 2. Data drift seen in Startup discussion dataset

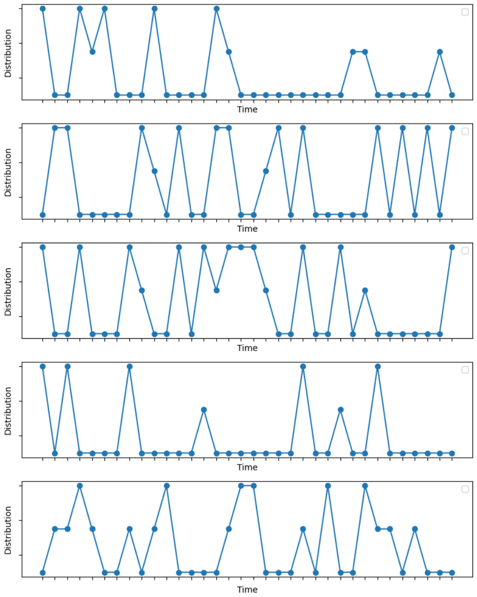

Fig. 3. Data drift seen in topical dataset

The experimental findings on diverse conversational datasets underscore the significance of continuous monitoring and adaptation strategies in maintaining robust and reliable models.

The insights gained from this work extend beyond the identification of data drift; they contribute to a deeper understanding of the evolving linguistic landscapes and the implications for NLP models. By addressing issues such as decreased model performance, bias amplification, and domain shift, our research lays the groundwork for the development of adaptive NLP models capable of navigating the challenges posed by dynamic conversational data.

VI. Conclusion

In conclusion, this work has delved into the critical realm of identifying data drift in conversational datasets over time. Through a comprehensive exploration of linguistic patterns, user preferences, and contextual nuances, our research has aimed to address the challenges posed by the dynamic nature of language use in natural language processing (NLP) applications.

The significance of ongoing research in this domain cannot be overstated. Continuous advancements in conversational AI necessitate vigilant monitoring and adaptive strategies to ensure the longevity and effectiveness of NLP models in real-world applications. This work serves as a stepping stone, offering insights and methodologies that pave the way for future investigations into the evolving challenges of data drift in natural language conversations.

References

[1] S. Ackerman, O. Raz, M. Zalmanovici, and A. Zlotnick, “Automatically detecting data drift in machine learning classifiers,” *arXiv preprint arXiv:2111.05672*, 2021.

[2] B. Sahiner, W. Chen, R. K. Samala, and N. Petrick, “Data drift in medical machine learning: implications and potential remedies,” *The British Journal of Radiology*, p. 20220878, 2023.

[3] A. Tsymbal, “The problem of concept drift: definitions and related work,” *Computer Science Department, Trinity College Dublin*, vol. 106, no. 2, p. 58, 2004.

[4] A. R. Masegosa, A. M. Martínez, D. Ramos-López, H. Langseth, T. D. Nielsen, and A. Salmerón, “Analyzing concept drift: A case study in the financial sector,” *Intelligent Data Analysis*, vol. 24, no. 3, pp. 665–688, 2020.

[5] A. L. Suárez-Cetrulo, A. Cervantes, and D. Quintana, “Incremental market behavior classification in presence of recurring concepts,” *Entropy*, vol. 21, no. 1, p. 25, 2019.

[6] R. Elwell and R. Polikar, “Incremental learning of concept drift in nonstationary environments,” *IEEE Transactions on Neural Networks*, vol. 22, no. 10, pp. 1517–1531, 2011.

[7] C. Alippi, G. Boracchi, and M. Roveri, “Just-in-time classifiers for recurrent concepts,” *IEEE transactions on neural networks and learning systems*, vol. 24, no. 4, pp. 620–634, 2013.

[8] J. B. Gomes, M. M. Gaber, P. A. Sousa, and E. Menasalvas, “Mining recurring concepts in a dynamic feature space,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 25, no. 1, pp. 95–110, 2013.

[9] A. Bifet and R. Gavalda, “Learning from time-changing data with adaptive windowing,” in *Proceedings of the 2007 SIAM international conference on data mining*. SIAM, 2007, pp. 443–448.

[10] J. Gama, P. Medas, G. Castillo, and P. Rodrigues, “Learning with drift detection,” in *Advances in Artificial Intelligence–SBIA 2004: 17th Brazilian Symposium on Artificial Intelligence, Sao Luis, Maranhao, Brazil, September 29-Ocotber 1, 2004. Proceedings 17*. Springer, 2004, pp. 286–295.

[11] R. S. Barros, D. R. Cabral, P. M. Gonçalves Jr, and S. G. Santos, “Rddm: Reactive drift detection method,” *Expert Systems with Applications*, vol. 90, pp. 344–355, 2017.

[12] I. Frias-Blanco, J. del Campo-Ávila, G. Ramos-Jimenez, R. Morales-Bueno, A. Ortiz-Díaz, and Y. Caballero-Mota, “Online and non-parametric drift detection methods based on hoeffding’s bounds,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 27, no. 3, pp. 810–823, 2014.

[13] M. Baena-Garcıa, J. del Campo-Ávila, R. Fidalgo, A. Bifet, R. Gavalda, and R. Morales-Bueno, “Early drift detection method,” in *Fourth international workshop on knowledge discovery from data streams*, vol. 6. Citeseer, 2006, pp. 77–86.

[14] S. Parasteh and S. Sadaoui, “A probabilistic approach for detecting real concept drift.” in *ICAART (2)*, 2024, pp. 301–311.

[15] Nursyahrina, “Chat data for sentiment and knowledge analysis.” [Online]. Available: https://www.kaggle.com/datasets/nursyahrina/chat-conversation-data-for-it-startup

[16] K. Gopalakrishnan, B. Hedayatnia, Q. Chen, A. Gottardi, S. Kwatra, A. Venkatesh, R. Gabriel, and D. Hakkani-Tür, “Topical-Chat: Towards Knowledge-Grounded Open-Domain Conversations,” in *Proc. Interspeech 2019*, 2019, pp. 1891–1895. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2019-3079