# Don’t Shoot The Breeze: Topic Continuity Model Using Nonlinear Naïve Bayes With Attenion

Shu-Ting Pi, Pradeep Bagavan, Yejia Li, Disha, Qun Liu
Amazon
{shutingp, deepgag, imyejia, disha, qunliu}@amazon.com

## Abstract

**Utilizing Large Language Models (LLM) as chatbots in diverse business scenarios often presents the challenge of maintaining topic continuity. Abrupt shifts in topics can lead to poor user experiences and inefficient utilization of computational resources. In this paper, we present a topic continuity model aimed at assessing whether a response aligns with the initial conversation topic. Our model is built upon the expansion of the corresponding natural language understanding (NLU) model into quantifiable terms using a Naive Bayes approach. Subsequently, we have introduced an attention mechanism and logarithmic nonlinearity to enhance its capability to capture topic continuity. This approach allows us to convert the NLU model into an interpretable analytical formula. In contrast to many NLU models constrained by token limits, our proposed model can seamlessly handle conversations of any length with linear time complexity. Furthermore, the attention mechanism significantly improves the model's ability to identify topic continuity in complex conversations. According to our experiments, our model consistently outperforms traditional methods, particularly in handling lengthy and intricate conversations. This unique capability offers us an opportunity to ensure the responsible and interpretable use of LLMs.**

## Introduction

Large-scale language models (LLMs) have transformed chatbot applications, enabling them to serve as office assistants, coding companions, and data explorers. However, deploying LLMs in business settings, particularly in customer service, presents significant challenges. One key issue is maintaining topic coherence, as off-topic responses can degrade user experience and waste resources. In customer service, chatbots must ensure each sentence logically follows the previous ones to keep conversations on track. This is critical because disjointed conversations lead to inefficiencies and frustration.

To address this, a natural language understanding (NLU) model can be used to assess whether each new sentence in a conversation remains on-topic. If a sentence diverges, the conversation should be redirected or concluded. This challenge is often tackled using BERT-based models, which evaluate the contextual relationship between sentences. However, BERT models face two significant challenges: token size limits and the complexity of real-world conversations. The token limit, typically around 512 tokens, can be problematic as conversations grow longer. Additionally, BERT models are trained on sentence pairs with close semantic relationships, while real conversations often have more distant connections, making it harder to assess the relevance of follow-up sentences.

To overcome these issues, a new topic continuity model is proposed. This model integrates logarithmic nonlinearity and sentence attention within a naive Bayes framework, offering a fully analytical solution. By addressing the token size limit and accommodating semantic leaps in conversations, this model significantly improves the ability of chatbots to maintain topic coherence, enhancing their effectiveness in customer service roles.

## Nolinear Naive Bayes With Attention Mechanism

When a user is engaged in a conversation with a chatbot, our goal is to identify topic shifts in new sentences, **assuming that the first N-1 sentences are on-topic**. As discussed in Section 1, we can define an NLU model for this problem as a conditional probability expressed as follows:

$$P(y|S_1, S_2, \ldots; S_N) \quad (1)$$

Naive Bayes assumption, where the variables $(S_1, \ldots; S_N)$ are considered independent of each other, and we expand Eq.(1) upon this assumption as follows:

$$P(y|S_1, S_2, \cdots; S_N) = \frac{P(S_1|y)P(S_2|y)\cdots P(S_N|y)P(y)}{P(S_1)P(S_2)\cdots P(S_N)} = \Pi_i^N \left[\frac{P(S_i|y)}{P(S_i)}\right] P(y) \quad (2)$$

We aim to incorporate an attention mechanism into Eq.(2). To achieve this, we have intentionally reformulated the equation to include pairwise probability. Consequently,

$$P(y|S_i, S_N) = \frac{P(S_i, S_N|y)P(y)}{P(S_i, S_N)} = \frac{P(S_i|y)P(S_N|y)P(y)}{P(S_i)P(S_N)}$$

Thus,

$$P(S_i|y) = \frac{P(y|S_i, S_N)P(S_i)P(S_N)}{P(S_N|y)P(y)}$$

Let’s plug this term into Eq.(2). We have,

$$P(y|S_1 \ldots; S_N) = \Pi_i^N \left\{\frac{P(y|S_i, S_N)P(S_i)P(S_N)}{P(S_N|y)P(y)} \frac{1}{P(S_i)}\right\} P(y)$$

So,

$$P(y|S_1 \cdots; S_N) = \Pi_i^N \{P(y|S_i, S_N)\} P^{-N}(S_N|y)P^N(S_N)P^{1-N}(y)$$

Take log on both side,

$$logP(y|S_1 \cdots; S_N) = \sum_{i=1}^{N}\{logP(y|S_i, S_N)\} - NlogP(S_N|y) + NlogP(S_N) + (1-N)logP(y)$$

Note that in the first summation, there exists a term $logP(y|S_N, S_N)$, which can be approximated as $logP(y|S_N, S_N) \approx logP(y|S_N) = logP(S_N|y) + logP(y) - logP(S_N)$. Additionally, the term $logP(y)$ is essentially a constant and does not affect any of the subsequent calculations, so we can safely disregard this term. Thus, we have:

$$logP(y|S_1 \cdots; S_N) = \sum_{i=1}^{N-1}\{logP(y|S_i, S_N)\} + (N-1)\left[logP(S_N) - logP(S_N|y)\right] \quad (3)$$

Based on the above discussion, a straightforward approach is to maintain the mathematical form but introduce more non-linear operations. This can be achieved by replacing $\sum \rightarrow \mathcal{F}$ and $(N-1) \rightarrow \alpha$ as shown below:

$$logP(y|S_1 \cdots S_N) = \mathcal{F}\left\{logP(y|\tilde{\mathbf{S}}, S_N)\right\} + \alpha(\mathbf{S})\left[logP(S_N) - logP(S_N|y)\right] \quad (4)$$

## Attention Term

### 3.1 Designing Attention Functional

In a conservation, sentences typically fall into three scenarios: **1). Normal Sentences** correspond to responses to the previous sentence, the most frequent scenario. **2). Leap Sentences** correspond to responses to earlier sentences in the conversation, constituting a “leap conversation". In the following, we use the term "*target sentence*" to denote the sentence that the current sentence $S_N$ responds to. **3). Topic Shift Sentences** indicate a shift in topic.

To capture these three scenarios, we define the notation $log\mathcal{P}_{max} = \max\{logP(y|\tilde{\mathbf{S}}, S_N)\}$ and $log\mathcal{P}_{avg} = avg\{logP(y|\tilde{\mathbf{S}}, S_N)\}$. Then the attention functional is defined as:

$$\mathcal{F}\left\{logP(y|\tilde{\mathbf{S}}, S_N)\right\} = [1 + \tanh(log\mathcal{P}_{max})]\, log\mathcal{P}_{max} - \tanh(log\mathcal{P}_{max})log\mathcal{P}_{avg} \quad (5)$$

## Residual Term

To fulfill these criteria, a straightforward mathematical form is a sine function:

$$P_{nlu} = P_{att} + \beta \sin(\pi P_{att})$$

,where $\beta \ll 0.5$. The condition $\beta \ll 0.5$ arises from the situation where the perturbation term attains its maximum value at $P_{att} = 0.5$ and $P_{nlu} = 0.5 + \beta$. Given its nature as a perturbation, $\beta$ must be $\ll 0.5$. By taking a logarithm on both side, we get:

$$logP_{nlu} = log[P_{att} + \beta \sin(\pi P_{att})] = log(P_{att}) + log[1 + \beta sin(\pi P_{att})/P_{att}]$$

. Since $\beta sin(\pi P_{att})/P_{att} \ll 1$, first order of Taylor expansion yields

$$logP_{nlu} \approx log(P_{att}) + \beta \sin(\pi P_{att})/P_{att}$$

.

Comparing this from with eq.(4), we assert $\alpha$ should be:

$$\alpha = \frac{\sin(\pi e^{\mathrm{F}\{P(y|\mathbf{S},S_N)\}})}{e^{\mathrm{F}\{P(y|\mathbf{S},S_N)\}}} \frac{\eta}{|log(\epsilon)|} \quad (6)$$

Here, $P_{att}$ is represented as its original form $e^{\mathrm{F}\{P(y|\mathbf{S},S_N)\}}$ and the term $\eta/|log(\epsilon)|$ serves as a scaling factor with $\eta \ll 0.5$

## Datasets

For the experiment, we used a dataset that was generated by professional customer service agents interacting with an LLM, simulating customers asking the LLM questions related to online video streaming. The dataset was entirely generated through simulation and did not use any real user data, with the purpose of protecting user privacy. We sampled data from the following four different categories:

- **Normal Conversation (on-topic)** 1000 data points where the current sentence responses to the preceding sentence.
- **Leap Conversation (on-topic)** 1000 data points in which the current sentence is a response to an earlier sentence in the conversation.
- **Out-of-Domain Topic Shift (off-topic)**: 1000 data points where the current sentence diverges completely from the main topic and is entirely unrelated to Amazon’s services.
- **In-domain Topic Shift (off-topic)** 1000 data points in which the current sentence diverges significantly from the main topic but remains relevant to Amazon’s services.

## Calculation of Likelihoods

The attention term $P(y|S_i, S_N)$ involves determining the contextual relationship between $S_i$ and $S_N$, which can be estimated using language models such as BERT. This task is commonly known as Next Sentence Prediction (NSP) in many machine learning studies [19, 20]. Numerous open-source NSP models are available on platforms like Hugging Face, making it unnecessary for us to retrain them.

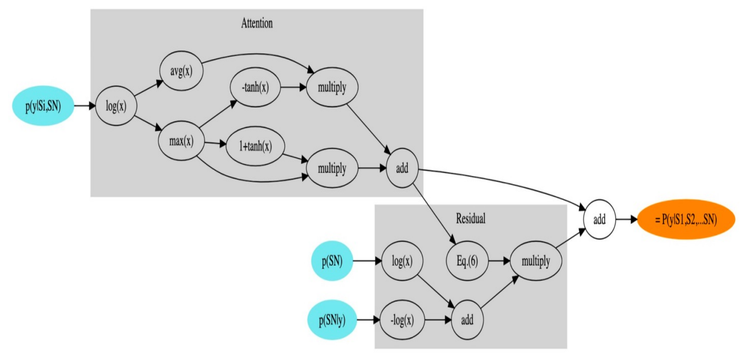
Estimating $P(S_N|y)$ and $P(S_N)$ involves context-dependent factors. In theory, these quantities should be calculated by integrating over all relevant variables. However, performing these integrals directly is impractical. Therefore, we employ an indirect approach.

Consider, for example, a customer service chatbot designed to respond to various product-related queries, such as those about "cell phones." To establish $P(S_N|y)$ for the "cell phone" topic, we randomly sample numerous sentences from historical conversations on this subject. The likelihood of a sentence appearing in the context of this topic can then be estimated using an out-of-distribution (OOD) method, such as Isolation Forest [21, 22]. The process is as follows:

- Encode each sentence using pre-trained models, such as Sentence BERT [23].
- Train an Isolation Forest on this dataset to generate anomaly scores for all sentences. We invert the sign compared to the original approach, so higher anomaly scores ($\theta$) indicate a greater likelihood of a sentence being part of the dataset.
- Once the distribution of $\theta$ is obtained, we estimate its probability density function $p(\theta)$. For a future sentence with a score $\theta = c$, the corresponding probability is determined by the Cumulative Distribution Function (CDF):

$$P(S_N|y) = \int_{-\infty}^{c} p(\theta)\, d\theta.$$

## Computation Graph




## Experiments

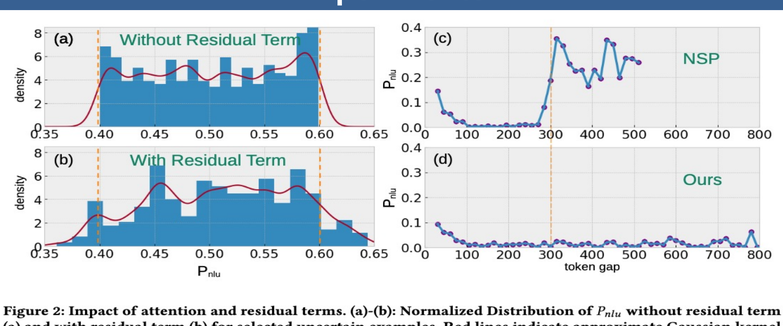



**Figure 2: Impact of attention and residual terms. (a)-(b): Normalized Distribution of $P_{nlu}$ without residual term (a) and with residual term (b) for selected uncertain examples. Red lines indicate approximate Gaussian kernel density fitting. (c)-(d): Average probability output per segmentation, categorized by token length, is shown in (c) for NSP and (d) for our model. The dashed lines denote 300 tokens. Data beyond 512 tokens were truncated in (c) due to NSP’s processing limit.**

| | $\Delta_T \leq 300$ | | $300 < \Delta_T \leq 512$ | | $\Delta_T > 512$ | |
|---|---|---|---|---|---|---|
| Metrics | NSP | Ours | NSP | Ours | NSP | Ours |
| Precision | **0.747** | 0.734 | 0.612 | **0.697** | 0.588 | **0.703** |
| Recall | 0.961 | **0.983** | **0.982** | 0.972 | 0.917 | **0.980** |
| Accuracy | **0.818** | 0.814 | 0.679 | **0.775** | 0.637 | **0.783** |
| F1 score | 0.840 | **0.841** | 0.754 | **0.812** | 0.717 | **0.819** |

**Table 1: Comparison among different models with varying token gap lengths $\Delta_T$. The differences between NSP and our model are minimal for narrow token gap but gradually increase as the token gap widens.**

## Conclusion

With the rapid development of large language models (LLMs), the effective utilization of LLMs in various business scenar- ios has become an important issue. In this paper, we propose a method that ensures user conversations with LLMs remain focused on fixed topics. This method is based on the intro- duction of non-linear transformations and attention mecha- nisms through an extension of Naive Bayes. Experimental results across various scenarios consistently demonstrate that our approach outperforms traditional methods. We be- lieve this method will be highly beneficial for using LLMs in topic-constrained scenarios