SI_CP_Paper (1).docx

By qwerty

WORD COUNT	3483	TIME SUBMITTED	03-OCT-2023 06:04AM
		PAPER ID	103178985

Quora Question Duplication Detection: An ML Approach for Identifying Semantically Equivalent Questions

Zareena Begum
Department of Artificial Intelligence and Data Science
Vishwakarma Institute of Technology.
Pune,India,411037
@vit.edu

Sahil Kalal
Department of Artificial Intelligence and Data Science
Vishwakarma Institute of Technology.
Pune,India,411037
sahil.kalal21@vit.edu

Aryan Kashyap
Department of Artificial Intelligence and Data Science
Vishwakarma Institute of Technology.
Pune,India,411037
aryan.kashyap21@vit.edu

Omkar Kamble
Department of Artificial Intelligence and Data Science
Vishwakarma Institute of Technology.
Pune,India,411037
omkar.kamble21@vit.edu

Saniya Mhamane
Department of Artificial Intelligence and Data Science
Vishwakarma Institute of Technology.
Pune,India,411037
saniya.mhamane22@vit.edu

Soham Pande
Department of Artificial Intelligence and Data Science
Vishwakarma Institute of Technology.
Pune,India,411037
soham.pande22@vit.edu

***Abstract*— This research paper delves into the domain of Natural Language Understanding (NLU) by tackling the task of duplicate question detection using the Quora dataset. Our study encompasses an exhaustive exploration of the dataset and the application of machine learning methodologies, specifically Random Forest and XGBoost models. Significantly, our findings illuminate the prominence of a straightforward Continuous Bag of Words neural network model, surpassing the performance of more intricate recurrent and attention-based models. In addition, we rigorously conduct error analysis, uncovering nuances and subjectivity embedded in the dataset's labeling process. This investigation underscores the effectiveness of neural network-based approaches in addressing the intricacies of duplicate question detection, contributing to the broader field of NLU research.**

Keywords*— *Natural Language Understanding, Duplicate Question Detection, Quora Dataset, Machine Learning, Neural Network Models, Error Analysis, Subjectivity Analysis.

I. Introduction

The Quora dataset presents a critical challenge in the field of Natural Language Understanding (NLU): the task of determining whether pairs of questions have identical meanings, thus classifying them as duplicates. Quora, a popular question-and-answer platform, serves as a valuable resource for users seeking answers and insights across a wide range of topics. However, the increasing volume of questions has led to a proliferation of duplicate inquiries, potentially impeding users from accessing high-quality responses. Furthermore, responders may hesitate to address the same question repeatedly.

Recognizing duplicate questions is instrumental in alleviating these issues and optimizing the user experience. It not only streamlines the burden on responders but also facilitates the redirection of users to the most pertinent responses, enhancing overall user satisfaction.

This task necessitates robust Natural Language Understanding (NLU) capabilities. Effective NLU involves the construction of meaningful representations of human language, a formidable challenge with implications extending to various Natural Language Processing (NLP) tasks, including translation, summarization, and reading comprehension. The core challenge here lies in determining whether two sentences convey the same meaning, requiring the model to grasp lexical and syntactic nuances such as quantification, tense, modality, and syntactic ambiguity.

Given the complexity of this task, the Quora dataset serves as a compelling problem for exploration. In this paper, we embark on a comprehensive investigation of machine learning models to evaluate their performance on this dataset. Our methodology deviates from previous approaches that incorporated complex models, such as Support Vector Machines (SVM), gradient boosted trees, and deep neural networks. Instead, we adopt a simpler baseline approach employing linear models. We meticulously analyze and discuss the performance of these models.

The essence of duplicate question detection lies in binary classification, where the objective is to classify questions of varying lengths as either duplicates or non-duplicates. The pivotal challenge is to transform sentences into numerical representations suitable for machine learning algorithms. The prevalent industry practice involves manual feature engineering, often combined with tree-based models, like random forests. This aligns with Quora's current approach (Dandekar, 2017) and can be further enhanced by incorporating bag-of-words-based models (Siu, 2016).

While traditional approaches have been effective, the surge in neural network research has introduced a plethora of deep learning methods for sentence classification and representation building (Sutskever et al., 2014; Collobert and Weston, 2008). Notably, substantial progress has been made in Natural Language Inference (NLI), a task involving the determination of entailment, contradiction, or neutrality between pairs of sentences. Inspired by work on the Stanford Natural Language Inference corpus (SNLI) (Bowman et al., 2015), our neural network explorations draw from these advancements.

In this paper, we present a comprehensive examination of duplicate question detection, spanning traditional feature engineering to state-of-the-art neural network approaches, aiming to shed light on their effectiveness and applicability.

II. LITERATURE REVIEW

[1]The research uses natural language processing technology to create an interactive project management platform, which represents a novel approach in the construction sector. The system automates contract administration procedures by merging the Progressive Scale Expansion Network (PSENet), Convolutional Recurrent Neural Network (CRNN), and Bi-directional Recurrent Neural Networks Convolutional Recurrent Neural Network (BRNN-CNN) toolboxes. This process effectively organizes papers, drastically reduces the amount of human mistakes, and clears up ambiguities through real-time exact communication. It acts as a cutting-edge remedy for huge real-time document flows, successfully fostering collaboration and communication among all contract stakeholders.

[2] This paper examines the traditional approaches to risk management in significant transport projects, concentrating on risk assessment. Expert comments gained through risk workshops have always been crucial. The uniqueness of this strategy hasn't received enough attention, though. This study intends to evaluate the similarities in project risks among 70 significant transport projects carried out using various techniques. The study thoroughly examines risk registers using natural language processing (NLP) and the deep learning technique Word2vec. Surprisingly, a considerable degree of similarity between risk registers for various projects was found, emphasizing the possibility of a data-driven strategy to establish a common risk register while taking into account the particular hazards associated with each project. The key contributions of this work are developing a relationship between risk distinctiveness and project delivery strategies in transport projects and taking a novel way to analyze risk registers at the project level.

[3] This paper delivers the relationship between a university and industry to create an AI-based simulation platform for social work education is explored in this conceptual paper. The study emphasizes Natural Language Processing (NLP) as a pedagogical innovation and critically assesses the ongoing project, assessing both the potentials and constraints of NLP in the context of social work education. The research offers "lessons learned" based on the case study and is grounded in the Community of Inquiry (CoI) paradigm. It promotes the active participation of social work educators in the development of pedagogies made possible by cutting-edge AI technologies, including artificial intelligence, natural language processing, and virtual simulation.

[4]This paper discusses the idea of technical debt, which is a compromise between developer shortcuts and programme quality. Prior studies, concentrating on design and requirement debt, found self-admitted technical debt through source code comments. This work provides an automated identification approach using Natural Language Processing (NLP), as opposed to manual methods. The research outperforms previous keyword-based techniques in detecting self-admitted technical debt after analyzing 10 open-source projects. Notably, the study uses very little comment usage to achieve amazing accuracy in identifying specific phrases connected to design and requirement debt. This method represents a significant development since it ensures precise technical debt detection with little data, demonstrating its effectiveness and promise for real-world use.

[5]This paper discusses the crucial problem of construction project scheduling quality, highlighting the difficulties encountered during the planning and design phases. Although schedule quality is frequently compromised by time restrictions, little has been done to review and maintain schedules throughout the building phase. Schedule quality can only be manually diagnosed, which takes time and is arbitrary. Using task ontology and dependency-based information schema, this study provides a unique semantic-based logic reasoning and representation technique. By automatically separating building techniques and tasks, this methodology maintains consistent project timetables. The system's effectiveness is demonstrated by the assessment, which provides academics and practitioners with an automated method to spot schedule flaws and keep high-quality schedules over the course of a project.

[6] This study explores how pharmacoepidemiological findings are communicated in the media representation of the CNODES isotretinoin research. The study uses natural language processing to analyze 26 news stories and 3 CNODES publications. The analysis pinpoints separate media coverage clusters, exposing differences in vocabulary and topics. Notably, all of the publications used more complicated vocabulary than what is advised for reading about health. The study emphasizes the significance of using NLP strategies to evaluate the efficacy of medication safety communication. This research provides crucial insights into the public's comprehension and highlights areas for development in upcoming communication strategies. It also throws light on the dissemination difficulties experienced by drug safety studies in the media.

III. METHODOLOGY

1. Exploratory Data Analysis:

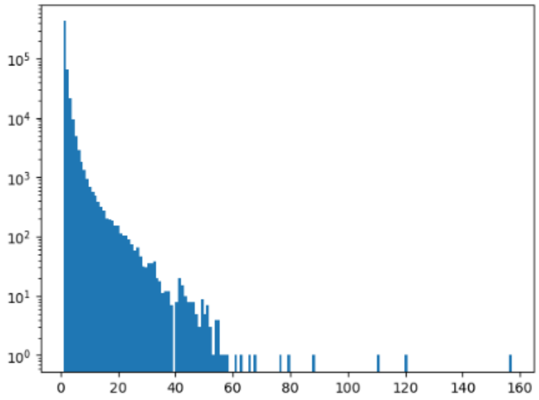
The dataset utilized in this research comprises the Quora Question Pairs dataset, consisting of a training set containing 404,290 question pairs and a test set comprising 2,345,795 question pairs, originally provided as part of a Kaggle competition [1].

To enable the calculation of additional performance metrics and facilitate further error analysis on our prediction models, we chose to construct our own test set using the provided training set. As such, our data exploration and model evaluation focused exclusively on the training set of 404,290 question pairs. Detailed information regarding the partitioning of this dataset into training, validation, and test subsets can be found in subsection 2.2.

The dataset includes various fields for each sample point:

- id: unique ID of each pair
- qid1: ID of first question
  qid2: ID of second question
- question1: text of first question
- question2: text of second question
- is_duplicate: are the questions duplicates of each other (0 indicates not duplicate, 1 indicates duplicate)

One notable aspect of the dataset is class imbalance. Specifically, 255,027 question pairs (63.08%) are labeled as non-duplicates (0), while 149,263 question pairs (36.92%) are labeled as duplicates (1). Managing this class imbalance played a crucial role in our model development and evaluation.

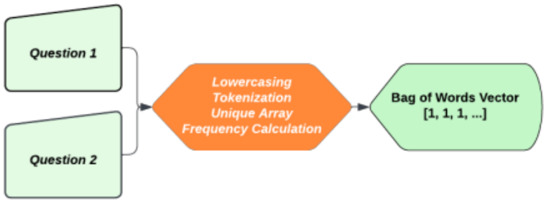


Furthermore, while each question pair in the dataset is unique, individual questions within these pairs may be repeated. Approximately 79.22% of the questions appear more than once, with some questions appearing up to 158 times across different pairs. The dataset comprises a total of 537,933 unique questions, and 111,780 of these questions occur in multiple question pairs. Visual representation of the distribution of question repetitions is provided in Figure 1.

The character set within our dataset is not strictly limited to ASCII characters. We identified 6,228 questions containing non-ASCII characters, which were distributed across 8,744 question pairs. Additionally, two pairs in the dataset contained empty strings for one of their questions.

As additional information, it's worth noting that the dataset includes 537,933 unique questions, with 111,780 questions appearing in multiple pairs. This comprehensive dataset description provides valuable insights into the dataset's characteristics, challenges, and the steps taken to address them throughout our analysis and modeling processes.

2. Bag of Words (BoW) approach



In this study, we adopt the Bag of Words (BoW) approach to address our research objectives efficiently. BoW simplifies text analysis by treating documents as unordered collections of words.

The BoW algorithm includes the following steps:

- Text Preprocessing: We preprocess textual data by tokenization, lowercasing, and removing punctuation, stop words, and noise.

- Feature Extraction: We convert text into a numerical matrix, with each unique word serving as a feature, creating a high-dimensional feature space.

- Vectorization: The BoW representation is typically large and sparse. Dimensionality reduction techniques like TF-IDF may be employed to reduce complexity.

Applying the BoW approach results in the inclusion of approximately 3000 new numerical features in the dataset. This addition enhances its analytical capacity and facilitates more robust analysis and modeling.

3. Data Pre-processing

In the data preprocessing phase, several techniques were applied to enhance the quality and consistency of the dataset. These preprocessing steps included:

- Lowercasing: Both questions were converted to lowercase to ensure uniformity in text case throughout the dataset.
- Whitespace Removal: Extraneous white spaces were removed to streamline the text and maintain data cleanliness.
- Special Character Conversion: Special characters were transformed into their respective string equivalents to eliminate any potential irregularities in the text.
- Decontracting Words: The decontraction process was applied to standardize contractions (e.g., "can't" to "cannot") for better text analysis.
- HTML Tag and Punctuation Removal: HTML tags and punctuation marks were removed from the text to improve text clarity and facilitate subsequent analysis.

In addition to data preprocessing, we undertook basic feature engineering by developing various functions that introduced seven new features into the dataset. These newly created features are as follows:

- q1_len: The length of the first question.
- q2_len: The length of the second question.
- q1_words_number: The number of words in the first question.
- q2_words_number: The number of words in the second question.
- words_common: The count of common words shared between both questions.
- total_words: The total number of words in both questions combined.
- word_share: The ratio of common words to the total number of words (common/total).

As a result of these data preprocessing steps and the incorporation of additional features, the dataset now encompasses a total of 6007 features, which collectively contribute to a more comprehensive and analytically rich dataset for our research analysis.

4. Advanced Feature Engineering

In this phase of advanced feature engineering, our aim was to further enhance the predictive power of our model beyond what basic feature engineering had initially provided. Through careful analysis and insights drawn from the Kaggle competition and the research community, we identified the need to incorporate additional parameters into our dataset. These new features fall into three distinct categories: token features, length-based features, and fuzzy features.

To better comprehend these feature additions, it is essential to establish some key terminology. Firstly, "token" refers to all the words present in a given question. "Stop words," on the other hand, denote words that contribute little semantic meaning to a sentence, such as "a," "the," or "of." Thus, "words" refer to tokens excluding stop words.

Building upon this foundation, we introduced a set of features inspired by successful approaches observed in other solutions within the Kaggle competition:

A. Token Features:

- cwc_min: The ratio of the number of common words to the length of the smaller question.
- cwc_max: The ratio of the number of common words to the length of the larger question.
- csc_min: The ratio of the number of common stop words to the smaller stop word count among the two questions.
- csc_max: The ratio of the number of common stop words to the larger stop word count among the two questions.
- ctc_min: The ratio of the number of common tokens to the smaller token count among the two questions.
- ctc_max: The ratio of the number of common tokens to the larger token count among the two questions.
- last_word_eq: 1 if the last word in the two questions is the same, 0 otherwise.
- first_word_eq: 1 if the first word in the two questions is the same, 0 otherwise.

B. Length-Based Features:

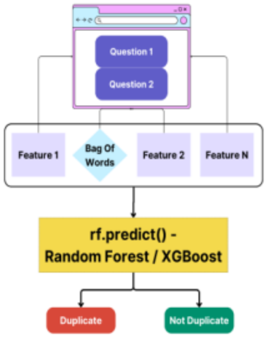
- mean_len: The mean length (number of words) of the two questions.
- abs_len_diff: The absolute difference between the lengths (number of words) of the two questions.
- longest_substr_ratio: The ratio of the length of the longest common substring between the two questions to the length of the smaller question.

C. Fuzzy Features:

- fuzz_ratio: Fuzzy ratio score derived using the fuzzywuzzy library.
- fuzz_partial_ratio: Fuzzy partial ratio obtained from fuzzywuzzy.
- token_sort_ratio: Token sort ratio from fuzzywuzzy.
- token_set_ratio: Token set ratio from fuzzywuzzy.

By incorporating these additional features, our dataset now boasts a total of 15 new attributes. When combined with the existing 6007 features, this comprehensive set of 6022 features equips our model with a richer and more informative dataset, enabling us to improve accuracy and enhance the predictive capabilities of our research.

5. Models



With the completion of our data preprocessing phase, we now possess a clean and structured dataset, primed for predictive modeling. Our approach involves utilizing machine learning models to forecast the desired output. As each new feature is extracted, it is seamlessly integrated into the evolving data frame, enriching our dataset with valuable information.

In this research, we have specifically opted to employ two machine learning models: Random Forest and XGBoost. These choices are motivated by their suitability for our research objectives.

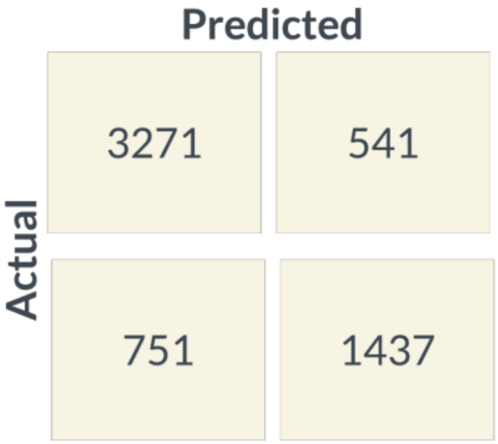
Random Forest is an ensemble learning method known for its robustness and ability to handle both categorical and numerical data effectively. It is capable of handling high-dimensional datasets, making it particularly well-suited for our feature-rich dataset. Moreover, its ensemble nature, which combines multiple decision trees, often results in improved accuracy and generalization.

XGBoost, short for Extreme Gradient Boosting, is another ensemble learning algorithm that has gained popularity for its exceptional performance in a variety of machine learning tasks. XGBoost excels in scenarios where feature engineering is critical, as it can handle complex relationships between features. Its efficient implementation and optimization for both classification and regression tasks make it a powerful choice for our research.

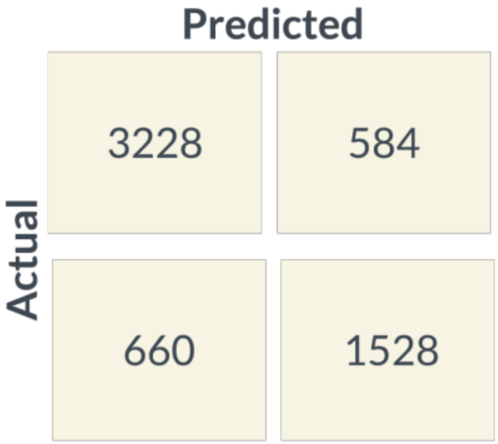
In conclusion, we have chosen Random Forest and XGBoost as our machine learning models due to their robustness, versatility, and ability to handle high-dimensional data effectively. These models are well-aligned with the complexities of our dataset and are expected to provide accurate and reliable predictions for our research objectives.

IV. Result And Discussion

In our initial experimentation with basic feature engineering for both Random Forest and XGBoost, we achieved an accuracy rate of 76 percent. However, recognizing the potential for improvement, we embarked on advanced feature engineering, which led to notable enhancements in predictive performance. Specifically, Random Forest achieved an accuracy of 78 percent, while XGBoost demonstrated an even higher accuracy of 79 percent. At first glance, this might suggest that XGBoost outperforms Random Forest in terms of accuracy.



	Predicted	
Actual	3271	541
	751	1437



	Predicted	
Actual	3228	584
	660	1528

Nonetheless, when evaluating the models beyond accuracy, we must consider the implications of a greater number of false positives. In the context of our problem, minimizing false positives is of paramount importance. When a non-duplicate question is erroneously marked as duplicate, it represents a significant drawback in terms of user experience, an outcome we aim to avoid. Therefore, despite its slightly lower accuracy, we argue that Random Forest may be a more suitable choice due to its potential to yield fewer false positives.

However, there is always room for improvement. In future research, strategies such as undersampling or oversampling could be explored to address the class imbalance within the dataset. Additionally, while we trained our models with a subset of 3000 data points due to computational limitations, augmenting the dataset with more examples could further enhance accuracy.

Furthermore, we recognize several avenues for potential enhancements. Implementing techniques like "Stemming" to reduce words to their root form and hyperparameter tuning to optimize model performance could yield substantial improvements. Expanding our scope beyond just Random Forest and XGBoost, we recommend exploring a broader range of machine learning algorithms, including Support Vector Machines (SVM) and Logistic Regression, to assess their effectiveness in this context.

Additionally, the creation of additional features could enhance model capabilities. Lastly, it is worth noting that while we employed the Bag of Words (BoW) technique for question vectorization, utilizing more advanced approaches like Word2Vec may yield even more accurate results. These considerations underscore the potential for ongoing research and refinement in pursuit of improved accuracy and model effectiveness.

V. Limitations

1. Data Size and Quality: The study relied on a relatively small dataset with potential data quality issues. A larger and cleaner dataset would enhance the reliability and generalizability of our findings.

2. Algorithm Selection: We limited our analysis to two machine learning algorithms due to resource constraints. Exploring a broader range of algorithms could uncover more effective approaches for the problem at hand.

VII. CONCLUSION

We extensively evaluated a wide array of machine learning models in our pursuit of addressing the challenge posed by duplicate questions within the Quora dataset. Surprisingly, our most effective model turned out to be a straightforward Continuous Bag of Words neural network. Our results underscore the potential of the Quora dataset as a valuable resource for advancing Natural Language Understanding tasks using machine learning methodologies.

REFERENCES

[1] Chen, Jieh-Haur, et al. "Smart project management: interactive platform using natural language processing technology." Applied Sciences 11.4 (2021).

[2]Erfani, Abdolmajid, Qingbin Cui, and Ian Cavanaugh. "An empirical analysis of risk similarity among major transportation projects using natural language processing." Journal of Construction Engineering and Management 147.12 (2021).

[3] Asakura, Kenta, et al. "A call to action on artificial intelligence and social work education: Lessons learned from a simulation project using natural language processing." Journal of Teaching in Social Work 40.5 (2020).

[4] da Silva Maldonado, Everton, Emad Shihab, and Nikolaos Tsantalis. "Using natural language processing to automatically detect self-admitted technical debt." IEEE Transactions on Software Engineering 43.11 (2017).

[5] Zhao, Xiaojing, Ker-Wei Yeoh, and David Kim Huat Chua. "Extracting construction knowledge from project schedules using natural language processing." The 10th International Conference on Engineering, Project, and Production Management. Springer Singapore, 2020.

[6] Mohammadhassanzadeh, Hossein, et al. "Using natural language processing to examine the uptake, content, and readability of media coverage of a pan-canadian drug safety research project: Cross-sectional observational study." JMIR formative research 4.1 (2020).

SI_CP_Paper (1).docx

ORIGINALITY REPORT

22%
SIMILARITY INDEX

PRIMARY SOURCES

1	arxiv.org Internet	283 words —	8%
2	medium.com Internet	112 words —	3%
3	Milind Rane, Sahil Kalal, Jenil Chandegara, Toshish Kakkad, Vishwas Jain, Simran Jagtap. "Career Prediction Website using Machine Learning", 2023 3rd International Conference on Intelligent Technologies (CONIT), 2023 Crossref	94 words —	3%
4	www.analyticsvidhya.com Internet	67 words —	2%
5	www.mdpi.com Internet	31 words —	1%
6	www.sersc.org Internet	29 words —	1%
7	www.tandfonline.com Internet	19 words —	1%
8	Abdolmajid Erfani, Qingbin Cui, Ian Cavanaugh. "An Empirical Analysis of Risk Similarity among Major Transportation Projects Using Natural Language	17 words —	< 1%

	Processing", Journal of Construction Engineering and Management, 2021 Crossref	
9	aclanthology.org Internet	12 words — < 1%
10	www.frontiersin.org Internet	11 words — < 1%
11	Gabriel Peres Nobre, Carlos H.G. Ferreira, Jussara M. Almeida. "A hierarchical network-oriented analysis of user participation in misinformation spread on WhatsApp", Information Processing & Management, 2022 Crossref	10 words — < 1%
12	Nadim, Mohammad. "Machine Learning for Empowering Community Applications and Security", The University of Texas at San Antonio, 2023 ProQuest	10 words — < 1%
13	naist.repo.nii.ac.jp Internet	10 words — < 1%
14	"Natural Language Processing – IJCNLP 2004", Springer Science and Business Media LLC, 2005 Crossref	9 words — < 1%
15	Mário André de Freitas Farias, Manoel Gomes de Mendonça Neto, Marcos Kalinowski, Rodrigo Oliveira Spínola. "Identifying self-admitted technical debt through code comment analysis with a contextualized vocabulary", Information and Software Technology, 2020 Crossref	8 words — < 1%
16	Everton da Silva Maldonado, Emad Shihab, Nikolaos Tsantalis. "Using Natural Language	7 words — < 1%

Processing to Automatically Detect Self-Admitted Technical Debt", IEEE Transactions on Software Engineering, 2017
Crossref

17 López, Néstor Nápoles. "Automatic Roman Numeral Analysis in Symbolic Music Representations.", McGill University (Canada), 2023
ProQuest

7 words — < 1%

18 Kenta Asakura, Katherine Occhiuto, Sarah Todd, Cedar Leithead, Robert Clapperton. "A Call to Action on Artificial Intelligence and Social Work Education: Lessons Learned from A Simulation Project Using Natural Language Processing", Journal of Teaching in Social Work, 2020
Crossref

6 words — < 1%

EXCLUDE QUOTES	OFF	EXCLUDE SOURCES	OFF
EXCLUDE BIBLIOGRAPHY	ON	EXCLUDE MATCHES	OFF