How to Cite:
Rani, K. P., Latha, P. V., & Rao, K. S. (2022). A comprehensive study based on societal question and answer system using natural language processing with topic modelling for job assistantship in the world industry. *International Journal of Health Sciences*, *6*(S5), 4101–4124.
https://doi.org/10.53730/ijhs.v6nS5.9523

A comprehensive study based on societal question and answer system using natural language processing with topic modelling for job assistantship in the world industry

K. Pushpa Rani
Research Scholar, CSE, Koneru Lakshmaiah Education Foundation, Vaddeswaram, AP, India
Email: rani536@gmail.com

Dr. P. Vidyu Latha
Professor, CSE, Koneru Lakshmaiah Education Foundation, Vaddeswaram, AP, India
Email: pvidyullatha@kluniversity.in

Dr. K. Srinivas Rao
Professor, Department of Computer Science and Engineering, MLR Institute of Technology, Hyderabad
Email: principal@mlrinstitutions.ac.in

***Abstract*---**This breakthrough study explores public opinion on technical words or topics using data from social media. Individuals seeking work assistance must consider public opinion on the existence of terminological or technical obstacles, as well as their impact on the environment and society. Public assistance is also required for the execution of legislation and mitigation programmes. A public opinion study is required to better understand the social environment and social processes. Researchers are paying special attention to social media data because it provides extremely significant information on public perceptions and responses to conflicting socio-technical terminology or issues, such as quorums, stack overflows, and yahoo! It also responds to Twitter and other social media platforms, and it's frequently used to monitor and analyze how society reacts to natural and social calamities. Search for keywords or topics in topic models to identify distinct topics to get the most out of social media data. In classic topic models, users may provide an erroneous number of subjects, resulting in unsatisfactory grouping results. For data retrieval and cluster trend identification in these scenarios, exact representations are essential. Unmarked and

incorrect texts or subjects are related to suitable ways to model themes utilizing NLP and NLTK models to tackle the existing challenge. As mentioned in the related works section on NLP and subject modelling, we can leverage your application with a list of references based on who is discussing in the public question and answer system to overflow the stack. This study discusses the country's public Q&A framework and examines the evolution of key issues and projects, with an emphasis on automatically disseminating relevant customer responses and determining what information items you require. The next generation of global empowerment technology has access to housing and work prospects. Finally, the results of the experiments suggest that employing theme models to process natural language improves accuracy in eliciting more acceptable responses for accommodations and interviews.

***Keywords*---**Stack-Overflow, Natural Language Processing, Machine Learning.

1 Introduction

The Annual Developer Survey by Stack Overflow is the major and most all-inclusive survey of coders in the world. Every year, we conduct a poll that covers anything from developer preferences to favourite technologies. We've announced the findings of our annual developer survey for the ninth year, and approximately 90K developers took the 20Min poll earlier this year. Despite the study's vast scope and capacity to yield useful conclusions, we acknowledge that our findings do not reflect everyone in the development community. Our research sample reflects the fact that we still have exertion to organize to make Stack Overflow the hospitable, inclusive, and diverse platform we want it to be. We intend to build on the progress we made in 2018 and continue to improve in this area this year and beyond. Some of the findings of these research are being used to direct these efforts. Check where we summarize results by nation or gender, highlight results for diminished ethnic/ethnic assemblies, or utilise survey weighting to compensate for demographic distortions to address the peculiarities of our data.

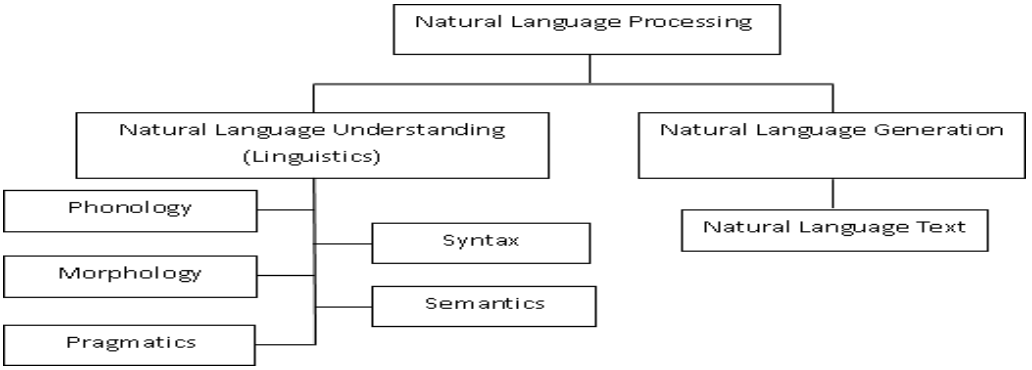


Fig: 1 The Broad Classification of NLP

Natural language processing appeared to make the user's job easier and to fulfill their need to express through the computer in natural language. Because not all employers are familiar with the machine's explicit language, NLP is available to individuals who prepare not have enough time to learn new languages or perfect them. A collection of rules or a set of characters can be used to define a language. To express information or transmit information, symbols are combined and employed. The Rules have complete control over the symbols Software engineers need developer forums to fix their difficulties with the support of professionals in such communities. However, occasionally the solutions (answers) to a problem are insufficient or pose a difficulty to the selection of a viable remedy. To choose the best answer, most people look through all of the responses in the question thread. Manually selecting the proper answers is a time-consuming and tiresome procedure. We offer an artificial categorization strategy for predicting correct answers in developer forums in this publication. We start by extracting the metadata and Q&A combo for each developer community post (Stack Overflow). The question-answer combinations from a specific data collection are then preprocessed using natural language processing techniques. Following that, for each question-and-answer combination, a keyword ranking algorithm is used to abstract phrases and their ranking outcomes. For each question/answer combination, a keyword-based feature vector is produced based on the keywords and their ranking outcomes. Word embedding is used to turn each pre-processed combination of questions and responses into a text-based feature vector. Finally, we use deep learning to predict the proper responses using metadata, keyword-based functions, and textual functions from the ensemble model. The findings of the 10-fold cross-validation reveal that the new method is more accurate and superior than the old one. By an average of 1.72 percent, 24.96 percent, 6.57 percent, and 16.62 percent, it improves accuracy, precision, recall, and f-measurement [1]. phases and level of NLP summarized in below.

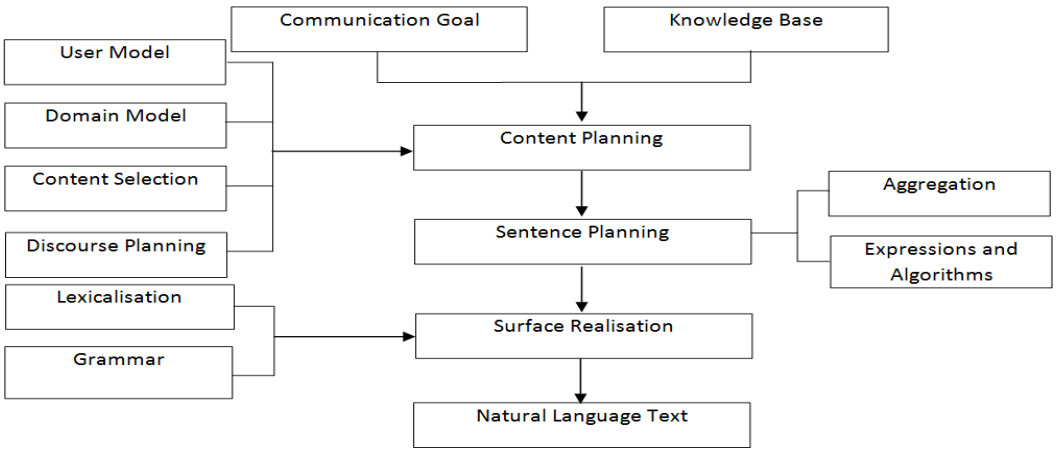


Fig: 2 The Phases of NLP Model

Linguistics is a science that studies the meaning of language, its environment, and its many forms. The following are some of the key terms in Natural Language Processing [2-5] Stack Overflow features a lot of duplicate questions that have already been asked and answered, unfluctuating if users are prompted to search a topic before situation a new question. To minimize the number of duplicate questions on Stack Overflow, authorized employees manually verify for duplicate questions, which earnings a lot of stretch and effort.

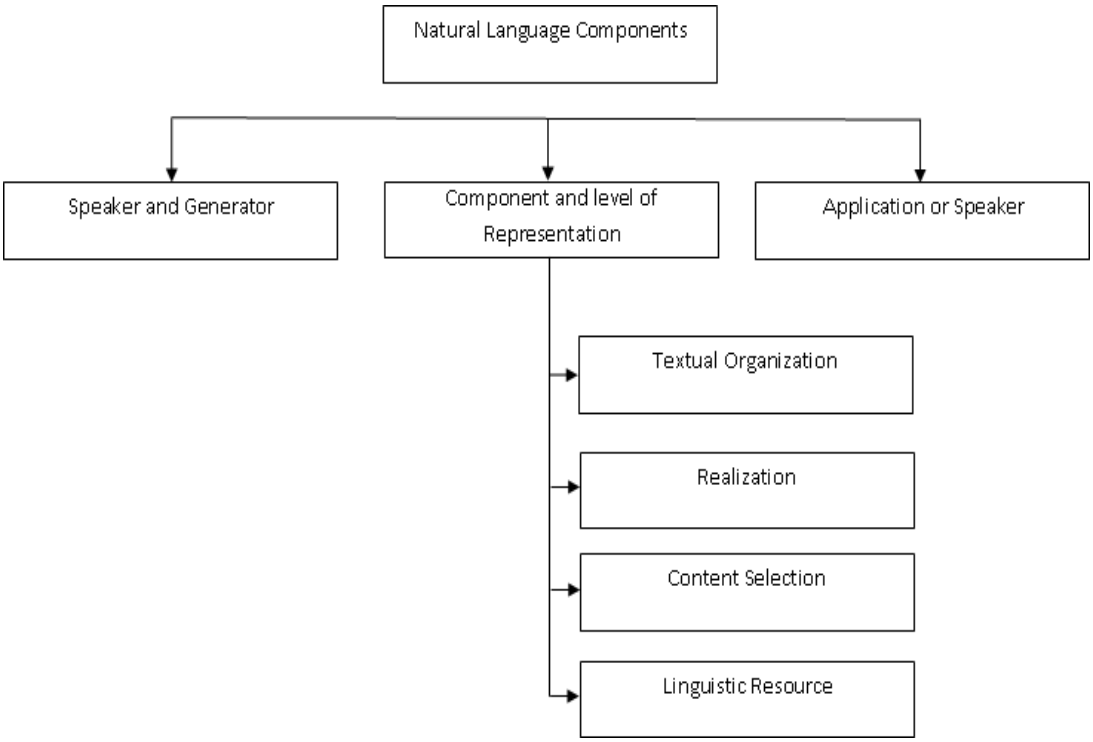


Fig: 3 The Components of NLG

On the Stack Overflow website, previous study looked examined certain approaches for automatically spotting duplicate concerns. The DupPredictor method for spotting identical questions on Stack Overflow, which considers the similarity characteristics of each pair of questions' themes, titles, descriptions, and tags [3]. Ahasanuzzaman et al. created the Dupe technique on Stack Overflow, which uses five routines to discover duplicate problems [2]. These five features include the cosine similarity value, term overlap, object overlap, object type overlap, and WordNet similarity characteristics. On Stack Overflow, to find duplicate issues [4], Silva et al. In NLP, responsibilities such as text categorization and mood analysis classical machine erudition techniques and deep learning approaches are being widely used. Deep learning approaches sometimes outperform traditional machine learning methodologies. However, several software engineering problems, such as detecting a code duplicate [7], detecting bug reports [8], predicting semantically relevant insights [9], and forecasting software defects, are solved using in-depth training. For some software engineering jobs, they have been found to be effective [10-11]. Deep

learning approaches outperform classical ML approaches in recognizing duplicate tasks and can fully capture semantics at the essay level, according to our findings.

2. Related Works

Amateur programmers and developers use question and answer websites like Stack Overflow for help during programming [12]. To assist Stack Overflow readers, researchers have proposed a range of proposals, including automatic tag suggestions and question templates for users asking questions [31], API abuse warnings [23], and API descriptions [28]. With the exception of pre-existing synonymous tag pairs offered by Stack Overflow Beyer et al. suggested a tool to provide synonymous tags for an input tag. Using a tag history of similar postings and user information, EnTagRec++ is proposed to suggest tags to be applied to queries asked [31]. FastTagRec classifies posts using neural networks and assigns tags to fresh Stack Overflow postings depending on the classification. To give an indication of the themes that are commonly covered, Beyer et al. proposed categorising the questions into seven groups based on the context in which they are asked. SOTagger was created to contextually mark concerns on Stack Overflow in three of the authors' six categories [?]. Using the LDA model and unconstrained ML techniques, Allam anis et al. classified stack overflow situations into five groups. Escudero et al. investigated the use of several question components on Stack Overflow and recommended which components should be included and which should be removed in order to publish high-quality posts. Using a logistic regression model trained on 46 different characteristics, Wang et al. presented a study aiming at understanding the elements that influence and inspire the fastest response to published questions. They also stressed the importance of bettering question-and-answer systems in order to motivate respondents [30]. Using semantic analysis, SOLinker is proposed to find associations between question tags and URLs and to mark questions correctly based on the URLs used in questions and responses [12]. Based on the user's search criteria, Example Overflow extracts samples of executable code entangled in the text. Example Check is a plugin for detecting API calls that are utilized incorrectly in stack overflow code snippets. It also shows how to utilise APIs correctly, including examples to help users learn how to use them properly [23]. It is proposed that StackDoc be used to enhance Stack Overflow with descriptions and examples of the Java APIs utilized in the inquiries [28]. Using Example Check [35], Zhang et al. identified APIs that are overused on Stack Overflow. When visitors visit a code example on Stack Overflow, Zhang et al. built Example Stack to display a list of similar code examples provided by GitHub and other Stack Overflow inquiries. They track this data to increase the trustworthiness of a stack overflow instance and to educate users about the various patterns of use and abuse of the instances [36].

The term "machine translation" didn't exist until the late 1940s, although work on it has begun. During this time, research is not totally localized. Although Russian and English are the most commonly utilized languages for TM, other languages, such as Chinese, have also been employed (Booth, 1967). According to an ALPAC report published in 1966, TM/NLP research was on the verge of extinction in 1966. Some MT manufacturing systems did, however, later offer

items to their clients (Hutchins, 1986) [11]. Work on using computers for literary and linguistic studies began around that time [14]. aimed to go beyond LUNAR by integrating big databases. Computational grammar theory, which is concerned with the logic of meaning and the ability of knowledge to cope with user beliefs and intents, as well as properties such as accents and stems, became a very active topic of study in the early 1980s.

Because it did not merely require data analysts, statistical language processing became popular in the 1990s (Manning and Schuetze, 1999) [24]. An technique was information retrieval and automatic aggregation (Mani and Maybury, 1999) [25]. Unsupervised and semi-supervised learning methods have been the subject of recent research. Many researchers have worked on NLP, developing tools and techniques that have helped to shape the field. NLP is an excellent field for research because of tools like mood analyzers, parts of speech (POS) markers, separation, named object recognition (NER), emotion detection, and semantic role labelling. The mood analyzer (Jeonghee et al., 2003) [26] extracts moods on a certain topic. Extracting topic-specific phrases, extracting moods, and associating them through relationship analysis constitute mood analysis. The Mood Lexicon and the Mood Model Database are used in mood analysis. Scrutinize the texts for positive and negative words and rate them on a scale of -5 to +5

Chunking: It operates by tagging sentence segments with syntactically relevant keywords, such as noun phrases and verb phrases, and is also known as Shadow Parsing (NP or VP). Each word has its own tag, which is usually referred to as the start-of-part (B-NP) or internal-part (I-NP) tag. The CoNLL 2000 shared task is frequently used to test part cutting. CoNLL 2000 gives fragmentation test data.

Because people do not utilise traditional or standard English, using name recognition in venues like the Internet is a difficulty. This has a big impact on how basic natural language processing techniques work. Annotating phrases or tweets, as well as developing solutions based on unmarked data both within and outside the domain (Alan Ritter., 2011) [33]. When compared to typical natural language processing technologies, it increases productivity. Similar to mood analysis, emotion detection (Shashank Sharma, 2016) works on social media platforms when two languages (English + any other Indian language) are blended. He divided the remarks into six categories based on their emotional content. They were able to determine the language of ambiguous terms, which are prevalent in Hindi and English, as well as indicate a lexical category or sections of speech in mixed script, indicating the speaker's primary language. SRL is a semantic role that is defined in a single sentence. The roles of words that are arguments of a verb in a sentence are assigned in the Prop Bank formalism (Palmer et al., 2005) . The exact arguments are determined by the verb frame, and if a sentence has numerous verbs, there may be multiple tags. Building a parse tree, recognizing which parse tree nodes represent a verb's arguments, and lastly sorting these nodes to compute the relevant SRL labels are all steps in most modern SRL systems. Using a graphical model to evaluate any social media stream to determine if it contains a person's name or a location name, time, etc. (Edward Benson et al., 2011), Edward Benson et al. Despite unsuitable noisy message noise and very terrible message grammar, this model was able to extract high-precision records by aggregating multiple information across several messages.

However, a wider range of factorial functions could be used to improve the results.

Machine Translation

Because the majority of the world is online, making data available and accessible to everyone is a difficult task. The language barrier is the most significant obstacle to data access. There are many languages, each with its own syntax and sentence structure. Computer translation uses a statistical machine-like Google Translate to translate phrases from one language to another. The difficulty with machine translation technology is not in translating words directly, but in preserving the sense of sentences, as well as grammar and tenses. Statistical machine learning gathers as much data as possible that appears to be parallel between two languages, then condenses it to determine the likelihood that something in language A matches something in language B. Google launched a new machine translation system based on artificial neural networks and deep learning in September 2016.

Text Categorization

Categorization systems organise enormous amounts of data, such as official papers, military casualty reports, market data, and news, into predetermined categories or indexes. An email filtering solution employs a collection of protocols to evaluate which incoming messages are spam and which aren't. Spam filters come in a variety of shapes and sizes. Content filters: Look at the message's content to see if it's spam or not. Header Filters Look for incorrect information in the email header. All emails from blacklisted recipients are blocked by common blacklist filters. User-defined criteria are used in rule-based filters. Stopping emails from a specific person or messages containing a specific word, for example. Permission filters: Require the recipient to approve anyone who sends a message. Challenge Response Filters Requires anyone sending an email to input a code before sending it.

Information Extraction

The purpose of information retrieval is to find interesting phrases in textual material. Extracting names, places, events, dates, times, and prices is a powerful approach to summarizing information that is important to user needs in many applications. Automatic identification of relevant information in the case of a domain-specific search engine can improve the accuracy and efficiency of targeted search. Hidden Mark Models (HMM) are used to extract the relevant study areas. These text segments are used to facilitate searching in specific fields, to efficiently display search results, and to compare document references. For example, beware of pop-up ads on any page showing current items from an online retailer at a bargain price. Two types of models have been used in information retrieval (McCallum and Nigam, 1998) [35]. A fixed vocabulary is required for both modules. However, in the first model, the document is created by selecting a subset of words from a dictionary and then repeating the process as many times as possible, at least once incorrectly. Bernoulli's multivariate

model is called that. Indicates which words are used in a document, regardless of the number or sequence in which they appear. The document is generated on the second model by selecting a collection of word matches and arranging them in random order. This model is known as a multinomial model because it collects information about how many times a word is used in a document in addition to Bernoulli's multivariate model. In recent years, knowledge discovery has been a popular research topic. Voice tagging (POS), parsing or shadow, stop words, keywords that are used and must be removed before the document is processed, interval (matching words with some basis for, there are two methods based on the dictionary and the initial Porter style (Porter, 1980) [35], the former has higher accuracy but higher implementation costs, while the latter has lower accuracy but lower implementation costs compound or statistical (index units for compound phrases and statistics with multiple characters , not individual tokens.) Put constraints on the meaning of words (Understanding the correct meaning of a word in context is the job of defining its meaning. In the document vector, terms are replaced by their senses.)

Dialogue System

Dialogue systems, which focus on tightly specified applications (such as refrigerators or home theatre systems), already incorporate phonetic levels and language lexicons, and are arguably the most desired application of the future in big end-user application providers' systems. These dialogue systems have the potential to be completely automated when all levels of language processing are applied. [27] (2001, Elizabeth D. Lydie). Either through text or speech. This could pave the way for technologies that allow robots to converse with humans using natural language.

The Duplicate Questions

Stack Overflow is one of the most well-known CQAs for software programming. Stack Overflow had over 18 million questions, 28 million answers, 76 million comments, and 56,000 bookmarks as of October 2019. Each question on Stack Overflow has a unique ID, title, body, tags, creation date, and closing date, among other things. n Members of Stack Overflow are recommended to search the site before starting a new topic, however duplicate queries do occur frequently. Questions that have already been asked and answered on Stack Overflow are considered duplicates. To reduce the number of duplicate queries, high-profile users are encouraged to manually mark duplicate questions on Stack Overflow.

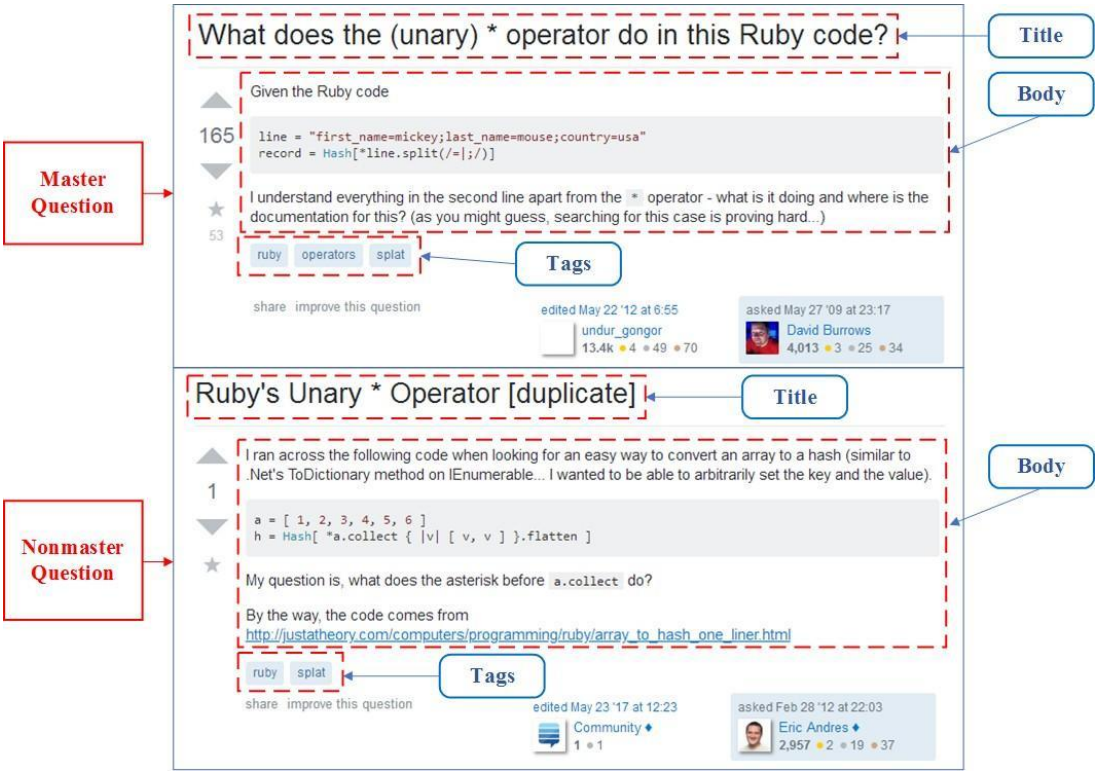


Fig: 4 question pairs with duplications

When two questions are asked at the same time, one of them is usually marked as a small question (duplicate question) and closed. The other issue will be designated as the primary issue. A duplicate question pair also consists of a main and non-primary question. In a duplicate question pair, the parent question is the one that was made first, while the non-parent question is the one that was created later. Figure 1 shows an example of a trustworthy user manually flagging a pair of duplicate questions. A main question with ID 918449 and a non-primary question with ID 9490456 make up the duplicate question pair. A non-primary question's title, body, and tags are all similar to the main questions. These two issues are linked to the answer to the "unary operator * Ruby" issue. A Stack Overflow user with a high reputation flags a minor issue as a duplicate question of the relevant larger issue [2]. Each duplicate problem has a [duplicate] in the title, and all duplicate issues are closed by trusted users. As a result, we create duplicate questions based on the word "[duplicate]" in each question title.

The Deep Learning

Deep learning is becoming more popular, and it's being applied in a variety of industries, including natural language processing and software engineering. Several deep learning models have been presented that use self-learning processing to uncover underlying patterns, basic dynamics, and semantic information in data. CNN [15], RNN [26], and LSTM [27], for example, are three prominent deep learning models. They're commonly employed to address natural

language processing problems including text classification, mood analysis [16], and other things. They're also utilized for software engineering jobs including code clone detection, mistake detection reporting [8], semantically linked knowledge prediction [19], and software defect prediction. CNN is a multilayer forwarding neural network with fully connected decision layers and one or more convolution and subsample layers. In sentence categorization and mood analysis tasks, CNN has greatly outperformed standard natural language processing algorithms. RNN is a useful design for sequential learning problems with data that is strongly correlated along a single axis. RNN has been effectively applied to a variety of applications, including language modelling and word embedding. Furthermore, LSTM can solve a significant number of jobs that earlier RNN training algorithms have failed to perform. Deep learning is increasingly being viewed as superior to traditional software engineering problem-solving methodologies. To answer the challenge of discovering duplicate issues on Stack Overflow, we investigate three deep learning algorithms based on the three common deep learning models mentioned above.

The Word Embedding

The Word Embedding4 refers to a set of language modelling techniques and procedures used in natural language processing to map words or phrases in a dictionary to real-number vectors The Neural Network Language Model (NNLM) [23], [24], Latent Semantic Analysis (LSA) [25], and the Latent Dirichlet Distribution (LDA) [26] are some of the earlier advancements in word embedding. Word2Vec has recently gained popularity. It can transform a word into a numeric vector and is used to learn how to embed words.

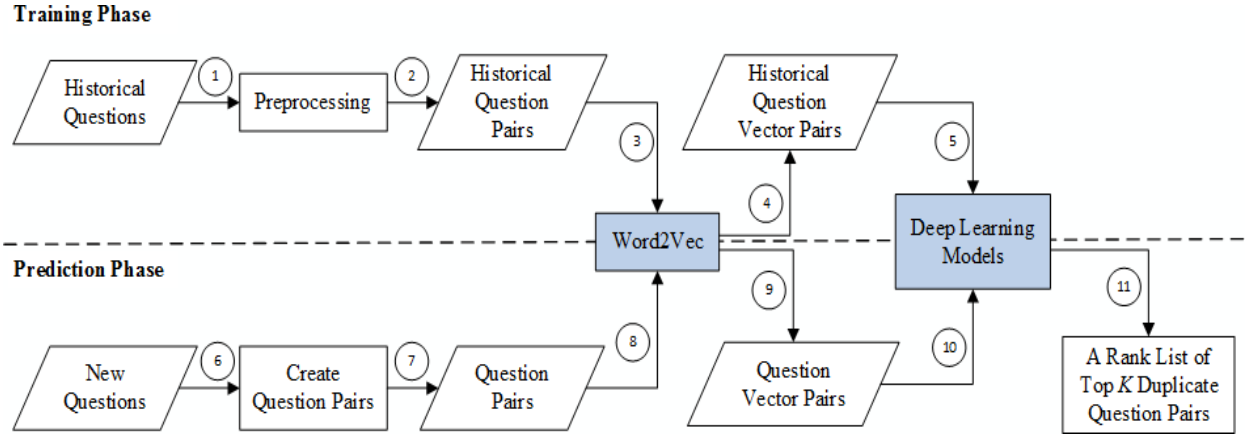


Fig 5. The Overall Framework of Our Approaches

Word2Vec, in particular, is capable of fully capturing word-level semantics. The Continuous Bag-of-Words (CBOW) and Skip-gram architectures are both included in Word2vec. Word vector representations are computed using the CBOW and Skip gram designs, which not only increase the quality of word vectors but also reduce computational complexity. 5 As a result, we employ Word2Vec to compute word vector representations and the CBOW architecture to learn high-quality text vectors in our research

The data collection

The Creative Commons Data Dump Service makes Stack Overflow data available to the public. 6 Stack Overflow initially gathered all edits made amongst August 2008 and September 2014. Second, we extract all minor editions that have been closed as duplicate editions and append the word "duplicate" to their titles. Third, we check the Linotyped on the post link table to determine whether these child concerns are related to their parent issues. We got 134,261 subsidiary questions and 88,476 core questions in the end. We derived the questions from six separate groups of questions labeled with Java, Html, Python, C, Ruby, and Objective-C, as in the prior work. Each question's textual content consists of the title, body, and tags. Furthermore, based on earlier work [4], pairs of questions based on the data sets of these six separate sets of questions were constructed as our experimental data sets.

Table: 1 Related works on NLP with Machine Learning Models

AUTHOR NAME	YEAR OF PUBLICATION	PAPER TITLE	PROPOSED MODEL	GAPS IDENTIFIED
B.J. Sowmya et al. [1]	2016	Large scale multi-label text classification of a hierarchical data set using Rocchio algorithm.	Rocchio Algorithm	Classification on hierarchical data
S. Bloehdorn et al. [2]	2004	Bloehdorn, S.; Hotho, A. Boosting for text classification with semantic features.	Boosting	AdaBoost for with semantic features
A. Genkin et al. [3]	2007	Bayesian logistic regression for text categorization. Technometrics	Logistic Regression	Logistic regression analysis of high-dimensional data
Kim, S.B et al. [4]	2006	Some effective techniques for naïve bayes text classification.	Naïve Bayes	Multivariate poisson model for text Classification
K. Chen et al. [5]	2016	Turning from TF-IDF to TF-IGM for term weighting in text classification.	SVM and KNN	Introduced TFIGM (term frequency & inverse gravity moment)
Z. Yang et al. [6]	2016	Hierarchical Attention Networks for Document Classification.	Deep Learning, Deep Belief Network	Hybrid text classification model based on deep belief network and SoftMax regression.
M. Jiang et al. [7]	2018	Text classification based on deep belief network and SoftMax regression.	Deep Belief Network	Hybrid text classification model based on deep belief network and softmax regression
X. Zhang et al. [8]	2015	Character-level convolutional networks for text classification.	CNN	Character-level convolutional networks (ConvNets) for text classification
K. Kowsari [9]	2018	Random Multimodel Deep Learning for Classification.	Ensemble deep learning Algorithm (CNN, DNN and RNN)	Solves the problem of finding the best deep learning structure and architecture
Verberne 2006 [10]	2006	Developing an approach for why-question answering.	Named Entity Recognition	Query formed by keywords may not be appropriate to retrieve documents
Per Holth 2013[11]	2013	Different sciences as answers to different why questions.	Proposed classification based on behavioural analysis of questions	Highlighted confusion in accepting varied explanations for similar kind of why-questions
Manvi [12]	2018	Analysis of Why-Type Questions for the Question Answering System.	Informational, Historical, Contextual/Situational, Opinionated	Classified on a small dataset. Ongoing refinements

AUTHOR NAME	YEAR OF PUBLICATION	PAPER TITLE	PROPOSED MODEL	GAPS IDENTIFIED
SARDAR [19]	2019	Deep Learning for Natural Language Parsing	BiLSTM parsing, deep learning, dependency parsing, natural language processing, parsers, shift-reduce parsing, syntactic parsing, transition-based parsing.	That our parsing architecture performs comparably to state-of-the-art parsers across a range of language families demonstrates that BiLSTM feature embeddings allow for effective multi-lingual parsing, setting our parser out from other state-of-the-art parsers which are often more specialized for certain language families.
HAFIZ UMAR IFTIKHAR [20]	2021	Deep Learning-Based Correct Answer Prediction for Developer Forums	keyword ranking algorithm	we pass the metadata, keywords-based features, and text-based features to the ensemble deep learning model for training to predict correct answers.
LITING WANG[21]	2020	Duplicate Question Detection With Deep Learning in Stack Overflow	Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN) and Long Short-Term Memory (LSTM)	to detect duplicate questions in Stack Overow. In addition, we use Word2Vec to obtain the vector representations of words. They can fully capture semantic information at document-level and word-level respectively.
ZEINAB SHAHBAZI[22]	2021	Fake Media Detection Based on Natural Language Processing and Blockchain Approaches	Natural language processing, blockchain, fake media, reinforcement learning.	To improve this platform in terms of security, the decentralized lockchain framework applied,which provides the outline of digital contents authority proof.
Javier Jorge[23]	2022	Live Streaming Speech Recognition Using Deep Bidirectional LSTM Acoustic Models and Interpolated Language Models	Automatic speech recognition, streaming, decoding, acoustic modeling, language modeling, neural networks	In this work an improved decoder based on the conventional hybrid ASR approach was proposed by adapting state-of-the-art models to the streaming setup. In particular, deep BLSTM acoustic models were adapted to the streaming conditions by using a sliding window of future context

AUTHOR NAME	YEAR OF PUBLICATION	PAPER TITLE	PROPOSED MODEL	GAPS IDENTIFIED
JINGXUAN ZHANG[25]	2018	Recommending APIs for API Related Questions in Stack Overflow	Application programming interfaces, information retrieval, recommendation system, stack over ow.	API speci_cations and historical resolved questions to detect correct APIs for new API related questions.
DARSHINI MAHENDRAN[26]		Review: Privacy-Preservation in the Context of Natural Language Processing	Data privacy, natural language processing, privacy preservation, privacy policy	we studied methods that prevent an adversary from listening to the latent representation in the middle and obtaining sensitive information Finally,we provide a tabular summary of related work and discuss future directions to help guide a path ahead.
HAO WU[27]	2021	SCADA-NLI: A Natural Language Query and Control Interface for Distributed Systems	Natural language interface, intent classi_cation, semantic parsing, scada systems, human computer interaction.	solve the problem of human-computer natural language interaction in the SCADA system, and the accuracy of intent recognition
SUSHANT SINGH[28]	2021	The NLP Cookbook: Modern Recipes for Transformer Based Deep Learning Architectures	Deep learning, natural language processing (NLP), natural language understanding (NLU), natural language generation (NLG), information retrieval (IR), knowledge distillation (KD), pruning, quantization.	we summarize and examine the current state-of-the-art (SOTA) NLP models that have been employed for numerous NLP tasks for optimal performance and ef ciency.We provide a detailed understanding and functioning of the different architectures, a taxonomy of NLP designs, comparative evaluations, and future directions in NLP.
ARSHAD AHMAD[29]	2019	Toward Empirically Investigating Non-Functional Requirements of iOS Developers on Stack Overflow	Non-functional requirements (NFRs), quality requirements, iOS, Latent Dirichlet allocation (LDA), Stack Overow.	Our Findings revealed that the highly frequent topics the iOS developers discussed are related to usability, reliability, and functionality followed by efficiency. Interestingly, the most problematic areas unresolved are also usability, reliability,and functionality though followed by portability.
QUANYI HU[30]	2020	Towards a Context-Free Machine Universal Grammar (CF-MUG) in Natural Language Processing	Semantic document exchange, natural language processing, universal grammar.	a novel Machine Universal Grammar provides a universal grammar that accepts all coming languages and improves semantic accuracy in natural language processing

3. Results and Discussions

A lot of research projects have looked into various aspects of Stack Overflow. Stack Overflow, by Ahmed et al. [15], is the first wide-ranging survey of the excavation literature. They divided the numerous Stack Overflow works on software development into two categories: OS Design and Usage and OS Applications, in a quick summary. Treude et al. [11] used quality coding to classify OS concerns and thoroughly analyzed OS's involvement in software development. Yang et al. [37] conducted a study that focused specifically on OS security vulnerabilities. They employed a genetic algorithm to categorize distinct security vulnerabilities based on their texts using an LDA-based model. The findings found that OS security challenges include a wide range of areas, including Web security, mobile security, encryption, software security, and system security. Pinto and Camei [38] looked through SO papers to see how they were related to refactoring tools. To determine the level of developer activity and popular subjects mentioned, Fontao et al. [39] examined 1,568,377 OS technical issues linked to Android, iOS, and Windows Phone platforms. Rather, we want to highlight the most important concerns and NFRs that are only covered in iOS postings. The detected NFRs are then compared to the ISO9126 quality model, which is missing from their work.

Selection Process for Stack Overflow Stack Overflow Stack Overflow is a prominent CQA website for software developers. We used Stack Overflow queries as our experimental data. We extract 134,261 sub questions and 88,476 main questions from Stack Overflow between August 2008 and September 2014 based on the data collected in Section II. We gathered questions from six different question groups based on these questions: Java, HTML, Python, C++, and Ruby. Table 1 shows data sets from six different question sets. For each question set, 80 percent of the duplicate question pairs were utilized as training data, while the remaining 20% were used as test data. as a source of test data Similarly, during training, 80 percent of unduplicated question pairs are employed. In addition, in our trials, all characteristics of the three deep learning techniques are set to their defaulting levels. The key parameters of CNN, for example, are all in channels and kernel size, which have values of 1 and 5, respectively. The basic parameters of RNN and LSTM are hidden state and num layers, which have values of 64 and 2, respectively andhas been the most popular forum for identifying programming difficulties for the past decade. Many people contributed to it, and many people read it. It frequently appears among the first search results for programming-related inquiries. Stack Overflow has a voting mechanism where users can give high-quality comments by adding positive voices and topical categories [6], such as Python and machine learning, to ensure that good information is visible. In this method, a reputation assessment is established, which identifies individuals with greater knowledge in specific areas and grants high-ranking users preferred privileges, such as community voting, editing, and moderator [6]. Only well-documented inquiries on the topic and accurately marked questions/answers are approved by community moderation. These capabilities assure the platform's content quality and provide a rich resource for social analytics. Only the super user who asked the question/request can accept the response on Stack Overflow. Consider the circumstance when a superuser and other users agreed with the answer and upvoted it. Later, the superuser posts a more dependable and

effective solution than the one that was received and voted on. The superuser should replace the most accurate answer in this scenario. In such circumstances, however, the most accurate solution is frequently disregarded. To avoid the human process of creating accurate answers through votes, badges, and reputation, Stack Overflow demands an automated way of determining the correct answer.

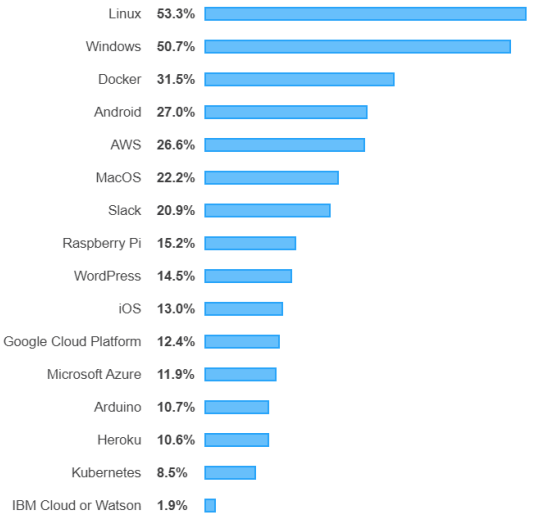
The voting and badge systems are used by Stack Overflow's Q&A system; however, Roy et al. [7] contend that the voting system is not a good means to check the quality of the response; occasionally a better answer, posted later, does not receive many upvotes and is placed at the end. They classified the data using an open-source stack exchange dataset and divided it into three categories based on voting. Answers with fewer than one vote were considered low-quality, those with one to two votes were considered medium-quality, and answers with more than two votes were considered high-quality. Miswords, code snippets, user reputation, readability, activity/topic entropy, subject reputation, question and answer similarity, and answer and answer similarity are among the 26 functions they derive. For classification, they trained three naïve Bayesian classifiers, random forest, and gradient gain, with the results showing that gradient gain performs best.

Consumer reputation is another essential element that improves the likelihood of an answer getting accepted on Stack Overflow [8], [9], and [10]. Bosu et al. [10] demonstrate how a user can quickly establish a positive rating on Stack Overflow. Answering less experienced inquiries, density tags, quick replies to questions, being the first official in control, becoming an active participant during peak times, and participating in various areas are only some of the findings. Zheng and Li [11] used AdaBoost to predict the reasonable responses to the stack overflow dataset, taking into account three factors (text, code snippets, and response history). Their classifier model obtains 63 percent accuracy, 59 percent recall, and 61 percent f-measures, respectively. Ponzanelli et al. [12] suggested an automated strategy that detects low quality stack overflow posts to refine the review queue in order to analyze the quality of the inquiries. They analyzed many data sets totaling nearly 5 million questions and separated them into two groups depending on the findings. The high-quality content class contains questions with a score greater than zero, while the low-quality content class comprises all other questions. His strategy is primarily based on two aspects: text-based features that include post content and community-based factors that also include user popularity.

In comparison to the fundamental techniques, the findings demonstrated that the naïve multiple-name Bayes achieves good accuracy. Ponzanelli and Xia worked together to anticipate quality issues. Stack Overflow has thousands of questions that have yet to be answered by users. Treude et al. [14] studied 15 days of Stack Overflow data and categorized problem kinds to determine the cause. They looked at the top 200 tagged terms in the data and discovered that they covered anything from 60 to 193 tags. They discovered that such markers are present in the vast majority of instances. Calefato et al. [15] looked into how users might improve their odds of getting their Stack Overflow replies accepted. To do so, they discovered four parameters that have a significant impact on response success: 1)

the presentation (URL, abbreviated code, length, capitalization); 2) the emotional reaction (good or negative); 3) the published response time; and 4) the reputation of the questions being asked. They created a dataset using the official Stack Overflow dump, which was published 30 days ago. There were 348,618 answers in all. They used regression analysis and got a 64% accuracy rate.

Platforms Oriented Terms with priority:

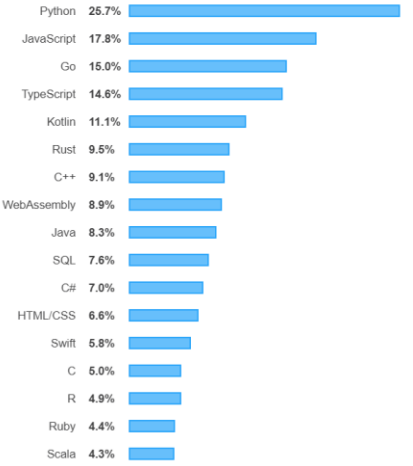


The most prevalent platforms on which our respondents have worked on development this year are Linux and Windows. Docker was the third most popular platform when we first asked about current uses this year.

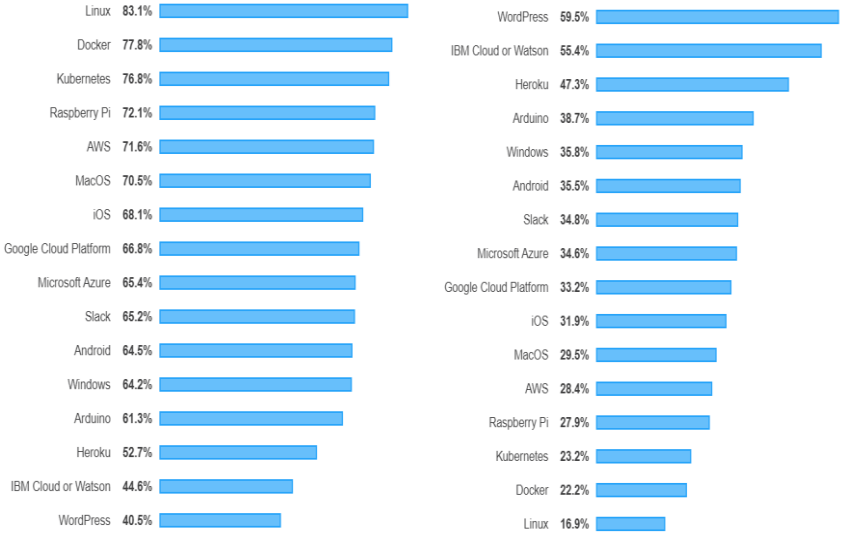
Dataset

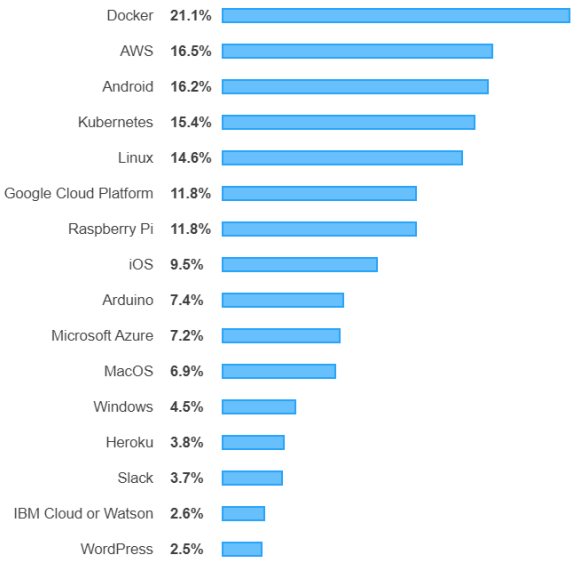
To test the proposed methodology, we used the open-source Stack Overflow question and answer set. We are retrieving the September 2019 Stack Overflow dataset. 2. There are 17 million multi-label questions and 26 million replies in the collection. The original data is skewed toward incorrect answers, with just 10% of the answers being labeled as right. As a result, we apply a five-marker filter (java,.net, php, ruby, and misc) to reduce the data set, which now contains 236,000 answers and 91,500 question threads, with 70,800 (30%) and 165,200 (70%) correct and incorrect answers, respectively.

The WANTED LANGUAGES with priority



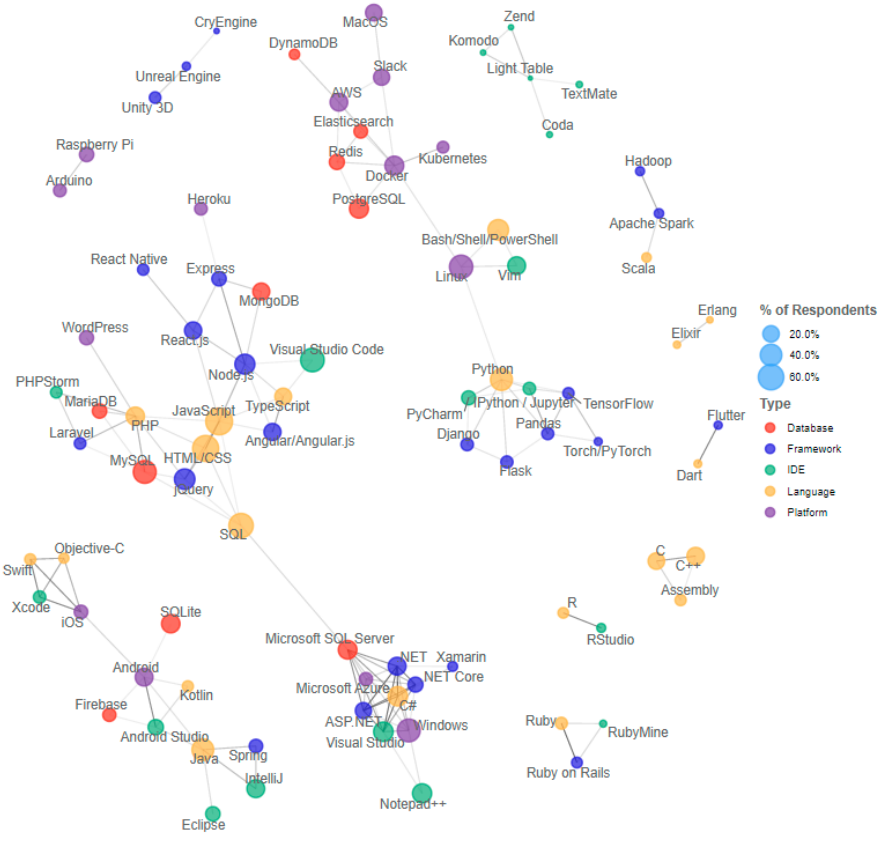
For the fourth consecutive, Rust seems to be the most popular programming language among our respondents, closely tracked by Python, the quickest foremost language today. As an outcome, more developers opt to stick with these languages rather than switching to another. VBA and Objective-C are the most dreaded languages this year. The most dreaded result is that the vast majority of developers who are currently using these technologies have stated that they have no desire to do so in the future. For the third year in a row, Python is the most desired language, with developers who haven't used it expressing an interest in learning it.





The Rust is the most widely used programming language among our defendants for the fourth year in a row, closely followed by Python, the fastest-growing major language today. As a result, rather than transferring to another language, more developers choose to continue with these. This year's most feared languages are VBA and Objective-C. The most dreaded outcome is that the vast majority of developers who are currently utilizing various technologies have shown no desire to do so in the future. Python is the most desired language for the third year in a row, with developers who have never used it showing an interest in learning it.

Consolidated Correlated Technologies:



Ecosystems are made up of technologies that are commonly employed by many of the same developers. This is shown in this network graph, which shows which technologies are by far the most strongly related. This study was grounded on a survey of 88,883 software developers after 179 countries. This is the number of deliberate "qualified" for diagnostic purposes founded on the length of time spent on the entire, completed survey; 400 additional responses were received but aren't included in the research since responders spent less than three minutes on the survey.

Location	Amount
Europe	36,073
North America	25,526
Asia	18,273
South America	3,459
Africa	2,850
Australia/Oceania	2,434
Other (country not listed)	268

The examination was retrieved from January 23 to May 5 ,2022. The overall time spent here on survey for eligible responses was 23.3 minutes. The majority of respondents were found using Stack Overflow-owned channels. The top five foundations of defendants were site messages, blog posts, email lists, Meta posts, banner ads, and social networking posts. Highly active Stack Overflow users are likely to see the encourage and click to start it since responses were obtained this way. As an incentive, responders who took the questionnaire were offered the option of getting a "Census" badge.For qualifying responses, the median time spent on the survey was 23.3 minutes. Respondents were mostly recruited using Stack Overflow-owned channels. Site message, blog posts, email lists, Meta posts, banner ads, and social networking posts were the top five sources of respondents. Because respondents were gathered this way, highly engaged Stack Overflow users were more likely to see the survey links and click to start it. Respondents who completed the survey were given the option of receiving a "Census" badge as an incentive [32].

Apparently, the survey wasn't really open to everyone this year. Our third-party survey software blocked traffic from Crimea, Cuba, Iran, North Korea, and Syria due to recent US transport/export sanctions, although some respondents utilized VPNs to get past it. Keep this unanticipated limitation of our study in mind when evaluating poll results. Professional developers are estimated based on what individuals read and do on Stack Overflow [33]. We gather data about user behavior to assist us in identifying jobs and questions that we estimate you can answer. At any moment, you can download and delete this information.

We inquired about response compensation. We began by asking each respondent about their favorite currency. Then we asked whether the respondent was paid weekly, monthly, or annually in that currency [34-35]. We converted salaries from user currencies to US dollars using the February 1, 2019 exchange rate, and then to yearly salaries based on 12 months and 50 weeks of labor. This was an optional question, as was the rest of the survey. We got data on wages from 55,945 persons (62.9 percent of qualifying responses). In the United States and overseas, the top 2% of salaries were slashed and replaced with threshold values. The threshold numbers were different inside and outside the United States.

Only a few questions were shown to respondents based on their past responses. Individuals who responded that they were employed were only presented questions about jobs and work. The questions were broken down into several blocks, each of which was allocated at random. The order of the answers to the majority of the questions was also randomized. We picked the technologies to include in this year's poll based on data from past years' surveys and Stack Overflow tag patterns. This year, we assessed which smaller or shrinking technologies we could eliminate, and we prioritized popular and fast-growing technology.

4. Conclusion

This study was based on a survey of 89,989 software developers from 279 countries. This is the number of responses we deem "qualified" for analytics purposes based on the time spent on the full and complete survey. This post lays out a comprehensive deep learning technique for predicting correct developer forum responses. We collect data from the developer community (Stack Overflow) and clean it up using natural language processing algorithms. After then, the Keyword Retrieval Tool is utilized to find and rank keywords. The textual information is then converted via word embedding. We train the suggested set-based classifier for deep learning using the vectors (metadata, keywords, and text). The results of the 10-fold cross-validation test reveal that the proposed method is accurate and superior to the previous one. Our research has a broader benefit in that it demonstrates that the question/response combination can be a valuable resource for predicting the correct answer. Our findings support further investigation into predicting the correct response. The results of the proposed ensemble model and other deep learning models are not yet comparable.

Question and answer websites like Stack Overflow are becoming increasingly popular in the programming sector as open-source knowledge sharing platforms gain favor. Many inexperienced coders use Stack Overflow to ask questions and find answers to challenges they encounter during the coding process. Professionals volunteer to answer questions on Stack Overflow based on their prior expertise or knowledge. The majority of these responses were accompanied by Stack Overflow user comments. Consumer feedback is included in Stack Overflow's questions, answers, and comments, which, when analyzed and displayed, may encourage users to read and respond to articles. However, the current Stack Overflow platform does not reflect the tone of these discussions. On social networking sites like Twitter, a lot of research has been done on interpreting emotions. Users are more likely to follow or comment to a post if the mood of the message is displayed. Despite the fact that there are several programmes that augment or annotate the Stack Overflow platform for developers and, in future enhancements, add new machine learning techniques and frameworks for generating good results, we are not aware of any solutions that deal with post mood. In addition, topic models will be introduced in the Hadoop ecosystem for storing large amounts of data, as well as AWS for automating work with high security using machine learning paradigms.

References

1. Sowmya, B.; Srinivasa, K. Large scale multi-label text classification of a hierarchical data set using Rocchio algorithm. In Proceedings of the 2016 International Conference on Computation System and Information Technology for Sustainable Solutions (CSITSS), Bangalore, India, 6–8 October 2016; pp. 291–296.
2. Bloehdorn, S.; Hotho, A. Boosting for text classification with semantic features. In International Workshop on Knowledge Discovery on the Web; Springer: Berlin/Heidelberg, Germany, 2004; pp. 149–166.
3. Genkin, A.; Lewis, D.D.; Madigan, D. Large-scale Bayesian logistic regression for text categorization.
   Technometrics 2007, 49, 291–304. [CrossRef]
4. Kim, S.B.; Han, K.S.; Rim, H.C.; Myaeng, S.H. Some effective techniques for naive bayes text classification.IEEE Trans. Knowl. Data Eng. 2006, 18, 1457–1466.
5. Chen, K.; Zhang, Z.; Long, J.; Zhang, H. Turning from TF-IDF to TF-IGM for term weighting in text
   classification. Expert Syst. Appl. 2016, 66, 245–260. [CrossRef]
6. Yang, Z.; Yang, D.; Dyer, C.; He, X.; Smola, A.J.; Hovy, E.H. Hierarchical Attention Networks for Document Classification. In Proceedings of the HLT-NAACL, San Diego, CA, USA, 12–17 June 2016; pp. 1480–1489.
7. Jiang, M.; Liang, Y.; Feng, X.; Fan, X.; Pei, Z.; Xue, Y.; Guan, R. Text classification based on deep belief network and softmax regression. Neural Comput. Appl. 2018, 29, 61–70. [CrossRef]
8. Zhang, X.; Zhao, J.; LeCun, Y. Character-level convolutional networks for text classification. Adv. Neural Inf.Process. Syst. 2015, 28, 649–657.
9. Kowsari, K.; Heidarysafa, M.; Brown, D.E.; Jafari Meimandi, K.; Barnes, L.E. RMDL: Random Multimodel Deep Learning for Classification. In Proceedings of the 2018 International Conference on Information System
   and Data Mining, Lakeland, FL, USA, 9–11 April 2018; doi:10.1145/3206098.3206111.
10. Verberne S., 2006. Developing an approach for why-question answering. In Conference Companion of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2006), Trento, Italy, pp. 39-46
11. Holth, P., 2013. Different sciences as answers to different why questions. European Journal of Behavior Analysis, 14(1), pp.165-170.
12. Breja, M., & Jain, S. K. (2018, September). Analysis of Why-Type Questions for theQuestion Answering System.In European Conference on Advances in Databases and In-formation Systems (pp. 265-273). Springer, Cham.
13. Harvey M., Hauff C., Elsweiler D.; 2015. Learning by Example: Training Users with High-quality Query Suggestions. ACM.
14. Xiao-Lin Zheng, Senior Member, Ieee, Chao-Chao Chen,
   Jui-Long Hung, Wu He, Fu-Xing Hong, And Zhen Lin, A Hybrid Trust-Based Recommender System For Online Communities Of Practice, Ieee Transactions On Learning Technologies, Vol. 8, No. 4, October-December 2015
15. ZHENG HU 1, JIAO LUO 1, CHUNHONG ZHANG 2, AND WEI LI 3, A Natural Language Process-Based Framework for Automatic Association Word Extraction, Digital Object Identifier 10.1109/ACCESS.2019.2962154

16. Andrea Galassi , Marco Lippi , and Paolo Torroni, Attention in Natural Language Processing, IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, VOL. 32, NO. 10, OCTOBER 2021
17. ANDREW HEPPNER, ATISH PAWAR , DANIEL KIVI, AND VIJAY MAGO, Automating Articulation: Applying Natural Language Processing to Post-Secondary Credit Transfer, Digital Object Identifier 10.1109/ACCESS.2019.2910145
18. Nicollas R. de Oliveira 1,† , Pedro S. Pisa 2,† , Martin Andreoni Lopez 3,† , Dianne Scherly V. de Medeiros 1,† and Diogo M. F. Mattos 1,*,†, Identifying Fake News on Social Networks Based on Natural Language Processing: Trends and Challenges, Information 2021, 12, 38. https://doi.org/10.3390/info12010038
19. SARDAR JAF 1 AND CALUM CALDER2, Deep Learning for Natural Language Parsing, Digital Object Identifier 10.1109/IEEE ACCESS.2019.2939687
20. Hafiz Umar Iftikhar 1, Aqeel Ur Rehman 2, Olga A. Kalugina3,QasimUmer 2, And Haris Ali Khan2, Deep Learning-Based Correct Answer Prediction For Developer Forums, Digital Object Identifier 10.1109/Ieee Access.2021.3108416
21. LITING WANG , LI ZHANG , AND JING JIANG, Duplicate Question Detection With Deep
    Learning in Stack Overflow, Received January 8, 2020, accepted January 16, 2020, date of publication January 21, 2020, date of current version February 11, 2020.Digital Object Identifier 10.1109/ACCESS.2020.2968391
22. ZEINAB SHAHBAZI AND YUNG-CHEOL BYUN, Fake Media Detection Based on Natural Language Processing and Blockchain Approaches, Received September 6, 2021, accepted September 9, 2021, date of publication September 14, 2021,
    date of current version September 23, 2021.Digital Object Identifier 10.1109/ACCESS.2021.3112607
23. Javier Jorge ,AdriàGiménez , Joan Albert Silvestre-Cerdà , Jorge Civera , Albert Sanchis ,
    and Alfons Juan, Live Streaming Speech Recognition Using Deep Bidirectional LSTM Acoustic Models and Interpolated Language Models, IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, VOL. 30, 2022
24. Marco Lippi , Marcelo A. Montemurro , Mirko DegliEsposti, and Giampaolo Cristadoro, Natural Language Statistical Features of LSTM-Generated Texts, IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, VOL. 30, NO. 11, NOVEMBER 2019
25. JINGXUAN ZHANG 1, (Student Member, IEEE), HE JIANG1, 2, 3, (Member, IEEE),ZHILEI REN1, AND XIN CHEN1, Recommending APIs for API Related Questions in Stack Overflow, Received September 7, 2017, accepted November 7, 2017, date of publication November 28, 2017, date of current version March 9, 2018.Digital Object Identifier 10.1109/ACCESS.2017.2777845
26. DARSHINI MAHENDRAN , CHANGQING LUO , (Member, IEEE), AND BRIDGET T. MCINNES, Review: Privacy-Preservation in the Context of Natural Language Processing, Received October 6, 2021, accepted October 24, 2021, date of publication October 28, 2021, date of current version November 8, 2021.Digital Object Identifier 10.1109/ACCESS.2021.3124163

27. HAO WU 1, CHUNSHAN SHEN 1, ZHUANGZHUANG HE 1, (Student Member, IEEE),
YONGMEI WANG 1, AND XINYUAN XU 2, SCADA-NLI: A Natural Language Query and
Control Interface for Distributed Systems, Received May 3, 2021, accepted May 19, 2021, date of publication May 25, 2021, date of current version June 3, 2021.Digital Object Identifier 10.1109/ACCESS.2021.3083540
28. SUSHANT SINGH, (Member, IEEE), AND AUSIF MAHMOOD , (Member, IEEE), The NLP Cookbook: Modern Recipes for Transformer Based Deep Learning Architectures, Received April 21, 2021, accepted April 30, 2021, date of publication May 4, 2021, date of current version May 14, 2021.
Digital Object Identifier 10.1109/ACCESS.2021.3077350
29. ARSHAD AHMAD 1,2, CHONG FENG 1, KAN LI 1, SYED MOHAMMAD ASIM3,AND TINGTING SUN1, Toward Empirically Investigating Non-Functional Requirements of iOS Developers on Stack Overflow, Received April 6, 2019, accepted April 28, 2019, date of publication May 2, 2019, date of current version May 22, 2019.Digital Object Identifier 10.1109/ACCESS.2019.2914429
30. QUANYI HU 1, JIE YANG1,2, PENG QIN1, (Graduate Student Member, IEEE),AND SIMON FONG 1, (Member, IEEE), Towards a Context-Free Machine Universal Grammar (CF-MUG) in Natural Language Processing, Received July 12, 2020, accepted August 31, 2020, date of publication September 8, 2020, date of current version September 22, 2020.Digital Object Identifier 10.1109/ACCESS.2020.3022674
31. Suryasa, I.W., Sudipa, I.N., Puspani, I.A.M., Netra, I.M. (2019). Translation procedure of happy emotion of english into indonesian in kṛṣṇa text. Journal of Language Teaching and Research, 10(4), 738–746
32. Liting Wang, Li Zhang, Jing Jiang. "Duplicate Question Detection With Deep Learning in Stack Overflow" , IEEE Access, 2020
33. Hafiz Umar Iftikhar, Aqeel Ur Rehman, Olga A. Kalugina, QasimUmer, Haris Ali Khan. "Deep Learning Based Correct Answer Prediction for Developer Forums" , IEEE Access, 2021
34. Liting Wang, Li Zhang, Jing Jiang. "Detecting Duplicate Questions in Stack Overflow via Deep Learning Approaches" , 2019 26th AsiaPacific Software Engineering Conference (APSEC), 2019
35. Mukhtar, A. U. S., Budu, B., Sanusi B, Y., Mappawere, N. A., & Azniah, A. (2022). The effect of reproductive health education with multimedia video learning on the improvement of fluor albus prevention behavior young woman pathologist. International Journal of Health & Medical Sciences, 5(1), 75-79. https://doi.org/10.21744/ijhms.v5n1.1841
36. Jiayan Pei, Yimin Wu, Zishan Qin, Yao Cong, Jingtao Guan. "Attention-based model for predicting question relatedness on Stack Overflow" , 2021 IEEE/ACM 18th International Conference on Mining Software Repositories (MSR), 2021
37. Zhifang Liao, Wenlong Li, Yan Zhang, Song Yu. "Detecting Duplicate Questions in Stack Overflow via Semantic and Relevance Approaches" , 2021 28th Asia-Pacific Software Engineering Conference (APSEC), 2021