News Text Sentiment Classification Method Based on Knowledge Graph

Yanting Xu[a], Tianxiao Ji[b] ,Zhi Li[c]*

[a]xuyanting@126.com,[b] jitianxiao@outlook.com, Corresponding author:[c] jasonleo_zhi@hotmail.com

Shanghai Branch of National Computer network Emergency Response technical Team/Coordination Center of China Shanghai, China

Abstract—With the rapid innovation of the Internet, news text sentiment classification is an urgent information processing problem. This paper proposes a news text sentiment classification method based on knowledge graphs, constructs news knowledge graphs based on the basic news text corpus, sets the emotional bias base, and calculates the emotional coefficients through the knowledge graphs to obtain the emotional bias of news texts. In the news text verification data, news text knowledge reasoning achieved an accuracy rate of 93.659%.

Keywords: news text sentiment classification, knowledge graph, news text corpus, emotional knowledge map

1 INTRODUCTION

The concept of Knowledge Graph was formally proposed by Google in 2012, aiming to realize a smarter search engine, and it has been popularized in academia and industry since 2013[1]. At present, with the continuous development of intelligent information service applications, knowledge graphs have been widely used in intelligent search, intelligent question and answer, personalized recommendation, intelligence analysis, anti-fraud and other fields. In addition, information, data, and link relationships on the Web can be aggregated into knowledge through knowledge graphs, making information resources easier to calculate, understand, and evaluate.

Using knowledge graphs for knowledge reasoning can be used to infer potential relationships, attributes and other information between entities. The core of the knowledge graph is the semantic network, which is a structured knowledge representation that can be used to represent various concepts, entities, attributes and the relationship between them. Knowledge reasoning is a computer technology that simulates the process of human reasoning and judgment. It can derive new knowledge from existing knowledge, thereby expanding our knowledge system. Knowledge reasoning can be used in many fields, such as natural language processing, artificial intelligence, expert systems, etc.

ICPDI 2023, September 01-03, Chongqing, People's Republic of China

DOI 10.4108/eai.1-9-2023.2338760

With the rapid development of the Internet, news text classification has become an important issue in the field of information processing. In recent researches, Koloski proposed knowledge graph informed fake news classification via heterogeneous representation ensembles[2]. Tang proposed a novel graph based neural network method, Gupdater, which is built upon graph neural networks (GNNs) with a text-based attention mechanism to guide the updating message passing through the KG structures[3]. Liu condisered to use An Enhanced Knowledge Graph for News Recommendation [4]. Rai classified the fake news using transformer basd enhanced LSTM and BERT [5]. Pan proposed the B-TransE model, to detecting fake news based on news content using knowledge graphs[6]. Mayank combined natural language processing (NLP) and tensor decomposition model to encode news content and embed Knowledge Graph (KG) entities, respectively[7]. Traditional rule-based or supervised learning methods often require a large number of manual annotations and lack sufficient generalization capabilities. Therefore, the technical research of using knowledge graphs to classify news texts has become increasingly important. The knowledge graph has rich semantic information, which can effectively assist machine learning algorithms and improve the accuracy and efficiency of news text sentiment analysis.

2 NEWS TEXT SENTIMENT CLASSIFICATION STEPS

Using knowledge graph to realize sentiment classification analysis of news text needs to use the relationship and connectivity in the graph to extract relevant information and make sentiment prediction. The following are the general steps for sentiment classification using knowledge graphs:

- Building a knowledge graph: first create a knowledge graph that represents various entities, concepts, and relationships related to news topics and sentiments. This can involve collecting data from reliable sources, using natural language processing techniques to extract entities and relationships, and organizing them into graph structures.
- Define Sentiment Labels: Identify the sentiment labels or categories you want to classify your news text into. For example, you might have positive, negative, and neutral sentiment categories, or you could use a finer-grained sentiment grading.
- Adding Sentiment Annotations to Knowledge Graphs: Assign sentiment labels to related entities and relations in knowledge graphs. This can be done manually by human annotators or by automated sentiment analysis techniques.
- Extracting features from news text: preprocessing news text, cleaning and word segmentation. Then, relevant features that can be used for sentiment classification are extracted from the text. This may involve techniques such as bag-of-words models, word embeddings (e.g. Word2Vec, GloVe), or more advanced contextual embeddings (e.g. BERT, GPT).
- Map news texts to knowledge graphs: analyze news texts and identify entities, concepts, or keywords that exist in the text. Use this information to search and traverse the knowledge graph to find relevant entities and relationships.

- Sentiment Classification Using Knowledge Graph: Once the news text is mapped into the knowledge graph, the sentiment annotations in the graph can be utilized for sentiment prediction. Sentiment labels can be aggregated from related entities, or the sentiment of a specific relationship can be used to determine the overall sentiment of a news text.
- Train and fine-tune the model: Use the sentiment-annotated data in the knowledge graph to train a sentiment classification model. You can employ various machine learning techniques such as supervised learning (e.g. SVM, Random Forest, Neural Networks) or even graph-based methods (e.g. Convolutional Networks). Fine-tuning a model with specific data can improve its performance.
- Evaluate and iterate: Evaluate the performance of your sentiment classification model using appropriate metrics such as accuracy, precision, recall, or F1-score. Iterate models and knowledge graphs as needed, optimize annotations and improve the training process to improve the accuracy of sentiment classification.

It is worth noting that building and maintaining a knowledge graph can be a challenging task. It needs to be constantly updated and curated to keep up with news topics and emotional patterns. Furthermore, the effectiveness of using knowledge graphs for sentiment classification depends on the quality and completeness of the data and relationships contained in the graphs.

3 CONSTRUCTION OF BASIC KNOWLEDGE GRAPH

The construction of the knowledge map mainly includes three stages, information extraction, knowledge fusion, knowledge processing.

1. Information extraction: (entity extraction, relationship extraction, and attribute extraction) extract entities, attributes, and interrelationships between entities from various types of data sources, and form ontological knowledge expressions on this basis;

2. Knowledge fusion (entity linking, knowledge merging): After acquiring new knowledge, it needs to be integrated to eliminate contradictions and ambiguities. For example, some entities may have multiple expressions, and a specific title may correspond to multiple different entities, etc.;

3. Knowledge processing (ontology construction, knowledge reasoning, and quality assessment): For the new knowledge that has been integrated, it needs to undergo quality assessment (some of which require manual identification) before adding qualified parts to the knowledge base to ensure that the knowledge base the quality of.

Specific steps are as follows:

3.1Basic data preparation

Basic data preparation includes obtaining basic news corpus data and labeling corpus sentiment according to the emotional tendency of news. First of all, it is necessary to collect a large

number of news articles as a corpus, and obtain news articles through web crawlers, API interfaces, etc.

According to the needs, determine the emotional labeling scheme, and use the label for emotional labeling. According to the emotional tendency conveyed by news articles, it is marked as the corresponding emotional category. Calculate the positive and negative values of the word according to the total sequence assignment of all the relationships around a word, then calculate the positive and negative values of the corpus according to the sum of the matched words in the corpus, and finally judge the positive and negative values of the corpus according to the discriminant median. Define positive coefficient ip, negative coefficient in,

if(relation=Positive) { ip=2, in=0}

if(relation= mayPositive) { ip=1, in=0}

if(relation= Even) {ip=0, in=0}

if(relation= mayNegative) { ip=0, in=1}

if(relation = Negative) { ip = 0, in = -2 }

3.2Calculation of word co-occurrence matrix

News titles are segmented into words, and word co-occurrence matrices of positive emotional news and negative emotional news are calculated respectively.

Word co-occurrence matrix (Co-occurrence Matrix) is a data structure used to measure the co-occurrence relationship between words in text. It records the co-occurrence frequency or count of each pair of words in a given text corpus.

When constructing the word co-occurrence matrix, we first need to define a fixed vocabulary, which contains all the different words that appear in the corpus. We then treat each document or sentence in the text as an observation unit and count the number of co-occurrences of each pair of words in these documents.

The rows and columns of the word co-occurrence matrix correspond to words in the vocabulary, and each element of the matrix represents the number or frequency of co-occurrence times or frequencies of words represented by the corresponding row and column in the corpus. The size of the matrix depends on the number of words in the corpus, so for larger corpora and larger vocabularies, the word co-occurrence matrix can be very large.

Word co-occurrence matrix is often used in text feature representation and semantic association analysis in the fields of natural language processing and information retrieval. It can be used to calculate the similarity between words, construct text representation vectors, and perform text clustering and other tasks. Common word co-occurrence matrix variants include point mutual information matrix (Pointwise Mutual Information Matrix) and orthogonal word co-occurrence matrix (Normalized Pointwise Mutual Information Matrix), etc. These variants can better capture the relevance between words .

A simple way to calculate the word co-occurrence matrix is to count the co-occurrence times of each pair of words by traversing the corpus sentence by sentence or document by document. Here is a basic step:

Define vocabulary: Based on the words in the corpus, build a vocabulary that contains all the different words.

Initialize the matrix: Create an empty word co-occurrence matrix with the number of rows and columns equal to the number of words in the vocabulary. In this paper, the number of words is 73,207.

Traverse the corpus: For each document or sentence in the corpus, the following steps are performed:

a. Traverse the words in the current document.

b. For each pair of words (word A and word B), determine whether they co-occur.

c. If word A and word B co-occur, increment the count of the corresponding element in the word co-occurrence matrix.

After completing the traversal, each element in the word co-occurrence matrix represents the number or frequency of co-occurrence of the corresponding word pair.

It should be noted that for large corpora, the word co-occurrence matrix may be very large, resulting in high storage and computational costs. In order to reduce storage space and computational complexity, a sparse matrix representation method can be used to store the word co-occurrence matrix, and only the positions and values of non-zero elements are recorded.

3.3Construction of graph association relationship

Through the number of positive and negative co-occurrence relationships of the same group of words, taking into account the overall co-occurrence ratio, construct a graph association relationship

Taking all parties into consideration, the final solution is:

Construct positive bilingual co-occurrence knowledge map and negative bilingual co-occurrence knowledge map through positive attachment of bilingual news corpus

For the news that has been marked with ip and in, the number of news is N. And for a specific group of words, the word frequency that appears in the k^{th} article is recorded as f_k. Then the graph relationship of a specific group of words can be expressed as (1)

$$(\Sigma ip_k * f_k \ , \ \Sigma in_k * f_k) \tag{1}$$

3.4News Text Classification Using Knowledge Graph

For the news text to be classified, its word co-occurrence relationship has a total of n items, which are recorded as R_1 to R_n. Then for the k^{th} word co-occurrence relationship R_k, the coefficient of the relationship in the positive and negative emotional knowledge map is recorded as ip_k and in_k. After that the normalization is processed: judging the positive and negative 2.4 through the magnitude difference relationship, the average value of Positive is 10.54 and the average value of Negative is 2.75, 4 times the difference, return the relationship

if(Σink==0){relation=Positive}

```
if(Σipk==0){relation=Negative}
if(Σipk>=4*Σink){
    if(Σipk>=16*Σink){relation= Positive}
    else if(Σipk>=8*Σink){relation= mayPositive}
    else if(Σipk>=4*Σink){relation= Even}
}
else if(Σipk<4*Σink){
    if(4*Σink>=4*Σipk){relation= Negative}
    else if(4*Σink>=2*Σipk){relation= mayNegative}
    else if(4*Σink>=Σipk){relation= Even}
```

3.5Knowledge graph update

The importance of building the knowledge map update process and updating the knowledge map is to ensure the quality of the knowledge base.The basic BKG constructed contains the basic corpus text data. In order to add positive and negative emotion-related knowledge to the knowledge graph, it is necessary to add new entities and new relationships to the previous graph based on the new corpus data and integrate them into a new knowledge graph. Using bilingual news website corpora in knowledge graph updates. The update of the knowledge map is helpful for the learning and maintenance of later knowledge, and improves the accuracy of knowledge reasoning.

4 RESULT

Through the above steps, 218,000 pieces of positive emotional news corpus are derived from news predictions to construct positive correlations; 240,000 pieces of negative emotional news corpus are derived to construct negative correlations; finally 233,400 positive co-occurrence relationships are constructed to construct negative co-occurrence relationships 499,700. Finally, a total of 73,207 nodes and 994,976 relationships were constructed in the knowledge map, which is shown in Figure 1.

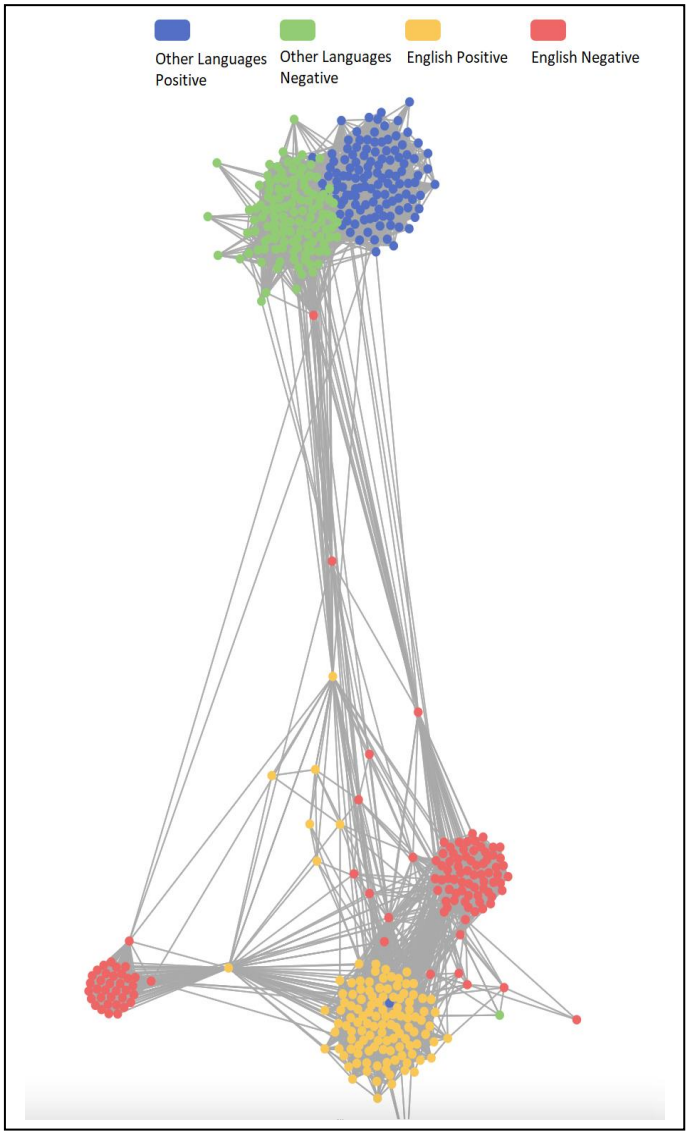

Figure 1. Result of knowledge map

It can be clearly seen that the positive and negative of other languages and the positive and negative of English are aggregated, so it is more reasonable to construct a knowledge map by language. A few words and keywords have cross-lingual or cross-emotional connections, which are worth studying.

Part of the knowledge map results are shown in the Figure 2:

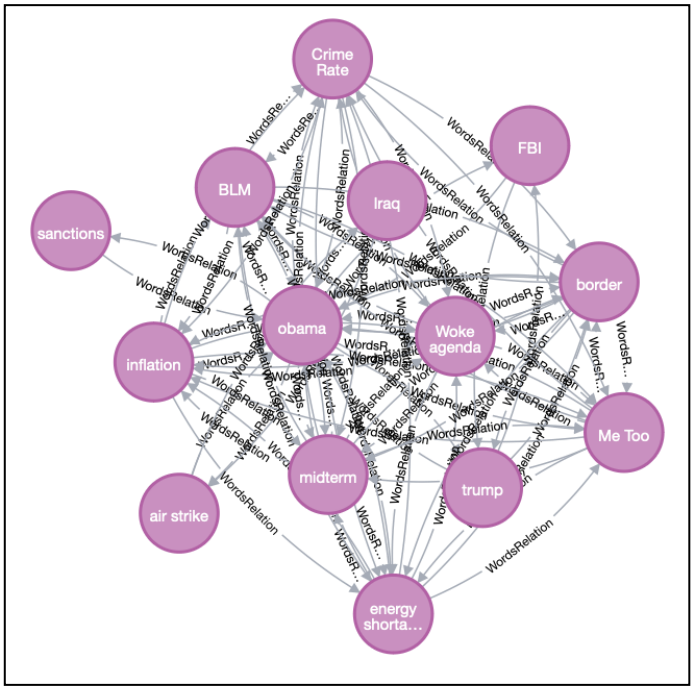

Figure 2. Part of the knowledge map results

We selected 50,000 test news predictions and tested the emotional results of knowledge graph inference news. This paper uses Accuracy, Precision, Recall, and Macro F1 to measure the accuracy of our model.

A confusion matrix, also known as an error matrix, is a standard format for expressing an accuracy estimate. The confusion matrix mainly contains the following four values.

True Positives, abbreviated TP: The number of features that are predicted to be positive samples and are actually positive samples

False Positives, abbreviated FP: The number of features predicted as positive samples and actually negative samples

True Negatives, abbreviated TN: The number of features predicted as negative samples and actually negative samples

False Negatives, abbreviated FN: The number of features predicted as negative samples and actually positive samples

Accuracy is defined as (2):

$$\text{Accuracy} = (TP+TN)/(P+N) \tag{2}$$

Precision is defined as (3):

$$P=TP/(TP+FP) \tag{3}$$

Recall is defined as (4):

$$R=TP/(TP+FN) \tag{4}$$

The macro F1 measure (macro F1) is defined as (5):

$$F1=2*Precision*Recall/(Precision+Recall) \tag{5}$$

The final experimental data test results are shown inThe results are as follows: TP=48748; FP=395; FN=2909; TN=52. Accuracy=93.659%; Precision=99.196%; Recall=94.369%; F1_Score=96.722%.

In future research, it can be considered based on the subgraph and the path of the order of the words in the sentence in the subgraph, calculate the weight of the path, use the subgraph to reason and calculate, and compare whether the results of the two are more accurate. From the perspective of words, weighting can be carried out according to their different positions in the sentence. From the perspective of graphs, a large amount of corpus can be used to analyze the weights of words or relationships in the training graphs in inference.

REFERENCES

[1] Fensel D, Şimşek U, Angele K, et al. Introduction: what is a knowledge graph [J]. Knowledge graphs: Methodology, tools and selected use cases, 2020: 1-10.

[2] Koloski B, Perdih T S, Robnik-Šikonja M, et al. Knowledge graph informed fake news classification via heterogeneous representation ensembles[J]. Neurocomputing, 2022, 496: 208-226.

[3] Tang J, Feng Y, Zhao D. Learning to update knowledge graphs by reading news[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). 2019: 2632-2641.

[4] Liu D, Bai T, Lian J, et al. News Graph: An Enhanced Knowledge Graph for News Recommendation[C]//KaRS@ CIKM. 2019: 1-7.

[5] Rai N, Kumar D, Kaushik N, et al. Fake News Classification using transformer based enhanced LSTM and BERT[J]. International Journal of Cognitive Computing in Engineering, 2022, 3: 98-105.

[6] Pan J Z, Pavlova S, Li C, et al. Content based fake news detection using knowledge graphs[C]//The Semantic Web–ISWC 2018: 17th International Semantic Web Conference, Monterey, CA, USA, October 8–12, 2018, Proceedings, Part I 17. Springer International Publishing, 2018: 669-683.

[7] Mayank M, Sharma S, Sharma R. DEAP-FAKED: Knowledge graph based approach for fake news detection[C]//2022 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM). IEEE, 2022: 47-51.