What issues are data scientists talking about? Identification of current data science issues using semantic content analysis of Q&A communities

Fatih Gurcan [Corresp. 1]

[1] Department of Management Information Systems, Faculty of Economics and Administrative Sciences, Karadeniz Technical University, Trabzon, Turkey

Corresponding Author: Fatih Gurcan
Email address: fgurcan@ktu.edu.tr

Background. Because of the growing involvement of communities from various disciplines, data science is constantly evolving and gaining popularity. The growing interest in data science-based services and applications presents numerous challenges for their development. Therefore, data scientists frequently turn to various forums, particularly domain-specific Q&A websites, to solve difficulties. These websites evolve into data science knowledge repositories over time. Analysis of such repositories can provide valuable insights into the applications, topics, trends, and challenges of data science.

Methods. In this article, we investigated what data scientists are asking by analyzing all posts to date on DSSE, a data science-focused Q&A website. To discover main topics embedded in data science discussions, we used latent Dirichlet allocation (LDA), a probabilistic approach for topic modeling.

Results. As a result of this analysis, 18 main topics were identified that demonstrate the current interests and issues in data science. We then examined the topics' popularity and difficulty. In addition, we identified the most commonly used tasks, techniques, and tools in data science. As a result, "Model Training", "Machine Learning", and "Neural Networks" emerged as the most prominent topics. Also, "Data Manipulation", "Coding Errors", and "Tools" were identified as the most viewed (most popular) topics. On the other hand, the most difficult topics were identified as "Time Series", "Computer Vision", and "Recommendation Systems". Our findings have significant implications for many data science stakeholders who are striving to advance data-driven architectures, concepts, tools, and techniques.

1 # What issues are data scientists talking about?
2 # Identification of current data science issues using
3 # semantic content analysis of Q&A communities

6 Fatih Gurcan[1]

8 [1] Department of Management Information Systems, Faculty of Economics and Administrative
9 Sciences, Karadeniz Technical University, 61080, Trabzon, TURKEY

11 Corresponding Author:
12 Fatih Gurcan[1]
13 Department of Management Information Systems, Faculty of Economics and Administrative
14 Sciences, Karadeniz Technical University, 61080, Trabzon, TURKEY
15 Email address: fgurcan@ktu.edu.tr

17 ## Abstract
18 **Background.** Because of the growing involvement of communities from various disciplines,
19 data science is constantly evolving and gaining popularity. The growing interest in data science-
20 based services and applications presents numerous challenges for their development. Therefore,
21 data scientists frequently turn to various forums, particularly domain-specific Q&A websites, to
22 solve difficulties. These websites evolve into data science knowledge repositories over time.
23 Analysis of such repositories can provide valuable insights into the applications, topics, trends,
24 and challenges of data science.
25 **Methods.** In this article, we investigated what data scientists are asking by analyzing all posts to
26 date on DSSE, a data science-focused Q&A website. To discover main topics embedded in data
27 science discussions, we used latent Dirichlet allocation (LDA), a probabilistic approach for topic
28 modeling.
29 **Results.** As a result of this analysis, 18 main topics were identified that demonstrate the current
30 interests and issues in data science. We then examined the topics' popularity and difficulty. In
31 addition, we identified the most commonly used tasks, techniques, and tools in data science. As a
32 result, "Model Training", "Machine Learning", and "Neural Networks" emerged as the most
33 prominent topics. Also, "Data Manipulation", "Coding Errors", and "Tools" were identified as
34 the most viewed (most popular) topics. On the other hand, the most difficult topics were
35 identified as "Time Series", "Computer Vision", and "Recommendation Systems". Our findings
36 have significant implications for many data science stakeholders who are striving to advance
37 data-driven architectures, concepts, tools, and techniques.

INTRODUCTION

The volume and variety of data produced and shared in today's digital life cycle, also known as the "age of big data", is increasing exponentially on a daily basis. Over the last decade, big data has been at the forefront of the technology-driven revolution in computing ecosystems. Big data applications, including but not limited to search engines, social networks, e-commerce, and multimedia streaming services, have achieved unprecedented success (Assunção et al., 2015; Donoho, 2017; Vicario & Coleman, 2019). As a result, data-driven services and applications are in high demand and interest across all industrial and social segments (Cao, 2017; Gurcan & Cagiltay, 2019). This data-driven digital transformation in information technologies has resulted in an increase in activity around data science, and it has thus become an important cornerstone of today's technology-oriented life cycle (Cao, 2017; Sarker, 2021). Data science is a multidisciplinary field that uses scientific approaches, processes, algorithms, and systems to extract information and insights from structured or unstructured data (Cao, 2017; Sarker, 2021). Data Science employs a wide range of interdisciplinary techniques and theories from mathematics, statistics, computer science, and information science. Data science fundamentals are based on the combined use of statistical models, data mining, machine learning, and computing methods to analyze and comprehend real-world events using data in various structures (Cao, 2017; Vicario & Coleman, 2019; Plotnikova, Dumas & Milani, 2020; Sarker, 2021).

The introduction of big data systems, in particular, has resulted in a significant transformation in the architectures, methodologies, and competencies used in data science (Donoho, 2017; Gurcan, 2019; Saltz & Krasteva, 2022). In other words, data-driven digital transformation has increased the responsibilities of data scientists (Cao, 2017). Modern data science applications and practices include a wide range of tasks, technologies, tools, and paradigms. This necessitates data scientists being proficient in a wide range of knowledge-domains and skill sets (Gurcan, 2019; Sarker, 2021). Even experienced data scientists may struggle to keep up with the rapid evolution of these data-driven paradigms (Gurcan, 2019).

For these reasons, data scientists frequently use domain-specific Q&A websites such as Data Science Stack Exchange (DSSE) to get help and advice from colleagues about technical challenges (Stack Exchange, 2022). DSSE seeks solutions to the problems of data scientists from various backgrounds by managing their thousands of posts (Stack Exchange, 2022). With the increase in data science activities in recent years, the DSSE platform has become an important source of information and a useful guide for data scientists (Karbasian & Johri, 2020). All of the posts shared on DSSE can be considered as an important data repository that keeps a history of data scientists' topics and thoughts.

Analyzing this repository can provide important insights into which tools and technologies data scientists prefer, the reasons for their choices, the types of work environments and motivations they require, and the technical challenges they face (Karbasian & Johri, 2020). From this perspective, researchers have recently conducted many empirical studies using Stack Overflow data to investigate the sub-contexts of software development, such as testing (Kochhar, 2016),

79   security (Yang et al., 2016), mobile development (Rosen & Shihab, 2016), requirements (Zou et
80   al., 2015), concurrency development (Ahmed & Bagherzadeh, 2018), chatbot development
81   (Abdellatif et al., 2020), and IOT development (Uddin et al., 2021). Despite these valuable
82   efforts, studies exploring the holistic landscape of data science using semantic content analysis
83   based on topic modeling from online Q&A platforms are still limited, with a few exceptions
84   (Bagherzadeh & Khatchadourian, 2019; Hin, 2020; Karbasian & Johri, 2020). In fact, we believe
85   that analyzing data from Q&A platforms will have significant implications for understanding the
86   current state of data science.
87   Taking into account the current background, this study aims to conduct an in-depth and
88   comprehensive analysis of the common issues and challenges faced by data scientists, as well as
89   to fill a gap in the literature. To that end, a semantic content analysis based on probabilistic topic
90   modeling was applied to posts shared on DSSE, a data science-specific Q&A platform, over the
91   last nine years between 2014 and 2022 (Stack Exchange, 2022). The semi-automatic
92   methodology proposed in this study includes a series of semantic topic modeling processes based
93   on Latent Dirichlet Allocation (LDA), a probabilistic topic modeling algorithm used to
94   automatically discover topics (Blei, Ng & Jordan, 2003). Using this methodology, we identified
95   the main issues and themes that data scientists are discussing, as well as their underlying
96   dependencies and indicators and long-term trends. In summary, we motivate and present the
97   following research questions for our empirical study.
98   **RQ1.** What topics are discussed by data scientists?
99   **RQ2.** How do the data science topics evolve over time?
100 **RQ3.** How do the popularity and difficulty of the topics vary?
101 **RQ4.** What are the most commonly used tasks, techniques, and tools in data science?
102 **RQ5.** How do data science topics relate to data-driven technologies?

103 # BACKGROUND AND RELATED WORK

104 To provide a better understanding of the contextual dimensions of data science, we base our
105 research on three pillars and discuss relevant studies here under the headings of data science,
106 Q&A communities, and topic modeling.

107 ## Studies related to the data science

108 In recent years, the rapid development of data-oriented services and applications has increased
109 the number of studies aimed at understanding the data science paradigms. The majority of
110 domain-specific research focuses on data science concepts, architectures, practices, application
111 areas, big data, and related fields (Donoho, 2017). These studies provided insights into the use of
112 data science in academia and industry, as well as data science issues and challenges (Vicario &
113 Coleman, 2019; Karbasian & Johri, 2020). Previous studies have discussed the use and benefits
114 of data science in various disciplines, ranging from business science to computer sciences,
115 finance to medicine, bioinformatics to natural sciences (Donoho, 2017; Vicario & Coleman,
116 2019). In today's data-driven world, common data science application areas include business
117 analytics, supply chain, social media analytics, smart cities, business logistics, business
118 intelligence, recommendation systems, decision support systems, natural language processing,

119   financial fraud detection, advertising, behavioral analytics, and manufacturing (Schoenherr &
120   Speier-Pero, 2015; Vicario & Coleman, 2019; Sarker, 2021). Several studies have emphasized
121   the close relationship of data science with big data, predictive analytics, and data-driven decision
122   making (Cao, 2017; Sarker, 2021).
123   A number of studies, similar to our current study, have been conducted based on the analysis of
124   posts from Q&A communities to investigate the issues and challenges of data science
125   (Bagherzadeh & Khatchadourian, 2019; Hin, 2020; Karbasian & Johri, 2020). An empirical
126   study was conducted in which data science-related posts on two Q&A websites, Stack Overflow
127   and Kaggle, were analyzed using the topic modeling approach, and 24 data science-related
128   discussion topics were identified (Hin, 2020). In another study, Karbasian and Johri used a topic
129   modeling-based approach to analyze data science-specific posts from two popular Q&A
130   communities, StackExchange and Reddit (Karbasian & Johri, 2020). The discussion topics
131   discovered on both platforms were compared using their analysis (Karbasian & Johri, 2020).
132   Furthermore, Bagherzadeh and Khatchadourian examined Stack Overflow posts to investigate
133   big data-related issues, and as a result, they identified 30 topics (Bagherzadeh & Khatchadourian,
134   2019). Our study differs from theirs in that we only analyze posts shared on DSSE, a data
135   science-specific Q&A platform, and we propose a semi-automated methodology based on
136   unsupervised machine learning for such analysis. Furthermore, our work complements theirs by
137   adding dimension and depth.

138   ## Studies on Q&A communities

139   Researchers have long studied online Q&A platforms such as Stack Overflow (Ahmad et al.,
140   2018), Stack Exchange (Karbasian & Johri, 2020), Quora (Maity, Sahni & Mukherjee, 2015),
141   Kaggle (Hin, 2020), Yahoo (Zhang et al., 2020), and Reddit (Karbasian & Johri, 2020). Stack
142   Overflow, in particular, is an important information and experience sharing platform for all
143   developers from various backgrounds under the umbrella of information technologies, and it has
144   been the subject of numerous studies (Ahmad et al., 2018).
145   More specifically, a number of these studies have used Stack Overflow data in order to classify
146   the posts (Barua, Thomas & Hassan, 2014; Beyer et al., 2020); to investigate descriptive
147   indicators (i.e., tags, view count, score, answer count, etc.) of the posts (Barua, Thomas &
148   Hassan, 2014; Yang et al., 2016; Uddin et al., 2021); to analyze the textual contents of posts
149   (Barua, Thomas & Hassan, 2014; Rosen & Shihab, 2016); to organize its settings, design, and
150   infrastructure (Treude & Wagner, 2019; Alrashedy et al., 2020); to explore the usability of
151   settings and configurations (Treude & Wagner, 2019), and to discover the relationship between
152   questions and their answers (Xu et al., 2018).
153   Stack Overflow data has also been used in a number of studies that revealed discussion topics in
154   specific subfields of software development such as testing (Kochhar, 2016), security (Yang et al.,
155   2016), mobile development (Linares-Vásquez, Dit & Poshyvanyk, 2013; Rosen & Shihab, 2016),
156   chatbot development (Abdellatif et al., 2020), IOT development (Uddin et al., 2021), and
157   machine learning (Alshangiti et al., 2019). As previously stated, a small number of studies have
158   also been conducted based on the analysis of posts shared on the Q&A platforms to explore the

159   issues and challenges related to data science (Bagherzadeh & Khatchadourian, 2019; Hin, 2020;
160   Karbasian & Johri, 2020). Our study expands on previous research by analyzing the
161   comprehensive dimensions of posts shared on the DSSE platform, which only includes data
162   science-related posts.

163   ## Topic modeling

164   Topic modeling is a generative approach for comprehending and summarizing the semantic
165   content of large collections of documents (Blei, 2012; Gurcan et al., 2022b). It ensures that the
166   word groups that best describe the semantic map of the documents are identified as a separate
167   topic (Blei, 2012). Topic modeling is becoming more popular for mining unstructured textual
168   data due to its capabilities and benefits for semantic content analysis (Gurcan & Cagiltay, 2022).
169   Moreover, topic modeling is a generative approach actively used to reveal emerging trends in
170   specific contexts of technical sciences such as software engineering, computer sciences, and
171   information sciences (Silva, Galster & Gilson, 2021; Gurcan et al., 2022a). The amount of data
172   produced in today's software repositories and social networks is growing at an exponential rate.
173   Because of the rapid increase in the amount of unstructured data, semantic content analysis of
174   this data has become difficult but important, even though it provides an opportunity for research
175   on this data (Silva, Galster & Gilson, 2021).
176   From its inception to the present, topic modeling procedures have been successfully applied to
177   various types of data, including textual documents, genetic data, web archives, log files, source
178   codes, images, videos, forums, blogs, Q&A platforms, software repositories, and social
179   networks. (Blei, 2012; Silva, Galster & Gilson, 2021). Furthermore, topic modeling was
180   employed to reveal the major trends in software engineering research (Mathew & Menzies, 2018;
181   Gurcan et al., 2022b); to explore essential competencies and skills for software engineers by
182   analyzing online jobs (Gurcan & Kose, 2017; Gurcan & Cagiltay, 2019); to identify what issues
183   are software developers are discussing about specific contexts of software engineering, such as
184   IOT development (Uddin et al., 2021), mobile development (Rosen & Shihab, 2016), testing
185   (Kochhar, 2016), chatbot development (Abdellatif et al., 2020) security (Yang et al., 2016) and
186   programming languages (Chakraborty et al., 2021).
187   In data science research, the LDA-based topic modeling method was also used to analyze data
188   scientists' discussions on Q&A platforms (Bagherzadeh & Khatchadourian, 2019; Hin, 2020;
189   Karbasian & Johri, 2020). Apart from these aforementioned studies, numerous studies based on
190   topic modeling procedures have been conducted in sub-contexts of various disciplines (Silva,
191   Galster & Gilson, 2021). In conclusion, the effectiveness and suitability of the topic model
192   approach for research in the technical sciences has increased our motivation to investigate data
193   science issues using topic models.

194   # METHOD

195   ## Data collection and extraction

196   To provide an objective methodology, we used a data dump that included all posts shared on
197   Data Science Stack Exchange (DSSE), a data science-specific Q&A platform (Stack Exchange,
198   2022). The datasets created and analyzed during this current study are publicly available in the

199 Internet Archive repository (Internet Archive, 2022) as a data dump in XML format. The first
200 step was to download the most recent XML data dump (last updated on June 8, 2022) and parse
201 it into a database. Each question post in the data dump contains various metadata elements such
202 as a title, a body, tags, answers, comments, and other indicators (Yang et al., 2016).
203 From May 2014 to June 2022, this parsed data dump contained 70283 posts (33492 questions
204 and 36791 answers). Figure 1 depicts the number of data science-related posts in our
205 experimental data set by year. According to Figure 1, the number of questions and answers
206 increased up until 2019 and then decreased. Although the number of answers is generally greater
207 than the number of questions, it has been lower since 2020. It was also discovered that an
208 average of 3721 questions and 4088 answers were shared on the DSSE platform each year.

209 ## Data preprocessing

210 Data preprocessing consists of a series of operations used to convert textual raw data into
211 structured datasets. We considered the preprocessing steps that have been widely preferred in
212 previous work to reduce noise and structure textual data in order to implement an efficient
213 preprocessing (Rosen & Shihab, 2016; Yang et al., 2016; Uddin et al., 2021). Initially, we
214 removed all non-semantic paragraphs and content from a post, such as code snippets marked
215 with <code></code> and non-text blocks marked with HTML tags like <p></p> and <a>. In
216 the second step, we removed stop words (e.g., "a", "an", "the", "with", etc.), numbers,
217 punctuation, and non-alphabetic characters (Řehůřek & Sojka, 2011). In this way, we have
218 removed words and characters from the posts that do not make sense on their own. As a result of
219 this process, we only kept the meaningful words required for semantic topic modeling in the
220 corpus. Following that, we used the lemmatization process to reduce the words to their roots
221 while keeping the various meanings derived from the same root (Řehůřek & Sojka, 2011; Gurcan
222 & Cagiltay, 2022).

223 ## Implementation of LDA-based topic modeling

224 Topic modeling is a statistical and computational approach that uses unsupervised machine
225 learning to extract latent semantic patterns from a large collection of documents. For text mining
226 and natural language processing research, several topic modeling algorithms have been
227 proposed, including Latent Semantic Indexing (LSI), Latent Dirichlet Allocation (LDA), Non-
228 Negative Matrix Factorization (NMF), Dirichlet Multinomial Regression (DMR), Hierarchical
229 Latent Dirichlet Allocation (HLDA), Hierarchical Dirichlet Process (HDP), Dynamic Topic
230 Model (DTM), and Correlated Topic Model (CTM) (Řehůřek & Sojka, 2011; Gurcan &
231 Cagiltay, 2022; Gurcan et al., 2022b). Unfortunately, the majority of these algorithms do not
232 provide a widely accepted method for calculating the consistency score, which is used to
233 estimate the optimal number of topics (Gurcan & Cagiltay, 2022). Among all topic modeling
234 algorithms, LDA provides the fundamental background of topic modeling and is thus more
235 widely used than the others. LDA is widely regarded as one of the most effective techniques for
236 discovering hidden semantic structures known as "topics" in natural language text documents
237 (Blei, 2012). In fact, LDA is the most widely used topic modeling algorithm in text mining and
238 natural language processing (Gurcan et al., 2022a). In this regard, in this study, we used Latent

239 Dirichlet Allocation (LDA), a probabilistic model, for topic modeling-based semantic content
240 analysis of a textual corpus of data science posts. The theoretical background and graphical
241 model representations of the LDA algorithm is presented in the study by Blei et al. (2003), which
242 introduced the LDA (see Figure 1and 7 in (Blei, Ng & Jordan, 2003)).
243 With the aim of implementing the LDA-based topic modeling procedures to our data science
244 corpus, we used Gensim (Řehůřek & Sojka, 2011), a comprehensive library for text mining and
245 topic modeling. The LDA topic modeling algorithm generates a list of topics from which to
246 group the input preprocessed posts based on the specified K (number of topics). Taking into
247 account previous work (Uddin et al., 2021), we used the C_V metric, which is included in the
248 Gensim package (Řehůřek & Sojka, 2011), to calculate a coherence score for each K value and
249 to determine the optimal number of K topics for our corpus. Consistent with previous work for
250 topic modeling of short texts (Zuo et al., 2016; Gurcan et al., 2022b), the prior parameters of
251 =0.1 and =0.01 were used to fine-tune the distribution of topics per document and distribution of
252 words per topic, respectively. The LDA model was implemented using these prior parameters,
253 with K values ranging from 10 to 40. (increasing one at a time). Concurrently, as shown in
254 Figure 2, we calculated a coherence score (C_V) for each topic model implemented for each K
255 (Řehůřek & Sojka, 2011). For the number of topics K = 18, the maximum consistency score (C_V
256 =0.52589) was obtained within these C_V metrics, revealing the optimal semantic consistency of
257 the topics. As a result, we determined that 18 was the optimal number of topics.

EMPIRICAL STUDY

259 In this section, we answer our five research questions:
260 **RQ1.** What topics are discussed by data scientists?
261 **RQ2.** How do the data science topics evolve over time?
262 **RQ3.** How do the popularity and difficulty of the topics vary?
263 **RQ4.** What are the most commonly used tasks, techniques, and tools in data science?
264 **RQ5.** How do data science topics relate to data-driven technologies?

What topics are discussed by data scientists? (RQ1)

Motivation

267 Because of the rapid development of data science paradigms, many innovative architectures,
268 techniques, and tools to support data science development have been developed. To keep up with
269 this rapid change, data scientists must first understand the issues and challenges they face when
270 employing these architectures, techniques, and tools. The number of empirical studies that
271 comprehensively investigate the issues facing data scientists is quite limited. As a result, this
272 type of analysis can aid in understanding the issues that data scientists are discussing.

Approach

274 We used LDA-based topic modeling to investigate the questions that data scientists are asking on
275 the DSSE platform. LDA is a probabilistic and generative topic modeling technique that
276 represents topics as a probability distribution over words in a corpus. In the method section, we
277 describe in detail how we adapted and applied LDA-based topic modeling to our corpus. We
278 discovered 18 topics as a result of this LDA analysis, with each topic represented by 30

279   descriptive keywords. Two independent domain-specific experts assessed the consistency of the
280   topics described with descriptive keywords. Experts evaluated 20 random samples of posts for
281   each of the 18 topics to see if the posts were consistent with the dominant topic to which they
282   were assigned. The topics were then given names based on their keywords. The total percentages
283   of each topic in the entire corpus were then calculated, taking into account the dominant topic to
284   which each post is assigned. For example, if a topic has a rate of 10%, it is the dominant topic in
285   10% of all question posts, and those posts are assigned to that topic.

286   ***Finding***

287   As a result, Table 1 shows the 18 topics discovered by LDA-based topic modeling, along with
288   their topic names, descriptive keywords, and rates. As shown in Table 1, the topics (issues) are
289   listed in descending order by percentage, and the keywords for each topic are also listed in
290   descending order. Because the topics in Table 1 reveal major issues specific to data science, the
291   terms topic and issue are used interchangeably throughout this paper. The topics indicated that
292   data scientists face a wide range of issues, from "Machine Learning" to "Model Training",
293   "Neural Networks" to "Time Series", "NLP" to "Computer Vision".
294   As a result, the top five most frequently asked topics were "Model Training", "Machine
295   Learning", "Neural Networks", "NLP", and "Time Series". The least asked topics, on the other
296   hand, were "Clustering", "Coding Errors" and "Dimensionality Reduction". Based on the issues
297   discovered (see Table 1), we can conclude that machine learning and its subtasks such as "Model
298   Training", "Neural Networks", "Feature Engineering", "Classification", and "Clustering" are
299   dominant topics for data science. Furthermore, the topics of "NLP", "Computer Vision", and
300   "Recommendation Systems" have emerged as the most common application areas of data
301   science.

302   **How do the data science topics evolve over time? (RQ2)**

303   ***Motivation***

304   Data science is a constantly evolving discipline that includes new or outdated tasks,
305   architectures, techniques, and tools. Therefore, the interests, recommendations, and experiences
306   of data scientists may change over time. From this perspective, we attempted to examine how the
307   topics discussed by data scientists have evolved over time. Such an analysis will help to improve
308   understanding of chronic and unsolvable data science issues, as well as generate new solutions to
309   them.

310   ***Approach***

311   At this stage, we examined how data science problems have evolved over time. To achieve this,
312   we looked at the distribution of the number of questions for each topic over time. The annual
313   number of questions for each topic was calculated. Then, for each year, we divided the number
314   of questions per topic by the total number of questions for that year. In this way, we normalized
315   the question distribution of each topic as a percentage for that year. In other words, we calculated
316   the percentage of questions for each topic per year. Then we subtract the previous year's
317   percentages from the current year's percentages. As a result, we calculated how much the topics

PeerJ Computer Science   Manuscript to be reviewed

318   changed in the current year compared to the previous year. Finally, we calculated the overall
319   temporal trend of the topics by adding the annual percentage changes for each topic.
320
321   ***Finding***
322   Table 2 shows the annual percentages, total trends, and trend directions for the topics, which are
323   listed in descending order by overall percentage. Table 2 shows the annual percentage changes
324   for each topic in each row, and a number of inferences can be drawn about how the topics have
325   evolved over time. Figure 3 depicts the total trend values of the topics in descending order to
326   provide a broader understanding. Figure 3 demonstrates 11 topics with an increasing trend and
327   seven topics with a decreasing trend. As seen in Figure 3, "Model Training", "Neural Networks",
328   "Regression Models", "Time Series", and "Computer Vision" are the top five topics with the
329   most increasing trend, while "Tools", "Machine Learning", The topics "Recommendation
330   Systems", "NLP" and "Clustering" have the most decreasing trend.

331   ## How do the popularity and difficulty of the topics vary? (RQ3)
332   ***Motivation***
333   Our findings at RQ1 revealed the broad scope and diversity of data science issues. Data scientists
334   discuss various technical issues at various levels in order to understand specific tasks,
335   techniques, and tools and how they operate effectively. As a result, these issues are unlikely to
336   have the same level of popularity and difficulty. Some questions are displayed numerous times,
337   while others are not displayed at all. Identifying the popularity and difficulty of data science
338   issues can help prioritize efforts to improve data science. Taking into account the implications of
339   this analysis, for example, data science beginners can concentrate on more popular topics. Tool
340   builders can create tools that provide solutions to more difficult data science problems.
341   ***Approach***
342   A question post on DSSE includes descriptive indicators such as the number of views, answers,
343   accepted answers, score, favorites, and comments. After a user posts a question on DSSE, other
344   users respond with answers. When a user realizes an answer that resolves their question, that user
345   marks that answer as the accepted answer for that question. In this way, the user indicates that
346   the problem has been resolved, and other users who encounter the same issue can view this
347   solution. At this stage, we performed a series of computational analyses using these indicators to
348   determine the difficulty and popularity of each topic. We began by calculating the number of
349   questions assigned to each topic. Next, we divided the total number of views for each topic by
350   the total number of questions for each topic to calculate the average number of views for each
351   topic. We calculated the popularity of the topics in this way. We then calculated the average of
352   the number of accepted answers for each topic by dividing the number of accepted answers for
353   that topic by the number of questions for that topic. We determined the difficulty of the topics in
354   this manner. Similarly, we calculated the average number of answers, favorites, comments, and
355   score for each topic. Since the average number of accepted answers ranges from 0 to 1, we
356   presented this indicator as a percentage.
357   ***Finding***

358 Table 3 shows the calculated averages of the questions, views, answers, scores, comments,
359 accepted answers, and favorites for each topic in order to highlight all dimensions of data
360 scientists' interest in domain-specific issues. The topics in this table are listed in descending
361 order of their percentages. The findings in Table 3 provide insight into the interests and
362 perspectives of data scientists on various issues. Moreover, we calculated the popularity of the
363 topics based on the average number of views and presented it in Figure 4. As shown in Figure 4,
364 the top five most viewed (most popular) topics are "Data Manipulation", "Coding Errors",
365 "Tools", "Regression Models", and "Neural Networks". On the other hand, the least viewed
366 (least popular) topics are "Recommendation Systems", "Time Series", and "Computer Vision".
367 Also, we determined the difficulty of the topics by considering the average number of accepted
368 answers and presented them in Figure 5. According to the difficulty metrics of the topics shown
369 in Figure 5, the most difficult topic is "Time Series", which has the lowest rate (21%) based on
370 the number of accepted answers. The other most difficult topics are "Computer Vision",
371 "Recommendation Systems", "NLP", and "Statistics".

372 ## What are the most commonly used tasks, techniques, and tools
373 ## in data science? (RQ4)

374 ### *Motivation*
375 The growing interest in data-driven applications and services has resulted in an expansion and
376 diversification of data science technologies. Innovative data technologies, which cover a wide
377 range of tasks, methods, and tools, are now widely used in today's data science environments. As
378 a result, the majority of data science issues and challenges are closely related to these tasks,
379 techniques, and tools. It is highly possible that trends in data science technologies will evolve in
380 concurrently with the technological transformations experienced in data-driven ecosystems.
381 Identifying the most commonly used tasks, techniques, and tools in data science can provide
382 important insights into various dimensions of data science issues. An analysis of this nature will
383 also help in understanding the relative popularity of data-driven technologies over time.
384 Revealing useful and popular data technologies can help data scientists choose the appropriate
385 tools to advance data science.

386 ### *Approach*
387 Each DSSE question post contains tags that provide context and background for that question.
388 These tags are chosen and added to the question by the user who asked it. The tags are
389 descriptive keywords that display data science-related themes, tasks, techniques, and tools that
390 users associate with their questions. To extract a definitive set of all tags used in data science, we
391 first separated the tags of each post into individual tags and calculated the tag frequencies across
392 all posts. Following that, we identified the tags with the highest frequency across the entire
393 corpus. Then, taking into account their annual frequency distributions, we calculated the annual
394 percentages of these tags. We calculated how each tag changed in that year compared to the
395 previous year by subtracting the percentage of each tag in the previous year from the percentage
396 in the current year. In this way, we determined the percentage increase or decrease of tag
397 frequencies for each year. Following that, we calculated the overall trend for each tag over time

398 by adding these annual changes. Finally, we identified the most commonly used tasks,
399 techniques, and tools in data science by categorizing the tags according to their functions and
400 contexts.
401 ***Finding***
402 We identified 665 unique tags used in data science by analyzing the tags of all posts in the
403 corpus. We calculated the annual percentages and total percentages of the top 50 tags with the
404 highest frequency and presented them in Table 4 in descending order of their percentages to
405 provide a clearer understanding of the tags. Table 4 demonstrates the top 50 tags' key findings,
406 including yearly rates (from 2014 to 2022), total rates, trend values, and trend directions. Figure
407 6 also shows the top 20 tags with the highest frequency and their percentages. According to
408 Figure 6, "machine-learning" is the most prominent tag, followed by "python", "deep learning",
409 "neural network", and "classification". Consistent with the data science topics discovered by the
410 LDA (see Table 1), machine learning-related tags such as "machine learning", "deep learning",
411 and "neural network" emerged as clearly dominant.
412 In order to demonstrate the temporal trends of these 50 tags, we also present the annual
413 percentages, trend values, and trend directions of the tags in Table 4. In Figure 7, we visualized
414 the top 15 tags with increasing and decreasing trends to better illustrate the tag trends. As seen in
415 Figure 7, the prominent tags with an increasing trend are "deep-learning", "python",
416 "tensorflow", "keras" and "lstm". On the other hand, top tags with decreasing trend are identified
417 as "data-mining", "machine-learning", "r", "text-mining" and "classification". Considering these
418 trends, we can predict that deep learning will become a widespread research and application area
419 for data scientists in the near future. It is an interesting finding that, while the trend of "data-
420 mining" and "machine-learning" tags is decreasing, the trend of "deep-learning" and related tags
421 such as "tensorflow", "keras", and "lstm" is increasing. This finding clearly shows a shift in data
422 science from "machine-learning" to "deep-learning".
423 The topics and tags discovered in the previous stage of our analysis revealed that data science
424 issues cover a wide range of data science tasks, techniques, and tools. When the context and
425 background of data science issues are thoroughly examined, it is clear that the vast majority of
426 these issues and challenges are related to how data-driven task techniques and tools are used in
427 data science. In order to provide an understanding of the tasks, techniques, and tools used in data
428 science, we categorized the tags based on their functions and contexts. In this way, we classified
429 tags into three categories: tasks, algorithms, and tools.
430 Table 5 shows the top 25 tags for each category. The tasks, algorithms, and tools listed in Table
431 5 can be defined as the three pillars of data science. As shown in Table 5, the top five data
432 science tasks are "machine-learning", "deep-learning", "classification", "nlp", and "time-series".
433 The dominance of "machine-learning" and "deep-learning" tasks in data science is especially
434 noteworthy. The most commonly used algorithms in data science are identified as "neural-
435 network", "cnn", "lstm", "random-forest", and "rnn". The top five data science tools are
436 discovered to be "python", "keras", "scikit-learn", "tensorflow", and "r."
437 **How do data science topics relate to data-driven technologies? (RQ5)**

438 ***Motivation***
439 Earlier stages of our analysis revealed that the issues discussed by data scientists cover a wide
440 range of tasks, techniques, and tools used in data science. Analyzing the connections between
441 data-driven technologies (tasks, techniques, and tools) and data science issues can lead to
442 significant discoveries. In this way, the findings of such an analysis will contribute to a better
443 understanding of data science issues and the advancement of data science. To achieve this, we
444 expanded our analysis at this point to investigate correlations between data science issues and
445 data-driven technologies.

446 ***Approach***
447 At this stage, we added another process to our current analysis and attempted to correlate our
448 findings in RQ1 and RQ3. We began by calculating the tag distribution for posts assigned to
449 each topic discovered in RQ1. In RQ3, we explained how the tags for each post are parsed. We
450 then identified the top 15 tags with the highest frequency for each topic. In this way, we
451 determined which tasks, techniques, and tools are closely related to which data science issues.

452 ***Finding***
453 As a result of this process, we identified the top 15 tags for each topic and presented them in
454 Table 6, where the topics are listed in descending order by percentage. Likewise, the top 15 tags
455 for each topic are sorted in descending order. As shown in Table 6, the tag "machine-learning"
456 has the highest frequency for the topic "Model Training", while "random-forest" has the lowest.
457 Figure 8 shows the first three topics and their tags to illustrate the significance of these tags for
458 each topic. In this way, we discovered a wide range of data-driven technologies (e.g., tasks,
459 algorithms, analysis tools, programming languages, and machine learning libraries) categorized
460 under the 18 issues. According to the results in Table 6, "machine-learning" is the first tag in 9 of
461 the 18 topics. In other words, machine learning is clearly a dominant concept in data science
462 issues. Python was seen as the first tag in three of these topics ("Data Manipulation", "Tools",
463 and "Coding Errors"). This finding highlighted Python as the most popular programming
464 language for data science. Apart from these, the tags "neural-network", "nlp", "time-series",
465 "deep-learning", "classification", and "clustering" appeared as the first tag in only one topic.
466
467

468 # DISCUSSION
469 Data science is comprised of dynamic and competitive working environments in which tasks,
470 paradigms, tools, technologies, skills, and experiences are constantly updating and progressing.
471 Our analysis identifies the most frequently asked questions by data scientists on the DSSE
472 platform and investigates the dimensions that indicate their importance. Our findings revealed
473 the data science issues and challenges as 18 distinct topics discovered by LDA. The findings of
474 our study have important implications for understanding data scientists' thoughts, interactions,
475 work flows, and issues in today's IT industry. We will now go over these findings in detail.
476 Initially, we found that the topic "Machine Learning" and its related processes are emerging as
477 the most dominant tasks and most common implementations for data scientists, so tools and

478   applications support for machine learning may need to mature further (Alshangiti et al., 2019;
479   Hin, 2020). Other machine learning-related topics with high percentages, such as "Model
480   Training", "Neural Networks", and "Feature Engineering", support our finding (see Table 1). We
481   frequently encountered machine learning-related keywords and tags in other topics as well. Table
482   6 shows that "machine-learning" is the first tag in 9 of 18 topics. Furthermore, "NLP",
483   "Computer Vision", and "Recommendation Systems" were identified to be prominent data
484   science application areas (see Table 1) (Hin, 2020; Karbasian & Johri, 2020).
485   We also analyzed the temporal evolution of data science issues and discovered a number of
486   noteworthy findings. "Model Training" is once again at the top of the list of topics with the
487   fastest growing trend. Following that, the data science topics with the highest increasing trend
488   were identified as "Neural Networks", "Regression Models", and "Time Series" (see Figure 3).
489   Furthermore, within application areas, "Computer Vision" is on the rise, while "NLP" and
490   "Recommendation Systems" are on the decline (Liu et al., 2017). The strong increasing trend of
491   "Model Training", "Neural Networks", and "Computer Vision" topics suggests that deep
492   learning will gain a more leading position in the near future (Karbasian & Johri, 2020). These
493   findings point to a significant shift from machine learning to deep learning (see Figure 3) (Hin,
494   2020).
495   We expanded our findings by including some indicators of the issues in our analysis to better
496   understand the most important, popular, and difficult issues for data scientists. The number of
497   questions, views, and answers for a topic on the DSSE platform reveal various insights about it
498   (Bagherzadeh & Khatchadourian, 2019). If a question has already been asked, it will not be
499   asked again; instead, the user will see the previously asked question and its related answers. As a
500   result, the number of views on the topics is an important indicator of the topic's popularity
501   (Uddin et al., 2021). We discovered that the top three most popular topics were "Data
502   Manipulation", "Coding Errors", and "Tools" (see Figure 4). These three topics are the most
503   widely discussed in data science. "Time Series" is the most difficult topic, with the lowest
504   percentage of accepted answers, followed by "Computer Vision", "Recommendation Systems",
505   "NLP", and "Statistics" (see Figure 5) (Sarker, 2021). Such common issues in data science that
506   have yet to be resolved should be investigated further and supported by data scientists (Cao,
507   2017).
508   Our study also found that data science issues and challenges are closely related to data-driven
509   technologies (tasks, techniques, and tools) (Cao, 2017; Sarker, 2021). The most frequently used
510   data-driven technologies are "machine-learning", "python", "deep-learning", "neural-network",
511   and "classification" (see Table 4 and Figure 6) (Hin, 2020).The prominent tags with an
512   increasing trend are "deep-learning", "python", "tensorflow", "keras", and "lstm" (see Figure 7)
513   (Karbasian & Johri, 2020). Deep learning will be the most focused area in data science in the
514   near future, according to the top five tags with the highest increasing trend (Karbasian & Johri,
515   2020). On the other hand, the top five tags with decreasing trend were identified as "data-
516   mining", "machine-learning", "r", "text-mining", and "classification" (see Figure 7).

517 Another significant finding is that, while the trend of "data mining" and "machine learning" tags
518 is decreasing, deep learning-related tags such as "deep-learning", "tensorflow", "keras", and
519 "lstm" are increasing (see Figure 7). This discovery highlights a significant shift in data science
520 from "machine learning" to "deep learning" (Hin, 2020; Karbasian & Johri, 2020). Based on our
521 findings, we can also speak of a significant shift from "R" to "Python", which are data-oriented
522 programming languages (see Figure 7). Furthermore, among the important findings of our study
523 are the most commonly used tasks (i.e., "machine-learning", "deep-learning", and
524 "classification"), algorithms (i.e., "neural-network", "cnn", and "lstm") and tools (i.e., "python",
525 "keras", and "scikit-learn") in data science (see Table 5) (Cao, 2017; Karbasian & Johri, 2020).

526 ## Implications of Findings

527 Knowledge and experiences shared on Q&A platforms like DSSE should motivate researchers,
528 practitioners, and developers to create documentation aimed at resolving common data science
529 issues. We have inferred remarkable implications and guidelines for data science stakeholders
530 based on the empirical background, methodology, and findings of this study, which will
531 contribute to their understanding of the field. We hope that our implications will help data
532 science communities with various profiles, such as developers, researchers, practitioners,
533 educators, and enthusiasts. Developers can lead data science innovation by creating more
534 specific tools and applications to address the current issues and needs of data scientists, which
535 are also highlighted in our findings. One of the most popular topics is "Data Manipulation". For
536 such widespread issues, tool developers can create useful libraries or tools. Developers can use
537 our findings to improve data-driven tools or to choose which frameworks and libraries to
538 support.
539 Identifying the questions that data scientists are asking on platforms like DSSE can help the
540 research community better understand the challenges of data science. While all of the issues
541 identified are important in their own right, our findings indicate that data science researchers
542 should prioritize the most visible and difficult issues. Researchers can also use our methodology
543 for experimental research and analysis in a variety of settings. Furthermore, data science
544 educators can create more tutorials to assist in the training of data scientists and candidates,
545 especially considering the difficulty of the topics as well as the most commonly used tasks,
546 algorithms, and tools. Educators can keep their training programs and curricula up to date with
547 current field trends, allowing them to provide up-to-date background to data scientists training.
548 Our findings can be used by DSSE and other Q&A platforms to develop new approaches for
549 contextually tagging posts and better categorizing user posts. Data science enthusiasts and
550 general readers may find our findings useful in keeping up with emerging developments and
551 trends in the data science industry and ecosystems. More researchers in this field can also
552 contribute to the improvement of data science processes. We hope that our findings, which
553 highlight the challenges that data scientists face, will help to guide future research in this area.
554
555 ## CONCLUSIONS

556 This research aims to shed light on common issues and challenges encountering data scientists.
557 To that end, all posts shared on the DSSE platform were analyzed using LDA-based semantic
558 topic modeling. Furthermore, the most commonly used data-driven technologies and their
559 connections to data science issues were investigated. Our research methodology is based on the
560 adaptation and implementation of LDA, an unsupervised generative approach for semantic topic
561 modeling that is widely used in textual content analysis.
562 As a result of the LDA analysis, 18 topics were identified that demonstrate the current landscape
563 of data science issues and trends. Among these topics, "Model Training", "Machine Learning",
564 and "Neural Networks" were the most frequently asked. Furthermore, the most viewed (most
565 popular) topics were "Data Manipulation", "Coding Errors", and "Tools". The most difficult
566 topics were identified as "Time Series", "Computer Vision", and "Recommendation Systems".
567 One of our key discoveries is that data science issues and challenges are inextricably linked to
568 data-driven technologies (tasks, techniques, and tools). While the trend for "data-mining" and
569 "machine-learning" tags is decreasing, the trend for deep learning-related tags such as "deep-
570 learning", "tensorflow", "keras", and "lstm" is increasing. As the findings show, there has been a
571 significant shift in data science from machine learning to deep learning. It was also determined
572 that the most commonly used algorithms are "neural-network", "cnn", "lstm", "random-forest",
573 and "rnn", and the most commonly used tools are "python", "keras", "scikit-learn",
574 "tensorflow", and "r". Thus, by analyzing the issues discussed by data scientists, this study
575 provides an in-depth understanding of this dynamic discipline.
576 These findings will help online communities with diverse profiles understand data science
577 focuses and issues. Our findings have significant implications for the various data science
578 stakeholders who are working to advance data science. Our findings can be used by tool builders
579 to improve support and documentation, by developers to create data applications and libraries,
580 and by educators to create modern training and curricula. By focusing on current issues,
581 researchers can provide more solutions for data science. Our methodology can also be applied to
582 other developer platforms like forums, blogs, and portals, as well as different Q&A platforms
583 like Kuggle, Reddit, and Quora.

References

585 Abdellatif A, Costa D, Badran K, Abdalkareem R, Shihab E. 2020. Challenges in Chatbot
586   Development: A Study of Stack Overflow Posts. In: *Proceedings - 2020 IEEE/ACM 17th*
587   *International Conference on Mining Software Repositories, MSR 2020*. DOI:
588   10.1145/3379597.3387472.
589 Ahmad A, Feng C, Ge S, Yousif A. 2018. A survey on mining stack overflow: question and
590   answering (Q&A) community. *Data Technologies and Applications*. DOI: 10.1108/DTA-
591   07-2017-0054.
592 Ahmed S, Bagherzadeh M. 2018. What do concurrency developers ask about?: A large-scale
593   study using stack overflow. In: *International Symposium on Empirical Software*
594   *Engineering and Measurement*. DOI: 10.1145/3239235.3239524.
595 Alrashedy K, Dharmaretnam D, German DM, Srinivasan V, Aaron Gulliver T. 2020. SCC++:
596   Predicting the programming language of questions and snippets of Stack Overflow. *Journal*
597   *of Systems and Software* 162. DOI: 10.1016/j.jss.2019.110505.

598 Alshangiti M, Sapkota H, Murukannaiah PK, Liu X, Yu Q. 2019. Why is Developing Machine
599   Learning Applications Challenging? A Study on Stack Overflow Posts. In: *International*
600   *Symposium on Empirical Software Engineering and Measurement*. DOI:
601   10.1109/ESEM.2019.8870187.
602 Assunção MD, Calheiros RN, Bianchi S, Netto MAS, Buyya R. 2015. Big Data computing and
603   clouds: Trends and future directions. *Journal of Parallel and Distributed Computing* 79–
604   80:3–15. DOI: 10.1016/j.jpdc.2014.08.003.
605 Bagherzadeh M, Khatchadourian R. 2019. Going big: A large-scale study on what big data
606   developers ask. In: *ESEC/FSE 2019 - Proceedings of the 2019 27th ACM Joint Meeting*
607   *European Software Engineering Conference and Symposium on the Foundations of*
608   *Software Engineering*. 432–442. DOI: 10.1145/3338906.3338939.
609 Barua A, Thomas SW, Hassan AE. 2014. What are developers talking about? An analysis of
610   topics and trends in Stack Overflow. *Empirical Software Engineering*. DOI:
611   10.1007/s10664-012-9231-y.
612 Beyer S, Macho C, Di Penta M, Pinzger M. 2020. What kind of questions do developers ask on
613   Stack Overflow? A comparison of automated approaches to classify posts into question
614   categories. *Empirical Software Engineering*. DOI: 10.1007/s10664-019-09758-x.
615 Blei DM. 2012. Probabilistic topic models. *Communications of the ACM* 55:77–84.
616 Blei DM, Ng AY, Jordan MI. 2003. Blei et al., 2003 - Latent Dirichlet Allocation. *Journal of*
617   *Machine Learning Research* 3:993–1022.
618 Cao L. 2017. Data science: A comprehensive overview. *ACM Computing Surveys* 50:1–42. DOI:
619   10.1145/3076253.
620 Chakraborty P, Shahriyar R, Iqbal A, Uddin G. 2021. How do developers discuss and support
621   new programming languages in technical Q&A site? An empirical study of Go, Swift, and
622   Rust in Stack Overflow. *Information and Software Technology* 137. DOI:
623   10.1016/j.infsof.2021.106603.
624 Donoho D. 2017. 50 Years of Data Science. *Journal of Computational and Graphical Statistics*
625   26:745–766. DOI: 10.1080/10618600.2017.1384734.
626 Gurcan F. 2019. Extraction of core competencies for big data: Implications for competency-
627   based engineering education. *International Journal of Engineering Education* 35:1110–
628   1115.
629 Gurcan F, Cagiltay NE. 2019. Big Data Software Engineering: Analysis of Knowledge Domains
630   and Skill Sets Using LDA-Based Topic Modeling. *IEEE Access* 7:82541–82552. DOI:
631   10.1109/ACCESS.2019.2924075.
632 Gurcan F, Cagiltay NE. 2022. Exploratory Analysis of Topic Interests and Their Evolution in
633   Bioinformatics Research Using Semantic Text Mining and Probabilistic Topic Modeling.
634   *IEEE Access* 10:31480–31493.
635 Gurcan F, Dalveren GGM, Cagiltay NE, Roman D, Soylu A. 2022a. Evolution of software
636   testing strategies and trends: Semantic content analysis of software research corpus of the
637   last 40 years. *IEEE Access* 10:106093–106109.
638 Gurcan F, Dalveren GGM, Cagiltay NE, Soylu A. 2022b. Detecting Latent Topics and Trends in
639   Software Engineering Research Since 1980 Using Probabilistic Topic Modeling. *IEEE*
640   *Access* 10:74638–74654. DOI: 10.1109/ACCESS.2022.3190632.
641 Gurcan F, Kose C. 2017. Analysis of software engineering industry needs and trends:
642   Implications for education. *International Journal of Engineering Education* 33:1361–1368.
643 Hin D. 2020. StackOverflow vs Kaggle: A Study of Developer Discussions About Data Science.

644   *arXiv preprint arXiv:2006.08334.*

645 Internet Archive. 2022.Internet Archive: Digital Library of Free & Borrowable Books, Movies,
646   Music & Wayback Machine. *Available at https://archive.org/download/stackexchange*
647   (accessed June 19, 2022).

648 Karbasian H, Johri A. 2020. Insights for curriculum development: Identifying emerging data
649   science topics through analysis of Q and A communities. In: *SIGCSE 2020 - Proceedings of*
650   *the 51st ACM Technical Symposium on Computer Science Education.* 192–198. DOI:
651   10.1145/3328778.3366817.

652 Kochhar PS. 2016. Mining testing questions on stack overflow. In: *SoftwareMining 2016 -*
653   *Proceedings of the 5th International Workshop on Software Mining, co-located with ASE*
654   *2016.* 32–38. DOI: 10.1145/2975961.2975966.

655 Linares-Vásquez M, Dit B, Poshyvanyk D. 2013. An exploratory analysis of mobile
656   development issues using Stack Overflow. In: *IEEE International Working Conference on*
657   *Mining Software Repositories.* 93–96. DOI: 10.1109/MSR.2013.6624014.

658 Liu W, Wang Z, Liu X, Zeng N, Liu Y, Alsaadi FE. 2017. A survey of deep neural network
659   architectures and their applications. *Neurocomputing* 234:11–26. DOI:
660   10.1016/j.neucom.2016.12.038.

661 Maity SK, Sahni JSS, Mukherjee A. 2015. Analysis and prediction of question topic popularity
662   in community Q&A sites: A case study of Quora. In: *Proceedings of the 9th International*
663   *Conference on Web and Social Media, ICWSM 2015.* 238–247.

664 Mathew G, Menzies T. 2018. Software Engineering's Top Topics, Trends, and Researchers.
665   *IEEE Software.* DOI: 10.1109/MS.2018.3571230.

666 Plotnikova V, Dumas M, Milani F. 2020. Adaptations of data mining methodologies: A
667   systematic literature review. *PeerJ Computer Science* 6:1–43. DOI: 10.7717/PEERJ-
668   CS.267.

669 Řehůřek R, Sojka P. 2011. *Gensim—Statistical Semantics in Python.*

670 Rosen C, Shihab E. 2016. What are mobile developers asking about? A large scale study using
671   stack overflow. *Empirical Software Engineering.* DOI: 10.1007/s10664-015-9379-3.

672 Saltz JS, Krasteva I. 2022. Current approaches for executing big data science projects—a
673   systematic literature review. *PeerJ Computer Science* 8. DOI: 10.7717/PEERJ-CS.862.

674 Sarker IH. 2021. Data Science and Analytics: An Overview from Data-Driven Smart Computing,
675   Decision-Making and Applications Perspective. *SN Computer Science* 2:1–22. DOI:
676   10.1007/s42979-021-00765-8.

677 Schoenherr T, Speier-Pero C. 2015. Data science, predictive analytics, and big data in supply
678   chain management: Current state and future potential. *Journal of Business Logistics*
679   36:120–132. DOI: 10.1111/jbl.12082.

680 Silva CC, Galster M, Gilson F. 2021. Topic modeling in software engineering research.
681   *Empirical Software Engineering* 26. DOI: 10.1007/s10664-021-10026-0.

682 Stack Exchange. 2022.Data Science Stack Exchange. *Available at*
683   *https://datascience.stackexchange.com/* (accessed October 3, 2022).

684 Treude C, Wagner M. 2019. Predicting good configurations for github and stack overflow topic
685   models. In: *IEEE International Working Conference on Mining Software Repositories.* 84–
686   95. DOI: 10.1109/MSR.2019.00022.

687 Uddin G, Sabir F, Guéhéneuc YG, Alam O, Khomh F. 2021. An empirical study of IoT topics in
688   IoT developer discussions on Stack Overflow. *Empirical Software Engineering.* DOI:
689   10.1007/s10664-021-10021-5.

690   Vicario G, Coleman S. 2019. A review of data science in business and industry and a future
691   view. *Applied Stochastic Models in Business and Industry*. DOI: 10.1002/asmb.2488.
692   Xu B, Shirani A, Lo D, Alipour MA. 2018. Prediction of relatedness in stack overflow: Deep
693   learning vs. SVM: A reproducibility study. In: *International Symposium on Empirical*
694   *Software Engineering and Measurement*. DOI: 10.1145/3239235.3240503.
695   Yang XL, Lo D, Xia X, Wan ZY, Sun JL. 2016. What Security Questions Do Developers Ask?
696   A Large-Scale Study of Stack Overflow Posts. *Journal of Computer Science and*
697   *Technology* 31:910–924. DOI: 10.1007/s11390-016-1672-0.
698   Zhang J, Chen Y, Zhao Y, Wolfram D, Ma F. 2020. Public health and social media: A study of
699   Zika virus-related posts on Yahoo! Answers. *Journal of the Association for Information*
700   *Science and Technology* 71:282–299. DOI: 10.1002/asi.24245.
701   Zou J, Xu L, Guo W, Yan M, Yang D, Zhang X. 2015. Which non-functional requirements do
702   developers focus on? An empirical study on stack overflow using topic analysis. In: *IEEE*
703   *International Working Conference on Mining Software Repositories*. DOI:
704   10.1109/MSR.2015.60.
705   Zuo Y, Wu J, Zhang H, Lin H, Wang F, Xu K, Xiong H. 2016. Topic modeling of short texts: A
706   pseudo-document view. In: *Proceedings of the ACM SIGKDD International Conference on*
707   *Knowledge Discovery and Data Mining*. 2105–2114. DOI: 10.1145/2939672.2939880.
708

Details about the topics generated by LDA

1   **Table 1** Details about the topics generated by LDA

Topic Name	Descriptive LDA keywords	%
Model Training	model training train set datum test accuracy dataset validation loss epoch learn split parameter different performance increase prediction run rate cross improve time best hyperparameter	8.88
Machine Learning	learn datum machine algorithm problem learning project deep example ml approach task want method best analysis technique research model lot book idea give field kind	7.81
Neural Networks	network layer neural output input weight number size lstm activation neuron convolutional function learn architecture cnn hide keras batch example sequence train convolution deep nn	7.76
NLP	word text document sentence model example topic want nlp extract embedding language name give tag entity sentiment bert token way embed similar task approach corpus	7.20
Time Series	time datum series predict model day prediction lstm value sequence event problem month forecast anomaly want point different example dataset future window time-series give input	6.04
Data Manipulation	column datum row value dataframe want dataset file name table number list create code way contain panda id python string csv convert format frame different example date output set multiple	6.03
Feature Engineering	feature variable value datum categorical dataset target correlation encode scale numerical continuous model column problem category way predict transform apply different binary independent want encoding	5.91
Regression Models	regression feature model tree linear random decision forest logistic xgboost algorithm importance prediction parameter method selection predict fit boost coefficient variable best sklearn different classifier	5.86
Statistics	datum distribution sample number point mean value probability set way variance standard outlier example test group parameter give case different estimate random method approach size	5.48
Tools	python run file datum code r package memory library gpu time google notebook spark script tool tensorflow way database version want create load process orange	5.39
Computer Vision	image object model cnn dataset detection train box want detect pixel different video color segmentation picture example recognition size label problem map face classify extract	5.35
Classification	class classification label dataset classifier problem datum binary model sample predict classify probability imbalance set example train training prediction positive give category accuracy case bayes	4.98
Recommendation Systems	user product customer item datum base system predict want problem model dataset sale recommendation category number rating approach give price buy example build purchase group	4.81
Loss Functions	function loss gradient weight amp value equation cost error descent update give calculate formula problem mean derivative optimization parameter compute square sum entropy term define	4.55
Evaluation	score plot value metric negative curve precision positive calculate matrix mean graph show recall line average threshold confusion code true compare evaluate evaluation interpret performance	3.63
Clustering	cluster algorithm action state graph point clustering reward node distance k-means game agent reinforcement problem policy player learn value give method number take different space	3.57
Coding Errors	error code array shape model keras tensorflow input return function import tensor run custom numpy line output valueerror fault call expect problem message pass create	3.54
Dimensionality Reduction	vector matrix dimension feature similarity distance pca space autoencoder kernel point representation datum attention component dimensionality embed example cosine represent encoder reduction reduce dimensional calculate	3.21

Temporal trends of the topics

1   **Table 2** Temporal trends of the topics

Topic Name	Total Rate	Yearly Rates (%)									Trend Value	Trend
		2014	2015	2016	2017	2018	2019	2020	2021	2022		
Model Training	8.88	2.33	3.45	5.44	6.41	8.26	9.74	10.56	10.39	9.87	7.54	⇑
Machine Learning	7.81	19.50	17.83	11.88	9.70	7.72	7.07	6.09	5.87	5.96	-13.54	⇓
Neural Networks	7.76	1.79	3.88	7.90	11.48	9.74	7.89	7.10	6.86	5.69	3.90	⇑
NLP	7.20	8.41	7.84	7.71	6.92	5.91	6.54	7.97	7.81	7.67	-0.74	⇓
Time Series	6.04	5.19	3.10	5.35	5.01	5.95	5.99	5.87	6.62	8.61	3.43	⇑
Data Manipulation	6.03	4.29	6.63	6.06	5.25	6.46	6.69	5.65	6.07	5.24	0.94	⇑
Feature Engineering	5.91	4.83	5.60	4.64	5.25	5.84	5.79	6.51	6.14	6.57	1.74	⇑
Regression Models	5.86	3.40	6.03	6.34	5.90	5.52	5.65	6.15	5.61	7.02	3.62	⇑
Statistics	5.48	5.37	5.94	6.34	5.59	5.37	4.71	6.00	5.58	5.24	-0.13	⇓
Tools	5.39	17.89	9.91	7.57	6.10	5.48	5.12	4.25	4.19	4.06	-13.83	⇓
Computer Vision	5.35	1.97	1.89	2.74	4.77	5.90	5.55	5.61	6.57	5.35	3.38	⇑
Classification	4.98	5.01	4.48	3.79	5.07	5.09	5.00	4.91	5.36	5.09	0.08	⇑
Recommendation Systems	4.81	8.23	6.55	7.62	5.62	4.32	4.01	4.48	4.47	4.74	-3.49	⇓
Loss Functions	4.55	2.15	3.79	3.74	4.77	4.79	5.12	4.96	4.03	4.10	1.95	⇑
Evaluation	3.63	1.25	3.19	3.41	3.12	3.00	3.55	4.51	3.78	4.21	2.96	⇑
Clustering	3.57	3.94	4.82	4.40	3.81	4.07	3.43	2.93	3.37	3.34	-0.60	⇓
Coding Errors	3.54	0.72	0.78	1.61	2.26	3.59	4.84	3.60	3.98	3.80	3.08	⇑
Dimensionality Reduction	3.21	3.76	4.31	3.45	2.98	2.99	3.31	2.85	3.30	3.45	-0.30	⇓

Descriptive indicators of the data science topics

1   **Table 3** Descriptive indicators of the data science topics

Topic Name	Question Count (%)	Question Count (#)	View Count (#)	Answer Count (#)	Comment Count (#)	Favorite Count (#)	Score (#)	Accepted Answer Count (%)
Model Training	8.88	2973	1656	1.14	1.23	0.57	1.89	33.10
Machine Learning	7.81	2615	1485	1.40	1.24	1.22	2.84	33.19
Neural Networks	7.76	2598	2060	1.04	0.93	1.02	2.72	37.34
NLP	7.20	2411	1211	1.07	0.96	0.55	1.84	29.45
Time Series	6.04	2023	828	0.90	1.10	0.52	1.51	20.56
Data Manipulation	6.03	2020	4491	1.23	1.05	0.40	1.61	41.39
Feature Engineering	5.91	1979	2007	1.23	1.10	0.88	2.15	35.17
Regression Models	5.86	1964	2079	1.16	1.00	0.76	2.29	38.19
Statistics	5.48	1835	983	1.02	1.15	0.53	1.79	29.92
Tools	5.39	1805	2308	1.15	1.06	0.73	2.48	32.02
Computer Vision	5.35	1791	911	0.87	0.96	0.42	1.57	25.52
Classification	4.98	1668	1923	1.16	1.22	0.75	2.29	33.75
Recommendation Systems	4.81	1611	554	1.06	1.02	0.37	1.37	27.25
Loss Functions	4.55	1525	1538	1.01	0.87	0.86	2.44	38.30
Evaluation	3.63	1217	1922	1.09	1.06	0.48	1.81	36.15
Clustering	3.57	1197	1245	0.99	1.14	0.57	1.93	32.25
Coding Errors	3.54	1186	3771	0.96	1.03	0.33	1.38	32.55
Dimensionality Reduction	3.21	1074	1299	1.04	0.80	0.81	2.44	33.05

Time-based trends in the top 50 data-driven technologies

1   **Table 4** Time-based trends in the top 50 data-driven technologies

Tags	Total Rate	Yearly Rates (%)									Trend Value	Trend
		2014	2015	2016	2017	2018	2019	2020	2021	2022		
machine-learning	12.28	15.63	16.61	13.19	13.77	13.08	12.33	11.53	10.80	11.13	-4.50	⇓
python	7.28	3.07	4.91	6.15	6.56	7.43	8.34	7.36	7.51	6.66	3.59	⇑
deep-learning	5.28	0.65	1.22	3.08	5.78	6.30	5.18	5.14	5.79	5.96	5.31	⇑
neural-network	4.85	2.33	3.98	5.53	6.66	6.60	4.75	4.30	3.65	3.54	1.21	⇑
classification	3.57	6.79	5.58	4.04	3.88	3.41	3.35	3.31	3.27	3.78	-3.01	⇓
keras	3.11	0.00	0.08	1.19	2.37	3.76	4.10	3.36	3.12	2.37	2.37	⇑
nlp	2.78	3.63	2.31	2.35	2.37	2.10	2.57	3.03	3.50	3.37	-0.26	⇓
scikit-learn	2.48	1.30	1.76	2.85	2.47	2.50	2.58	2.68	2.36	2.11	0.81	⇑
tensorflow	2.46	0.00	0.17	1.37	2.36	2.57	2.39	2.76	3.07	2.50	2.50	⇑
time-series	1.97	2.14	1.09	1.93	1.76	1.68	1.92	1.76	2.38	2.92	0.78	⇑
regression	1.68	1.40	1.80	1.63	1.90	1.58	1.56	1.89	1.49	1.94	0.55	⇑
r	1.67	4.93	6.38	4.20	2.79	1.42	1.11	1.15	0.95	1.21	-3.72	⇓
dataset	1.60	4.00	2.85	1.89	1.72	1.39	1.43	1.49	1.67	1.33	-2.67	⇓
cnn	1.53	0.00	0.00	0.02	0.43	1.90	2.19	1.73	1.70	1.30	1.30	⇑
clustering	1.53	3.44	3.10	2.59	1.79	1.43	1.33	1.29	1.30	1.42	-2.02	⇓
pandas	1.41	0.74	0.34	1.01	1.16	1.44	1.81	1.48	1.41	1.24	0.50	⇑
data-mining	1.34	6.79	5.62	3.06	2.19	1.25	1.00	0.77	0.65	0.66	-6.13	⇓
lstm	1.32	0.00	0.00	0.02	0.55	1.70	1.67	1.44	1.26	1.90	1.90	⇑
predictive-modeling	1.31	1.58	2.98	3.00	1.59	1.13	1.22	1.08	0.92	1.04	-0.54	⇓
statistics	1.21	3.63	2.52	1.45	1.08	0.94	0.99	1.30	1.22	1.27	-2.36	⇓
feature-selection	1.05	1.77	1.01	1.27	1.19	0.90	1.01	1.05	1.08	0.90	-0.86	⇓
random-forest	0.93	0.93	1.59	1.05	1.11	0.78	0.77	1.07	0.87	0.87	-0.06	⇓
data	0.90	0.00	0.96	1.13	0.94	0.72	0.93	0.89	0.94	1.00	1.00	⇑
image-classification	0.85	0.00	0.71	0.80	0.67	0.91	0.93	0.85	0.95	0.70	0.70	⇑
rnn	0.84	0.00	0.17	0.60	0.89	1.35	0.94	0.78	0.57	0.80	0.80	⇑
machine-learning-model	0.82	0.00	0.00	0.02	0.00	0.68	1.00	0.96	1.20	1.47	1.47	⇑
decision-trees	0.82	0.09	1.01	1.00	0.93	0.91	0.68	0.89	0.82	0.63	0.53	⇑
data-cleaning	0.81	1.21	1.30	1.27	0.78	0.76	0.68	0.77	0.85	0.66	-0.55	⇓
linear-regression	0.80	0.00	0.96	0.96	0.58	0.90	0.82	0.85	0.66	0.92	0.92	⇑
convolutional-neural-network	0.79	0.00	0.08	1.11	1.50	0.96	0.50	0.32	1.18	0.84	0.84	⇑
logistic-regression	0.75	1.40	1.38	0.84	0.83	0.60	0.70	0.73	0.70	0.92	-0.48	⇓
xgboost	0.73	0.00	0.29	0.86	1.04	0.49	0.79	0.89	0.65	0.70	0.70	⇑
training	0.71	0.00	0.00	0.38	0.64	0.68	0.63	0.92	0.86	0.86	0.86	⇑
visualization	0.69	1.77	1.22	1.41	1.00	0.54	0.60	0.58	0.62	0.43	-1.34	⇓
cross-validation	0.68	0.47	0.34	0.76	0.55	0.70	0.72	0.86	0.59	0.57	0.10	⇑
reinforcement-learning	0.68	0.19	0.38	0.38	0.64	0.90	0.83	0.53	0.66	0.70	0.52	⇑
feature-engineering	0.68	0.00	0.21	0.58	0.65	0.56	0.76	0.73	0.80	0.73	0.73	⇑
pytorch	0.68	0.00	0.00	0.00	0.18	0.26	0.69	0.81	1.35	1.15	1.15	⇑
computer-vision	0.67	0.09	0.29	0.40	0.43	0.85	0.52	0.69	0.83	1.01	0.92	⇑
text-mining	0.65	3.35	1.93	1.43	1.22	0.67	0.45	0.40	0.32	0.28	-3.07	⇓
data-science-model	0.63	0.00	0.00	0.00	0.01	0.44	0.85	0.77	0.87	1.03	1.03	⇑
svm	0.63	1.40	1.47	0.68	1.11	0.52	0.57	0.59	0.46	0.43	-0.97	⇓
loss-function	0.59	0.00	0.08	0.28	0.29	0.52	0.70	0.69	0.73	0.72	0.72	⇑
class-imbalance	0.58	0.47	0.38	0.36	0.53	0.54	0.61	0.67	0.72	0.44	-0.02	⇓
multiclass-classification	0.58	0.00	0.76	0.46	0.57	0.55	0.59	0.56	0.63	0.72	0.72	⇑
preprocessing	0.56	0.00	0.59	0.44	0.50	0.52	0.55	0.57	0.64	0.70	0.70	⇑
optimization	0.52	1.02	0.71	0.62	0.42	0.43	0.50	0.50	0.57	0.61	-0.41	⇓
word-embeddings	0.52	0.00	0.42	0.48	0.37	0.48	0.56	0.56	0.53	0.69	0.69	⇑
recommender-system	0.51	2.05	0.96	0.92	0.57	0.42	0.37	0.44	0.46	0.58	-1.46	⇓
unsupervised-learning	0.51	0.00	0.29	0.64	0.60	0.49	0.48	0.55	0.50	0.46	0.46	⇓

Top 25 most commonly used tasks, algorithms, and tools in data science.

1   **Table 5** Top 25 most commonly used tasks, algorithms, and tools in data science.

Tasks	%	Algorithms	%	Tools	%
machine-learning	12.28	neural-network	4.85	python	7.28
deep-learning	5.28	cnn	2.32	keras	3.11
classification	3.57	lstm	1.32	scikit-learn	2.48
nlp	2.78	random-forest	0.93	tensorflow	2.46
time-series	1.97	rnn	0.84	r	1.67
regression	1.68	decision-trees	0.82	pandas	1.41
clustering	1.53	linear-regression	0.80	pytorch	0.68
data-mining	1.34	logistic-regression	0.75	numpy	0.40
predictive-modeling	1.31	xgboost	0.73	word2vec	0.39
statistics	1.21	svm	0.63	dataframe	0.37
feature-selection	1.05	loss-function	0.59	apache-spark	0.27
image-classification	0.85	word-embeddings	0.52	orange	0.20
data-cleaning	0.81	k-means	0.50	python-3.x	0.20
training	0.71	gradient-descent	0.47	matplotlib	0.19
visualization	0.69	autoencoder	0.38	apache-hadoop	0.11
cross-validation	0.68	pca	0.38	matlab	0.09
reinforcement-learning	0.68	bert	0.35	gensim	0.07
feature-engineering	0.68	backpropagation	0.33	kaggle	0.04
computer-vision	0.67	gan	0.26	gridsearchcv	0.03
text-mining	0.65	naive-bayes-classifier	0.23	grid-search	0.03
class-imbalance	0.58	graphs	0.22	jupyter	0.03
multiclass-classification	0.58	k-nn	0.15	vgg16	0.03
preprocessing	0.56	yolo	0.03	opencv	0.02
optimization	0.52	lda	0.03	sql	0.02
recommender-system	0.51	smote	0.02	orange3	0.02

Topics and related data-driven technologies

1   **Table 6** Topics and related data-driven technologies

Topic Name	Related Tags
Model Training	machine-learning deep-learning neural-network keras python cross-validation tensorflow scikit-learn classification training cnn overfitting accuracy hyperparameter-tuning random-forest
Machine Learning	machine-learning deep-learning python neural-network classification data-mining nlp dataset data statistics bigdata predictive-modeling scikit-learn clustering algorithms
Neural Networks	neural-network deep-learning machine-learning keras cnn tensorflow python lstm convolutional-neural-network rnn convolution pytorch classification backpropagation regression
NLP	nlp machine-learning python text-mining word-embeddings deep-learning word2vec bert classification neural-network named-entity-recognition sentiment-analysis text-classification nltk data-mining
Time Series	time-series machine-learning lstm python deep-learning forecasting predictive-modeling keras neural-network rnn regression classification anomaly-detection prediction tensorflow
Data Manipulation	python pandas machine-learning r dataframe data-cleaning dataset scikit-learn data data-mining numpy clustering classification visualization time-series
Feature Engineering	machine-learning python scikit-learn feature-selection categorical-data feature-engineering regression classification correlation preprocessing feature-scaling data-cleaning neural-network r clustering
Regression Models	machine-learning decision-trees scikit-learn python random-forest regression feature-selection classification xgboost linear-regression logistic-regression r predictive-modeling neural-network deep-learning
Statistics	machine-learning statistics python classification regression neural-network deep-learning clustering probability data-mining dataset distribution scikit-learn r outlier
Tools	python machine-learning tensorflow r deep-learning bigdata gpu dataset apache-spark keras orange scikit-learn neural-network data apache-hadoop
Computer Vision	deep-learning machine-learning cnn image-classification neural-network computer-vision python tensorflow keras object-detection image-recognition convolutional-neural-network classification gan dataset
Classification	classification machine-learning class-imbalance python multiclass-classification scikit-learn deep-learning neural-network multilabel-classification keras dataset naive-bayes-classifier nlp svm tensorflow
Recommendation Systems	machine-learning recommender-system python predictive-modeling classification data-mining clustering statistics dataset regression r nlp deep-learning time-series machine-learning-model
Loss Functions	machine-learning neural-network gradient-descent deep-learning loss-function optimization backpropagation python classification linear-regression logistic-regression mathematics regression svm tensorflow
Evaluation	machine-learning python classification visualization scikit-learn evaluation r matplotlib confusion-matrix metric regression deep-learning plotting seaborn keras
Clustering	clustering machine-learning reinforcement-learning python k-means deep-learning q-learning graphs neural-network scikit-learn unsupervised-learning dqn data-mining classification dbscan
Coding Errors	python keras tensorflow machine-learning deep-learning scikit-learn neural-network numpy lstm cnn pytorch pandas nlp classification r
Dimensionality Reduction	machine-learning dimensionality-reduction deep-learning pca neural-network nlp python svm autoencoder clustering similarity classification transformer scikit-learn word-embeddings

Figure 1

The number of questions and answers by year
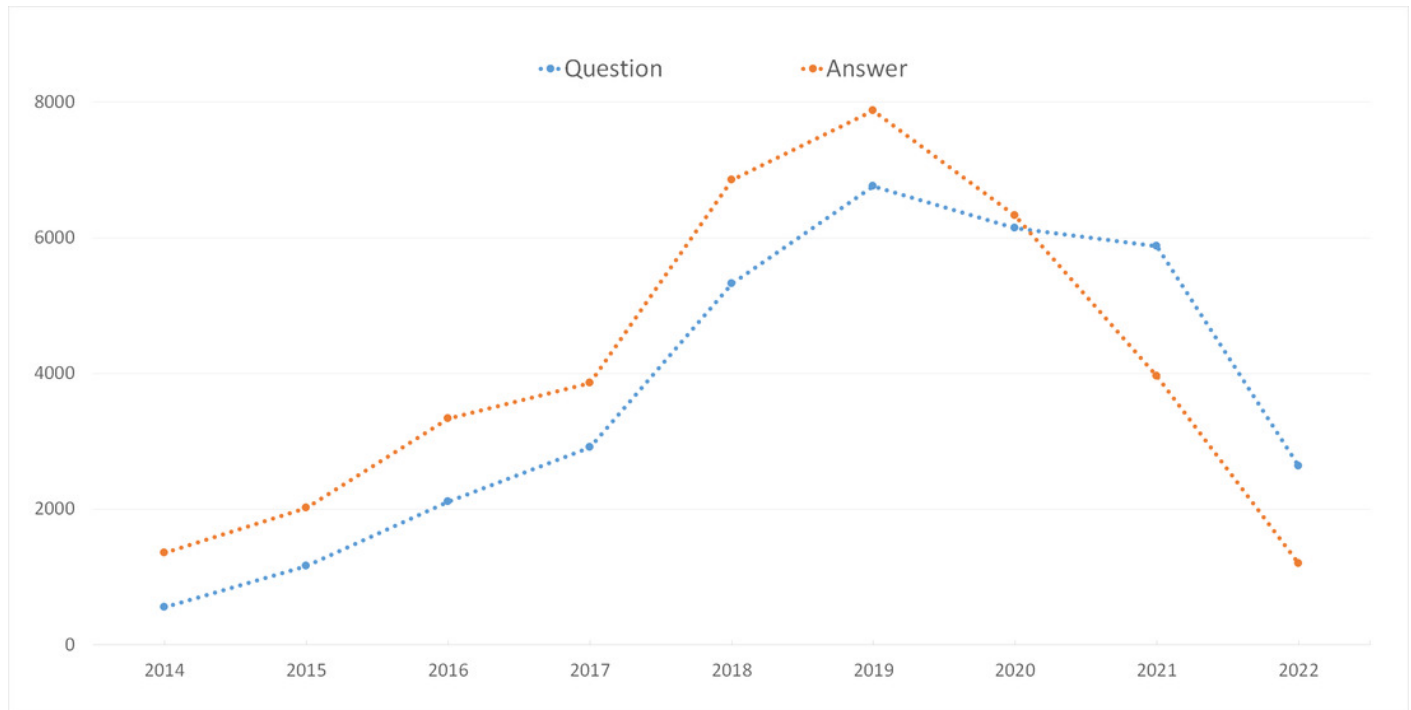
Figure 2

Coherence scores calculated for each number of topics
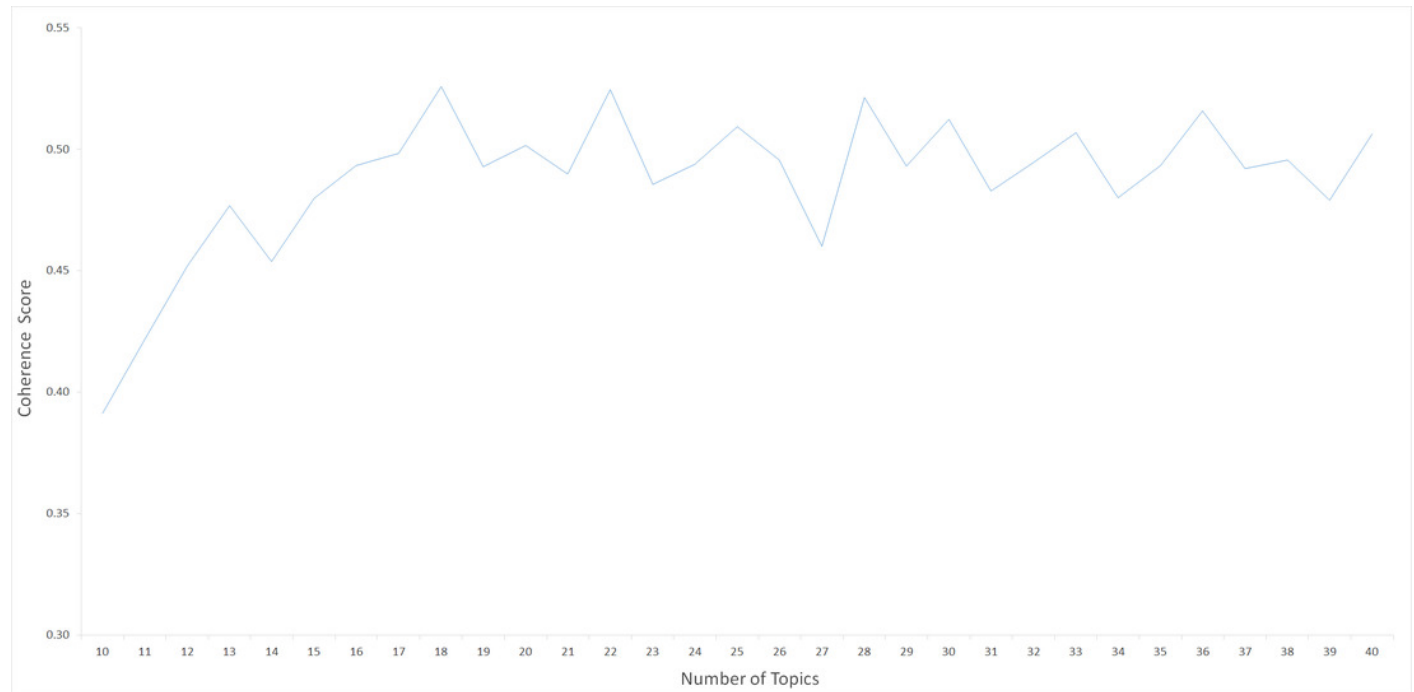
Figure 3

The topics with increasing and decreasing trends
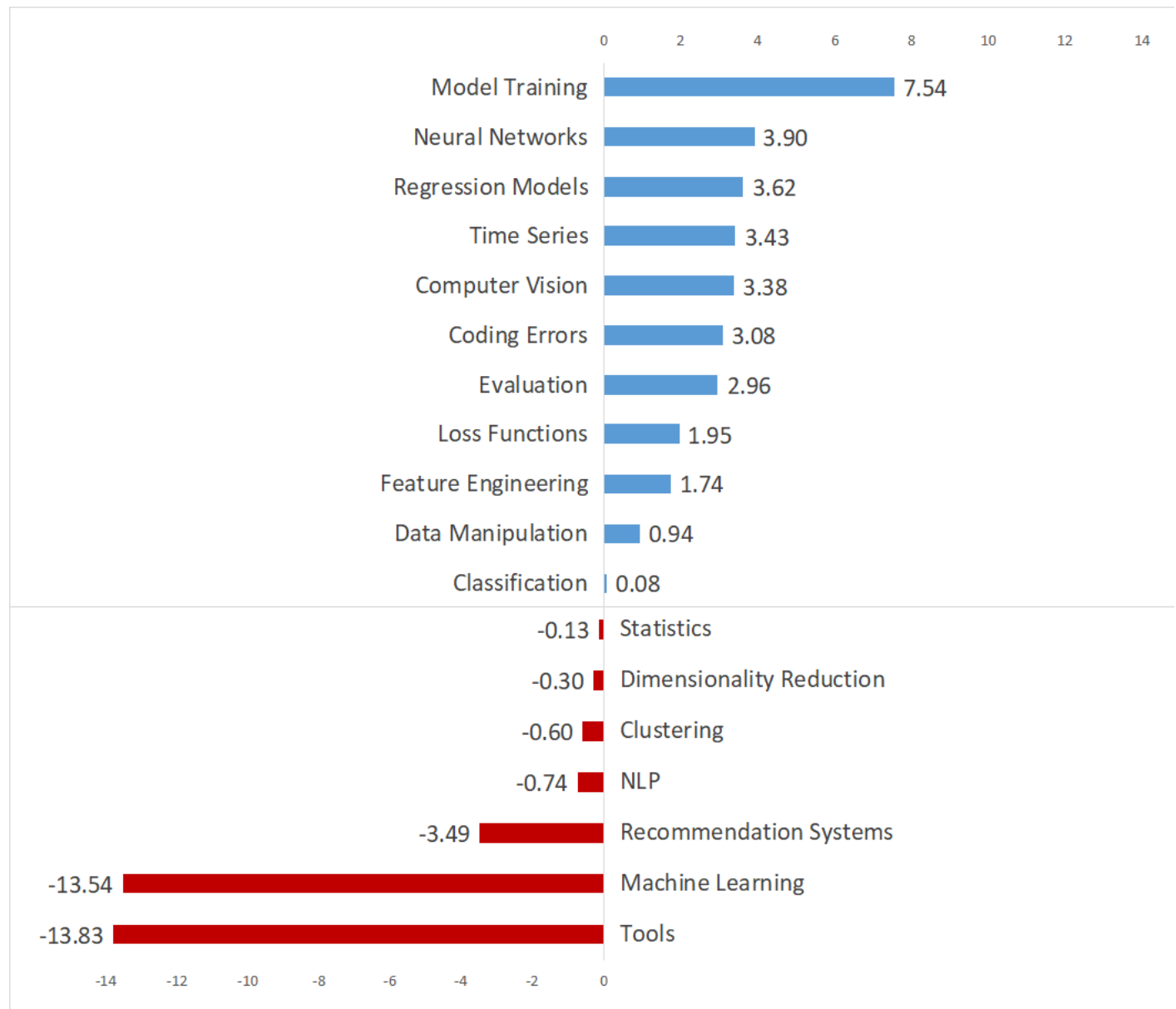
Figure 4

Average number of views of the topics
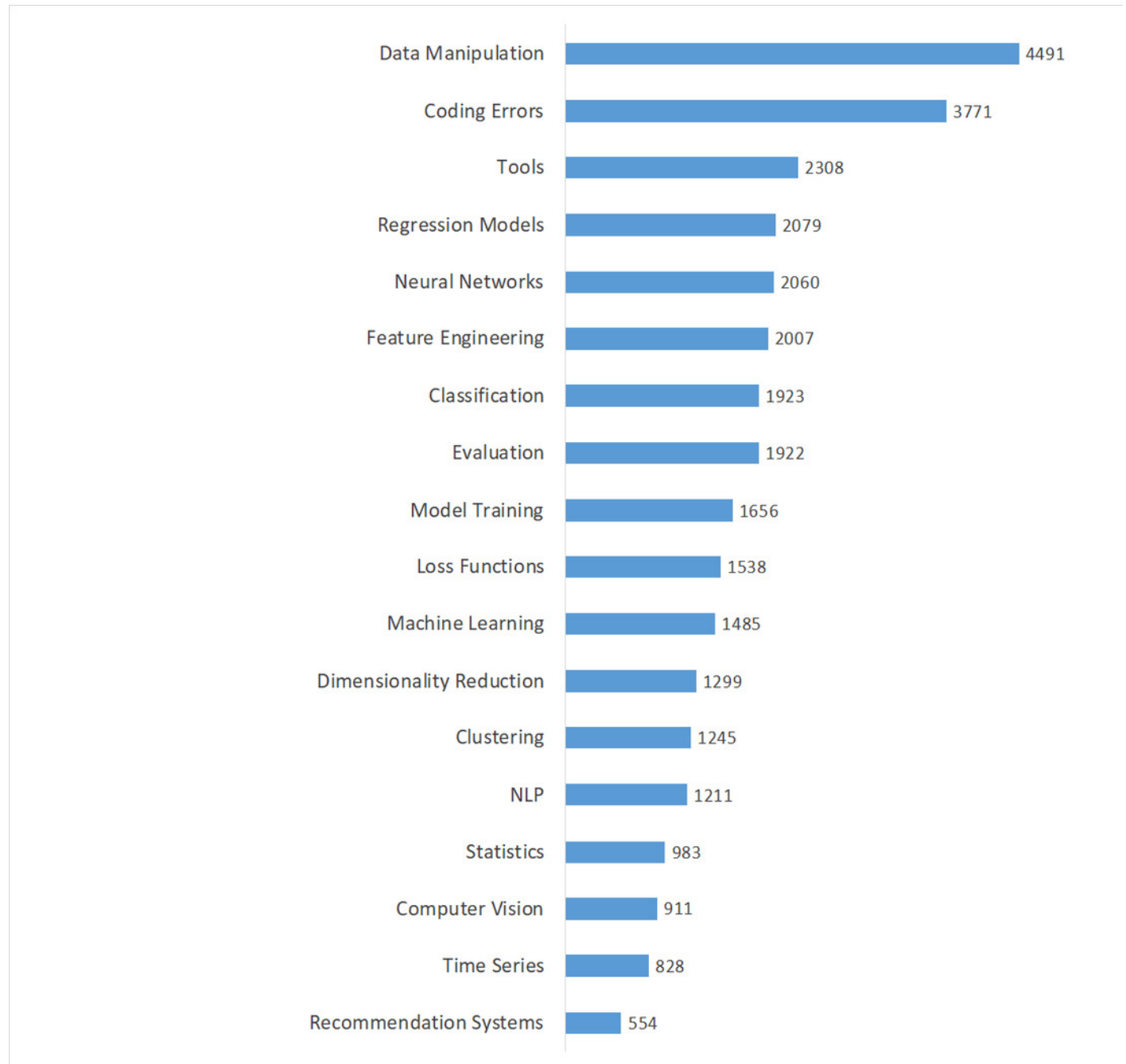
Figure 5

Percentage of accepted answers of the topics

Topic	Value
Data Manipulation	41
Loss Functions	38
Regression Models	38
Neural Networks	37
Evaluation	36
Feature Engineering	35
Classification	34
Machine Learning	33
Model Training	33
Dimensionality Reduction	33
Coding Errors	33
Clustering	32
Tools	32
Statistics	30
NLP	29
Recommendation Systems	27
Computer Vision	26
Time Series	21
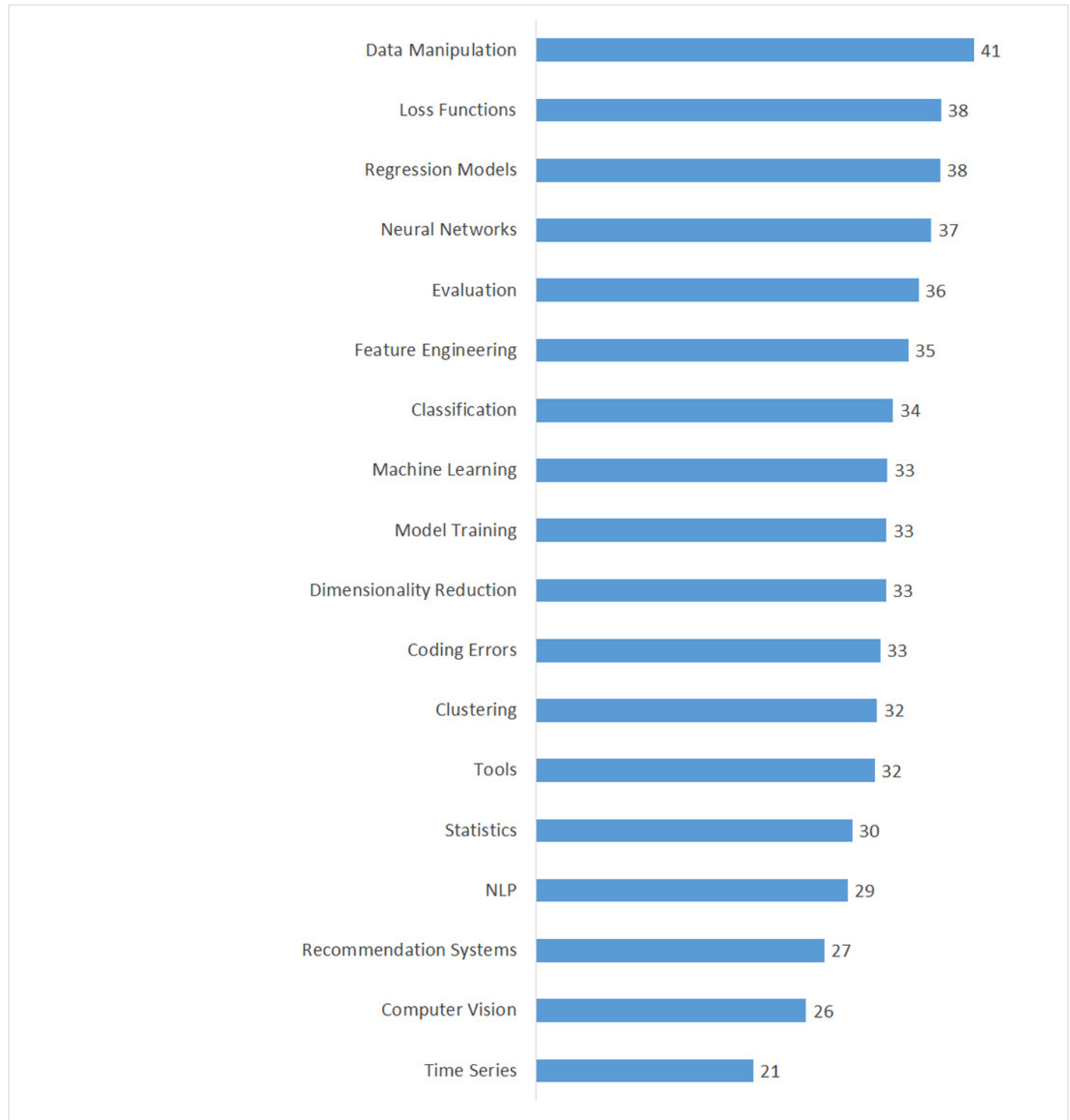
Figure 6

Percentages of top 20 data-driven technologies
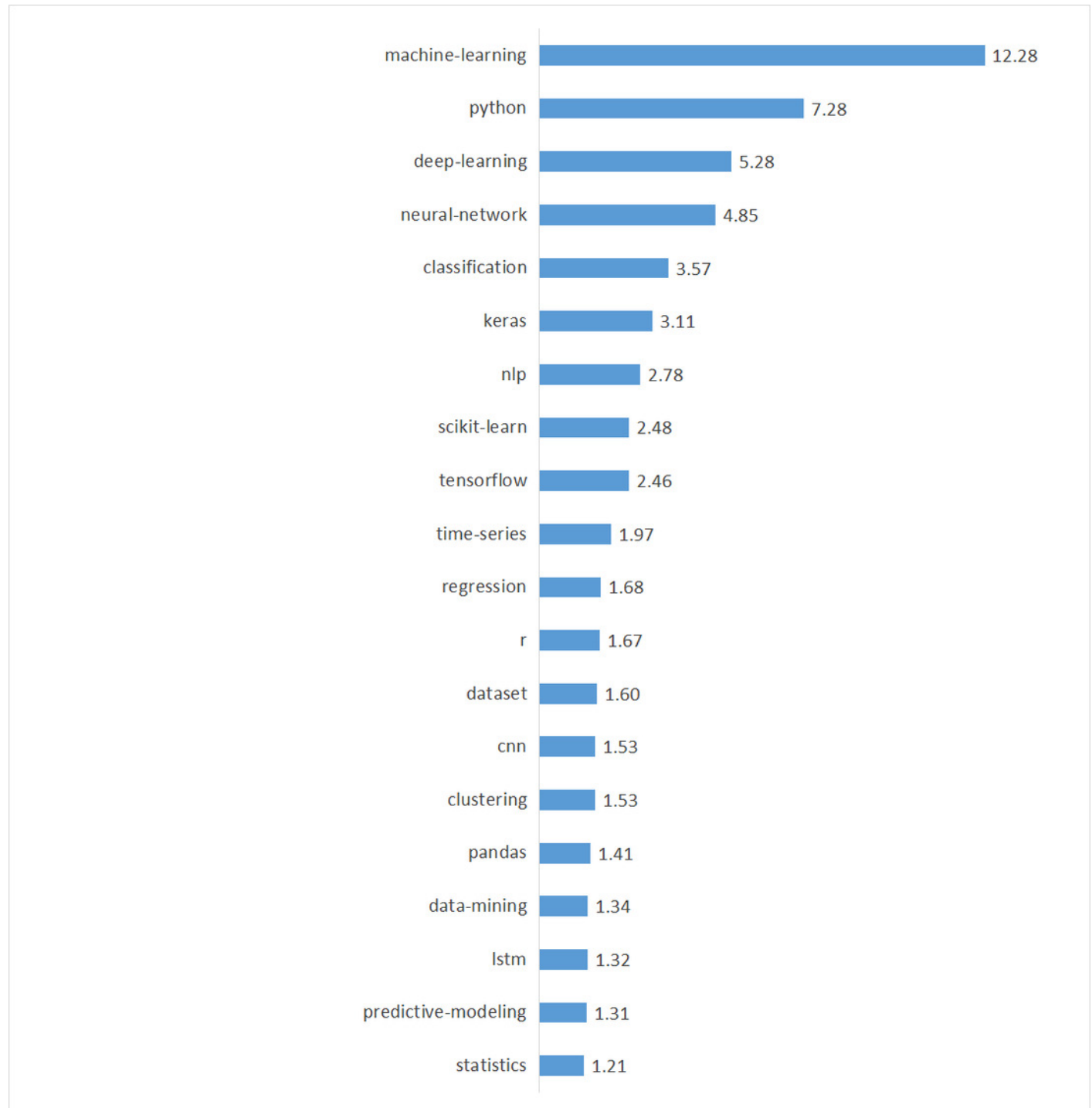
Figure 7

The top 15 data-driven technologies with increasing and decreasing trends
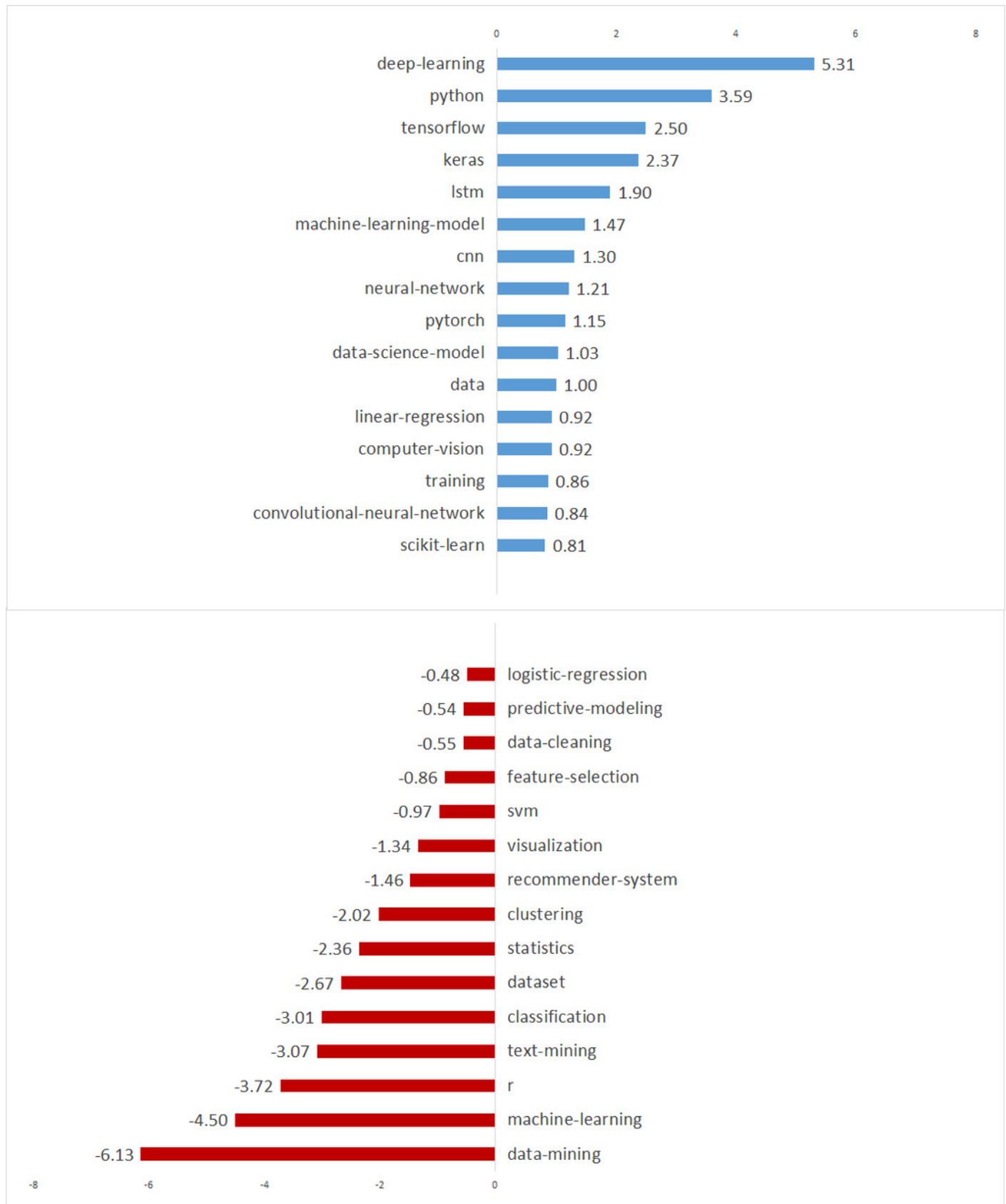
Figure 8

Percentages of data-driven technologies for the top three topics