proceedings

MDPI

Proceedings

A Jaccard Similarity-Based Model to Match Stakeholders for Collaboration in an Industry-Driven Portal †

İnanç Kabasakal * and Haluk Soyuer

Faculty of Economics & Administrative Sciences, Department of Business Administration, Ege University, İzmir 35040, Turkey; haluk.soyuer@ege.edu.tr

* Correspondence: inanc.kabasakal@ege.edu.tr; Tel.: +90-0506-601-1077

† Presented at the 7th International Management Information Systems Conference, Online, 9–11 December 2020.

Abstract: The industry–university collaboration has been emphasized for innovation and economic development in the Triple Helix Model. To facilitate this collaboration often necessitates implementing interfaces between stakeholders. EGEVASYON, an industry-driven platform coined from the combination of innovation and Ege for the Aegean Region of Turkey, has been proposed to foster collaboration by involving researchers in industry projects. Moreover, the platform has a portal project under development, where researchers can receive recommendations among ongoing projects. Our study presents the use of Jaccard similarity measure in this recommendation model. Moreover, recommendation selection is demonstrated using a sample dataset of EU projects.

Keywords: Jaccard similarity; university–government–industry collaboration; triple helix model

1. Introduction

The collaboration among universities and the industry is often emphasized as an essential factor that facilitates innovation. The triple helix model implies the necessity of communication and interaction across plural parties that paves the way to knowledge-intensive industries and innovation [1]. Moreover, it is a necessity for universities to understand the challenges and problems being dealt with by the industry and prepare graduates accordingly [2]. There is substantial work in prior research that covers the role of government to establish convenient environments for collaboration between university and industry. Despite the opportunities, a variety of factors inhibit university–industry collaboration. In 2010, those barriers were extensively explored in [3], and one of the primary issues noted was the difference of focus in the stakeholders. The study argued that academic researchers are orientated towards pure science in the long term, while the industry expects more practical outputs in the short run. Furthermore, [4] pointed out significant differences between the university-driven and industry-driven innovation ecosystems. In either way, the outputs of collaboration are mostly complex and intangible, and their benefits are often observed in the long run [5], which complicates performance evaluation in university–industry collaboration.

A variety of actors take place in establishing relations across stakeholders to increase university–industry collaboration, including the Technology Transfer Offices. As another stakeholder in the ecosystem, EGEVASYON Center has been planned as an industry-driven interface to foster innovation and improve the degree of collaboration among researchers and the industry in Turkey. The principal qualification of the center is its industry-driven orientation, which aims to attract academic researchers in R&D projects conducted in the industry. Thereby, a priority for the center is to increase the degree of cooperation between the industry and researchers in developing solutions towards the challenges in industrial projects, while facilitating the commercialization of solutions brought up from academicians. Additionally, the EGEVASYON project is planned to host

Citation: Kabasakal, İ.; Soyuer, H. A Jaccard Similarity-Based Model to Match Stakeholders for Collaboration in an Industry-Driven Portal. *Proceedings* **2021**, *74*, 15. https://doi.org/10.3390/proceedings2021074015

Published: 11 March 2021

a web portal that aims to establish an interface among the stakeholders. When launched, researchers and businesses will be able to sign up and exhibit their expertise by filling out details in their profile, such as projects conducted or training programs participated in. Moreover, the users will be able to search and browse other public profiles, contact other users, and announce calls for collaboration.

A feature of the web portal for researchers is to facilitate finding relevant projects regarding their expertise and background. On the other hand, such function is also beneficial for the industry since their R&D projects might attract attention from researchers among portal visitors, who potentially could provide valuable contributions. Such a feature might be described as a match-making process that uncovers opportunities for cooperation, which are mutually beneficial. As a preliminary work towards achieving this functionality, this study adopts a method to match relevant projects with researchers represented as a list of keywords. Such a problem might be defined as the following: "Among a list of projects defined by a list of tokens, what is the most relevant one given a list of particular keywords (tokens)?"

This study follows a similarity-based approach to figure out which item is more relevant to a given set of terms describing researchers and research items/projects. The following section introduces the method being employed for measuring similarity. Subsequently, the third section demonstrates which projects are more similar to each other, and provides an example where a set of keywords are matched to the most relevant project, using the similarity measure. Finally, the potential benefits of this approach will be discussed, along with its limitations.

2. Materials and Methods

As noted, our study addresses a requirement in a university–industry collaboration web portal project and deals with the problem in a similarity-based model. In particular, measuring the similarity between researchers and research items helps to generate recommendations and provide opportunities for collaboration. This sort of a "matchmaking" described in our context corresponds to the problem of picking the most relevant research items for a researcher—or vice versa.

Term-based similarity measures in text mining involve the Euclidian distance, cosine similarity, Manhattan distance, Dice's coefficient, Jaccard similarity, and the simple matching coefficient [6]. Jaccard and cosine measures incorporate different approaches in measuring similarity [7], yet both provide similar results when employed [8]. Our case requires the calculation of similarity among keywords in corpus and might be dealt with as a text-mining problem.

A collaborative filtering approach in recommender systems often computes the similarity of user ratings with measures/methods including the Jaccard coefficient, cosine similarity, Pearson's correlation coefficient, mean square distance, factored item-based similarity, and weighted regularized matrix factorization [9]. In data mining, grouping similar data points is an essential step for clustering that requires the use of similarity measures [10] (p. 116). Among the measures listed above, the Jaccard similarity is often used in document clustering for calculating similarity with respect to the words included [11]. Moreover, text documents represented as document matrices can be classified using the k-nearest neighbors algorithm, where the occurrences of words can be computed by Jaccard coefficient [12] (p. 107).

The next subsection presents a brief overview of this similarity measure. Moreover, our study involves mapping project descriptions into a global list of tokens before the analysis. Accordingly, data collection, tokenization, and preprocessing will also be detailed separately in subsections before similarity calculation.

2.1. Jaccard Similarity Coefficient

Jaccard coefficient has been introduced in [13] and has been extensively used to measure the similarity of sets. The coefficient is often used as a similarity measure for sparse binary data sets since it takes the common or disjoint elements in two sets [14] (p. 76).

The use of the Jaccard coefficient is ubiquitous in various applications. The coefficient is occasionally used to measure the similarity of text, which corresponds to quantitative, multidimensional data when handled as a bag of words [14] (p. 75). In addition, the measure might be used in social network analyses [15] to find out the similarity of authors in terms of focal research areas, through a comparison of keywords in publications. Moreover, the measure was used to group machines regarding their components in cellular manufacturing [16]. Another exciting use of the Jaccard coefficient was to evaluate the lesion detection performance of a computer-aided diagnosis system, in comparison with the manual detections from radiologists [17]. Furthermore, a study by Lu et al. [18] improved the scalability of a news recommendation system by integrating the coefficient into k-means clustering when measuring the similarity (distance) between users.

Given its frequent use for similar tasks in prior research and considering its ease of use in implementation, the Jaccard coefficient was found suitable to calculate project–researcher similarity in this preliminary study. The formal definition of Jaccard similarity measure (or the Jaccard's coefficient) is given below:

For sets S_X and S_Y with binary representations $\overline{X}$ and $\overline{Y}$, Jaccard's coefficient is a symmetric overlap measure [19] which equals:

$$J(\overline{X},\overline{Y}) = \frac{|S_x \cap S_Y|}{|S_x \cup S_Y|} \quad (1)$$

The formula in (1) might also been generalized for multiple sets as in the following:

$$J(S_{i=1..k}) = \frac{|\cap\ S_i|}{|\cup\ S_i|} \quad (2)$$

Calculating the Jaccard index for disjoint sets results in 0 since the count of common elements is zero. In contrast, the calculation of this measure for two sets involving the same members would result in 1. The Jaccard coefficient for any two sets is within the range [0, 1]. A higher score indicates a large number of common items in comparison with the total count of elements in sets. Accordingly, high scores for the Jaccard coefficient signify a high degree of similarity. However, specifying thresholds might depend on the problem domain.

Prior research involves various types of problems where the coefficient has been in use. Jaccard coefficient is applicable to measure the similarity of words in search engines by comparing letters to handle mistyped phrases [20]. Park and Kim [21] utilized the measure for comparing keywords obtained from travelers with those already present in websites. Accordingly, the authors calculated dissimilarities between both lists and utilized the differences to improve recommendations. Additionally, Yu et al. [22] proposed a novel method to generate recommendations in both a content-based and collaborative-filtering-based approach, where the Jaccard distance was utilized to compare recommendations represented by sets including both items and target users. The measure has also been adopted by Fletcher and Islam [23] to measure the similarity of patterns obtained through data mining. Moreover, the authors demonstrated how the interpretability and usability of association rules might be improved through this approach.

The Jaccard coefficient might also be used in data mining when measuring the similarity of transactions in market basket analysis, since such data involves Boolean attributes [24]. Furthermore, Singh et al. [25] proposed the similarity calculation for encrypted data points using Jaccard distance and demonstrated the use of the measure in privacy-preserving data mining.

2.2. Dataset

Our analysis utilizes a list of Horizon 2020 projects involved in a dataset [26] publicly shared over the Kaggle platform. Although there is a topic-based classification of projects, such categorization was made based on the predefined programs related to the call. Moreover, the dataset did not involve a selection of keywords that describe the projects. Even so, a variety of attributes were present in the dataset, as listed below:

- Total budget;
- Start and end dates;
- Relevant programs in Horizon 2020;
- Project acronym;
- Project identifier;
- Project coordinator;
- List of participants;
- Coordinator country;
- Title;
- Objective;
- URL for project website.

Despite the availability of attributes that describe projects in various aspects, virtually none of them summarizes what the project is about since there were no keywords provided. On the other hand, project details are involved extensively in the objective attribute, and partially in the title. Among 30,084 research projects in the dataset, the average length of values for the "objective" attribute is 1832 characters, while the same average is 80 for project titles. With such consideration, the objective statements were selected as the primary source to generate tokens.

2.3. Obtaining Tokens

The tokens in objective statements were generated in RStudio using the "tokenizers" package [27]. The R script executed in this step is below:

```
> library(tokenizers)                                          (1)
> filename <- "d:/projects.txt"                                (2)
> my_data <- readChar(filename, file.info(filename)$size)      (3)
> words <- tokenize_words(my_data)                             (4)
> listt <- words[[1]]                                          (5)
> sink(file = "d:/output.txt")                                 (6)
> listt_u <- unique(listt)                                     (7)
> listt_u[1:1000]                                              (8)
> listt_u[1001:2000]                                           (9)
> listt_u[2001:2132]                                           (10)
> sink(file = NULL)                                            (11)
```

In our dataset, the script generated 2132 unique tokens. The lines 8–11 simply write the list into a text file. The next phase in our methodology involved the preprocessing of data before the analysis.

2.4. Data Preprocessing

Accordingly, the list of tokens obtained was manually inspected to eliminate common words such as "a", "will", "be", "an", and "with". Moreover, a variety of irrelevant words that did not specify a particular term were also excluded. Additionally, plural forms of tokens have been excluded in this phase. Initially, the dataset included 2110 tokens. The eliminated data involved 456 common or irrelevant words and 149 plural words. As a result, 1505 tokens remained in our dataset. The list of projects and the count of tokens are given in Table 1.

Table 1. The research items listed with the token counts.

Project No.	Tokens	Project No	Tokens	Project No.	Tokens
1	26	16	18	7	13
18	25	23	18	31	13
20	25	32	18	33	13
19	24	4	17	49	13
22	24	9	17	12	12
13	23	11	17	36	12
2	22	15	17	44	11
21	22	17	17	50	11
29	22	25	17	37	10
39	21	26	17	38	10
6	20	35	17	46	10
24	20	30	16	48	10
28	20	3	15	10	8
40	20	34	15	42	7
14	19	41	15	43	7
27	19	45	15	47	7
5	18	8	14		

An average project description includes a high number of characters (1802 chars) at first glance. However, the elimination of redundant words has led to lower token counts. Also, the number of tokens listed ignore the repetitive tokens in text. Thereby, the projects listed above have an average of 16.34 distinct tokens. Table 2 reports the most referred tokens in project descriptions:

Table 2. Top 50 of most frequent tokens in project descriptions.

Token	Frequency	Token	Frequency	Token	Frequency
disease	13	sector	7	initiative	4
stakeholder	11	data	6	observation	4
process	10	infrastructure	6	structure	4
system	10	medicine	6	academia	3
team	10	patient	6	action	3
material	9	programme	6	cell	3
mechanism	9	europe	5	chair	3
application	8	expert	5	collaboration	3
environment	8	innovation	5	communication	3
institution	8	institute	5	community	3
leader	8	management	5	competitive	3
society	8	network	5	complex	3
excellence	7	partnership	5	deep	3
goal	7	region	5	ecosystem	3
group	7	relation	5	effort	3
position	7	economic	4	gene	3
disease	13	industry	4		

An SQLite database has been created to store the entities (projects and tokens) and the relations between projects and tokens.

3. Results

The calculation of the Jaccard similarity coefficient essentially requires a comparison of two sets of tokens in our case. In this way, seeking for the most relevant project against a set of keywords requires calculating the measure by the Formula (1), iteratively.

Our first scenario for recommendation selection assumes the availability of projects before the membership of researchers. The motive for such an inquiry has the intent to find similar R&D projects and to provide opportunities for cooperation across organizations. Although this objective seems to be irrelevant from the industry–university collaboration, facilitating cooperation within an industry is necessarily a positive action that complies to the industry-driven aspect of the center. In this case, the most similar projects might be found through pairwise comparison based on the Jaccard coefficient measure. As listed in Table 3, Project#29 and Project#35 had six common items in a total of 33 tokens. Accordingly, the most similar project pair had an 18.18% similarity.

Table 3. Projects paired based on relevancy.

Projects		**∩ [1]**	**∪ [2]**	**Similarity [3]**	**Projects**		**∩ [1]**	**∪ [2]**	**Similarity [3]**	**Projects**		**∩ [1]**	**∪ [2]**	**Similarity [3]**
29	35	6	33	0.1818	8	19	4	34	0.1176	44	48	2	19	0.1053
34	40	5	30	0.1667	8	22	4	34	0.1176	5	8	3	29	0.1034
6	11	5	32	0.1563	23	24	4	34	0.1176	8	32	3	29	0.1034
31	37	3	20	0.1500	10	44	2	17	0.1176	11	34	3	29	0.1034
24	50	4	27	0.1481	16	44	3	26	0.1154	9	45	3	29	0.1034
20	34	5	35	0.1429	32	44	3	26	0.1154	2	29	4	40	0.1000
29	44	4	29	0.1379	32	50	3	26	0.1154	16	41	3	30	0.1000
11	26	4	30	0.1333	8	18	4	35	0.1143	41	43	2	20	0.1000
46	47	2	15	0.1333	18	41	4	36	0.1111	4	9	3	31	0.0968
27	43	3	23	0.1304	27	50	3	27	0.1111	9	35	3	31	0.0968
18	50	4	32	0.1250	37	38	2	18	0.1111	14	41	3	31	0.0968
12	45	3	24	0.1250	38	46	2	18	0.1111	31	38	2	21	0.0952
2	22	5	41	0.1220	43	49	2	18	0.1111	31	46	2	21	0.0952
21	22	5	41	0.1220	38	39	3	28	0.1071	6	34	3	32	0.0938
29	41	4	33	0.1212	19	43	3	28	0.1071	16	35	3	32	0.0938
7	34	3	25	0.1200	6	44	3	28	0.1071	20	38	3	32	0.0938
35	44	3	25	0.1200	8	43	2	19	0.1053					

[1] Count of common items; [2] Count of distinct items contained in projects; [3] Jaccard similarity score.

The second type of similarity is across researchers and projects in our case study. Considering a project description filled in by a researcher, or a set of terms obtained from a researcher's profile, it is possible to recommend ongoing industry projects by performing simple subset calculations. Accordingly, a number of projects having high similarity scores might be selected for recommendation for an individual.

As an example, the set (S) of following tokens was assumed to represent a researcher's (i) profile:

Let S_i = {academia, activities, ageing, agriculture, application, change, climate, conduct, cooperation, ecosystem, expectancy, group, initiative, member, mia, opportunities, sites, transition}.

Considering this set (S_i), Table 4 lists the most relevant projects selected based on Jaccard similarity.

The high similarity scores in the table signify higher degrees of relevancy. Resultantly, the most relevant project for the list of tokens given was Project #5, which includes the following tokens:

{ageing, ccdrc, disease, excellence, expectancy, gender, independence, intervention, job, living, member, mia, opportunities, portugal, services, sites, stepping, team}

Notably, a total of six common tokens (ageing, expectancy, member, mia, opportunities, sites) between the Project #5 and the examplary researcher profile results in a similarity score of 0.2. Accordingly, among a total of 30 distinct tokens in both sets, there were six common (20%) items.

Table 4. Projects relevant to a given subset of tokens (S_i).

Project	∩	∪	Similarity [1]	Project	∩	∪	Similarity [1]
5	6	30	0.200	31	1	30	0.033
41	4	29	0.138	8	1	31	0.032
18	5	38	0.132	45	1	32	0.031
17	3	32	0.094	30	1	33	0.030
16	3	33	0.091	9	1	34	0.029
10	2	24	0.083	32	1	35	0.029
46	2	26	0.077	14	1	36	0.028
22	2	40	0.050	24	1	37	0.027
42	1	24	0.042	40	1	37	0.027
38	1	27	0.037	29	1	39	0.026
44	1	28	0.036	20	1	42	0.024
7	1	30	0.033				

[1] Indicates the Jaccard similarity measure.

4. Discussion

Our study demonstrates the use of the Jaccard similarity measure to match a list of projects with a relevant subset of keywords. The motive for that process is to match relevant projects and researchers in a web portal, and hopefully uncover opportunities for cooperation to facilitate the university–industry collaboration. The portal has been conceptually designed and will be developed as an essential part of the EGEVASYON Project, which has been recently under preparation in coordination with the Aegean Region Chamber of Industry.

Finding similar subsets of keywords has various use-cases in accordance with the requirements of the web portal. In particular, it is required to process a given set of keywords that might describe another project, a researcher's fields of interest, or simply a user-provided query. In all three scenarios mentioned, the problem is analogous to selecting more relevant sets based on an input set. As mentioned in [28], Jaccard's index is among the easiest methods (p. 172) with which to measure similarity. Intuitively, recommendations should originate from the projects with the highest similarity to the visitor's profile.

A limitation of our study is the unavailability of data collection on the web portal, which is to be implemented after the kick-off of the EGEVASYON Project. Instead, a small dataset of EU Horizon Projects has been used for obtaining tokens and measuring similarity. The initial steps required the extraction of tokens and the elimination of redundant words before proceeding to match relevant items. An alternative approach is to start with a predefined list of terms. Even so, the maintenance of the list would still occasionally require human intervention. Another limitation of the study is the use of monolingual project descriptions, all of which have been in English. However, a platform for university–industry collaboration might host researchers from various countries and necessitate multilingual statements.

5. Conclusions

Our study demonstrates the use of a well-known similarity measure in a portal that has been designed as a part of EGEVASYON project. As the primary interface of the pro-

ject, the portal will link the researchers with research projects and facilitate industry–university collaboration. Our methodology in recommendation selection utilizes a similarity-based minimization approach that makes use of the Jaccard similarity measure.

The use of the Jaccard similarity measure in our case helped to select relevant projects with an input of 14 tokens, considering project descriptions. Moreover, a pairwise comparison of all projects was easily implemented through an SQL (Structured Query Language) query with a few join operations. However, as mentioned in data preprocessing, human intervention was required to eliminate redundant tokens. Except for this phase, our study demonstrated that using the Jaccard similarity might be a practical idea for matching researchers with projects in the portal. Additionally, the similarity scores might be used to select appropriate recommendations for members, including both businesses and researchers. However, the use of multilingual project descriptions and keywords would go beyond our limitations in this study.

Author Contributions: Conceptualization, İ.K. and H.S.; methodology and analysis, İ.K.; writing—original draft preparation, İ.K.; writing—review and editing, İ.K. and H.S.; project administration, H.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Acknowledgments: The authors would kindly present thanks to Atilla SEVİNÇLİ, board member in the Aegean Region Chamber of Industry, for his support for this study.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Leydesdorff, L.; Etzkowitz, H. The triple helix as a model for innovation studies. *Sci. Public Policy* **1998**, *25*, 195–203.
2. Azaroff, L.V. Industry—University Collaboration: How to make it work. *Res. Manag.* **1982**, *25*, 31–34.
3. Bruneel, J.; D'Este, P.; Salter, A. Investigating the factors that diminish the barriers to university–industry collaboration. *Res. Policy* **2010**, *39*, 858–868.
4. Levchenko, O.; Kuzmenko, H.; Tsarenko, I. The role of universities in forming the innovation ecosystem. *IEM* **2018**, *5*, 10–16.
5. Perkmann, M.; Neely, A.; Walsh, K. How should firms evaluate success in university–industry alliances? A performance measurement system. *R D Manag.* **2011**, *41*, 202–216.
6. Vijaymeena, M.K.; Kavitha, K. A survey of similarity measures in text mining. *Mach. Learn. Appl. Int. J.* **2016**, *3*, 19–28.
7. Leydesdorff, L. On the normalization and visualization of author co-citation data: Salton's Cosine versus the Jaccard index. *J Assoc. Inf. Sci. Technol.* **2008**, *59*, 77–85.
8. Schneider, J.W.; Borlund, P. Matrix comparison, Part 1: Motivation and important issues for measuring the resemblance between proximity measures or ordination results. *J. Assoc. Inf. Sci. Technol.* **2007**, *58*, 1586–1595.
9. Bag, S.; Kumar, S.K.; & Tiwari, M.K. An efficient recommendation generation using relevant Jaccard similarity. *Inf. Sci.* **2019**, *483*, 53–64.
10. Igual, L.; Seguí, S. *Introduction to Data Science, A Python Approach to Concepts, Techniques and Applications*; Springer: Cham, Switzerland, 2017.
11. Saxena, A.; Prasad, M.; Gupta, A.; Bharill, N.; Patel, O.P.; Tiwari, A.; Joo, E.M.; Weiping, D.; Lin, C.T. A review of clustering techniques and developments. *Neurocomputing* **2017**, *267*, 664–681.
12. Kotu, V.; Deshpande, B. *Data Science: Concepts and Practice*, 2nd ed.; Morgan Kaufmann: Burlington, MA, USA, 2018.
13. Jaccard, P. The distribution of the flora in the alpine zone. *New Phytol.* **1912**, *11*, 37–50.
14. Aggarwal, C. *Data Mining: The Textbook*, 1st ed.; Springer: Cham, Switzerland, 2015; pp. 75–76.
15. Öztemiz, F.; Karcı, A. Akademik Yazarların Yayınları Arasındaki İlişkinin Sosyal Ağ Benzerlik Yöntemleri İle Tespit Edilmesi. *Uludağ Univ. J. Fac. Eng.* **2020**, *25*, 591–608.
16. Seifoddini, H.; Djassemi, M. The production data-based similarity coefficient versus Jaccard's similarity coefficient. *Comput. Ind. Eng.* **1991**, *21*, 263–266.
17. Osman, F.M.; Yap, M.H. The effect of filtering algorithms for breast ultrasound lesions segmentation. *Inform. Med. Unlocked* **2018**, *12*, 14–20.
18. Lu, M.; Qin, Z.; Cao, Y.; Liu, Z.; Wang, M. Scalable news recommendation using multi-dimensional similarity and Jaccard-Kmeans clustering. *J. Syst. Softw.* **2014**, *95*, 242–251.

19. Egghe, L. New relations between similarity measures for vectors based on vector norms. *J. Assoc. Inf. Sci. Technol.* **2009**, *60*, 232–239.
20. Niwattanakul, S.; Singthongchai, J.; Naenudorn, E.; Wanapu, S. Using of Jaccard coefficient for keywords similarity. In Proceedings of the International Multiconference of Engineers and Computer Scientists, Hong Kong, China, 13–15 March 2013; pp. 380–384.
21. Park, S., Kim, D.Y. Assessing language discrepancies between travelers and online travel recommendation systems: Application of the Jaccard distance score to web data mining. *Technol. Forecast Soc. Chang.* **2017**, *123*, 381–388.
22. Yu, C., Lakshmanan, L.V., Amer-Yahia, S. Recommendation diversification using explanations. In Proceedings of the 2009 IEEE 25th International Conference on Data Engineering, Shanghai, China, 29 March–2 April 2009; pp. 1299–1302.
23. Fletcher, S.; Islam, M.Z. Comparing sets of patterns with the Jaccard index. *Australas. J. Inf. Syst.* **2018**, *22*, 1–17.
24. Han, J.; Kamber, M. *Data Mining Concepts and Techniques*, 2nd ed.; Morgan Kaufmann: Burlington, MA, USA, 2006.
25. Singh, M.D., Krishna, P.R., Saxena, A. A privacy preserving Jaccard similarity function for mining encrypted data. In Proceedings of the TENCON 2009: 2009 IEEE Region 10 Conference, Singapore, 23–26 November 2009; pp. 1–4.
26. Gültekin, H. EU Research Projects. Available online: https://www.kaggle.com/hgultekin/eu-research-projects (accessed on 27 August 2020).
27. Mullen, L. Introduction to the Tokenizers Package. Available online: https://cran.r-project.org/web/packages/tokenizers/vignettes/introduction-to-tokenizers.html (accessed on 19 August 2020).
28. Simske, S. *Meta Analytics: Consensus Approaches and System Patterns for Data Analysis*, 1st ed.; Elsevier: Cambridge, MA, USA, 2019.