Analisis Tren Perangkat Lunak dan Data pada Jurnal *International Conference on Software Maintenance and Evolution*(ICSME) menggunakan *Topic Modelling*

Mohammad Daffa Gashandy[1] Faris Qanit[2], Hanyel Daryus Bancin[3], Rizky Anugrah Suharto[4], Philipus Hans Christian[5]

[1]Department of Computer Science and Electronics, Universitas Gadjah Mada

***Abstract*— Salah satu cabang ilmu dari NLP adalah Topic Modelling. Dengan memanfaatkan Topic Modelling, kami telah mencari tahu tren terkini dari perangkat lunak berdasarkan analisa yang dilakukan pada 100 judul jurnal ICSME. Dari 100 jurnal yang telah dianalisa tersebut, dihitung kata penting serta similaritas dari masing-masing kata. Didapati kata penting tersebut dibagi menjadi beberapa topik. Untuk kasus 5 tahun terakhir, kata penting yang paling sering muncul adalah *code, developer, dan software* . Ditemukan juga dari 100 paper, negara terbanyak publisher paper adalah USA sebanyak 27 judul, lalu diikuti dengan Canada sebanyak 12 judul, dan China sebanyak 10 judul.**

Keywords— NLP, Topic Modelling, ICSME, Jurnal

I. PENDAHULUAN

*International Conference on Software Maintenance and Evolution (*ICSME) adalah sebuah bidang konferensi untuk yang digunakan untuk memperkenalkan tema-tema dan hasil dari riset yang diperoleh di tahun tersebut. Konferensi tersebut dilaksanakan setiap tahun setiap bulan november hingga awal oktober. Ada berbagai-bagai macam pembahasan yang dilakukan setiap tahun tetapi ada enam tema pokok yang selalu menjadi bahan pembahasan di konferensi tersebut. Tema-tema tersebut diklasifikasikan sebagai sebuah *track* riset. Track dapat dianggap sebagai jalur riset yang dilakukan oleh setiap pihak yang menjalani dan selalu memiliki ketua dan wakil masing-masing. Track yang terdapat pada konferensi ICSME adalah *track riset, track nier, track industri, track tool demo, track artifak, dan track doctoral symposium.* Setiap track memiliki cara review tersendiri dan jenis konten yang bermacam-macam.

Track riset adalah sebuah track yang didedikasikan untuk melakukan riset mendalam mengenai sebuah topik dan memahami inti dari sebuah permasalahan. Track tersebut memiliki konten-konten yang berkaitan dengan pembelajaran mengenai cara melakukan atau memperbaiki sebuah testing method yang berkaitan dengan cara melakukan, cara memperbaiki dan cara mengoptimasi/otomatisasi sebuah parameter yang akan digunakan untuk menjalankan model yang dibuat. Tentunya dengan adanya perkembangan di sisi testing dapat mengurangi hasil defect yang dihasilkan oleh model-model tersebut. Tentunya tidak semua defect dapat diatasi dan maka dari itu ada beberapa peneliti yang memfokuskan riset mereka ke bidang bug retrieval/debugging dan error prevention/analysis. Dengan adanya riset topik tersebut, kemungkinan error dan bug yang ditemukan dapat dikurangi atau diperbaiki. Ada sebagian riset yang di dalam research track tersebut yang tidak berhubungan dengan testing melainkan cara-cara yang berhubungan dengan analisis untuk mempelajari perilaku sebuah program terhadap aksi yang diambil dan mengenai program ekonomi dalam kemasyarakatan sehari-hari perihal pembayaran, mempermudah akses, sistem keuangan, dll. Hal-hal tersebut dapat diklasifikasikan sebagai analisis mengenai program yang telah terbentuk dan sedang di test di sebuah situasi ekstrim. Dengan hasil analisis yang dilakukan, para peneliti tersebut dapat kemudian membuat masukan atau rekomendasi untuk perbaikan. Perlu diketahui untuk bidang-bidang yang dipelajari berjangka dari program software sampai aplikasi yang sudah jadi, termasuk maintenance, organisasi dan evaluasi.

Track Nier adalah singkatan dari new ideas and emerging results dan adalah sebuah track yang berhubungan langsung dengan mempresentasikan ide-ide baru yang sedang diriset oleh para peneliti. Dikarenakan track tersebut adalah sebuah introduksi terhadap tema-tema baru yang dapat dijadikan riset dan bila ditekuni, maka track tersebut dapat dijadikan atau dikumpulkan sebagai materi untuk track riset di tahun berikutnya. Tema-tema dan konten track tersebut tidak perlu melalui empirical evaluation dan dapat langsung di submitkan.

Track Industry adalah sebuah track yang mengambil hasil terbaik dari tema-tema maintenance dan evaluasi yang diimplementasikan di bidang industri dalam tahun tersebut. Bidang-bidang yang termasuk dalam tema industri adalah bidang otomotif, kesehatan, dan juga metode transaksi. Metode tersebut dites dalam kehidupan sehari-hari dan dievaluasi dengan menyertakan informasi atau keterangan tentang bidang yang diimprovisasi serta menyertakan bukti bahwa metode yang diimplementasikan membuahkan hasil yang baik.

Track tool demo adalah sebuah track yang berkaitan dengan maintenance dan evolusi sebuah tool yang dapat digunakan untuk membuat atau memodifikasikan sebuah software. Tool tersebut dipresentasikan dan dijelaskan kemajuan-kemajuan yang dialami dan dijalankan.

Track artefak adalah sebuah track yang mendalami, mengevaluasi dan menerima hasil dari artefak yang dapat digunakan untuk riset dan memberikan hasil yang baik. Evaluasi terdiri dari pengalaman yang dialami dari hasil yang diperoleh dari track riset sedangkan Individual adalah pengalaman yang dialami dari hasil riset yang belum pernah di publish di ICSME sebelumnya.

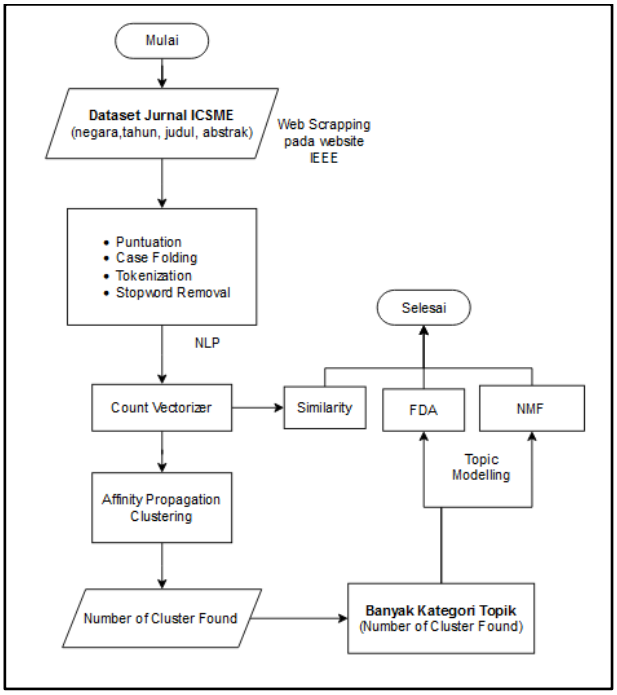
Track doctoral symposium adalah track yang menunjukan hasil riset doktoral sebuah pihak. Topik-topik yang digunakan tentunya berhubungan dengan topik yang sedang dibahas dengan pendalaman yang lebih mendalam.

Disini akan dilakukan analisis menggunakan berbagai metode NLP untuk mencoba menentukan Tren terkait IT yang ada pada jurnal ICSME tersebut berdasarkan Track-track yang ada.

II. METODE

Metode yang digunakan dalam melakukan analisis jurnal adalah dengan metode Topic Modeling. Topic modelling merupakan sebuah algoritma NLP yang mengekstraksi topik untuk kumpulan dokumen. Algoritma topik modelling juga termasuk teknik reduksi dimensi yang digunakan untuk data numerik. Teknik ini dapat dianggap sebagai proses mendapatkan fitur yang dibutuhkan dari kumpulan kata-kata. Hal itu sangat penting karena di NLP setiap kata yang ada dalam korpus dianggap sebagai fitur. Dengan demikian pengurangan fitur dapat membantu proses ekstraksi sehingga lebih fokus pada konten yang tepat daripada membuang-buang waktu dengan semua teks dalam data.

Berikut adalah *Flowchart* proses dari metode yang digunakan untuk melakukan analisis tren pada jurnal ICSME tersebut :



Gambar 1. *Flowchart* Metode Topic Modelling

A. *Web Scraping*

Web scraping merupakan proses pengumpulan data dan informasi yang ada dalam sebuah website secara otomatis dan spesifik. Dikatakan spesifik karena data yang diambil hanya di bagian tertentu saja sesuai dengan kebutuhan. Dengan adanya metode scraping, pengguna akan dimudahkan dalam mengambil data dari suatu situs[3].

Pada jurnal kali ini, kami melakukan scraping manual pada website IEEE explore, untuk mencari 100 judul dari jurnal ICSME dari tahun 2017-2021. Dalam 5 tahun, kami merepresentasikan 20 paper dalam satu tahun. Untuk dianalisis lebih lanjut.

B. *Natural Language Processing (NLP)*

NLP adalah sebuah cabang dari artificial intelligence (AI) yang mengizinkan komuter untuk berhubungan dengan manusia dengan menggunakan bahasa natural, sebuah bahasa yang digunakan manusia secara umum untuk berkomunikasi dalam kesehariannya. NLP digunakan untuk mengukur sentimen dan juga untuk menentukan kepentingan sebuah bahasa dalam kalimat yang dikatakan. Dengan mengubah bahasa tersebut menjadi sebuah komputasi dan mengimplementasikan machine learning, deep learning dan bahkan statik model untuk mempelajari dan memahami arti mendalam sebuah kata.

Pada analisis topik kali ini, digunakan berbagai langkah dan metode dalam NLP pada abstrak berbagai jurnal untuk mendapatkan berbagai tren terkait IT :

- *Punctuation*, Menghilangkan Symbol dari teks
- *Case Folding*, Menyeragamkan karakter pada teks
- *Tokenization*, membagi teks yang dapat berupa kalimat, paragraf atau dokumen, menjadi token-token/bagian-bagian tertentu[1]
- *Stopword removal*, proses filtering, pemilihan kata-kata penting dari hasil token yaitu kata-kata apasaja yang digunakan untuk mewakili dokumen[2]
- *Lemmatization*, pemotongan kata dalam bahasa tertentu menjadi bentuk dasar pengenalan fungsi setiap kata dalam kalimat[4].
- *Count Vectorizer*, mengubah fitur teks menjadi sebuah representasi vector[5].

C. *Clustering : Affinity Propagation*

Clustering adalah proses untuk mengelompokan data ke dalam beberapa *cluster* atau kelompok sehingga data dalam satu *cluster* memiliki tingkat kemiripan yang maksimum dan data antar *cluster* memiliki kemiripan yang minimum[6]. Dalam *machine learning*, *clustering* merupakan U*nsupervised learning* yang digunakan untuk menarik kesimpulan dari dataset.

Salah satu algoritma *clustering* adalah *Affinity Propagation*. Algoritma *Affinity Propagation* merupakan salah satu algoritma clustering yang berbasis exemplar, pada algoritma ini semua titik data dianggap sebagai calon exemplar, lalu nilai *Preference* dihitung berdasarkan nilai median dari keseluruhan data pada tabel *Similarity Matrix*[7].

Terkait implementasi, setelah dilakukannya *CountVectorizer*, dibuat beberapa cluster menggunakan representasi vector yang ada berdasarkan abstrak yang telah di proses menggunakan NLP. Setelahnya akan ditemukan beberapa cluster, yang nantinya jumlah cluster tersebut akan digunakan untuk membuat topik pada model LDA.

D. Topic Modelling

Topic modeling adalah sebuah branch dari NLP yang difungsikan sebagai metode pengklasifikasian berdasarkan relevansi terhadap sebuah topik. Topik tersebut dipilah-pilah berdasarkan konten dari kata-kata yang dimiliki masing-masing dokumen. Bagian dari masing-masing dokumen yang disendirikan tersebut kemudian di scan dan di bagi-bagi menjadi sebuah klasifikasi topik. Tujuan akhir dari topic modeling adalah untuk mengklasifikasikan seluruh dokumen menggunakan topik-topik klasifikasi yang telah disediakan.

Pada paper kali ini disajikan beberapa model topic modelling yang digunakan, yaitu :

- *Non-Negative Matrix Factorization*(NMF)

Non-Negative Matrix Factorization (NMF) adalah teknik tanpa pengawasan sehingga tidak ada pelabelan topik yang akan dilatih oleh model. Cara kerjanya adalah, NMF menguraikan (atau memfaktorkan) vektor dimensi tinggi menjadi representasi dimensi yang lebih rendah. Vektor berdimensi lebih rendah ini non-negatif yang juga berarti koefisiennya non-negatif[8].

- *Latent Dirichlet Allocation*(LDA)

LDA dapat digunakan untuk meringkas, melakukan klasterisasi, menghubungkan maupun memproses data yang sangat besar karena LDA menghasilkan daftar topik yang diberi bobot untuk masing-masing dokumen. Topik yang muncul dari pengolahan data tersebut selanjutnya akan dilakukan uji koherensi topik, yaitu keterkaitan dari uraian probabilitas kata-kata yang ditemukan satu sama lain dalam menyusun suatu topik[9].

III. HASIL

Setelah dilakukan Scrapping Manual dari website IEEE Explorer, diambil data Judul, Negara, Abstrak, dan Tahun(2017-2022) lalu dikumpulkan menjadi sebuah Dataset "dataPaper".

Nomor	Judul	Tahun	Negara	Abstrak
1	Artifacts for Dynamic Analysis of Android Apps	2017	USA	We describe a set of artifacts for dynamic a
2	TraceLab Components for Generating Extractive Summari	2017	USA	This artifact is a reproducibility package for
3	Flattening Code for Metrics Measurement and Analysis	2017	Japan	When we measure code metrics or analyze
4	Keynote abstracts	2017	China	Provides an abstract for each of the keynot
5	CCLearner: A Deep Learning-Based Clone Detection Appr	2017	USA	Programmers produce code clones when d

Gambar 2. Preview Dataset

Dataset tersebut kemudian di preprocessing pada Google Collab dengan metode NLP, dilanjutkan dengan Topic Modelling untuk menghitung Top 30-Most Salient Term pada 5 Tahun terakhir berdasarkan banyaknya Cluster yang terbentuk, beserta negara penyumbang paper terbanyak(per-100 judul) dan juga didapat beberapa topic yang populer berdasarkan *topic modelling LDA* dan *NMF*. Dan diakhiri dengan perbandingan similaritas antar baris pada dataset.

```
Negara
USA        27
Canada     12
China      10
```

Gambar 3. Negara publisher jurnal terbanyak(5 tahun)

Pada paper ini kami membandingkan dan menganalisis :

- Jurnal ICSME dalam rentang 5 tahun
- Jurnal ICSME dari negara USA
- Jurnal ICSME dari negara Canada
- Jurnal ICSME dari negara China

Berikut adalah hasil dari perbandingan tersebut :

A. Number of Cluster Found

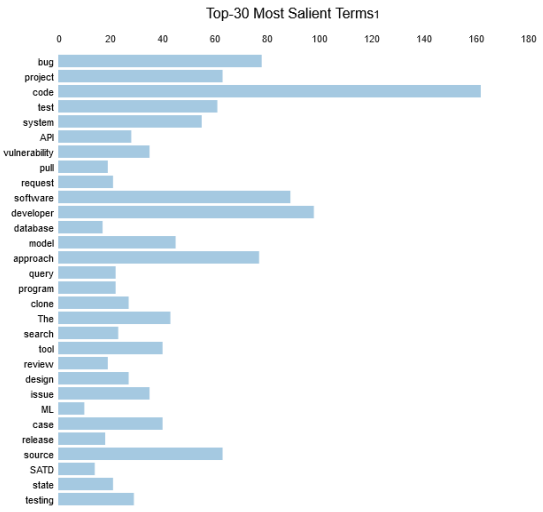
Dengan menggunakan *Affinity Propagation* pada representasi vector abstrak, ditemukan jumlah cluster yang nantinya akan menjadi acuan untuk menentukan banyaknya kategori topik yang akan dibuat.

- Jurnal ICSME rentang 5 tahun = 22
- Jurnal ICSME dari negara USA = 1
- Jurnal ICSME dari negara Canada = 3
- Jurnal ICSME dari negara China = 5

B. Top 30-Most Salient Term(FDA)

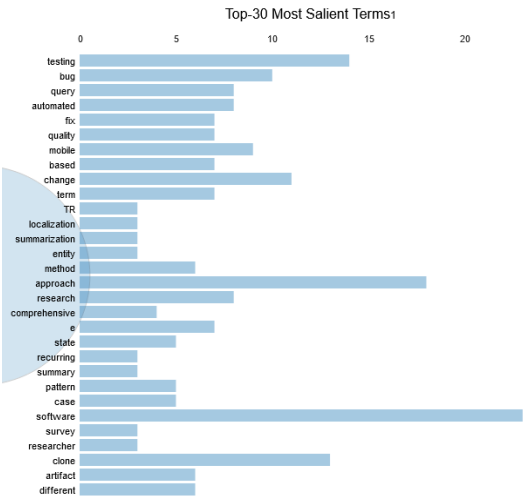
Berikut adalah *Top 30-Most Salient Term* berdasarkan banyak dari topik yang dibuat, dengan ketentuan index topik ke-0 dengan relevance matrix = 1.0

- Jurnal ICSME dalam rentang 5 tahun



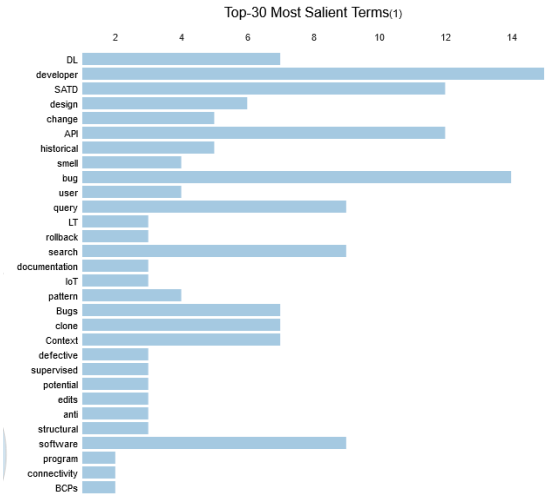
Gambar 4. FDA ICSME rentang 5 tahun

- Jurnal ICSME dari negara USA



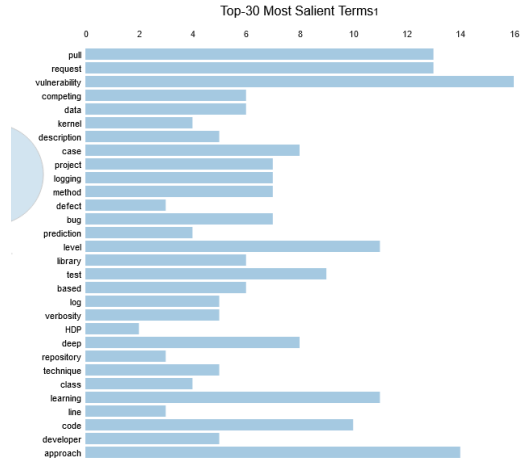
Gambar 6. FDA ICSME pada USA

- Jurnal ICSME dari negara Canada



Gambar 7. FDA ICSME pada Kanada

- Jurnal ICSME dari negara China



Gambar 8. FDA ICSME pada China

C. NMF Model

Yang akan ditampilkan adalah hanya topik pada index ke-0 dengan top word = 15.

- Jurnal ICSME dalam rentang 5 tahun

```
Topic 0:
Top Words:  the of and to in code we for that on is software developers this by
Context-Aware Software Documentation
Beyond Metadata: Code-Centric and Usage-Based Analysis of Known Vulnerabilities in Open-Source Software
Multimodal Representation for Neural Code Search
Confusion Detection in Code Reviews
```

Gambar 9. NMF ICSME rentang 5 tahun

- Jurnal ICSME dari negara USA

```
Topic 0:
Top Words:  and to the of in we as that software queries is for systems environments from this components such tracelab reviews
Beyond Metadata: Code-Centric and Usage-Based Analysis of Known Vulnerabilities in Open-Source Software
A Large-Scale Empirical Study on Linguistic Antipatterns Affecting APIs
On the Relation of Test Smells to Software Code Quality
TraceLab Components for Generating Extractive Summaries of User Stories
```

Gambar 10. NMF ICSME pada USA

- Jurnal ICSME dari negara Canada

```
Topic 0:
Top Words:  the of and to we in satd that dl api on bug design is information
How do Developers Test Android Applications?
A Tale of CI Build Failures: An Open Source and a Financial Organization Perspective
On-demand Developer Documentation
Flattening Code for Metrics Measurement and Analysis
```

Gambar 11. NMF ICSME pada Kanada

- Jurnal ICSME dari negara China

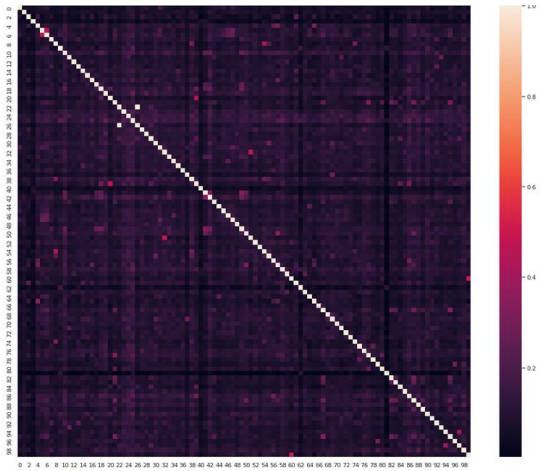
```
Topic 0:
Top Words:  the to approach and code based of in we our tree kernel logging prediction on
Learning to Predict Severity of Software Vulnerability Using Only Vulnerability Description
Continuous, Evolutionary and Large-Scale: A New Perspective for Automated Mobile App Testing
Flattening Code for Metrics Measurement and Analysis
On-demand Developer Documentation
```

Gambar 12. NMF ICSME pada China

D. Similarity

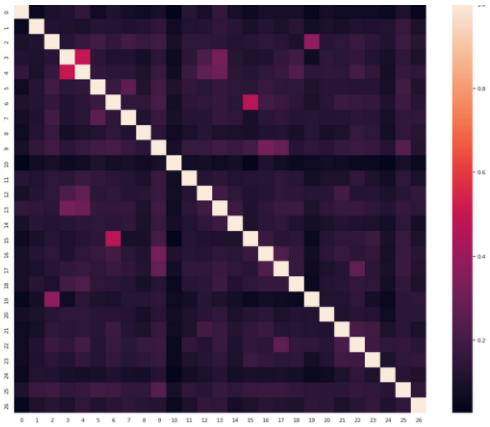
Visualisasi menggunakan *Heatmap*, dengan ketentuan semakin mendekati 1.0(warna putih) semakin terdeteksi sama dan sebaliknya.

- Jurnal ICSME dalam rentang 5 tahun



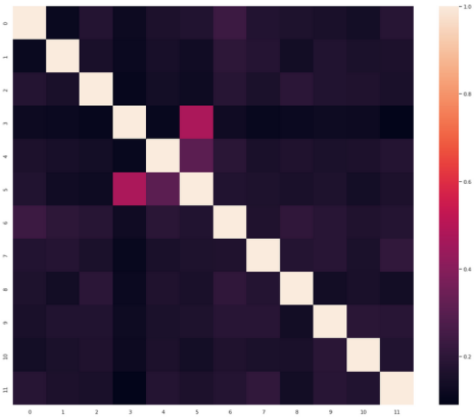
Gambar 13. Similaritas rentang 5 tahun

- Jurnal ICSME dari negara USA



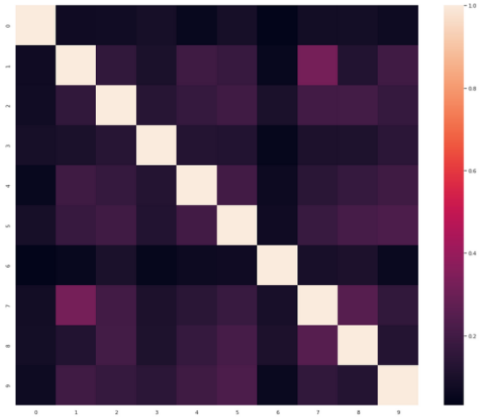
Gambar 14. Similaritas USA

- Jurnal ICSME dari negara Canada



Gambar 15. Similaritas Kanada

- Jurnal ICSME dari negara China



Gambar 16. Similaritas China

IV. DISKUSI DAN ANALISIS

Dari hasil yang didapatkan dan dapat dilihat di BAB III, kita dapat mengasumsikan bahwa jumlah negara yang berpartisipasi akan berhubungan dengan topik yang dipresentasikan. Topik yang berasal dari negara yang lebih modern akan mempresentasikan hasil yang lebih relevan terhadap situasi saat ini. Dengan adanya berbagai negara yang berpartisipasi, akan dipastikan variasi yang ada untuk konferensi tersebut.

Menggunakan metode-metode yang disebutkan di seksi sebelumnya, kita dapat melihat hasil evaluasi akhir dari scraping yang dilakukan. Scrapping tersebut dijalankan untuk melakukan koleksi sejumlah paper-paper dari 5 tahun yang berbeda. Hasil perhitungan similaritas antar abstrak paper juga mengindikasikan bahwa banyaknya penelitian serupa yang dilakukan antar *publisher*. Serupa bukan berarti sama melainkan berbeda alur penelitian dengan topik yang sama.

Dari pihak analisis dapat terlihat bahwa terdapat beberapa persamaan dengan penggunaan kata kunci yang disebutkan meskipun memiliki judul judul dan konten yang berbeda. Dikarenakan ini adalah riset dalam kategori ICSME, hal tersebut dapat diklasifikasikan sebagai sebuah norm yang tentu.

2017	2018	2019	2020
Mobile, Vulnerability, Testing	API, Repository, Back-End	Website, Algorithm, Eficiency, Debugging	ML, Cloud, Testing

Gambar 17. Kata Kunci per Tahun

2021	5 Tahun	USA	Kanada	China
Cryptography, crypto, ML, System Architecture, Testing	Code, Cryptography Crypto, Debugging, ML, Website	App, Developer, Android	API, Debugging, Back-End	Repository, Vulnerability, Debugging

Gambar 18. Kata Kunci Keseluruhan

Berdasarkan gambar-gambar 17 dan 18 yang terlihat diatas, kita dapat ,mengasumsikan relevansi kata yang paling berkesan di setiap pertemuan di tahun tersebut.

Hasil menunjukan bahwa setiap tahun memiliki kata kunci yang berbeda dan berkorelasi dengan permasalahan yang dianggap paling bermasalah di tahun tersebut. Gunakan tahun 2017 sebagai contoh, di mana masa perkembangan mobile sangat menarik perhatian orang, dapat terlihat kenaikan di topik mobile atau di tahun 2020-2021 terdapat banyaknya serangan dan karena itu topik seperti cloud, yang digunakan sebagai backup seandainya terjadi penyerangan dan cryptography untuk meningkatkan pertahanan terhadap serangan. Tetapi ada juga kata kunci yang statis di masing-masing negaranya seperti: *Development* untuk USA, *Debugging* untuk Kanada dan *Vulnerability* untuk China.

V. KESIMPULAN

Melalui keseluruhan hasil yang diperoleh diatas, kita dapat mengatakan bahwa konferensi ICSME lebih condong mengarah kepada riset track yang tetapi secara keseluruhan memiliki bidang-bidang riset yang berhubungan dengan kata kunci kode, software dan developer. Tiga kata kunci tersebut dapat ditemukan di berbagai hasil riset dalam bermacam-macam kategori track. Hasil tersebut dapat terlihat dari bagian BAB IV yang menunjukan hasil yang diperoleh dari mengklasifikasikan dan pendalaman masing-masing konten.

Untuk memperdalam hasil riset, mungkin membutuhkan kita untuk memperluas parameter yang ditentukan. Seperti hasil yang sudah disampaikan, kita dapat mengatakan bahwa topik diskusi mungkin tidak banyak perubahan. Karena topik tersebut mengandalkan resource yang sama untuk memastikan bahwa *maintenance* tetap berevolusi bersama/mengikuti perubahan yang terjadi dengan perangkat lunak.

VI. REFERENSI

[1] Tokenisasi - Wikipedia bahasa Indonesia, ensiklopedia bebas. "Tokenisasi - Wikipedia Bahasa Indonesia, Ensiklopedia Bebas," Diaksen 27 November 2022. https://id.wikipedia.org/wiki/Tokenisasi.

[2] M. Saiful A., Imam M. S., Sri M., 2019. Sistem Pencarian E-Journal menggunakan Metode Stopword Removal dan Stemming Berbasis Android, *KIMU*. pp : 3

[3] Hans, Rizal. "Kenali Web Scraping, Salah Satu Teknik Pengumpulan Data Sekunder!" Kenali Web Scraping, Salah Satu Teknik Pengumpulan Data Seku... Diakses November 27, 2022. https://www.dqlab.id/kenali-web-scraping-salah-satu-teknik-pengumpulan-data-sekunder.

[4] School of Computer Science. "Natural Language Processing," Diakses 27 November 2022. https://socs.binus.ac.id/2013/06/22/natural-language-processing/.

[5] Munawar, Yosua R., 2019. Sistem Pendeteksi Berito Hoax di Media Sosial dengan Teknik Data Mining Scikit Learn, JIK. pp : 3

[6] Tan, P.N., Steinbach, M., Kumar, V. 2006. *Introduction to Data Mining*. Boston:Pearson Education.

[7] Suwilo S., Situmorang Z., 2018. Analisis Kinerja Modifikasi Algoritma Affinity Propagation, *Repositoy USU.* pp : 1(Abstrak)

[8] Salgado, Rob. "Topic Modeling Articles with NMF." Medium, Diakses 27 April 2022. https://towardsdatascience.com/topic-modeling-articles-with-nmf-8c6b2a227a45.

[9] Listari. "Topic Modelling Menggunakan Latent Dirchlect Allocation (Part 1): Pre-Processing Data Dengan Python." Medium, Diakese 27 November 2022. https://medium.com/@listari.tari/topic-modelling-menggunakan-latent-dirchlect-allocation-part-1-pre-processing-data-dengan-python-87bf5c580923.