Adaptive Window Strategy for Topic Modeling in Document Streams

Pierre-Alexandre Murena
LTCI - Télécom ParisTech
Paris, France
murena@telecom-paristech.fr

Marie Al-Ghossein
LTCI - Télécom ParisTech
Paris, France
marie.alghossein@telecom-paristech.fr

Talel Abdessalem
LTCI - Télécom ParisTech
UMI CNRS IPAL NUS
Paris, France
talel.abdessalem@telecom-paristech.fr

Antoine Cornuéjols
UMR MIA518 - AgroParisTech INRA
Paris, France
antoine.cornuejols@agroparistech.fr

***Abstract*—Extracting global themes from a written text has recently become a major issue for computational intelligence, in particular in Natural Language Processing communities. Among all proposed solutions, Latent Dirichlet Allocation (LDA) has gained a vast interest and several variants have been proposed to adapt to changing environments. With the emergence of data streams, for instance from social media, the domain faces a new challenge: topic extraction in real time. In this paper, we propose a simple approach called Adaptive Window based Incremental LDA (AWILDA) originating from the cross-over between LDA and state-of-the-art methods in data stream mining. We train new topic models only when a drift is detected and select training data on the fly using ADWIN algorithm. We provide both theoretical guarantees for our method and experimental validation on artificial and real-world data.**

I. Introduction

The abundance of text sources provided by online platforms and social networks offers new opportunities and introduces new challenges in the domain of text modeling. Two classes of methods have emerged, based either on n-gram language models [1] or probabilistic topic modeling [2]. While the first class focuses on semantic modeling of languages based on the order of words, probabilistic topic modeling describes documents as an unordered bag of words drawn from mixtures of word distributions called topics. Even if the human interpretation of topics remains hard to achieve [3], these frameworks are used for a large variety of tasks ranging from text analysis [4], [5], recommendation [6], [7], sentiment analysis [8] to image annotation [9]. Among topic models, Latent Dirichlet Allocation (LDA) [10] has gained more and more attention for its simplicity and its modularity. Several variants of the original model have been developed to achieve new tasks that cannot be performed with the original model (see for instance [5], [7]).

The base model of LDA infers topic distributions from a given batch of documents. This setting is not adapted to evolving environments, including text mining on documents generated continuously at high rates or streams of documents. Nevertheless, solutions have been proposed to adapt LDA to temporal frameworks where the data distribution varies over time. Dynamic Topic Models (DTM) [11] are an attempt to include a dynamic behavior into LDA. DTM models the *word-topic distribution*, i.e., the distribution of words inside a topic, as an evolving parameter. The distribution of this parameter at time t is defined with respect to its distribution at time $t-1$. A closely related idea is developed by SeqLDA [12], but it is applied at the level of a book where the time parameter is associated to the index of the paragraph. An alternative is offered by continuous-time models [13] which assume that the distribution over topics is influenced by word co-occurrences (such as in standard LDA) and by the document date. The major disadvantage of this method is its offline nature: the model can only be learned once we have the whole corpus. It is thus inefficient in the context of stream mining. A frequent strategy for stream mining with LDA consists in grouping documents by time slices (see for instance [11], [14]). On the other hand, online incremental LDA offers an interesting alternative since it does not require storing previous data and relies only on the new received documents [15]. The major problem of this method is the difficulty of defining time slices. In particular, modifications in topics might occur on a time period significantly smaller than the chosen time slice. This scale-dependency is taken into account by some continuous-time methods [13], [16].

Our approach takes a completely different direction. We propose to use change detection methods to estimate change of topics in document streams. Learning with distribution changes, also called *concept drift*, has been widely investigated in a supervised setting [17], [18], [19]. Two global classes of methods emerge from this domain. *Passive algorithms* update the model for each received observation, no matter whether a drift actually occurred or not. On the contrary, *active algorithms* focus on detecting the drift and update the model only when a drift is found. All the methods presented for streaming topic models belong to the passive class of methods.

Knowing when the topic changes remains a crucial question in several domains like event detection [20]. Among active methods, sliding windows are intuitive approaches which consist in storing recent data only in memory. The width of the window can be fixed like in FLORA [21] or adaptive like in ADWIN [22].

The approach we propose is a combination of online LDA [23] and ADWIN. The idea is to detect drifts by changes of likelihood and to alternate online refinement of the model when no drift is found with batch training of the model when it is needed. A main advantage of our method is that it does not rely on pre-calculated time slices but it is a real online algorithm. The data used to train the model after drift detection is automatically selected by the algorithm.

The remainder of this article is organized as follows. In Section II, we propose a reminder of LDA and ADWIN, on which our method is based. In Section III, we present our method and the algorithm. An experimental validation is presented in Section IV and a conclusive discussion is proposed in Section V.

II. Reminder: LDA and ADWIN

A. Latent Dirichlet Allocation

Latent Dirichlet Allocation [10] is a probabilistic graphical model designed to provide a definition of documents based on latent features called *topics*. A topic corresponds to a word distribution and a document is modeled as a weighted mixture of topics. The generative process can be described as follows:

1) Choose $\theta \sim Dirichlet(\alpha)$
2) For each word W_n in document:
   a) Choose a topic $z_n \sim Mult(\theta)$.
   b) Choose a word w_n for the multinomial $p(w_n|z_n, \beta)$.

The corresponding generative model is given in figure 1.

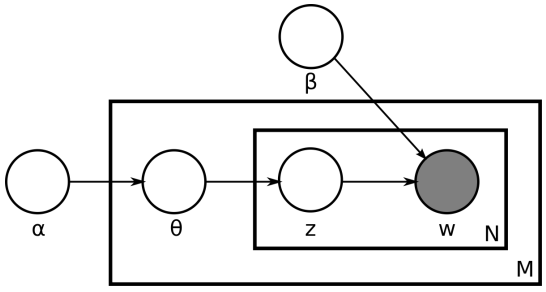

Fig. 1. Generative model of Latent Dirichlet Allocation

In this model, several parameters have a direct interpretation in terms of document analysis. First, the α parameter (hence the parameter of the Dirichlet distribution) influences the parameter of the multinomial topic distribution and corresponds to the mean value of a topic distribution θ inside a document. For instance, when $\alpha = (1, \ldots, 1)$, the topics are uniformly represented in documents. This parameter is important in document stream analysis since it depicts the topic trends. The parameter β is a word-topic distribution: the t-th column of matrix β is the vector of probabilities for a word to be drawn inside t-th topic.

LDA is trained either offline [10] or online [23], following Maximum Likelihood Principle. The algorithms used for the optimization are usually based on variational inference or Gibbs sampling. These methods will not be discussed in this paper.

B. Adaptive Sliding Window

Adaptive Sliding Window [22] is an algorithm developed for active mining of data streams. The idea of ADWIN is to keep a sliding window W with the most recent observations of a stream of real-valued elements x_t. At every time step, the algorithm adds the new element to the window W and decides whether the new W contains a drift or not.

The principle of ADWIN can be summed up as follows. The algorithm compares the mean of elements of sub-windows W_1 and W_2 of W. If the difference of means μ_{W_1} and μ_{W_2} of the two sub-windows is large enough and the size of the windows is large enough, then a drift is detected. The non-rigorous notions of "large enough" is defined through the choice of a statistical test for the detection.

Besides its real simplicity, ADWIN admits a couple of interesting theoretical properties. Among them, bounds are given for the probability of *incorrectly* splitting the current window (false positive rate bound) and *correctly* splitting the window (false negative rate bound).

III. Adaptive windowing for topic drift detection

A. Principle

The proposed method for topic change detection is based on the use of ADWIN combined with a training of LDA. We propose a framework in which documents arrive one by one in the form of a data stream. A document received at time step t is denoted by $\mathbf{w}_t$. Given parameters (α, β) and known latent variables (z, θ), the likelihood of the model is given by:

$$\mathcal{L}(\mathbf{w}) = \frac{\Gamma\left(\sum_i \alpha_i\right)}{\prod_i \Gamma(\alpha_i)} \int \left(\prod_{i=1}^{k} \theta_i^{\alpha_i - 1}\right) \left(\prod_{n=1}^{N} \sum_{i=1}^{k} \prod_{j=1}^{V} (\theta_i \beta_{ij})^{w_n^j}\right) d\theta \quad (1)$$

where Γ is the gamma function and w_n^j measures the quantity of word j in document n.

A change in the stream of likelihood corresponds to a change in the data distribution and can be detected by ADWIN algorithm. The selected indexes by ADWIN correspond to the documents received after the drift.

The principle of our method relies on a couple of intuitive guarantees:

- The likelihood measures the generative quality of the model with regards to observed data. When a model is not adapted, the likelihood decreases.

- ADWIN is sensitive to changes in the mean value of a time series. Thus, it will detect a change in the likelihood caused by a change of the model.
- ADWIN will select large sub-windows to train a new LDA model. The drift will be predicted with a better accuracy for large window sizes, which is also optimal to train a LDA model.

Following this idea, our method can be described as follows. At time step t, the system has access to a LDA model M_t which describes the data. When the system gets a new document, we compute the likelihood of observing the document, adds it to the current window, and inspects it with ADWIN to check if a drift occurred. When a drift is detected, the current LDA model is trained on the documents selected by the kept sub-window.

B. Algorithm

The algorithm we present is a direct implementation of these ideas. It is based on the idea of separating the tasks of document modeling and topic drift detection by associating a different model for each task. The LDA model used for document modeling is denoted by *LDA$_m$* and the LDA model used for drift detection is denoted by *LDA$_d$* .

The LDA model used for document modeling, *LDA$_m$* , is updated with each received document and retrained when a drift is detected.

For each received document, our approach, called *Adaptive Window based Incremental LDA* (AWILDA), computes the associated likelihood of the model *LDA$_d$* and adds it to ADWIN. If a drift is detected, the model *LDA$_m$* is retrained on the sub-window selected by ADWIN. Besides, *LDA$_m$* is updated with each received document based on Online LDA algorithm. We note that to initialize the model, we train it on a relatively small chunk of documents before starting the detection.

Whereas the LDA model used for document modeling, *LDA$_m$*, is updated with each received document and retrained when a drift is detected, the LDA model used for topic drift detection, *LDA$_d$*, is retrained on the sub-window selected by ADWIN for each detected drift. It is not updated as more documents are received.

C. Theoretical guarantees

Since it is based on theoretically trusted algorithms, AWILDA presents interesting theoretical properties which guarantee the quality of its results regarding drift detection.

We introduce the same notations as presented in [22]. We consider a window W of length n which is divided into two sub-windows W_0 and W_1 of respective sizes n_0 and n_1. Let m be the harmonic mean of n_0 and n_1 (hence $\frac{1}{m} = \frac{1}{n_0} + \frac{1}{n_1}$). We suppose that, in ADWIN, the drift is detected for $|\hat{\mu}_{W_1} - \hat{\mu}_{W_0}| \geq \epsilon_{cut}$ (where $\hat{\mu}_{W_0}$ designates the mean value over sub-window W_0). Let δ be such that:

$$\epsilon_{cut} = \sqrt{\frac{1}{2m} \ln \frac{4n}{\delta}} \tag{2}$$

With these parameters, Theorem 3.1 in [22] ensures both false positive rate bound and false negative rate bound. These results can be adapted to our setting.

Theorem 1. *At every time step, if documents are generated by a single LDA model in time period covered by W, the probability that AWILDA detects a drift at this step is at most δ.*

Proof. On the covered window, ADWIN gets a time series $X_t = \mathcal{L}(D_t)$ where D_t are equally distributed (for a single LDA model) and $\mathcal{L}$ represents the likelihood of *LDA$_d$* which is constant on W for AWILDA. Thus the mean of the variables remains constant on W. The conclusion follows from the properties of ADWIN. □

Following the same direction, the following theorem can be proven for false negative rate bound.

Theorem 2. *Suppose that, at a time step t, window W can be split in two parts W_0 and W_1 and documents are independent and identically distributed by a LDA distribution LDA_0 (resp. LDA_1) on sub-window W_0 (resp. W_1). If $|\mathbb{E}_{D\sim LDA_d}[p_{LDA_1}(D) - p_{LDA_0}(D)]| \geq 2\epsilon_{cut}$, then with probability $1-\delta$ AWILDA detects a drift inside sub-window W_1.*

Proof. The idea of the proof is the same. The mean value of $X_t = \mathcal{L}(D_t)$ on sub-window W_0 is:

$$\mu_t = \mathbb{E}_{D\sim LDA_0}[p_{LDA_d}(D)] = \mathbb{E}_{D\sim LDA_d}[p_{LDA_0}(D_t)]$$

An equivalent result can be found for W_1, and the theorem comes directly. □

Unlike for theorem 1, a simple interpretation of theorem 2 is not direct. For instance, two LDA models can be distinct and not share the targeted property. Finding conditions on the parameters of the three distributions is an interesting task that we will not address in this paper. However, it has to be noticed here that the guarantee on the false negative rate depends on the choice of *LDA$_d$* .

D. Nature of the drift

In practice, concept drift can happen in different ways. A drift is called *abrupt* when it happens at a given time step at any amplitude. On the other hand, a drift is called *gradual* when small distribution variations are happening at each time step on a certain period of time.

The case of abrupt drift has been explicitly studied with the setting of theorem 2. It corresponds to the case where the document distribution changes from one given state to another between sub-windows W_0 and W_1. Results given in [22] show that the detection delay can be estimated by $O(\mu \ln(1/\delta)/\epsilon^2)$ where μ is the mean of the distribution before drift. In our case, this delay is of critical importance since it defines the size of the chunk for retraining the model. AWILDA faces a trade-off between predicting a drift as early as possible (in order to maximize the likelihood) and collecting as many data

as possible to get a good estimator of the underlying LDA model.

The case of gradual drift is less adapted to the developed framework. Properties of ADWIN have been shown in the case of a linear gradual drift, but these results are difficult to translate directly into our setting where the time series tracked by ADWIN has a complex mathematical definition. Understanding the behavior of AWILDA in the case of gradual drift is a task that would come together with a proper study of theorem 2.

In our experiments, we will consider abrupt drifts only. A related discussion will be proposed in the conclusion.

IV. Experimental results

In this Section, we present the experiments we conducted in order to prove the effectiveness of our approach. We show in particular how it performs when addressing the problems of topic drift detection and document modeling, using a set of synthetic and real datasets.

A. Datasets

Synthetic data. To demonstrate the ability of detecting drifts, we generate synthetic datasets where we artificially insert drifts at random moments throughout the sequence of documents. Synthetic datasets are denoted by Sd_r, where r is the number of simulated drifts. Documents observed between two consecutive drifts are generated by one LDA model following its generative process. At each occurring drift, we draw uniformly the hyperparameters α and β. The number of topics is fixed for all the models used to generate one dataset.

We present experiments performed on the following two synthetic datasets: Sd_4 and Sd_9, containing 4 and 9 drifts respectively. Handling document streams is a very common task in environments where short texts are generated and shared, e.g., newswires, tweets. Thus, we choose to generate documents containing 100 words, and we fix the vocabulary size to 10,000 words and the number of topics, k, to 15. Following the setting in [10], α and β are first set to $50/k$ and 0.1 respectively, and are then changed at each drift. In Sd_4, we generate exactly 2,000 documents from each distribution, separating two consecutive drifts by the same number of documents. In Sd_9, we vary the number of documents generated by each model between 500 and 1,000 documents.

Real data. We also conduct experiments on real-world data. We use the dataset *Reuters-21758* [1] consisting of newswire articles classified by categories and ordered by their date of issue. The ApteMod version of this database contains 12,902 documents and each document is classified in multiple categories for a total of 90 categories. In the procedure of data preprocessing, we down-cased and stemmed all words in the articles.

Our approach is designed to detect topic drifts in document streams and to adapt the model accordingly. To demonstrate this functionality on real data, we reorder the newswire articles based on their categories. We artificially ensure an emergence of topics at specific points of the document stream and we try to provoke a drift in the topic distributions.

We derive from the initial ordered dataset two sets of articles that we use in our experiments. In the first set, denoted by $Reuters_1$, we select the articles belonging to the category "acq" followed by the articles belonging to the category "earn". We expect the algorithm to detect the sudden change in topics mentioned in the documents. In the second set, denoted by $Reuters_4$, we select articles classified in a specific category and add them consecutively to the dataset. This is done for the five following categories: "interest", "trade", "crude", "grain", and "money-fx".

B. Setting of AWILDA

As defined in equation 1, the likelihood of a LDA model is not computable. Thus, we relied on an upper-bound $\mathcal{L}'$ proposed in variational inference (see equation 1 in [23]). In practice, the results observed with this upper-bound are not satisfying due to a lack of precision: the probabilities to observe data are very low and the method fails at discriminating them with enough accuracy. In order to overcome this difficulty, we considered the logarithm of $\mathcal{L}'$ (hence an upper-bound of log-likelihood). This quantity is theoretically unbounded, which is a problem for ADWIN, but in practice it is observed that the values vary only in a small interval (the width of which depends on the dataset). In our experiments, we prevented the quantity to decrease too much by fixing a minimal bound so that the quantity of interest becomes bounded. A reasonably low value for this threshold was never reached in the scope of the presented experiments. We do not have any way to evaluate an optimal value for this bound in a general case though.

C. Evaluation

Our evaluation concerns the tasks of topic drift detection and document modeling.

Topic drift detection. We evaluate the ability of detecting drifts by checking the latency between the moment when the real drift happens and the moment it is detected.

Document modeling. Given a LDA model trained on a set of documents, the goal in document modeling is to maximize the likelihood on unseen documents. For the evaluation, we use the measure of perplexity, which is defined by:

$$perplexity(D_{test}) = 2^{-\frac{\sum_{d=1}^{M} \log_2 p(w_d)}{\sum_{d=1}^{M} N_d}} \quad (3)$$

Perplexity is the tool used by default in language modeling to measure the generalization capacity of a model on new data. Since we are considering document streams, the perplexity is computed for each received document using the current model.

The performance of our approach is compared to the online version of LDA [23]. In the standard version of LDA [10], the model is learned in batch. In comparison, online LDA can analyze document collections arriving in a stream and is therefore more adapted to the setting we adopt in this

[1] http://archive.ics.uci.edu/ml/

work. In our experiments, online LDA is considered to process documents one by one as they are received and updates the underlying model at each step.

We also compare AWILDA to three other variants. In these variants, the model LDA_m is updated in a similar way as for AWILDA, but the methods differ in the way the detection model LDA_d is updated:

- **AWILDA-2.** LDA_d is trained on a first small chunk of documents that is used to initialize all the models. It is not updated as more documents are received.
- **AWILDA-3.** LDA_d is updated for each received document and is equivalent to a classic online LDA model.
- **AWILDA-4.** LDA_d is updated for each received document using online LDA algorithm. When a drift is detected, the model is retrained on the sub-window selected by ADWIN.

Regarding the theoretical study, it can be easily verified that theorem 1 and theorem 2 hold true for AWILDA-2, but not for AWILDA-3 and AWILDA-4. In particular, it is noticeable that we do not have guarantees for the performances of AWILDA-3 and AWILDA-4 since the model LDA_d is updated at each step. *A priori*, there is no chance that the means remain constant when the likelihood function $\mathcal{L}$ varies.

D. Results

1) Comparison of AWILDA and its variants on Sd_4: In the first set of experiments, we compare the performance of AWILDA and its variants when performing the task of topic drift detection on the synthetic dataset Sd_4. The results are presented in Figure 2. The LDA model used to compute perplexity is learned and updated differently depending on the method considered (see Section III). We represent the perplexity as a moving average with a sliding window of 100 observations. The exact occurrence of drifts is marked by a green dashed vertical line and the detection of drifts is marked by a blue dotted vertical line.

AWILDA and AWILDA-2 detect only true positive drifts, while AWILDA-3 and AWILDA-4 detect false and true positive drifts. AWILDA is also more reactive than AWILDA-2 and spots drifts faster. Updating the LDA_d model with each received document in AWILDA-3 and AWILDA-4 modifies the underlying distribution of topics, leading ADWIN to detect false positive drifts.

AWILDA performs best for all the studied datasets, and we present in the following the results for the other datasets.

2) Performance of AWILDA on all the datasets: As shown in Figure 3, AWILDA is able to detect all the drifts occurring in the datasets Sd_4, Sd_9, and $Reuters_1$ after receiving only a few observations from the new distribution. Concerning the $Reuters_4$ dataset, our approach spots two drifts and misses the two others. We note that in this particular dataset, we switch from a topic to another relatively fast, i.e., around 500 documents per category. Topics in articles can also be interconnected which makes the task even more complicated.

3) Comparing AWILDA with online LDA: In the last set of experiments, we compare our approach with online LDA [23] using the dataset $Reuters_1$. We show, in Figure 4, how the perplexity evolves throughout the set of documents before and after the drift occurs. The perplexity is computed using LDA_m of AWILDA.

While processing the first set of arriving documents, online LDA and AWILDA are trained in the same fashion and lead to the same performance. When the drift occurs, AWILDA retrains its model on the relatively small sub-window selected by ADWIN, which explains the temporary increase in perplexity. As documents continue arriving, the model is adapted to the new data and AWILDA outperforms online LDA. We note that AWILDA performs better than online LDA when averaging on all the observed documents of the dataset.

E. Discussion

We notice that the observed properties of the four variants of the algorithm are close to the predictions which were given by theorems 1 and 2. In particular, it has been shown that AWILDA and AWILDA-2 would perform better than AWILDA-3 and AWILDA-4 with regards to false positives.

The superiority of AWILDA over the other variants raises interesting questions. It is noticeable that the best algorithm in terms of drift detection is also the only one which detection model is actively updated at each drift and not passively, at each step or for each observation. This property is particularly interesting: it means that the best algorithm in terms of drift detection is also the most efficient one in terms of computation time. However, in some examples, it might not be the optimal algorithm for the accuracy of the predicted model: there is no theoretical guarantee that the documents selected by ADWIN are a good representative set for the new distribution.

Moreover, the non-updating property of AWILDA is of particular interest: it illustrates the idea that good drift prediction does not require to have good modeling properties, which may be counter-intuitive in a way. The extreme case, AWILDA-2, shows rather good performance as well whereas the detection model is never updated, which means that it does not encode any information relative to the underlying distribution. A random LDA model could also work for this task. Having a completely unrelated detection model might produce false-negative errors though: if the detection model is too different from the actual model, there is a chance that the likelihood change, when the underlying model varies,is not important enough to be detected.

V. Conclusion

In this paper, we have presented a novel method for topic modeling on document streams. The proposed approach, called *Adaptive Window based Incremental LDA* (AWILDA), is a combination of Latent Dirichlet Allocation (LDA) and Adaptive Windowing (ADWIN). The algorithm combines two LDA models, one for topic modeling and the other for detecting drifts based on an adaptive sliding window. Despite its simplicity, the method has several advantages. First, theoretical

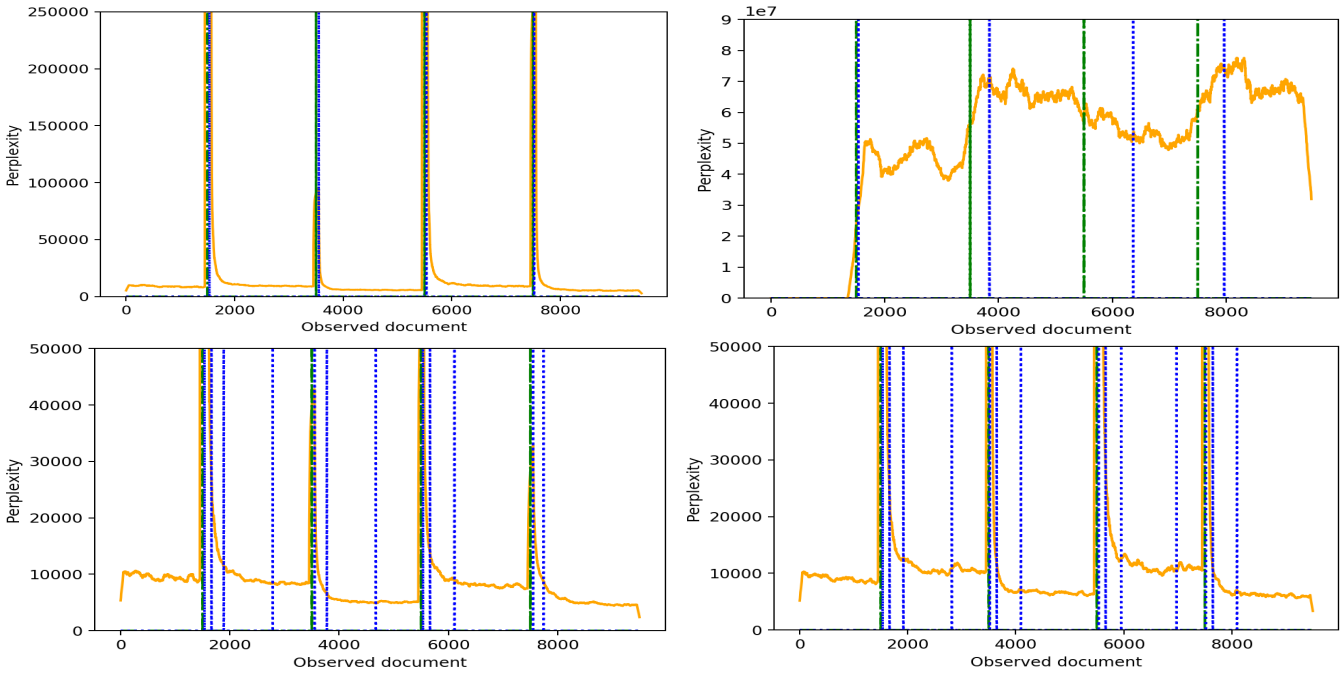

Fig. 2. Topic drift detection on the Sd_4 dataset using AWILDA (first figure), AWILDA-2 (second figure), AWILDA-3 (third figure), and AWILDA-4 (fourth figure).

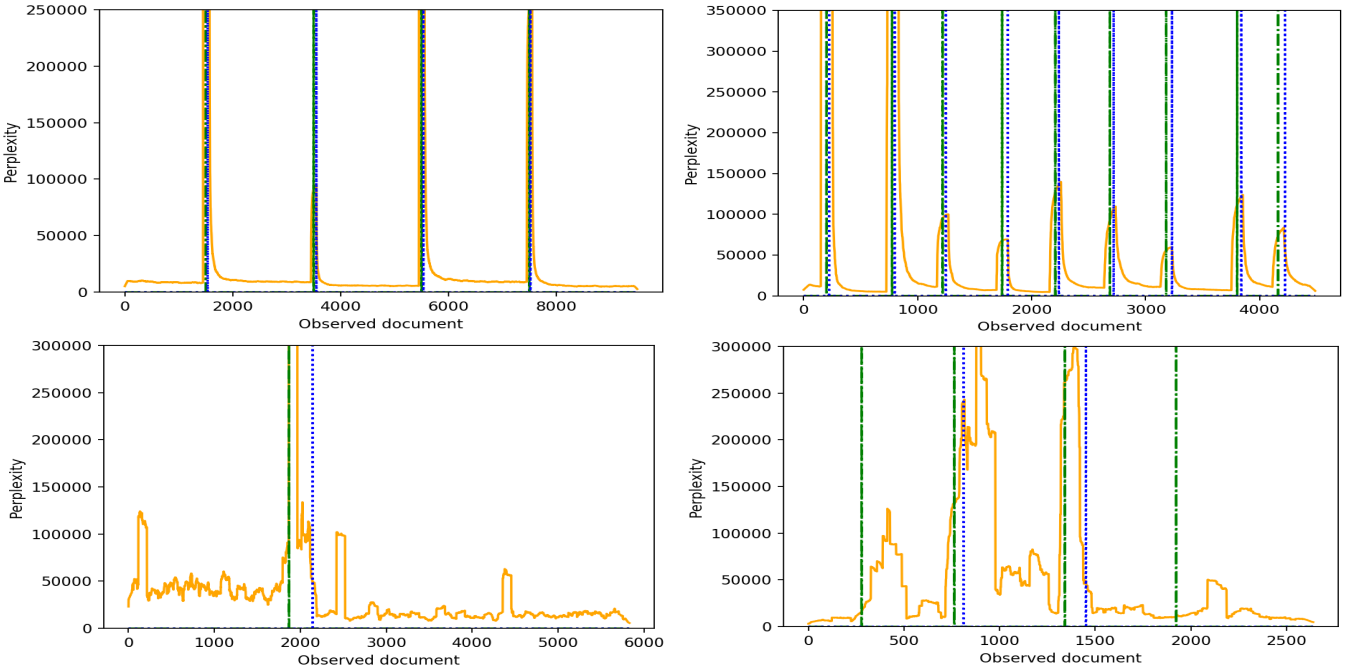

Fig. 3. Topic drift detection using AWILDA and applied on the Sd_4 dataset (first figure), Sd_9 dataset (second figure), $Reuters_1$ dataset (third figure), and $Reuters_4$ dataset (fourth figure).

guarantees can be inferred directly from the theories of ADWIN. Secondly, the method overcomes several disadvantages of concurrent frameworks. The algorithm selects by itself the document it will use, which avoids defining arbitrary time slices and document chunks as in most methods, and which guarantees a true online setting. Our algorithm can work on real-time streaming since the model is retrained only when necessary. AWILDA framework is well-designed for abrupt drifts and can also work when gradual drifts occur. In practice, the nature of the drift highly depends on the task. Gradual concept drift corresponds to slow evolution of topics that can be found for instance in news articles, while abrupt

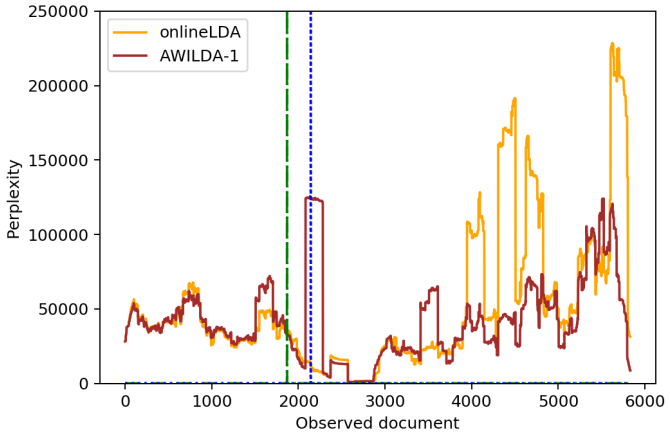

Fig. 4. Comparison of online LDA and AWILDA for the task of document modeling.

drifts corresponds to sudden changes in the distribution, hence to important events. Using AWILDA on event detection on Twitter can be an interesting application of our model.

Other applications are numerous since LDA has become a major technique in the recent years. In particular, text mining on larger datasets and recommender systems are obvious domains of application for AWILDA. Refinements of our model would be of great interest too. Instead of detecting one global change using one ADWIN module, it might be possible to detect changes inside the same topic by running one ADWIN module per topic. Finally, theoretical issues regarding the detection performance of the proposed algorithms are still open.

Acknowledgment

This research is supported by the program Futur & Ruptures (Institut Mines-Télécom). The authors would like to thank Albert Bifet and Jacob Montiel for their valuable help.

References

[1] S. F. Chen and J. Goodman, "An empirical study of smoothing techniques for language modeling," in *Proceedings of the 34th annual meeting on Association for Computational Linguistics*. Association for Computational Linguistics, 1996, pp. 310–318.

[2] T. Hofmann, "Probabilistic latent semantic analysis," in *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc., 1999, pp. 289–296.

[3] J. Chang, S. Gerrish, C. Wang, J. L. Boyd-Graber, and D. M. Blei, "Reading tea leaves: How humans interpret topic models," in *Advances in neural information processing systems*, 2009, pp. 288–296.

[4] X.-H. Phan, L.-M. Nguyen, and S. Horiguchi, "Learning to classify short and sparse text & web with hidden topics from large-scale data collections," in *Proceedings of the 17th international conference on World Wide Web*. ACM, 2008, pp. 91–100.

[5] M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth, "The author-topic model for authors and documents," in *Proceedings of the 20th conference on Uncertainty in artificial intelligence*. AUAI Press, 2004, pp. 487–494.

[6] C. Wang and D. M. Blei, "Collaborative topic modeling for recommending scientific articles," in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2011, pp. 448–456.

[7] B. Hu and M. Ester, "Spatial topic modeling in online social media for location recommendation," in *Proceedings of the 7th ACM conference on Recommender systems*. ACM, 2013, pp. 25–32.

[8] Y. Rao, Q. Li, X. Mao, and L. Wenyin, "Sentiment topic models for social emotion mining," *Information Sciences*, vol. 266, pp. 90–100, 2014.

[9] Y. Feng and M. Lapata, "Topic models for image annotation and text illustration," in *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 2010, pp. 831–839.

[10] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.

[11] D. M. Blei and J. D. Lafferty, "Dynamic topic models," in *Proceedings of the 23rd international conference on Machine learning*. ACM, 2006, pp. 113–120.

[12] L. Du, W. Buntine, H. Jin, and C. Chen, "Sequential latent dirichlet allocation," *Knowledge and information systems*, vol. 31, no. 3, pp. 475–503, 2012.

[13] X. Wang and A. McCallum, "Topics over time: a non-markov continuous-time model of topical trends," in *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2006, pp. 424–433.

[14] T. L. Griffiths and M. Steyvers, "Finding scientific topics," *Proceedings of the National academy of Sciences*, vol. 101, no. suppl 1, pp. 5228–5235, 2004.

[15] L. AlSumait, D. Barbará, and C. Domeniconi, "On-line lda: Adaptive topic models for mining text streams with applications to topic detection and tracking," in *Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on*. IEEE, 2008, pp. 3–12.

[16] T. Iwata, T. Yamada, Y. Sakurai, and N. Ueda, "Online multiscale dynamic topic models," in *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2010, pp. 663–672.

[17] M. Sayed-Mouchaweh and E. Lughofer, *Learning in non-stationary environments: methods and applications*. Springer Science & Business Media, 2012.

[18] J. Gama, I. Žliobaitė, A. Bifet, M. Pechenizkiy, and A. Bouchachia, "A survey on concept drift adaptation," *ACM computing surveys (CSUR)*, vol. 46, no. 4, p. 44, 2014.

[19] G. Ditzler, M. Roveri, C. Alippi, and R. Polikar, "Learning in non-stationary environments: A survey," *IEEE Computational Intelligence Magazine*, vol. 10, no. 4, pp. 12–25, 2015.

[20] J. H. Lau, N. Collier, and T. Baldwin, "On-line trend analysis with topic models:\# twitter trends detection topic model online," *Proceedings of COLING 2012*, pp. 1519–1534, 2012.

[21] G. Widmer and M. Kubat, "Learning in the presence of concept drift and hidden contexts," *Machine learning*, vol. 23, no. 1, pp. 69–101, 1996.

[22] 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.

[23] M. Hoffman, F. R. Bach, and D. M. Blei, "Online learning for latent dirichlet allocation," in *advances in neural information processing systems*, 2010, pp. 856–864.