IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Vol. 12, Issue 3, March 2023

DOI: 10.17148/IJARCCE.2023.12369

Duplicate Question Pairs detection using NLP

Kumar Saksham[1], Dr. Pawan Kumar[2], Dr. Mir Aadil[3]

Student, MCA Department (School of CS & IT), Jain (Deemed-to-be University), Bengaluru, India[1]

Assistant Professor, MCA Department (School of CS & IT) Jain (Deemed-to-be University), Bengaluru, India[2,3]

Abstract*:* Natural language processing programs that detect duplicate queries have a requirement for an efficient way to determine how similar two very brief texts or sentences are. An illustration of this is a conversational agent or dialogue system with script methods, where the implementation of comparable sentences is crucial. Sentence similarity will also be used in internet-related contexts. Sentence similarity has proven to be one of the most effective methods for increasing the effectiveness of retrieving Web pages.

Keywords: Duplicate, Question, Paper, Detection, and NLP.

I. INTRODUCTION

In-text mining uses sentence similarity as a criterion to extract hidden knowledge from textual collections. By using duplicate question identification, a question-answering service that is becoming more automated can examine previous interactions between clients and human support staff over a chat channel. This also holds true for identifying questions that have the same semantic meaning, which led to the creation of the Duplicate Question Detection project (DQD). If there are duplicate questions, the system warns you. A DQD project's major aim is to anticipate if two questions can be regarded as equivalent in order to ascertain whether this kind of question has ever been asked before. In DQD, a question is retrieved as an input, coupled with other questions, processed to collect important attributes, and then fed to our algorithm to make the final prediction. The amount of words that match between the two questions, the length of the two questions, the number of words and characters, and the number of stop words are all plainly visible text-based features.

Nevertheless, employing TF-IDF scores is both computationally expensive and not very beneficial. In the context of customer service via a chat channel, determining replies to inquiries is a crucial step. Given the resurgence of interest in conversational interfaces, automatic recognition of semantically identical questions is a language processing task of the utmost relevance. A new input query that is identical to one that has previously been recorded can automatically receive the same response as its recorded counterpart. Stack Overflow is a popular online question and answer site for software developers to share their experience and expertise, where two or more questions may express the same point and thus are duplicates of one another. Duplicate questions make Stack Overflow site maintenance harder, waste resources that could have been used to answer other questions, and cause developers to unnecessarily wait for answers that are already available. To reduce the duplicate question problem, Stack Overflow allows you to manually mark a question as a duplicate of another question. Yet, because so many questions are submitted to Stack Overflow every day, it is challenging to manually spot duplicate queries. Consequently, it is necessary to use an automated method to find these repeating queries. An example of a well-researched NLP issue called paraphrase identification, which uses Natural Language Sentence Matching (NLSM) to evaluate whether two sentences are identical, is the problem of spotting repeated queries.

II. RELATED WORK

This paper was published in IEEE Xplore in 2020 by L. Wang, L. Zhang mention that They explore the use of powerful deep learning techniques, including Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN) and Long Short-Term Memory (LSTM), to detect duplicate questions in Stack Overflow. Therefore, we construct three deep learning approaches WV-CNN, WV-RNN and WV-LSTM, which are based on Word2Vec, CNN, RNN and LSTM, to detect duplicate questions in Stack Overflow. Furthermore, the experimental results indicate that our approaches WV-CNN, WV-RNN, and WV-LSTM outperform four machine learning approaches based on Support Vector Machine, Logic Regression, Random Forest and eXtreme Gradient Boosting in terms of recall-rate@5, recall-rate@10 and recall-rate@20.

This paper was published in IEEE Xplore in 2020 by D. A. Prabowo and G. Budi Herwanto Mention that online forums are platforms for gathering, sharing information, and discussing between users on a particular topic. Users in online forums can ask questions about a topic, then other users who are experts on that question would answer the question. Therefore, a model is needed to detect the semantic similarity of questions in online forums. In this study, we are using Convolutional Neural Networks (CNN) to detect the semantic similarity of questions. To capture the semantic similarity between

DOI: 10.17148/IJARCCE.2023.12369

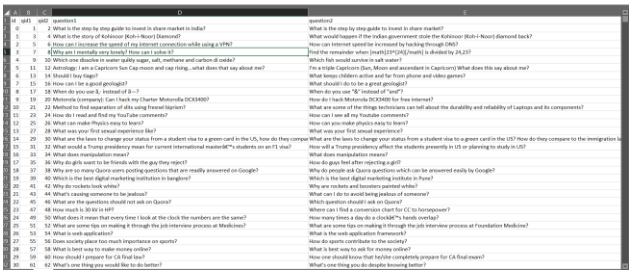
questions, we are using Glove pre-trained word embedding. This word embedding vector is used as an input for CNN, then the output is compared with Siamese Neural Networks

Wang, L., and Jiang, J. (2020) say in their research paper that in Stack Overflow, each question contains many attributes, such as an ID, title, body, tags, creation date, closed date, etc. Although users are reminded to search for the forum before creating a new question, duplicate questions frequently appear in Stack Overflow. Duplicate questions refer to questions that were previously created and answered on Stack Overflow. In order to reduce the number of duplicate questions, users with high reputation are encouraged to manually mark duplicate questions in Stack Overflow. Moreover, a pair of duplicate questions consists of a master question and a nonmatter question.

Paraphrase identification is a well-studied task in NLP (Das and Smith, 2009; Chang et al., 2010). Here, we focus on an instance that of finding questions with identical meaning. With the renaissance of neural networks, several neural-based frameworks have been proposed for the task of paraphrase identification. The first framework is based on a siamese neural network consisting of two subnetworks joined at their outputs, where the sub-networks share the same weights at all levels and are responsible for extracting features from the input, and the output level computes the distance between the two feature vectors generated by the sub-networks (3). The shortcoming of this approach is that there is no interaction between two sentences during the training process, which might cause information loss. The "compare-aggregate" approach is proposed (4), which captures the interation between two sentences by performing a word-level matching and aggregating the results into a vector the final classification. However, this approach fails to account for other types of matchings such as phrase-by-sentence and only performs matching in a single direction, thus neglecting information in the sentence pairs Wang et al. (2017) present the bilateral multi-perspective matching model (BiMPM) to tackle the limitations of neural-based frameworks. This approach uses a character-based LSTM at its input representation layer, a layer of bi-LSTMs for computing context information, four different types of multi-perspective matching layers, an additional bi-LSTM aggregation layer, followed by two-layer feedforward network for prediction. In contrast, the decomposable attention model uses four simple feedforward networks to attend, compare and predict, leading to a more efficient architecture

III. SYSTEM ANALYSIS

a. DATASET



The given sentence indicates that there is a dataset called Contin which contains more than 20,000 questions. The system is currently working on this dataset, meaning that it is processing, analyzing, or performing some kind of operation on the questions in the dataset. Additionally, the sentence mentions that the dataset contains different types of questions, which could refer to questions on various topics or in different formats.

b. Steps to perform system analysis

- Data collection: Gather a dataset of question pairs that are labeled as either duplicates or not duplicates.
- Data pre-processing: Clean the text data by removing any unnecessary characters, stop words, and converting text to lowercase.
- Feature engineering: Convert the text into numerical features that can be used by an algorithm. One common approach is to use word embeddings such as Word2Vec, GloVe, or BERT.
- Train a model: Train a machine learning model such as logistic regression, random forest, or neural networks on the labeled data. The model should be able to take in a pair of questions and output a score that indicates how similar they are.
- Evaluation: Evaluate the performance of the model using metrics such as accuracy, precision, recall, F1 score, and AUC-ROC.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

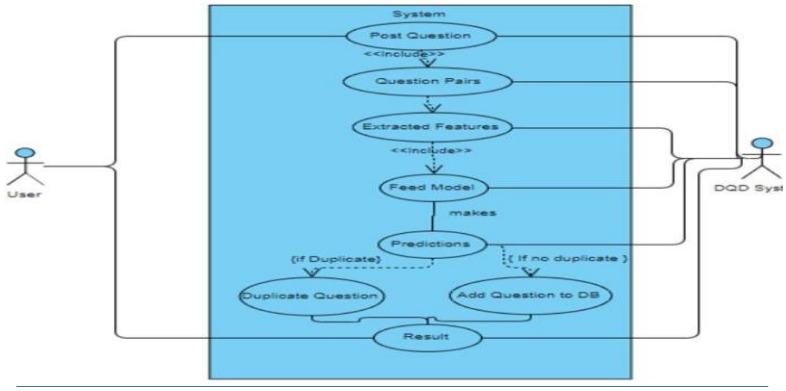
ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Vol. 12, Issue 3, March 2023

DOI: 10.17148/IJARCCE.2023.12369

- Model improvement: If the model is not performing well, try adjusting the hyperparameters, changing the feature representation, or using a different algorithm.
- Deployment: Once the model is performing well, deploy it in a production environment to detect duplicate question pairs in real-time

IV. SYSTEM DESIGN & ARCHITECTURE

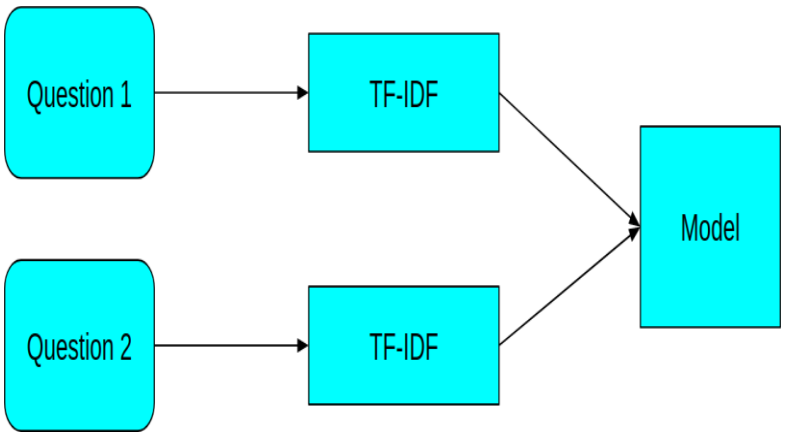
a. USECASE:



Explanation of the use case diagram:

- The user/client submits a question to the system to check for duplicate question.
- The system processes the question using NLP techniques to extract the semantic meaning and linguistic structure of the question.
- The system compares the processed question with a database of previously asked questions to check for any duplicates.
- The system generates a result indicating whether the submitted question is a duplicate or not.
- The result is displayed to the user/client.

b. DATAFLOW DIAGRAM



Explanation of the data flow diagram:

- The user/client submits a question to the system for duplicate detection.
- The submitted question is processed using NLP techniques to extract its semantic meaning and linguistic structure.
- The processed question is compared with a database of previously asked questions to check for any duplicates.
- The system generates a result indicating whether the submitted question is a duplicate or not.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Vol. 12, Issue 3, March 2023

DOI: 10.17148/IJARCCE.2023.12369

V. IMPLEMENTATION & RESULT

```
In [4]: new_df.head()
```

Out[4]:

	id	qid1	qid2	question1	question2	is_duplicate
398782	398782	496695	532029	What is the best marketing automation tool for...	What is the best marketing automation tool for...	1
115086	115086	187729	187730	I am poor but I want to invest. What should I do?	I am quite poor and I want to be very rich. Wh...	0
327711	327711	454161	454162	I am from India and live abroad. I met a guy f...	T.I.E.T to Thapar University to Thapar Univers...	0
367788	367788	498109	491396	Why do so many people in the U.S. hate the sou...	My boyfriend doesnt feel guilty when he hurts ...	0
151235	151235	237843	50930	Consequences of Bhopal gas tragedy?	What was the reason behind the Bhopal gas trag...	0

```
In [5]: def preprocess(q):
```

Figure 1: According to the given graphic, if two questions have the same meaning when compared, the result will be 1, signifying similarity. In contrast, if the questions' meanings disagree, the outcome will be 0, signifying dissimilarity. This implies that the diagram might be a depiction of a method for determining how semantically similar two queries are. A number of areas, including information retrieval and natural language processing, could benefit from the usage of such a system. We can increase the accuracy of automated language processing systems by comparing the similarity of the queries in order to better grasp the text's purpose.

VI. CONCLUSION

Duplicate question detection is a crucial task in the field of natural language processing that aims to determine whether a given pair of questions have similar meanings or not. This task is essential in information retrieval and question answering systems, especially with the growing use of online forums, Q&A websites, and virtual assistants. NLP techniques such as word embedding, semantic similarity, and deep learning models like Siamese Neural Networks have exhibited significant success in identifying duplicate questions, even when they differ in phrasing and context. Despite these achievements, there is still ongoing research in this field, as there is a need for more efficient and accurate models. The future direction of this research will focus on enhancing the scalability and robustness of these models to handle large data volumes, and exploring the potential of incorporating domain-specific knowledge to improve their performance. Overall, the task of duplicate question detection using NLP is challenging yet significant, as it has practical applications in improving the quality of information retrieval and question answering systems.

REFERENCES

[1] Liang, D., Zhang, F., Zhang, W., Zhang, Q., Fu, J., Peng, M., & Huang, X. (2019, July). Adaptive multi-attention network incorporating answer information for duplicate question detection. In Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 95-104

[2] Prabowo, D. A., & Herwanto, G. B. (2019, July). Duplicate question detection in question answer website using convolutional neural network. In 2019 5th International Conference on Science and Technology (ICST) (Vol. 1, pp. 1-6). IEEE.

[3] Shah, D. J., Lei, T., Moschitti, A., Romeo, S., & Nakov, P. (2018). Adversarial domain adaptation for duplicate question detection. arXiv preprint arXiv:1809.02255.

[4] Wang, L., Zhang, L., & Jiang, J. (2020). Duplicate question detection with deep learning in stack overflow. IEEE Access, 8, 25964-25975

[5] Xu-hang Wu, Xia Li, Shuo Kong, Yu Zhao, Lin Peng: Application of EfficientNetV2 and YoloV5 for tomato leaf disease identification, At IEEE Xplore 2022

[6] Zhu, X.K.: Research on Tomato Disease Identification Based on Convolutional Neural Network. Beijing University of Technology, Beijing, China (2020)

[7] Sardogan, M., Tuncer, A., Ozen, Y.: Plant leaf disease detection and classification based on CNN with LVQ algorithm. In: Proceedings of the 2018 3rd International Conference on Computer Science and Engineering (UBMK), Sarajevo, Herzegovina, pp. 382–385 (2018)

[8] Shah, D. J., Lei, T., Moschitti, A., Romeo, S., & Nakov, P. (2018). Adversarial domain adaptation for duplicate question detection. arXiv preprint arXiv:1809.02255.

[9] C. Saedi, J. Rodrigues, J. Silva, A. Branco and V. Maraev, "Learning Profiles in Duplicate Question Detection," 2017 IEEE International Conference on Information Reuse and Integration (IRI), San Diego,CA, USA, 2017, pp. 544-550, doi: 10.1109/IRI.2017.39

[10]Zhou, Q., Liu, X., & Wang, Q. (2021). Interpretable duplicate question detection models based on attention mechanism. Information Sciences, 543, 259-272.