Information Engineering Express
International Institute of Applied Informatics
2024, Vol. 10, No. 2, IEE802

Generating a Technical Trend Map by Analyzing the Structure of U.S. Patents Using Patent Families

Jun Nakamitsu[*], Satoshi Fukuda[*], Hidetsugu Nanba[*]

Abstract

Researchers and developers search for patents in fields related to their own research to obtain information on issues and effective technologies in those fields for use in their research. However, it is impossible to read through the full text of many patents, so a method that enables patent information to be grasped briefly is needed. In this study, we analyze the structure of U.S. patents with the aim of extracting important information. Using Japanese patents with structural tags such as "field", "problem", "solution", and "effect", and corresponding U.S. patents (patent families), we automatically created a dataset of 81,405 U.S. patents with structural tags. Furthermore, using this dataset, we conduct an experiment to assign structural tags to each sentence in the U.S. patents automatically. For the embedding layer, we use a language representation model BERT pretrained on patent documents and construct a multi-label classifier that classifies a given sentence into one of four categories: "field", "problem", "solution", or "effect". We are able to classify sentences with precision of 0.6994, recall of 0.8291, and F-measure of 0.7426. We have analyzed the structure of U.S. patents using our method and generated a technological trend map, which confirms the effectiveness of the proposed method.

Keywords: patent, document structure analysis, machine translation, machine learning, technical trend map

1 Introduction

When researchers and company engineers consider new research or development, utilizing patent information is important for grasping the latest technical trends. On the other hand, it is difficult to read through all the patents published around the world. Under such circumstances, a method that enables an efficient overview of technical trends is needed. To overview technical trends, it is effective to classify patents according to the viewpoints of technologies and problems, etc. However, to do so, it is necessary to extract the description part of technologies and problems from each patent. Therefore, this study aims to analyze the structure of U.S. patents.

Unlike U.S. patents, Japanese patents have explicit items such as "Field of Technology" (hereinafter referred to as "field"), "Problem to Be Solved by the Invention" (hereinafter referred to as "problem"), "Solution for Solving the Problem" (hereinafter referred to as "solution"), and

[*] Chuo University, Tokyo, Japan

"Effect of the Invention" (hereinafter referred to as "effect"). As a result, researchers and company engineers have to spend more time reading U.S. patents. Therefore, in this study, we perform a structural analysis of U.S. patents and automatically extract sentences that provide clues for classification.

To achieve this, we analyze the structure of U.S. patents by using patent families. Generally, patent rights are granted independently in each country. To obtain patent rights in each country, an applicant needs to apply for patents for the same invention in several countries. Such a group of patent documents with the same content is called a patent family. Although the language and structure of patents in the patent family differ, the texts that compose 'the application documents' closely correspond to each other. In this study, we analyze the structure of U.S. patents by assigning the same structural tags to untagged sentences in U.S. patents that share the same meaning as sentences in 'Japanese patents' to which the structural tags were manually assigned. We first find from U.S. patents the bilingual sentences described in the "field", "problem", "solution", and "effect" sections in Japanese patents. Then, we construct a U.S. patent dataset with a clear structure by assigning the same structural tags as those of the Japanese patent to the found sentences. Finally, by applying machine learning using the created dataset, we construct a system that can automatically extract sentences related to "field", "problem", "solution", and "effect", even for U.S. patents that do not have patent families.

By extracting key sentences about technology trends from U.S. patents, we can obtain clues for classifying each patent. This also allows researchers and company engineers to understand the current technology of their competitors and to conduct efficient research and development that meets global demand.

2 Related Work

2.1 Structural Analysis of Technical Documents

Although it is important to analyze technical documents such as patents and research papers, it is difficult to analyze the entire text. In such cases, it is useful to clarify initially the structure of the document to narrow down the target sentences before conducting the analysis. In this section, we introduce a study that used machine learning to analyze the structure of technical documents.

Prabhakaran et al. [1] analyzed the structure of abstracts with the aim of predicting the growth and decline of scientific topics. They constructed a classifier that applies seven different labels to sentences ("background", "objective", "data", "design", "method", "result," and "conclusion") using manually labeled abstracts, in which the authors assigned labels to sentences, as training data. They applied Conditional Random Field to these data and parsed approximately 2.4 million abstracts to investigate the relationship between labels and topics. The results showed that the technologies discussed in "conclusion" sentences tended to decline, while the technologies discussed in "method" sentences were in the early stage of growth.

Li et al. [2] assumed that evidence plays an important role in biomedical research and extracted evidential descriptions of the figures and tables from biomedical articles. They constructed a model consisting of embedding, attention, and tagging layers. For embedding, they used BioGloVe [3], BioBERT [4], and SciBERT [5], which were pretrained on biomedical texts, and Recurrent Neural Network (RNN) and Long Short Term Memory (LSTM) for attention.

Experiments were conducted on two datasets, PubMed-RCT [6] and SciDT [7], and the models using SciBERT and LSTM performed the best.

Given this background, the purpose of this study is to analyze the structure of U.S. patents with the aim of classifying patents. To achieve this, we analyze the structure of U.S. patents by assigning four types of structural tags, “field”, “problem”, “solution”, and “effect”, to each sentence in a U.S. patent. We then construct training data with the Japanese–U.S. patent family and use Bidirectional Encoder Representations from Transformer (BERT) [8] for embedding.

2.2 Translation of Patent Documents

In this study, we create a dataset using Japanese and U.S. patent families. In the process of creating the dataset, translation from Japanese to English is required. In this section, we introduce efforts required for the translation of patent documents.

One of the tasks in NTCIR-10 is the Patent Machine Translation Task [9]. This task provides a large test collection containing training, development, and test data for Chinese/English and Japanese/English patent machine translation. The collection contains a bilingual Japanese–English patent translation corpus of about 3.2 million pairs. We constructed a Japanese–English machine translation system based on state-of-the-art Transformer architecture [10].

3 Analyzing the Structure of U.S. Patents

3.1 Analyzing the Structure of U.S. Patents Using Patent Families

As described in Section 1, U.S. patents do not explicitly include items describing “field”, “problem”, “solution”, and “effect”. Therefore, we propose a method to construct a structural analysis system of U.S. patents by creating a dataset with structural tags using patent families and performing machine learning using the dataset as training data.

Patent rights must be obtained individually in each country, and a group of patent applications with identical content across different countries is known as a patent family. In this study, we analyzed the structure of U.S. patents by tagging sentences with similar meanings to those in Japanese patents and used this structured data for machine learning to develop a structural analysis system for U.S. patents.

The procedure for the structural analysis of U.S. patents is shown below.

1) Translate each sentence with a structure tag in the Japanese patent into English.

2) Represent all the sentences in the U.S. patent and the translated sentences in 1) as vectors. The sentence in the U.S. patent that has the highest cosine similarity with the translated sentence in 1) is assigned the same structure tag as that of the Japanese patent.

3) Using the data in 2) as training data, perform machine learning to extract automatically sentences related to “field”, “problem”, “solution”, and “effect” from U.S. patents.

Steps 1) and 2) are described in detail in Section 3.2, and step 3) is described in Section 3.3.

3.2 Dataset Creation

First, a machine translator built using the NTCIR-10 Patent Machine Translation Test Collection and the sequence modeling tool FAIRSEQ [11] was used to translate sentences with structural tags in Japanese patents into English. FAIRSEQ is a tool that can be used to train text generation models such as machine translation. The translator built in this experiment uses Transformer and achieved a BLEU score of 44.11.

Next, we extracted sentences from the U.S. patent that share the same meaning as the translated sentences. First, the sentences in the translation result and the full text of the U.S. patent are vectorized using PatentSBERTa [12], and then the cosine similarity between the sentences in the translation result and the full text of the U.S. patent, which is a patent family of the Japanese patent, is calculated. The sentence in the U.S. patent with the highest score is considered to have the same meaning as that in the Japanese patent, and is therefore assigned the same structure tag as the Japanese patent. Using this method, a dataset of U.S. patents with structural tags was created. Since the patent families were faithfully translated, the sentence with the highest score was expected to be almost a bilingual sentence. Note that a sentence may be assigned more than one structure tag because it may have the highest score for more than one resulting translation. In addition, not all structure tags are necessarily present in Japanese patents. For example, some Japanese patents do not include “effect”.

Steps 1) and 2) were used to assign structure tags to 81,405 U.S. patents that had Japanese patents as their patent families. Of the total 22,016,132 sentences, 1,366,165 sentences were assigned at least one of the four types of structure tags. In this experiment, we did not use sentences that had not been assigned any structure tags; we only classified sentences that had been assigned one or more structure tags. If the classifier is used for the full text of patents, there is a high possibility that structure tags will be assigned to ordinary sentences as well, but considering that patent classification will be performed in the future, we believe that the presence of some noise is not a significant problem.

3.3 Analyzing the Structure of U.S. Patents by Machine Learning

Machine learning is performed using the dataset created in the previous section as training data. We used BERT to assign structural tags to sentences. BERT is a Transformer-based pretrained model that can be applied to any task in natural language processing. By fine-tuning BERT to the structural tag classification task, we built multi-label classifiers that automatically classified a given sentence into one of four categories: “field”, “problem”, “solution”, or “effect”.

We pretrained the BERT model using 3.5 million sentences in the detailed description sections and claims of U.S. patents. We then constructed a classifier using the patent-specific BERT based on the pretrained BERT model. Due to the unbalanced number of data for each structural tag in the dataset, we also tested under-sampling and weighted loss function methods.

4 Experiment

To confirm the validity of the structural analysis methods for U.S. patents proposed in Section 3, we conducted experiments under various conditions.

4.1 Experimental Setup

Experimental Data

Of the 1,366,165 sentences that were automatically assigned structural tags according to Section 3.2, 60% were used for training, 20% for validation, and 20% for testing. Table 1 shows a breakdown of the structural tags assigned.

Table 1: Breakdown of structure-tagged sentences

	Number of Sentences			
	Training	***Validation***	***Testing***	***Total***
Field	56,486	18,969	18,727	94,182
Problem	243,606	81,320	81,336	406,262
Solution	464,662	155,017	154,825	774,504
Effect	106,648	35,324	35,706	177,678

Method

We examined the following four methods: patent-specific BERT, under-sampling, and weighted loss function. Furthermore, to confirm the effectiveness of our methods, we compared them with a classifier using BERT-base-uncased, which is a standard BERT model for English texts. For all classifiers, the sigmoid function and binary cross-entropy loss were used to calculate the loss. The training parameters were: maximum number of tokens = 128, batch size = 256, and number of epochs = 10. The details of each method are described below.

- Patent-specific BERT (our method): BERT was pretrained using 3.5 million sentences from U.S. patents. The classifiers were trained using the dataset in Table 1. The learning rate was 1e-6.
- Under-sampling (our method): Due to the disproportionate number of data for each structural tag, we under-sampled the training data. We matched the number of sentences in each structural tag to the number of training sentences in the "field" with the lowest number of sentences (56,486). The validation and testing numbers are shown in Table 1. Patent-specific BERT was used for embedding. The learning rate was 1e-10.
- Weighted loss function (our method): The dataset had an unbalanced number of positive and negative examples for each structural tag. Therefore, we weighted the losses when calculating them. We increased the weight of positive examples by multiplying the loss by the ratio of negative to positive examples. Patent-specific BERT was used for embedding. The classifiers were trained using the dataset in Table 1. The learning rate was 1e-6.
- BERT-base-uncased (baseline method): As a baseline method, we used BERT-base-uncased instead of patent-specific BERT.

4.2 Results

We evaluated the classification of each structural tag. As shown in Table 2, the results of the experiment showed that one of our methods, "patent-specific BERT", obtained an F-measure of 0.7426, which outperformed the others.

4.3 Discussion

In this experiment, we estimated bilingual sentences based on the assumption that sentences in a patent family are translated in a one-to-one relationship. However, there are cases in which a single sentence in a Japanese patent is divided into two sentences in a U.S. patent. Conversely, two sentences in a Japanese patent can be combined into a single sentence in a U.S. patent. Although the method used in this experiment produced highly accurate translations, we believe that some issues need to be addressed in the subsequent discovery of bilingual sentences. In the future, it will be necessary to consider the possibility of determining multiple bilingual sentences depending on the similarity score.

In this experiment, sentences without any structural tags were not used as training data. The reason for this is that untagged sentences are far more numerous than tagged sentences. We believe that eliminating untagged sentences would increase the likelihood that structural tags would be assigned to sentences that should not be assigned structural tags. It is therefore necessary to investigate the accuracy of classification for sentences that should not be assigned structural tags in the future.

Table 2: Evaluation of classification results using patent-specific BERT, under-sampling, weighted loss function, and BERT-base-uncased

		Precision	***Recall***	***F-measure***
Patent-specific BERT	Field	0.9457	0.8494	**0.8920**
	Problem	**0.8654**	0.7548	0.8052
	Solution	0.8297	**0.8877**	**0.8573**
	Effect	0.6757	0.3056	**0.4158**
	Average	0.8291	0.6994	**0.7426**
Under-sampling	Field	0.0838	**0.9084**	0.1528
	Problem	0.2976	**0.9962**	0.4575
	Solution	0.5603	0.7666	0.6467
	Effect	0.1308	**0.9998**	0.2307
	Average	0.2681	**0.9178**	0.3719
Weighted loss function	Field	0.6506	0.9378	0.7636
	Problem	0.7644	0.8341	0.7967
	Solution	0.8517	0.8359	0.8433
	Effect	0.3387	0.7276	0.4601
	Average	0.6514	0.8339	0.7159
BERT-base-uncased (baseline)	Field	**0.9525**	0.8434	0.8915
	Problem	0.8464	0.7714	**0.8061**
	Solution	**0.8346**	0.8758	0.8543
	Effect	**0.7045**	0.2766	0.3921
	Average	**0.8345**	0.6918	0.7360

5 Generation of Technical Trend Maps

To confirm the effectiveness of our method described in section 4, we generated a technical trend map using our method, patent-specific BERT. The Japanese Patent Office (JPO) selects about 10 technology themes each year, focusing on technological fields that should be promoted as national policy, conducts a technology trend survey, and publishes survey reports. These reports are written by analyzing domestic and foreign papers and patents related to each theme. In writing these reports, several technical perspectives, called analysis axes, are defined, and each paper and patent in the list is described as to which analysis axis it corresponds to. In this study, the sentences extracted by the proposed method are evaluated by the degree to which the analysis axes can be reproduced by clustering them.

5.1 Procedures for Generating Technological Trend Maps

Data

In this experiment, 94 U.S. patents in the high-barrier film field were used to generate a technological trend map. There are 314 axes of analysis in the high-barrier film field, consisting of up to four levels. In this study, three axes related to solution were manually selected from the top two levels and used in the experiment. The reason for selecting three axes is that when evaluating the clustering results using some evaluation methods described below, it becomes difficult when one patent belongs to multiple axes, so we selected axes so that one patent belongs to one axis.

Analysis of U.S. Patents

Using patent-specific BERT, we extracted sentences related to effect, field, problem, and solution, respectively, from U.S. patents. Each extracted sentence is transformed into a 300-dimensional vector using PatentSBERT [12]. By averaging each element of the sentence vector for each category of effect, field, problem, and solution, four types of document vectors were created. These are used for clustering.

Clustering Methods

Gaussian Mixture Model, K-Means, and Mean Shift [13] are used for clustering.

Evaluation Methods

The following four methods are used for evaluation.

- Mutual Information (MI)
- Normalized Mutual Information (NMI)
- Adjusted Mutual Information (AMI)
- v-Measure (VM) [14]

The smaller the values of the evaluation results by these methods, the more consistent the clustering results are with the analysis axes.

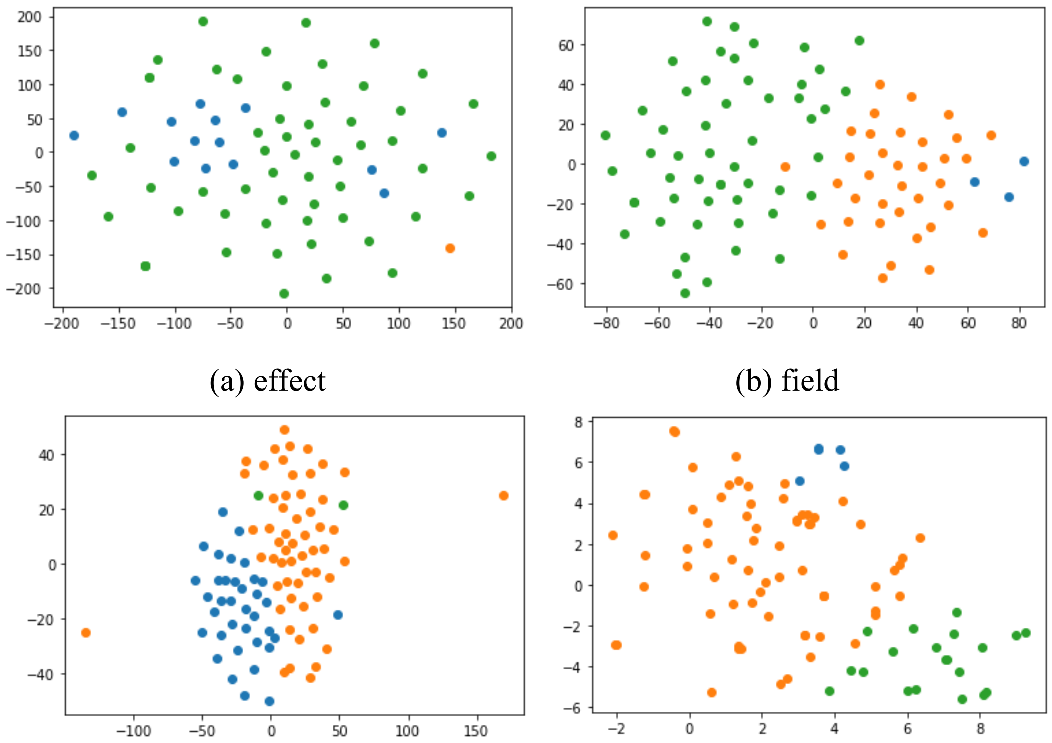
5.2 Results

The evaluation results are shown in Table 3. As can be seen from the table, the results of clustering by solution sentences matched the analysis axes almost perfectly. The results of projecting the clustering results using GMM onto a 2-dimensional plane with t-SNE [15] are shown in Figure 1. As can be seen from Figure 1, clustering using solution sentences is the most appropriate way to separate the analysis axes.

Figure 2 shows the abstract of US patent US20160254487A1 "Permeation barrier system for substrates and devices and method of making the same." Figures 3, 4, and 5 show part of the text concerning the problem, field, and solution extracted from U.S. patent US20160254487A1, respectively. From these figures, our method is able to extract sentences of different aspects from the same patent. This affects the clustering results, as shown in Figure 1.

Table 3: Evaluation of clustering results using GMM, K-Means and Mean shift.

Clustering method	Evaluation measure	Structure tags			
		effects	field	problem	solution
GMM	MI	0.100	0.010	0.021	**0.000**
	NMI	0.153	0.014	0.035	**0.000**
	AMI	0.117	-0.003	0.017	**0.000**
	VM	0.153	0.014	0.035	**0.000**
K-Means	MI	0.040	0.016	0.055	**0.000**
	NMI	0.050	0.020	0.073	**0.000**
	AMI	0.014	0.006	0.059	**0.000**
	VM	0.050	0.021	0.073	**0.000**
Mean Shift	MI	0.131	0.054	0.049	**0.000**
	NMI	0.177	0.112	0.115	**0.000**
	AMI	0.039	0.026	0.040	**0.000**
	VM	0.177	0.112	0.115	**0.000**



(a) effect

(b) field

(c) problem

(d) solution

Figure 1: Clustering results using 94 U.S. patents in the high-barrier film field. Orange, blue and green dots indicate the three axes of analysis: film material, film-making material and film-making technology, respectively.

> Disclosed is a novel moisture permeation barrier system for substrates and devices and method of making the same. The permeation barrier system includes two barrier layers. The first barrier layer is disposed over the substrate or an electronic device. The second barrier layer is then disposed over the first barrier layer. This system has relatively low permeability to moisture and is flexible. It may cover particles and provide moisture protection with a relatively small width edge seal.

Figure 2: Abstract of US patent US20160254487A1

> Opto-electronic devices that make use of organic materials are becoming increasingly desirable for a number of reasons. OLEDs make use of thin organic films that emit light when voltage is applied across the device. OLEDs are becoming an increasingly interesting technology for use in applications such as flat panel displays, illumination, and backlighting.

Figure 3: Part of the sentences extracted from US patent US20160254487A1 regarding the problem

> The present invention relates to permeation barriers for devices such as organic light emitting diodes and other devices, and devices including the same. Several OLED materials and configurations are described in U.S. Pat. Nos. 5,844,363, 6,303,238, and 5,707,745, which are incorporated herein by reference in their entirety.

Figure 4: Part of the sentences extracted from US patent US20160254487A1 regarding the field

> The permeation barrier system includes two barrier layers. The first barrier layer is disposed over the substrate or an electronic device. The second barrier layer is then disposed over the first barrier layer. This system has relatively low permeability to moisture and is flexible. It may cover particles and provide moisture protection with a relatively small width edge seal.

Figure 5: Part of the sentences extracted from US20160254487A1 regarding the solution

6 Conclusion

In this study, we created a dataset containing 81,405 U.S. patents with structural tags. Using this dataset, we conducted an experiment to classify automatically sentences from the U.S. patents that are important for classifying each patent onto a technical analysis axis. The experimental results showed that one of our methods, patent-specific BERT, obtained an F-measure of 0.7426, which outperformed the others. We have analyzed the structure of U.S. patents using the proposed method and generated a technology trend map, which confirms the effectiveness of the proposed method.

References

[1] Vinodkumar Prabhakaran, William L. Hamilton, Dan McFarland, and Dan Jurafsky, "Predicting the rise and fall of scientific topics from trends in their rhetorical framing," In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pp.1170–1180, 2016.

[2] Xiangci Li, Gully Burns, and Nanyun Peng, "Scientific discourse tagging for evidence extraction," In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, pp.2550–2562, 2021.

[3] Gully A Burns, Xiangci Li, and Nanyun Peng, "Building deep learning models for evidence classification from the open access biomedical literature," Database, 2019.

[4] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang, "BioBERT: a pre-trained biomedical language representation model for biomedical text mining," Bioinformatics, pp.1234–1240, 2020.

[5] Iz Beltagy, Arman Cohan, and Kyle Lo, "SciBERT: pretrained contextualized embeddings for scientific text," In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing,

pp.3615–3620, 2019.

[6] Franck Dernoncourt and Ji Young Lee, “Pubmed 200k RCT: a dataset for sequential sentence classification in medical abstracts,” In Proceedings of the 8th International Joint Conference on Natural Language Processing, pp.308–313, 2017.

[7] Gully APC Burns, Pradeep Dasigi, Anita de Waard, and Eduard H Hovy, “Automated detection of discourse segment and experimental types from the text of cancer pathway results sections,” Database, 2016.

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, “BERT: pre-training of deep bidirectional transformers for language understanding,” arXiv preprint arXiv:1810.04805, 2017.

[9] Isao Goto, Ka Po Chow, Bin Lu, Eiichiro Sumita, and Benjamin K. Tsou, “Overview of the patent machine translation task at the NTCIR-10 workshop,” In Proceedings of the 10th NTCIR Conference, pp.260–286, 2013.

[10] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin, “Attention is all you need,” In Proceedings of Advances in Neural Information Processing Systems, pp.6000–6010, 2017.

[11] Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli, “FAIRSEQ: a fast, extensible toolkit for sequence modeling,” In Proceedings of North American Association for Computational Linguistics (NAACL): System Demonstrations, pp.48–53, 2019.

[12] Hamid Bekamiri, Daniel S. Hain, and Roman Jurowetzki, “PatentSBERTa: a deep NLP based hybrid model for patent distance and classification using augmented SBERT,” arXiv preprint arXiv:2103.11933, 2021.

[13] Dorin Comaniciu and Peter Meer, “Mean shift: A robust approach toward feature space analysis,” IEEE Transactions on Pattern Analysis and Machine Intelligence, Volume 24, Issue 5, pp. 603-619, 2002.

[14] Andrew Rosenberg and Julia Hirschberg, “V-Measure: A conditional entropy-based external cluster evaluation measure,” In Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 410–420, 2007.

[15] Laurens van der Maaten and Geoffrey Hinton, “Visualizing High-Dimensional Data Using t-SNE,” Journal of Machine Learning Research 9, pp. 2579-2605, 2008.