information

MDPI

Article

Num-Symbolic Homophonic Social Net-Words

Yi-Liang Chung [1,*], Ping-Yu Hsu [1] and Shih-Hsiang Huang [2]

[1] Department of Business Administration, National Central University, No. 300, Zhongda Rd., Zhongli Dist., Taoyuan City 320317, Taiwan; pyhsu@mgt.ncu.edu.tw
[2] PwC Business Consulting Services Taiwan Ltd., 27F., No. 333, Sec. 1, Keelung Rd., Xinyi Dist., Taipei City 110208, Taiwan; steven.sh.huang@pwc.com
* Correspondence: richard@g.ncu.edu.tw

Abstract: Many excellent studies about social networks and text analyses can be found in the literature, facilitating the rapid development of automated text analysis technology. Due to the lack of natural separators in Chinese, the text numbers and symbols also have their original literal meaning. Thus, combining Chinese characters with numbers and symbols in user-generated content is a challenge for the current analytic approaches and procedures. Therefore, we propose a new hybrid method for detecting blended numeric and symbolic homophony Chinese neologisms (BNShCNs). Interpretation of the words' actual semantics was performed according to their independence and relative position in context. This study obtained a shortlist using a probability approach from internet-collected user-generated content; subsequently, we evaluated the shortlist by contextualizing word-embedded vectors for BNShCN detection. The experiments show that the proposed method efficiently extracted BNShCNs from user-generated content.

Keywords: text analysis; homophonic; user-generated content; net-word

check for updates

Citation: Chung, Y.-L.; Hsu, P.-Y.; Huang, S.-H. Num-Symbolic Homophonic Social Net-Words. *Information* **2022**, *13*, 174. https://doi.org/10.3390/info13040174

Academic Editor: Ralf Krestel

Received: 18 December 2021
Accepted: 16 March 2022
Published: 29 March 2022

1. Introduction

In the era of mobile access to the internet, the number of instant messages exchanged and the amount of user-generated content (UGC) are increasing rapidly, becoming significant in communities, such as PTT (批踢踢) [1,2], which includes many existing and new neologisms. Social media has become the principal means of adopting and using new words [3–5]. In lexicology, neologism refers to new vocabulary that accompanies changes in the social environment [6,7]. Unlike other writing, such as journalism, user-generated content is informal and colloquial [8–10]. The real corpus must contain neologisms that the word segmentation program cannot identify [11].

Neologisms are significant in Chinese, as they can directly affect the results of word segmentation, sentiment analyses and word semantics in the processing and analysis of the text. A simple fact currently barring these analyses is that Chinese has no spaces, obvious separators, or explicitly marked boundaries between words [12,13]; thus, words semantics depends on the locations in consecutive context.

Several excellent reviews for identifying neologisms are divided into supervised and unsupervised approaches based on their contents, depending on whether this is from a specific subject domain or not. The supervised approach depends on prior labeling of the different positions of a character within a word (for example, label "B" means the beginning of a word, and label "E" means the end of the word) [14].

However, using the same approach will significantly drop the segmentation accuracy to user-generated content with blended numeric and symbolic homophony Chinese neologisms from the internet. The method commonly used to remove invalid words, numbers and symbols is to preprocess the language library or text [15]. Consequently, colloquially as Chinese homonyms, using numbers or symbols with literal numerical meanings causes inaccurate labeling.

Take some examples: “是不是” (yes or no) is often replaced with “484” (a homonym of “yes or no”), “生氣/憤怒” (angry) is often replaced with “377” (a homonym of “angry/angrily”), and “好兄弟/八家將” (brothers) is often replaced with “8+9” (a homonym of “brothers”) in user-generated contexts.

Although “484”, “ 377” and “8+9” meanings are in numerical initially, they can also be homophones to have Chinese semantics. In Chinese, “484” (four-eight-four) is pronounced as sì bā sì, which is similar to the pronunciation of “是不是” (shì bù shì); similarly, the “377” (three–seven–seven) pronounced as sān qī qī, which is similar to the pronunciation of “生氣” (sēn qì qì). These new words have real semantic meaning and are different from stop words.

However, these words have a relatively high chance of being deleted in the preprocessing stage, which changes the overall semantic meaning of the text. For example, “你484還沒吃飯？(Have you eaten yet?)” becomes “你還沒吃飯 (You have not eaten yet)” after data preprocessing. Furthermore, “他在家377 (He at home being angry)” becomes “他在家 (He is at home)”. We call this type of vocabulary blended numeric and symbolic homophones in Chinese neologisms (BNShCNs). See the Sample sentence in Table 1.

Table 1. Sample sentence after pre-processing by default regulation.

Example	BNShCN	Original Sentence	Sentence Removed BNShCN
Sentence 1	484 *Yes or No*	你484還沒用餐？ *Have you eaten yet?*	你還沒用餐 *You have not eaten yet.*
Sentence 2	377 *Angry*	他在家裡377 *he is at home being angry.*	他在家裡 *he is at home.*

Due to the rising costs of labeling, effort-wise, we hope to propose a near-hands-off approach to net-word mining that requires less manual intervention. This study used the PTT Gossiping board of Traditional Chinese’s most extensive online forum to source its experimental data. We focused on blended numeric and symbolic homophony Chinese neologism (BNShCN) detection. First, we applied various statistical approaches with a predefined threshold for entropy and mutual information. Then, we compared initial dictionaries to find the shortlist of out-of-vocabulary (OOV) BNShCN candidates. Furthermore, using the position of the word vector in the semantic space can best reflect its relationship with other words in the context [6].

2. Literature Review

Most of the identifying neologism reviews in Chinese are also divided into supervised and unsupervised approaches. The character distribution and the probability of a character being present within the context are features of unsupervised approaches. In contrast, the supervised approach has higher requirements for tuning and optimizing texts and models, and the versatility between different texts and models is also limited.

2.1. Supervised Chinese Neologism Discovery

The supervised approach uses machine learning, with a function that maps an input to an output based on example input–output pairs and is often applied in the pre-detection of neologisms and for solving the central problem of Chinese word segmentation by labeling sequences beforehand. Meanings are inference by a function learning from labeled training data consisting of examples [11].

The condition-random-field (CRF) with long short-term memory or bidirectional long short-term memory (LSTM/Bi-LSTM) for recurrent neural networks (RNNs) has been the most popular methodology used in supervised Chinese word segmentation in recent years. However, some issues with this method make a neural-network-based method detect out-of-domain neologisms imprecisely.

The first issue is that the performances of neural-network-based methods rely on the training set being of high quality. The second issue is that high-quality domain-specific training sets are challenging to obtain continuously [14]. Although the supervised (machine) learning approach has excellent performance, it has many potential problems in terms of versatility. These problems include good-quality training data sets, the versatility of trained inference models across domains and the preliminary work and subsequent maintenance operations required. We have doubts about the ability of past data sets and trained machine-learning models to recognize new internet words in the future.

In the literature, Reference [16] presented a method for identifying adjective–noun pairs as neologisms and demonstrated that the use of pre-trained language models improved significantly over other baselines. Reference [17] proposed a variety of neural network architectures by combining long short-term memory (LSTM) networks with a conditional random field (CRF) layer. Reference [18] produced candidate sequences by improving a priori knowledge of and identifying the boundary of words using LSTM.

Reference [19] compared the reasonable of CRF and BiLSTM-CRF (bidirectional LSTM-CRF) as validating fine-grained annotation. Reference [20] used BiLSTM and convolutional neural network (CNN) to extract document and boundary features to construct a CRF to train an end-to-end Chinese word segmentation (CWS) model. Reference [21] used LSTM and the word-to-vector (Word2Vec) model to achieve Chinese word segmentation.

Reference [22] proposed a weakly supervised training framework for domain-specific CWS with only dictionary-based deep learning. Reference [12] proposed a probabilistic topic modeling method based on the latent Dirichlet allocation (LDA) of user-generated content. Reference [23] proposed a unified model for multi-criteria Chinese word segmentation by leveraging the transformer encoder and by utilizing a self-attention mechanism to model the criterion-aware context for each character neatly.

Reference [24] proposed a framework with BiLSTM, semi-CRF and a fusion layer for Chinese word segmentation. Reference [25] proposed four group features to build an identification model for cybersecurity words and four sets of features to identify cybersecurity neologisms.

2.2. Unsupervised Chinese Neologism Discovery

An unsupervised approach is a data-driven approach based on the original attribute of the content. The probability of a character sequence is a valid word is evaluated based on the frequency distribution of the character sequence, thus requiring less human involvement [14]. Analysis with an unsupervised method makes the following assumptions.

(1) The characters in the sequence are interdependent. Suppose the likelihood that the characters in the string will occur together is high. In other words, the higher the interdependence of characters in the string, the higher the chance of forming a word. For example, in “蝙蝠 (bats)”, “蜘蛛 (spiders)”, “徬徨 (hesitation)” and “忐忑 (anxiety)”, the characters are interdependent (internally solidified) because they always appear together.

(2) Entropy of a sequence and its left and right neighbors: The degree of free use of a character sequence is also an essential criterion for judging whether a string of characters is a word. The more independently formed the word is, the better matched and the higher entropy with more external characters. Information entropy can reflect how much information a specific event brings, on average, after obtaining the result. Assuming that the probability of an event occurring is p when an event occurs, we get $log(p)$ of information. In other words, the smaller the probability of an event occurring, the greater the amount of information obtained.

(3) The frequency of the character sequence in the text: Calculate the frequency of possible character sequences in the text.

The unsupervised approach uses the probabilities of characters in the text as the basis for computation. The detection efficiency of new words will be affected by the following. (1) Under different threshold settings, strings of different numbers and qualities are screened out. (2) The parameters and thresholds need to be adjusted and optimized in advance. (3) New neologism in the early stages of being adopted tend to fail to exceed the threshold due to a low frequency of appearance in the text.

Reference [26] found new words based on existing dictionaries and probabilities. In contrast, Reference [27] found new word candidates using their frequency and determined new words using their real-world frequency. Reference [28] extracted candidate terms using a frequency and n-gram analysis. Reference [29] analyzed the relationship between new internet words and public internet opinion. Reference [30] used an n-gram and constructed an objective function to identify new words.

Reference [6] proposed a novel method that combines word embedding and frequent n-gram string mining to discover new words from a domain corpora. Reference [7] presented an unsupervised approach to detecting neologisms and then normalized them to canonical words without relying on parallel training data. Reference [31] proposed an in-depth investigation based on the n-gram approach with finite-context models of characters.

Reference [32] proposed a new-word-discovery algorithm based on the internal solidification and frequency of multiple characters found using an n-gram. Reference [33] utilized recent advancements in unsupervised machine learning methods, word embedding, and latent Dirichlet allocation to research testbeds encompassing 29 exclusive, underground market QQ groups with 23,000 members.

In addition, Reference [34] proposed a domain-specific unsupervised approach based on the lexical features and statistic features and methodology used in mechanical design and manufacturing. Reference [10] proposed a domain-specific method using user-invented new words and converted sentiment words utilizing the assembled mutual information. References [14,35] used out-of-domain unsupervised as well as domain-independent Chinese new word detection methods.

The supervised approach is subject to the limitations of the training corpus from the field of interest, and its degree of generality is limited. However, it can obtain better efficiency after conducting a complete study. In contrast, an unsupervised approach analyzes the text in a probabilistic way. This approach has a better general-purpose ability but must recalculate the probability each time will be its disadvantage.

The form of neologisms from the internet community is constantly updated, and neologism recognition should also keep pace with the times. The omission or misjudgment of neologisms with real semantic meanings may result in semantic changes. It will necessitate trade-offs between the efficiency and correctness of neologism discovery.

Chinese characters do not have natural separators, and thus require performing Chinese word segmentation or related pre-processing before analyzing the text. Traditional methods filter out numbers or symbols in a specific Unicode range, resulting in a semantic variant of the text. For example, Wikicorpus removes numbers and symbols when fetching text from Wikipedia.

However, with increasing community users using blended numeric and symbolic characters instead of formal language to send texts, neologisms with real semantic meanings are being created and used in user-generated content from the community. We can reduce the amount of manual intervention, prior work and domain knowledge required in automated text analyses.

This paper proposes a near-hands-off hybrid approach. The method first obtains candidate words based on inter-character probabilities and dictionary comparisons for user-generated content from internet communities. Then, it calculates the semantic similarity with a pre-trained language model for determining blended numeric and symbolic Chinese neologisms.

3. Materials and Methods

This study uses the probability between characters as the basis for word segmentation to ensure the text's integrity regarding blended numeric and symbolic Chinese neologisms. The characters are used to build a retrieval tree based on different n-grams, frequencies and parent–child associations. Moreover, it determines the character boundary based on the entropy and mutual information between characters. Finally, it calculates the similarity between neologism via a fine-tuned pre-trained model.

3.1. N-Gram

An n-gram is a sequence of n consecutive items from a given text or speech. An n-gram calculates the highest probability that the consecutive string sequence will appear adjacent to other words in the collection within a specific context.

P(s) is an unknown probability for each given string and is represented by a mathematical model. The probability for a string words or n-grams can be represented as *w* = *w1w2...wn* or as the product below [36].

$$p(w) = \prod_{i=1}^{n} p(w_i \mid w_1 w_2 ... w_{i-1})$$

To simplify this mathematical model, we only consider the first *n-1* characters each time, and the probability relates to *n-1* characters before it. We call this the n-gram language model and the first-order Markov chain. If we use the maximum likelihood estimation (MLE) methodology, the probability of a character appearing in the string can also be represented as a product [37]. The threshold filters which strings are neologisms after removing information noise or interfering strings.

3.2. Trie (Tree)

Trie comes from retrieval, also known as a prefix tree or a dictionary tree and is an ordered tree used to store associative arrays, the keys of which are usually strings. In [38], Trie was used in a string-structure review using word frequency statistics in natural language processing. A single node of an English Trie-tree has 26 child nodes at most, while one for Chinese can generate 5000 child nodes based only on commonly used Chinese characters, thus, affecting the retrieval efficiency.

Take an example from user-generated content: We first use the natural statistical approach of dividing the phrase "你484在377 (Are you angry)" into four words: “你” (you), “484” (yes or no), “在” (in the state of...) and “377” (*angry*). Then, we check for the character node named “你” in existing root nodes and count its occurrences or create a new one under the root node. We continue to check the next character or create a new node for a child root until the end of the phrase. See the Chinese Trie-tree structure in Figure 1.

3.3. Information Entropy (IE)

The information Entropy of *X* and *Y* is defined as follows:

$$H(X) = \sum_{i} P(x_i) I(x_i) = -\sum_{i} P(x_i) log P(x_i)$$

The greater instability of the relationship between the segment and the left and right neighboring characters, the greater entropy and the more likely the string can be used as an independent segment. *I(x)* after the first equal sign in the formula represents the self-information of *x* [32].

3.4. Point(Wise) Mutual Information (PMI)

The point(wise) mutual information of *X* and *Y* is defined by

$$PMI(X;Y) = log(\frac{p(x,y)}{p(x)p(y)}) = log\frac{p(x \mid y)}{p(x)} = log\frac{p(y \mid x)}{p(y)}$$

In probability theory and information theory, the point(wise) mutual information (PMI) of two random variables measures the mutual dependence between the two variables [39]. Mutual information is intimately linked to the entropy of a random variable and quantifies the expected “amount of information” held in a random variable [11]. Point mutual information is mainly applied to measure the probability of whether a combination of two or more words appears together is a good neologism.

A higher mutual information value means a higher probability that x and y form a word, usually applied in the discovery and disambiguation of new words in natural language processing. The research results of [40] suggested that the processing of traditional Chinese and the filtering of special symbols are important reasons for reducing the accuracy of the experiment. They also indicated that word strings are partially composed of stop words should not be classified as illegal strings.

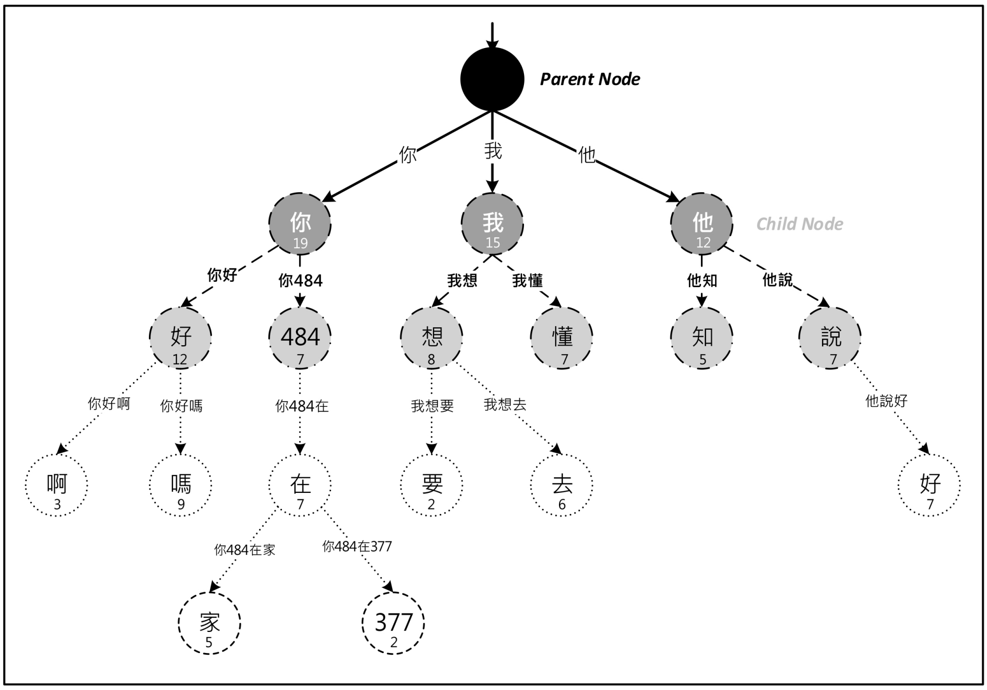

Figure 1. A sample Chinese retrieval (Trie)-tree.

3.5. Bidirectional Encoder Representations from Transformers (BERT)

Vector semantics are models that use a formal mathematical structure (i.e., vectors) to represent how lexical meanings of words are used in a vector space [7]. The early methods commonly used are "Word2vec" and "GloVe", both context-free word embeddings. However, they cannot resolve the ambiguity of words, that is, different meanings in different contexts. For example, a bank can be a financial institution or a riverbed.

BERT [41] is a transformer-based NLP pre-trained model developed by Google in 2018. The BERT-based pre-trained model has 12 layers with a hidden size of 768 and 12 self-attention heads with deep bidirectional representations (repr.) from an unlabeled text that jointly conditions each layer's left and rights contexts. Since BNShCNs consider multiple simultaneously semantics, a traditional single-semantic word embedding method faces the problem of reduced accuracy. See the study of environment and hyperparameter in Table 2.

BERT's performance in downstream tasks depends heavily on fine-tuning. This study thus uses a fine-tuned BERT model to evaluate blended numeric and symbolic homophones in Chinese neologisms. See the Overview of the input representation of BERT in Figure 2.

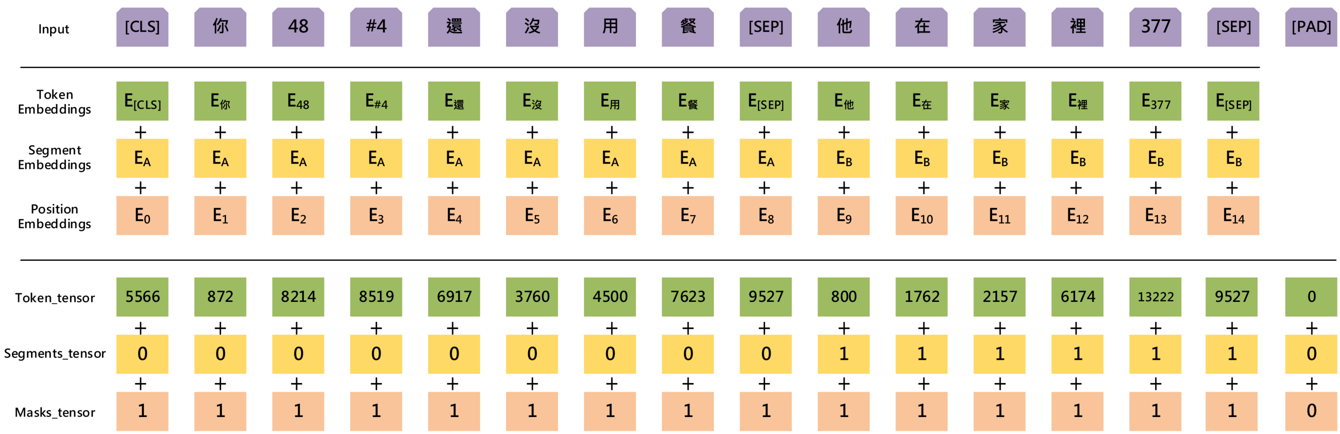

Figure 2. Overview of the input representation of BERT.

Table 2. Environment and hyperparameters.

Environment	Intel i9-11900K DDR4 3200 96G nVidia GeForce RTX3070Ti
Tokenizer Model	bert-base-chinese ckiplab/albert-tiny-chinese
Hyperparameters	Max Sequence Length = 128 Learning Rate = 5×10^{-5} Batch Size = 16 Epochs = 2

3.6. Cosine Similarity

The cosine similarity measures were similar between two non-zero vectors of an inner product space. The same vectors' inner products can both have a length of 1 by normalization.

$$S_c(A, B) = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \cdot \sum_{i=1}^{n} B_i^2}$$

In this study, the fine-tuned BERT pre-trained model is transformed in the last hidden state to regenerate dense vectors and compute the cosine similarity between sentences containing BNShCNs to validate the sentence.

4. Experiment

4.1. Algorithm Overview

This paper proposes a semi-supervised transfer learning-based approach. First, our approach ensures completeness and authenticity with minimal distortion in the data thread by removing unnecessary user-generated content, such as text-trimmed hyperlinks, quote replies and information sharing. Then, it implements unsupervised methods to analyze the user-generated content using the Trie-tree, point(wise) mutual information, information entropy and word frequency.

It randomly selects test data sets from these texts, tests for different thresholds and compares the differences between the thresholds. Then, it performs a text analysis according to the preset threshold to obtain the candidate data set to find blended numeric and symbolic homophones in Chinese neologisms (BNShCNs). This obtains the blended numeric and symbolic homophones in Chinese neologisms (BNShCNs) by comparing meanings from existing dictionaries and those gathered by machine learning.

The specific algorithm steps are as follows:

1. Set initial dictionary *Di*.
2. Parsing the user-generated content *UGCi* from the social web page, filter out the hyperlinks, quote replies, news sharing and announcements.
3. Calculate the statistical characteristics of the probability *Pi* of the Trie-tree, mutual information *PMIi*, point(wise) information entropy *IEi* and word frequency *WFi* as a data pipeline for the user-generated content with an unsupervised approach and retain and build up the set of candidate words found to be above the threshold as *SETi*.
4. Screen out out-of-vocabulary *OOVi* words by comparing the candidate word set *SETi* and the initial dictionary *Di*.
   Above steps please refer to the Algorithm 1, and following steps please refer to the Algorithm 2.
5. Comparing the sentences include Ni and another sentences for comparison by mean pooling.
6. Calculate the cosine similarity for the sentences via teh included angle.
7. Update the vocabulary dictionary *Di+1*.

```
Algorithm 1 Neologisms.
Input:
  Di—Initial Dictionary.
  RAW—PTT Post Set trims out hyperlinks, quote replies, sharing and announcements.
  SETi—Character portfolio candidate set.
  OOVi—SETi does not exist in the part of Di.
Output:
  Ni—Neologisms.
   1. Set initial dictionary Di.
   2. Maintain pure UGCi by trimming out the noise from RAW.
   3. SETi by calculating the character portfolio of probability Pi by Trie Tee.
   4. For SETi in UGCi
   5. –Point(wise) Mutual Information PMIi >= Predefined threshold.
   6. –Information Entropy IEi >= Predefined threshold.
   7. –Term Frequency TFi >= Predefined threshold.
   8. Return UGCi
   9. Obtain Ni by mapping SETi and Di
```

```
Algorithm 2 Similarity verification.
Input:
  Ni—Neologisms.
  SentNi—Sentences with Ni.
  RSentNi—Randomly Extract Sentence with Similarity SentNi.
Output:
  Similarity.
   1. Padding the sentences to max length.
   2. Comparing the sentences include Ni and another sentences for comparison by mean pooling.
   3. Calculate the cosine similarity for the sentences via included angle.
```

4.2. Algorithm Flowchart

This research proposes a semi-supervised approach for text analyses for a near-automated process. First, we use a web parser to collect user-generated text from specific Chinese social networking sites and filter out only hyperlinks and replies to previous quotes to avoid excessive text trimming. Next, through an unsupervised method, the candidate new words in the text that meet the predefined threshold are analyzed, which is supplemented by manual inspection, to determine if they have actual semantics. Then, the output layer is trained through the pre-trained model, and semantic similarities are used to verify the blended numeric and symbolic homophones in Chinese neologisms (BNShCNs). See the overview flowchart in Figure 3.

4.3. Dataset

We obtained the dataset as user-generated content from the most extensive traditional Chinese web community via PTT-Gossiping. PTT-Gossiping is an online web community with no restrictions on the specific domains and topics discussed. We conducted three large-scale user-generated content collections in 2019, 2020 and March 2021, collecting around 1,487,980 posts, with the raw data being over 5 GB, all in UTF-8 plain text format. An average of around 60,000 posts were made a month last year, with replies being around 30 to 300 words in each post, which is equivalent to generating around 1,800,000 to 18,000,000 words a month.

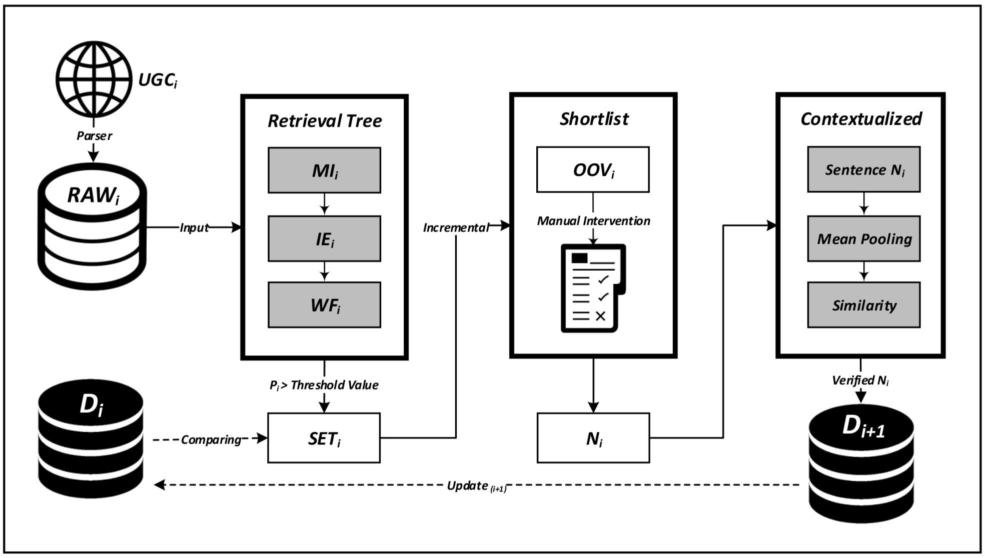

Figure 3. Semi-supervised algorithm flowchart.

4.4. Evaluation Index

Precision refers to the proportion of *TP* (true positives) in all recognized BNShCNs. The formula is shown below. TP represents the number of positive samples that were correctly identified as positive samples, while *FP* (false positives) represents the number of negative samples incorrectly identified as positive samples.

$$Precison = \frac{TP(TruePositives)}{TP(TruePositives) + FP(FalesPositives)}$$

Recall is expressed as the proportion of all positive samples that are identified correctly as positive samples.

$$Recall = \frac{TP(TruePositives)}{FN(FalesNegatives)}$$

FN (false negatives) represents the number of positive samples that were mistakenly identified as negative samples. *F-score* is the weighted average of precision and recall.

$$F\text{-}score = \frac{2PR}{P+R}$$

4.5. BNShCN Detection

We applied the probabilistic detection method described above to the user-generated texts collected in this study and performed different threshold detection methods by limiting the total number of posts to determine the final threshold. Table 3 shows the collected number of user-generated content.

Table 4 shows the number of formed words using a probabilistic method via threshold detection. Based on our observations, the probabilistic segmentation of text with IE > 2, frequency > 3, PMI > 4 works best without pre-pruning. Then, we compare the initial dictionary, gain unknown words and manually check their semantics.

Table 3. User-generated content list.

Years	**2015**	**2016**	**2017**	**2018**	**2019**	**2020**	**2021**
Posts	2640	2380	1240	1620	448,600	314,280	189,340

Table 4. The BNShCN probability detection results.

Frequency = 5	**IE = 0.001**	**IE = 0.01**	**IE = 0.1**	**IE = 1**	**IE = 2**	**IE = 3**
PMI = 6	7889	7888	7869	4316	900	179
PMI = 5	7969	7969	7936	5066	1611	327
PMI = 4	12,809	12,809	12,757	8097	2551	478
PMI = 3	19,275	19,351	19,284	12,090	3602	620
PMI = 2	28,436	28,436	28,356	17,636	4930	818
PMI = 1	41,527	41,527	41,431	25,868	6701	1038
Frequency = 4	**IE = 0.001**	**IE = 0.01**	**IE = 0.1**	**IE = 1**	**IE = 2**	**IE = 3**
PMI = 6	5812	5812	5796	3588	900	179
PMI = 5	10,018	10,018	9986	6141	1611	327
PMI = 4	16,074	16,074	16,022	9800	2551	478
PMI = 3	24,690	24,690	24,623	14,890	3602	620
PMI = 2	36,657	36,657	36,577	22,144	4930	818
PMI = 1	53,769	53,769	53,673	32,870	6701	1038
Frequency = 3	**IE = 0.001**	**IE = 0.01**	**IE = 0.1**	**IE = 1**	**IE = 2**	**IE = 3**
PMI = 6	4650	4650	4631	2953	901	179
PMI = 5	13,808	13,808	13,775	7438	1611	327
PMI = 4	22,353	22,353	22,300	12,002	2551	478
PMI = 3	34,732	34,732	34,665	18,428	3602	620
PMI = 2	52,061	52,061	51,982	27,789	4930	819
PMI = 1	76,230	76,230	76,134	41,446	6701	1038

4.6. Contextualized Evaluation

For the text sources used in this study, the length of each line of text is under 40 Chinese characters, and thus the length of each line does not exceed the length limit of the BERT token size. We embed sentences containing BNShCNs using the pre-trained model and manually select random sentences to calculate the cosine similarity formed between the two sentences.

4.7. Comparative Experiment

We referred to the methodology in related studies as a baseline by the collected corpus for evaluation and obtained significant differences in the results, as shown in Table 5. Among them, the bold words represent the best experimental results. Tables 6 and 7 show the cos θ (cosine similarity) values by comparing the semantics in positive and negative correlation.

Table 5. Comparison of different evaluation methodologies in BNShCNs.

Methodology	**Precision**	**Recall**	**F-Score**
Transformer Encoder (BERT)	**87.3**	**89.2**	**88.2**
ELMO [42]	81.6	83.8	82.7
Word2Vec [6]	66.2	58.2	61.9

Table 6. BNShCNs and positive sentences.

BNShCN	Sentence with BNShCN	Positive Sentence	Similarity
484 *Yes or No*	你484在生氣 *Are you angry*	你是在生氣嗎 *Are you angry*	0.85592383
377 *Angry*	你是不是在377 *Are you angry*	你是在生氣嗎 *Are you angry*	0.8032131
8+9 *good friend*	我跟他是8+9 *I am good friends with him.* 我們是8+9 *We are good friends.*	我跟他是好兄弟 *I am good friends with him.*	0.84317076 0.7407087

Table 7. BNShCNs and negative sentences.

BNShCN	Sentence with BNShCN	Positive Sentence	Similarity
484 *Yes or No*	你484在生氣 *Are you angry?*	今天天氣真好 *The weather is nice today.*	0.5758477
377 *Angry*	你是不是在377 *Are you angry?*	今天天氣真好 *The weather is nice today.*	0.56190294
8+9 *good friend*	我跟他是8+9 I am good friends with him. 我們是8+9 *We are good friends.*	我不認識他 *I do not know him.*	0.699402 0.45311478

4.8. Analysis of Results

In the experiments, we compared Word2Vec with static single vector and Elmo with contextualized word embedding and the contextualized similarities in positive and negative correlations. Our experiments aimed to analyze the importance of discriminating ambiguity in BNShCNs within context.

The experimental results show that user-generated content frequently uses mixed texts with semantic numbers and symbols. Therefore, since numbers or symbols have both literal numerical and Chinese neologistic meanings, the possibility of erroneous semantic changes caused by previous automatic text processing methods was high.

5. Discussion

Neologism constantly introduces and renews new words with innovative meanings; Reference [33] used word embedding to analyze the similarity of such static semantics. However, evaluating a transformer/a contextualized BERT is better suited for dynamic semantics considering entropy and context. Reference [8] suggested that Chinese has a particular use in society: expressing criticism through the use of non-offensive characters with the same or similar pronunciation and politically sensitive wording, which is different from the entry point for word searches without a predetermined position in this research.

Reference [43] used a subject-based search method for related information, while we used user-generated text collection in a subject-free manner. Reference [44] collected a large number of English online texts routinely and collected candidate words via dictionary matching. Their method was similar to the concept in this research; however, the Chinese language has no natural separators, which makes text analysis more challenging.

Reference [10] combined mutual-information new-word detection and word dissemination in a specific field, which is different from the text collection method used in this research, which did not limit the field of interest. Reference [6] used the method of mixing word embeddings and frequent n-gram string mining to discover new words from the domain corpus, which is also a different method than that in this research, which mixed different methods and did not limit the domain of interest.

In [32,45], the unsupervised method of calculating the probability of occurrence for unmarked text was similar to the concept in the first half of our hybrid method presented in this research. The method used in the second half of this research is comparable with their use of a preset dictionary and supplementation with one-time manual assistance in semantic judgment.

Reference [25] used unsupervised learning on underground market jargon and proposed four sets of features to construct a cybersecurity word-identification model and four sets of features to identify new network-security words. The first half of this research used an unsupervised calculation of word group probability. The second half used the pre-training model as the basis for similarity judgment. Reference [42] used contextual word embedding and spherical K-means clustering to detect homophones among neighboring vectors; their method was similar to our method in that they calculated the similarity of the vector angle of a specific sentence; however, their entry point was different.

The method in this study pursued a near-hands-off and low-manual-intervention approach but had a high dependence on the performance of the computation. A high memory usage rate was required when performing probabilistic calculations and fine-tuning our pre-trained model. The weights during model fine-tuning are much lower than those during initialization, making fine-tuning pre-trained models more time. Our approach also required an enormous amount of data for the calculations. It will be a conflicting issue in pursuing both efficiency and quality performance.

6. Conclusions

In this paper, we proposed a near-hands-off approach for text analyses for blended numeric and symbolic homophony Chinese neologism findings. First, widely identified neologism candidates were determined from the user-generated context from an unsupervised approach based on probability. We subsequently screened with a near-hands-off approach with manual assistance. We verified the similar semantics of BNShCNs by the pre-trained language model. The experimental results proved that our approach can improve the detection of new Chinese words with mixed numbers and symbols used as homophones.

Author Contributions: Conceptualization, Y.-L.C. and P.-Y.H.; methodology, P.-Y.H.; software, Y.-L.C.; validation, Y.-L.C., P.-Y.H. and S.-H.H.; formal analysis, Y.-L.C.; investigation, Y.-L.C.; resources, Y.-L.C.; data curation, Y.-L.C.; writing—original draft preparation, Y.-L.C.; writing—review and editing, P.-Y.H.; visualization, Y.-L.C.; supervision, P.-Y.H.; project administration, Y.-L.C.; funding acquisition, Y.-L.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: A secure and private dataset was analyzed in this study.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. PTT Web Forum. Available online: https://www.ptt.cc/bbs/index.html (accessed on 20 December 2021).
2. Liu, T.-J.; Hsieh, S.-K.; Prévot, L. Observing features of PTT neologisms: A corpus-driven study with N-gram model. In Proceedings of the 25th Conference on Computational Linguistics and Speech Processing (ROCLING 2013), Kaohsiung, Taiwan, 4–5 October 2013; pp. 250–259.
3. Huang, L.-F.; Liu, X.; Ng, V. Associating sentimental orientation of Chinese neologism in social media data. In Proceedings of the 2015 IEEE 19th International Conference on Computer Supported Cooperative Work in Design (CSCWD), Calabria, Italy, 6–8 May 2015; IEEE: New York, NY, USA, 2015; pp. 240–246.
4. Cole, J.R.; Ghafurian, M.; Reitter, D. Is word adoption a grassroots process? An analysis of Reddit communities. In Proceedings of the International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation, Washington, DC, USA, 5–8 July 2017; Springer: Berlin/Heidelberg, Germany, 2017; pp. 236–241.
5. Muravyev, N.; Panchenko, A.; Obiedkov, S. Neologisms on facebook. *arXiv* **2018**, arXiv:1804.05831.

6. Qian, Y.; Du, Y.; Deng, X.; Ma, B.; Ye, Q.; Yuan, H. Detecting new Chinese words from massive domain texts with word embedding. *J. Inf. Sci.* bf 2019, *45*, 196–211. [CrossRef]
7. Zalmout, N.; Thadani, K.; Pappu, A. Unsupervised neologism normalization using embedding space mapping. In Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019), Hong Kong, China, 4 November 2019; pp. 425–430.
8. Chu, Y.; Ruthrof, H. The social semiotic of homophone phrase substitution in Chinese netizen discourse. *Soc. Semiot.* **2017**, *27*, 640–655. [CrossRef]
9. Xu, J. Interpretation of Metaphorical Neologisms in Cognitive Linguistics under "Internet Plus". *Front. Soc. Sci. Technol.* **2019**, *1*, 67–74. [CrossRef]
10. Li, W.; Guo, K.; Shi, Y.; Zhu, L.; Zheng, Y. DWWP: Domain-specific new words detection and word propagation system for sentiment analysis in the tourism domain. *Knowl.-Based Syst.* **2018**, *146*, 203–214. [CrossRef]
11. Wang, K.; Wu, H. Research on neologism detection in entity attribute knowledge acquisition. In Proceedings of the 5th International Conference on Computer Science, Electronics Technology and Automation, Hangzhou, China, 15 April 2017.
12. Ma, B.; Zhang, N.; Liu, G.; Li, L.; Yuan, H. Semantic search for public opinions on urban affairs: A probabilistic topic modeling-based approach. *Inf. Process. Manag.* **2016**, *52*, 430–445. [CrossRef]
13. Liu, Y.-C.; Lin, C.-W. A new method to compose long unknown Chinese keywords. *J. Inf. Sci.* **2012**,*38*, 366–382. [CrossRef]
14. Liang, Y.; Yang, M.; Zhu, J.; Yiu, S.-M. Out-domain Chinese new word detection with statistics-based character embedding. *Nat. Lang. Eng.* **2019**, *25*, 239–255. [CrossRef]
15. Roll, Uri and Correia, Ricardo A and Berger-Tal, Oded. *Conserv. Biol.* **2018**, *32*, 716–724.
16. McCrae, J.P. Identification of adjective–noun neologisms using pretrained language models. In Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019), Florence, Italy, 2 August 2019; pp. 135–141.
17. Wang, M.; Li, X.; Wei, Z.; Zhi, S.; Wang, H. Chinese word segmentation based on deep learning. In Proceedings of the 2018 10th international Conference on Machine Learning and Computing, Macau, China, 26–28 February 2018; pp. 16–20.
18. Xie, T.; Wu, B.; Wang, B. *New Word Detection in Ancient Chinese Literature, Asia-Pacific Web (APWeb) and Web-Age Information Management (WAIM) Joint Conference on Web and Big Data*; Springer: Berlin/Heidelberg, Germany, 2017; pp. 260–275.
19. Xiong, Y.; Wang, Z.; Jiang, D.; Wang, X.; Chen, Q.; Xu, H.; Yan, J.; Tang, B. A fine-grained Chinese word segmentation and part-of-speech tagging corpus for clinical text. *BMC Med Inform. Decis. Mak.* **2019**, *19*, 66. [CrossRef]
20. Li, X.; Zhang, K.; Zhu, Q.; Wang, Y.; Ma, J. Hybrid Feature Fusion Learning Towards Chinese Chemical Literature Word Segmentation. *IEEE Access* **2021**, *9*, 7233–7242. [CrossRef]
21. Wang, X.; Wang, M.; Zhang, Q. Realization of Chinese Word Segmentation Based on Deep Learning Method. In *AIP Conference Proceedings*; AIP Publishing LLC: College Park, MD, USA, 2017; p. 020150.
22. Qiu, Q.; Xie, Z.; Wu, L.; Li, W. DGeoSegmenter: A dictionary-based Chinese word segmenter for the geoscience domain. *Comput. Geosci.* **2018**, *121*, 1–11. [CrossRef]
23. Qiu, X.; Pei, H.; Yan, H.; Huang, X. A Concise Model for Multi-Criteria Chinese Word Segmentation with Transformer Encoder. *arXiv* **2019**, arXiv:1906.12035.
24. Qun, N.; Yan, H.; Qiu, X.-P.; Huang, X.-J. Chinese word segmentation via BiLSTM+ Semi-CRF with relay node. *J. Comput. Sci. Technol.* **2020**, *35*, 1115–1126. [CrossRef]
25. Li, Y.; Cheng, J.; Huang, C.; Chen, Z.; Niu, W. NEDetector: Automatically extracting cybersecurity neologisms from hacker forums. *J. Inf. Secur. Appl.* **2021**, *58*, 102784. [CrossRef]
26. Sarna, G.; Bhatia, M.P.S. A probalistic approach to automatically extract new words from social media. In Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), San Francisco, CA, USA, 18–21 August 2016; IEEE: New York, NY, USA, 2016; pp. 719–725.
27. Wang, X.; Sha, Y.; Tan, J.-L.; Guo, L. Research of New Words Identification in Social Network for Monitoring Public Opinion. In Proceedings of the International Conference on Trustworthy Computing and Services, Beijing, China, 28 May–2 June 2012; Springer: Berlin/Heidelberg, Germany, 2012; pp. 598–603.
28. Breen, J.; Baldwin, T.; Bond, F. The Company They Keep: Extracting Japanese Neologisms Using Language Patterns. In Proceedings of the 9th Global Wordnet Conference, Singapore, 8–12 January 2018; pp. 163–171.
29. Cheng, K.; Wen, X.; Zhou, K. A Survey of Internet Public Opinion and Internet New Words. In *DEStech Transactions on Social Science, Education and Human Science*; DEStech Publishing Inc.: Lancaster, PA, USA, 2017.
30. Zhou, Q.; Chen, Y. New words recognition algorithm and application based on micro-blog hot. In Proceedings of the 2015 Seventh International Conference on Measuring Technology and Mechatronics Automation, Nanchang, China, 13–14 June 2015; IEEE: New York, NY, USA, 2015; pp. 698–700.
31. Zeng, H.-L.; Zhou, C.-L.; Zheng, X.-L. A New Word Detection Method for Chinese based on local context information. *J. Donghua Univ. (Engl. Ed.)* 2010. Available online: https://www.researchgate.net/publication/291707984_A_new_word_detection_method_for_chinese_based_on_local_context_information (accessed on 17 December 2021).
32. Li, X.; Chen, X. New Word Discovery Algorithm Based on N-Gram for Multi-word Internal Solidification Degree and Frequency. In Proceedings of the 2020 5th International Conference on Control, Robotics and Cybernetics (CRC), Wuhan, China, 16–18 October 2020; IEEE: New York, NY, USA, 2020; pp. 51–55.

33. Zhao, K.; Zhang, Y.; Xing, C.; Li, W.; Chen, H. Chinese underground market jargon analysis based on unsupervised learning. In Proceedings of the 2016 IEEE Conference on Intelligence and Security Informatics (ISI), Tucson, AZ, USA, 28–30 September 2016; IEEE: New York, NY, USA, 2016; pp. 97–102.
34. Chen, Q.; Cheng, G.; Li, D.; Zhang, J. Closeness Based New Word Detection Method for Mechanical Design and Manufacturing Area. *J. Comput. Comput. Soc. Repub. China (CSROC)* **2017**, *28*, 210–219.
35. Yang, C.; Zhu, J. New Word Identification Algorithm in Natural Language Processing. In Proceedings of the 2020 2nd International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI), Taiyuan, China, 23–25 October 2020; IEEE: New York, NY, USA, 2020; pp. 199–203.
36. Brown, P.F.; Della Pietra, V.J.; Desouza, P.V.; Lai, J.C.; Mercer, R.L. Class-based n-gram models of natural language. *Comput. Linguist.* **1992**, *18*, 467–480.
37. Gao, Y.; Zhou, L.; Zhang, Y.; Xing, C.; Sun, Y.; Zhu, X. Sentiment classification for stock news. In Proceedings of the 5th International Conference on Pervasive Computing and Applications, Hualien Taiwan, 10–13 May 2010.
38. Liang, F.M. *Word Hy-phen-a-tion by Com-put-er*; Department of Computer Science, Stanford University: Palo Alto, CA, USA, 1983.
39. Wang, J.; Ge, B.; He, C. Domain Neural Chinese Word Segmentation with Mutual Information and Entropy. In Proceedings of the 2019 7th International Conference on Information Technology: IoT and Smart City, Shanghai, China, 20–23 December 2019; pp. 75–79.
40. Shang, G. Research on Chinese New Word Discovery Algorithm Based on Mutual Information. In Proceedings of the 2019 2nd International Conference on Algorithms, Computing and Artificial Intelligence, Sanya, China, 20–22 December 2019; pp. 580–584.
41. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
42. Lee, Y. Systematic Homonym Detection and Replacement Based on Contextual Word Embedding. *Neural Process. Lett.* **2021**, *53*, 17–36. [CrossRef]
43. Chen, W.; Cai, Y.; Lai, K.; Yao, L.; Zhang, J.; Li, J.; Jia, X. WeiboFinder: A topic-based Chinese word finding and learning system. In Proceedings of the International Conference on Web-Based Learning, Cape Town, South Africa, 20–22 September 2017; Springer: Berlin/Heidelberg, Germany, 2017; pp. 33–42.
44. Kerremans, D.; Prokić, J. Mining the web for new words: Semi-automatic neologism identification with the NeoCrawler. *Anglia* **2018**, *136*, 239–268. [CrossRef]
45. Wang, F. Statistic Chinese New Word Recognition by Combing Supervised and Unsupervised Learning. In Proceedings of the 2019 IEEE Intl Conf on Parallel & Distributed Processing with Applications, Big Data & Cloud Computing, Sustainable Computing & Communications, Social Computing & Networking (ISPA/BDCloud/SocialCom/SustainCom), Xiamen, China, 16–18 December 2019; IEEE: New York, NY, USA, 2019; pp. 1239–1243.