International Journal of Scientific and Technical Research in Engineering (IJSTRE)
www.ijstre.com Volume 9 Issue 4 ǁ Sept-Oct 2024
ISSN: 2581-9941

News Classification Based On Statistical N-Gram Language Modeling

Hoa Tran Thi Thieu
Ha Tinh University, Viet Nam
tranthithieuhoa@gmail.com

Abstract: *Nowadays, media content is increasingly digitized and popularized as the digital transformation is taking place strongly. With the convenience of digital media and journalism, all information about areas of social life is transmitted quickly and readers can access it easily. News classification has become a necessary need to meet the needs of management agencies in quickly grasping emerging issues for timely handling.*

In this study, we rely on the n-gram statistical language model approach to solve the problem of news classification in Vietnamese. This approach has the advantage of avoiding two difficulties when performing the classification problem for Vietnamese, which is not having to perform the word segmentation problem and solving the problem of feature selection, which are challenges in processing Vietnamese that affect the classification results.

Keywords *- Texts classification, language modeling, n- gram.*

I. INTRODUCTION

Nowadays, the explosion and rapid development of Internet services, technology and applications on the Web platform have been thoroughly exploited by users to support their activities and work. Network information management is an important issue today when media content is increasingly digitized and popularized. Thanks to the convenience of digital media and journalism, economic and social information is quickly transmitted to readers. This issue for managers in capturing network information is important and urgent. Managers must be able to quickly capture information to have timely handling. Therefore, capturing and classifying network information that has occurred on network media is an important task for information management agencies. This task helps authorities to come up with timely response and handling plans in news management. To do this, managers must collect articles, evaluate, classify and synthesize them to have a basis for information processing. However, with the explosion of online news channels today, capturing and processing information manually is ineffective, time-consuming and costly in terms of human resources. With the development of natural language processing models, collecting, evaluating and classifying articles by news topics is a solvable task and gives feasible results. Researchers have solved this problem with the news classification problem, which is a specific case of the text classification problem. News classification is the task of extracting information and classifying it according to given topics.

News classification is a special case of text classification, this problem belongs to the field of natural language processing research, so the techniques and approaches are basically based on natural language processing methods or combine natural language processing methods with other methods. Currently, there are 3 main approaches to text classification: Bag of Words (BOW) approach [1], N-gram statistical language model approach [2] and hybrid approach of the above 2 methods [3]. With the characteristic of Vietnamese as a language with many features, researchers have used the Bag of Words (BoW) approach, this method is simple and can represent a sentence or a part of text. There are many machine learning techniques to apply to classification. However, one of the difficulties of this approach is the process of selecting features for the text. Vietnamese is a language with many features, when using this method for Vietnamese sentiment classification, some words are used frequently, while some words are used less, which can cause bias towards certain words in terms of word count. Meanwhile, the nature of each feature is an aspect that users comment on in the text contained in a group of sentences in text format. The text classification process is often represented by a feature

vector, which is a group of words. Another problem is that some disadvantages of using only single words as features are not addressed, negative points such as "not bad" or "not good" will not be taken into account. This shows that relying on single word features can lead to misclassification. Unlike English, word recognition in Vietnamese does not follow whitespace, so the BoW approach to the Vietnamese sentiment classification problem also encounters another difficulty, which is the word segmentation problem. The effectiveness of the classification depends heavily on the two steps of feature selection and word segmentation mentioned above. To solve the above difficulties, researchers have proposed a second approach, which is the N-gram statistical language model approach. This approach has the advantage of the word-based language model, which is that it does not have to solve the very difficult word segmentation problem for Vietnamese mentioned above and the feature selection problem is also solved. In addition, the word-based language model has a smaller size than the word-based model and also reduces the problem of sparse data. However, using statistical terminology also has disadvantages such as some terms will not be recognized well, or the recognition will be wrong. The effect of n-gram applications in text classification is to determine the frequency of occurrence in the first group of the sorted list. In practice, the number of k-grams will increase significantly because for every k-gram that appears, there will be 2(k+1)-grams that appear with it. Using feature filtering, the model will remove the grams that have a frequency lower than the average k-gram value. After obtaining the gram list, the model will vectorize the texts in the training set. Next, the model will build a training set of examples from the vectorized texts. The training set process can apply different machine learning methods to solve the problem. This implementation will be presented in the next section.

II. LANGUAGE MODELS AS TEXT CLASSIFIERS

Statistical language modeling is concerned with determining the probability of naturally occurring word sequences in a language. This is a common task in natural language processing. The n-gram model is a simple and effective way to perform language modeling, and in this model, a word is assumed to depend only on the previous n-1 words. In fact, many research works have proposed language models to improve the basic n-gram models. However, basic language modeling research is still a difficult problem. Improvements obtained from these complex language models often arise difficulties or increase their complexity. Therefore, up to now, in many situations, the n-gram language model is still the best choice in practice. Although basic language modeling research is difficult, language modeling is gaining increasing attention because it has been successfully applied to many real-world problems.

In this section, we use the statistical n-gram language modeling approach to classify documents. The approach of this method is that the most basic unit to be described is the word [2]. We will apply the syllable-based language model, that is, consider the text as a sequence of consecutive syllables. The reason for this is to take advantage of the syllable-based language model to avoid the word segmentation problem which is very difficult for Vietnamese and thus avoid feature selection (which can create duplicate or missing features as pointed out in section 1). The syllable-based language model is smaller than the word-based language model in terms of size and it also reduces the problem of data sparsity. The purpose of language modeling is to anticipate the probability of natural word sequences; or in a simpler way, to put high probability on word sequences which really occur (and the other probability that is low on word sequences which never occur). A word sequence $w_1w_2...w_T$ is given to be used as a test corpus and on this corpus, by the empirical perplexity (or entropy) the language model quality can be measured.

$$Perplexity = \sqrt[T]{\prod_{i=1}^{T} \frac{1}{P(w_i \mid w_1..w_{i-1})}} \quad (2.1)$$

$$Entropy = \log_2 Perplexity \quad (2.2)$$

The goal of language modeling is to obtain a small perplexity.

1. N-Gram Language Modeling

The n-gram model is the simplest and the most effective basis for language modeling. It is noted that the probability of any sequence can be written as follow by using the chain rule of probability:

$$P(w_1 w_2 \ldots w_T) = \prod_{i=1}^{T} P(w_i \mid w_1 \ldots w_{i-1}) \quad (2.3)$$

Any n-gram model approximates this probability by supposing that the only words relevant to forecasting $P(w_i | w_1 \ldots w_{i-1})$ are the previous n-1 words; in addition, it assumes the Markov n-gram independence assumption $P(w_i | w_1 \ldots w_{i-1}) = P(w_i | w_{i-n+1} \ldots w_{i-1})$

A straightforward maximum likelihood estimate of n-gram probabilities from a corpus is given by the observed frequency

$$P(w_i \mid w_{i-n+1} \ldots w_{i-1}) = \frac{\#(w_{i-n+1} \ldots w_i)}{\#(w_{i-n+1} \ldots w_{i-1})} \quad (2.4)$$

where #(.) is the number of occurrences of a specified gram in the training corpus.
Due to the heavy tailed language nature (Zipf's law) is probably likely to encounter novel n-grams which were never witnessed during training. Thus, some mechanism for assigning non-zero probability to novel n-grams is a central and unavoidable issue. One standard approach to smoothing probability estimates to cope with sparse data problems (and to cope with potentially missing n-grams) is to use some sort of back-off estimator [4].

In practice, the process of building an n-gram language model may encounter uneven distribution, that is, when using an n-gram model according to the "raw probability" formula, the uneven distribution of the training corpus may lead to inaccurate estimates. When the n-gram distribution is sparse, n-gram clusters do not appear or only appear a small number of times, the estimation of sentences containing n-gram clusters will have lower results. Assuming S is the size of the vocabulary, we will have Sn n-gram clusters generated from the vocabulary. However, in practice, the number of meaningful and frequently occurring n-gram clusters is small. When calculating the probability of a sentence, in many cases we will have n-gram clusters that have never appeared in the training data. This makes the probability of the sentence 0, while the sentence may be completely correct in terms of grammar and semantics. To overcome this situation, we use some smoothing methods to process and will be presented in the next section.

2. Smoothing Models

Linear interpolation models involve an Expectation Maximum procedure to optimize the weight for each component. Meanwhile, the back-off model is relatively simpler and quite suitable with models combined with naive Bayes model.

$$P(w_i \mid w_{i-n+1} \ldots w_{i-1}) = \begin{cases} \hat{P}(w_i \mid w_{i-n+1} \ldots w_{i-1}) & if \ \#(w_{i-n+1} \ldots w_i) > 0 \\ \beta(w_{i-n+1} \ldots w_{i-1}) * P(w_i \mid w_{i-n+2} \ldots w_{i-1}) & otherwise \end{cases} \quad (2.5)$$

Where $\hat{P}(wi \mid wi-n+1 \ldots wi-1) = \frac{discount\ \#(w_{i-n+1} \ldots w_i)}{\#(w_{i-n+1} \ldots w_{i-1})}$

is the discounted probability, and $\beta(w_{i-n+1} \ldots w_{i-1})$ is a normalization constant calculated to be:

$$\beta(w_{i-n+1} \ldots w_{i-1}) = \frac{1 - \sum_{x:\#(w_{i-n+1} \ldots w_{i-1} x)>0} \bar{P}(x \mid w_{i-n+1} \ldots w_{i-1})}{1 - \sum_{x:\#(w_{i-n+1} \ldots w_{i-1} x)>0} \bar{P}(x \mid w_{i-n+2} \ldots w_{i-1})} \quad (2.6)$$

3. Discounting Methods [2, 5]

The principle of the discounting algorithm is to reduce the probability of the n-gram clusters with probability greater than 0 to compensate for the n-gram clusters which have never appeared in the training set. These algorithms will directly alter the occurrence frequency of all n-gram clusters.

An n-gram is first matched against the language model to see if it has been observed in the training corpus. If that fails, then-gram is then reduced to an (n-1)-gram by shortening the context by one word. The discounted probability can be calculated by using different smoothing approaches including absolute smoothing, linear smoothing, Witten-Bell smoothing and Good-Turing smoothing. We used smoothing techniques to perform our calculations.

To describe the smoothing techniques, let ni denote the number of events which occur exactly i times in training data.

+ Absolute Smoothing

Frequency of a word is deducted by a constant b so that the discounted probability is calculated as

$$\hat{P}(w_i \mid w_{i-n+1}\ldots w_{i-1}) = \frac{\#(w_{i-n+1}\ldots w_i) - b}{\#(w_{i-n+1}\ldots w_{i-1})} \quad (2.7)$$

Where b is often defined as the upper bound $b = \frac{n_1}{n_1 + 2n_2}$

+ Linear Smoothing

The discounted probability is calculated as:

$$\hat{P}(w_i \mid w_{i-n+1}\ldots w_{i-1}) = (1 - \frac{n_1}{T}) * \frac{\#(w_{i-n+1}\ldots w_i)}{\#(w_{i-n+1}\ldots w_{i-1})} \quad (2.8)$$

Where T is the number of uni-grams, which corresponds to the number of words in the training data.

+ Good-Turing smoothing

The discounted probability is calculated as:

$$\hat{P}(w_i \mid w_{i-n+1}\ldots w_{i-1}) = \frac{GT_{\#(w_{i-n+1}\ldots w_i)}}{\#(w_{i-n+1}\ldots w_{i-1})} \quad (2.9)$$

Where $GT_r = (r+1)\frac{n_r + 1}{n_r}$

+ Witten-Bell smoothing

The discounted probability is calculated as:

$$\hat{P}(w_i \mid w_{i-n+1}\ldots w_{i-1}) = \frac{\#(w_{i-n+1}\ldots w_i)}{\#(w_{i-n+1}\ldots w_{i-1}) + D(w_{i-n+1}\ldots w_{i-1})} \quad (2.10)$$

Where $D(w_{i-n+1}\ldots w_{i-1})$ is the number of distinct words that can follow $w_{i-n+1}\ldots w_{i-1}$ in the training data.

4. Using n-Gram Language Models as Text Classifiers

Text classifiers attempt to determine attributes that distinguish documents in different categories. These attributes can include word average length, vocabulary terms, global syntactic, local n-grams and semantic properties. Also, language models make an attempt to capture such regularities, and supply another natural avenue to building text classifiers. An n-gram language model can be applied to text classification in a similar manner to a naive Bayes model [5].

Assume that we want to categorize a document d = $w_1w_2....w_n$ into a category c ∈ C= {c_1,c_2, ...,$c_{|c|}$}. Picking the category c with the largest posterior probability given the document is a natural choice. That is,

$$c^* = \arg\max\{p(c)P(d \mid c)\}$$
$$= \arg\max_{c \in C}\left\{P(c)\prod_{i=1}^{T} P(w_i \mid w_{i-n+1}\ldots w_{i-1}, c\right\} \quad (2.11)$$
$$= \arg\max_{c \in C}\left\{P(c)\prod_{i=1}^{T} P_c(w_i \mid w_{i-n+1}\ldots w_{i-1}\right\}$$

Here, c* is the best class for new document d. P(c) can be calculated from the training set.

$P_c(w_i \mid w_{i-n+1}\ldots w_{i-1})$ is computed using a back-off model introduced in previous section.

Thus, to know the topic of a new document d we work out the perplexity of this document as compared to the language model (trained from the training set)

Based on the research of [2], the authors pointed out that in the naive Bayes text classifier, the attributes (words) are considered to be independent of each other in the category. whereas in the language model-based approach, this is enhanced by considering the Markov dependence between adjacent words. Due to that similarity, the research team called the n-gram-augmented naive Bayes classifier a chain-augmented naive Bayes classifier (CAN). The characteristic is that the news classification problem is a special case of the text classification problem, with the task of classifying text content belonging to news groups according to pre-determined topics. In this study, we conduct experiments on such a basis.

III. VERIFICATION EXPERIMENT

To conduct the experiment, we used a dataset collected from popular online newspapers in Vietnam such as VNExpress [6], Vietnamnet [7], Dan Tri [8], Bao Moi [9], Nhan dan [10] and many other online newspapers [11]. The dataset was processed and manually labeled with a total of more than 5,000 articles in the fields of Economic News, Social News, Education News, Medical News, Science News, World News.

First, the collected data is pre-processed text, because our data is collected from websites, the data needs to be cleaned, the data needs to be processed such as spelling errors, spacing errors, special characters, hashtags, URLs, abbreviations, and many other issues. These factors can cause difficulties in processing and analyzing the data. We have conducted pre-processing steps to clean and standardize the data before applying the model. They are then passed through segmentation and sentence separation units to prepare them for further processing. Next step, we model and classify according to n-gram model. News classification results obtained with the n-gram model by 4 discounting smoothing methods: Good Turing, Absolute, Witten Bell and Linear, we choose n=1, 2, 3, 4 and other default parameters.

1. With n = 1

Table 1: Results with n = 1

Discounting smoothing methods	Recall	Precision	F1
Good Turing	81.97%	77.24%	79.55%
Absolute	81.83%	77.10%	79.41%
Witten Bell	81.81%	77.08%	79.39%
Linear	81.71%	76.91%	79.25%

2. With n = 2

Table 2: Results with n = 2

Discounting smoothing methods	Recall	Precision	F1
Good Turing	94.97%	90.24%	92.55%
Absolute	94.83%	90.10%	92.41%
Witten Bell	94.81%	90.08%	92.39%
Linear	94.71%	89.91%	92.25%

3. With n = 3.

Table 3: Results with n = 3

Discounting smoothing methods	Recall	Precision	F1

Good Turing	94.00%	89.20%	91.54%
Absolute	93.85%	89.19%	91.46%
Witten Bell	93.43%	89.06%	91.19%
Linear	93.87%	89.14%	91.45%

4. With n = 4

Table 4: Results with n = 4

Discounting smoothing methods	Recall	Precision	F1
Good Turing	88.2300%	82.9300%	85.5100%
Absolute	88.1800%	82.7700%	85.4000%
Witten Bell	88.1700%	82.3700%	85.1800%
Linear	87.3700%	82.3500%	84.7900%

IV. CONCLUSION

The classification results obtained on the 4 smoothing methods of discounting Good Turing, Absolute, Witten Bell and Linear with n=1, 2, 3, 4 show that the approach with n=2 gives the best results compared to the remaining values. This is completely consistent with the characteristics of the Vietnamese language, which is that the number of words with 2 syllables is larger than the number of words with 1, 3, 4 syllables and is suitable for practice. In the case of n=1, the value is the lowest, because the number of words with 1 syllable is very limited. For n=3, the result is the second highest after n=2 and higher than n=4, which is also consistent with the fact that in Vietnamese, the number of words with 3 syllables is more than 4 syllables, because most words in Vietnamese are compound words and in practice, the number of compound words is more than single words. For the results according to the smoothing method, we found that the Good Turing technique gave the best results, followed by the Absolute technique, the technique giving the lowest results was Linear for all 4 cases n = 1, 2, 3, 4. This result is completely consistent with the studies that have been conducted on English and other languages. This once again confirms that our experimental process is suitable for Vietnamese.

Text classification is still a problem with many problems to handle and has the characteristics of a natural language processing problem. N-gram is a statistical approach and thus n-gram will not directly use the results of the word segmentation stage like other methods, and also limits the difficulties in selecting text features. However, the results of the n-gram approach bring equivalent value to the results of traditional classification. This result has proven the superiority of the statistical model and it is also a way to solve the problem of text classification which has many challenges.

REFERENCES

Journal Papers:

[1] Wisam Abdulazeez Qader, Musa M.Ameen, Bilal I. Ahmed. An Overview of Bag of Words;Importance, Implementation, Applications, and Challenges. Conference: 2019 International Engineering Conference (IEC). June 2019. DOI:10.1109/IEC47844.2019.8950616.

[2] Fuchun Peng and Dale Schuurmans, Combining Naive Bayes and n-Gram Language Models for Text Classification, ECIR 2003, LNCS 2633, pp. 335–350, 2003..

[3] Maria Fernanda Caropreso, Stan Matwin, Fabrizio Sebastiani (2001). A Learner-Independent Evaluation of the Usefulness of Statistical Phrases for Automated Text Categorization, Text Databases and Document Management: Theory and Practice, Idea Group Publishing, Hershey, US, pp. 78--102

[4] Fuchen Peng, Dale Schuurmans, Shaojun Wang. (2004). Augmenting Naïve Bayes Classifiers with Statistical Language Models, Information Retrieval, 7, 317-345.

[5] 14.Fuchun Peng, Xiangji Huang. Machine learning for Asian language text classification. May 2007 Journal of Documentation 63(3). DOI:10.1108/00220410710743306.

[6] Https://vnexpress.net

[7] Https://vietnamnet.vn

[8] Https://dantri.com.vn

[9] Https://baomoi.com

[10] Https://nhandan.vn

[11] Https://tuoitre.vn