



Efficient Few-Shot Classification of Indonesian Academic Correspondence Using SetFit

Yudi Aditya Pratama^{1*}, Bagus Satrio Waluyo Poetro², Rizqi Agung Fauzi³

¹⁻³ Universitas Islam Sultan Agung, Indonesia

email: yudiaditya@std.unissula.ac.id¹

Article Info :

Received:
20-08-2026
Revised:
02-09-2026
Accepted:
15-09-2026

Abstract

Digital transformation in higher education has increased the need for intelligent systems capable of organizing and classifying institutional correspondence efficiently, particularly when labeled training data are limited. This study develops and evaluates an Indonesian academic-correspondence classification system using SetFit within an efficient few-shot learning framework. A quantitative experimental design was applied to 145 official PDF documents collected from five units at Universitas Islam Sultan Agung. Native digital documents were extracted using PyMuPDF, whereas scanned documents were processed using Doctr OCR, followed by correspondence-specific preprocessing, manual ground-truth labeling, and structured dataset preparation. Model training employed intfloat/multilingual-e5-base with incremental N-shot experiments from 1-shot to 6-shot and contrastive representation learning using CosineSimilarityLoss. The 6-shot configuration, using six training documents per class, produced the best reported performance, achieving an accuracy of 85.87% and a weighted F1-score of 0.85. Training-loss and error-rate patterns also indicated stable model adaptation under the reported experimental conditions. The selected model was subsequently integrated into a lightweight Streamlit prototype. The findings demonstrate that SetFit provides an efficient approach for academic-document classification when institution-specific labeled datasets are scarce, while consistent label definition and real-time OCR integration remain important considerations for further system development.

Keywords: Academic Correspondence, Few-Shot Learning, SetFit, Text Classification, Natural Language Processing.



©2026 Authors.. This work is licensed under a Creative Commons Attribution-Non Commercial 4.0 International License.
(<https://creativecommons.org/licenses/by-nc/4.0/>)

INTRODUCTION

Digital transformation has reshaped administrative information flows in higher education, where official correspondence is increasingly produced, distributed, retrieved, and archived through electronic systems rather than conventional paper-based procedures. Information systems have become strategically important for improving the operational efficiency and accessibility of institutional services in higher-education environments (Nugroho et al., 2025). Within this broader transformation, Natural Language Processing has evolved from rule-based and statistical approaches toward neural architectures capable of learning contextual and semantic relationships from large-scale textual data (Zhou et al., 2020). Transformer architectures have accelerated this development by enabling models to capture long-range dependencies and contextual interactions through attention-based representations rather than sequential recurrence (Lin et al., 2022). The growing maturity of transformer-based systems has consequently moved document management beyond passive digital storage toward intelligent classification, retrieval, and decision-support functions (Amatriain et al., 2023). Academic correspondence presents a particularly demanding application because formal institutional documents often contain repetitive lexical structures, standardized administrative expressions, and category-specific functions that may differ only through subtle semantic cues. In this setting, automated document classification becomes relevant not merely as a computational convenience but as an engineering mechanism for reducing manual sorting, improving archive organization, and supporting scalable institutional information management.

Previous studies demonstrate that pretrained language models and contextual embeddings have substantially improved Indonesian text classification, although their performance depends strongly on the linguistic domain and the availability of representative labeled data. IndoBERT established a

language-specific foundation for Indonesian NLP by providing pretrained representations adapted to Indonesian linguistic characteristics and downstream language-understanding tasks (Koto et al., 2020). The broader development of Indonesian NLP resources has also shown that language-specific benchmarks are critical for evaluating whether pretrained models genuinely capture local grammatical and semantic structures rather than merely transferring performance from high-resource languages (Cahyawijaya et al., 2021). In practical classification settings, IndoBERT has produced competitive results for Indonesian news categorization, supporting the effectiveness of contextual representations for distinguishing semantically related document classes (Juarto & Yulianto, 2023). Similar evidence emerges from economic-activity classification, where an IndoBERT-based approach achieved strong performance in mapping Indonesian textual descriptions to predefined categories (Syazali & Yulianti, 2025). Transformer-based classification has also been successfully applied to Indonesian public-opinion and emotion-labeling tasks, indicating that contextual modeling can discriminate linguistically complex categories beyond simple keyword matching (Anwar et al., 2021). Taken together, these studies establish a strong technical basis for Indonesian document classification, yet they primarily reflect domains with clearer benchmark structures or larger labeled resources than institution-specific academic correspondence.

The main limitation in extending these advances to university correspondence lies in the scarcity of labeled institutional documents, because conventional supervised learning and standard fine-tuning strategies typically benefit from substantially larger training corpora. Few-shot learning addresses this constraint by exploiting pretrained knowledge to generalize from only a small number of labeled examples, making it particularly relevant when domain-specific annotation is expensive or difficult to scale (Wang et al., 2021). Existing reviews characterize sample efficiency as the central advantage of few-shot learning, while also emphasizing that performance may vary considerably with task structure, example selection, and inter-class similarity (Parnami & Lee, 2022). SetFit provides a more specialized solution for text classification by combining contrastive fine-tuning of sentence representations with a lightweight classification head, reducing dependence on prompt engineering and large labeled datasets (Tunstall et al., 2022). Experimental evidence further indicates that contrastive learning within SetFit can outperform conventional fine-tuning under restricted-label conditions because the model learns discriminative relationships between semantically similar and dissimilar examples (Kilic et al., 2023). Research applying SetFit to Arabic scholarly-paper classification suggests that this architecture can remain effective in formal textual domains where labeled examples are limited (Khaled & Manayer, 2024). Yet the literature still provides little empirical evidence on whether this advantage persists for Indonesian academic correspondence, where repetitive institutional language and closely related administrative functions may reduce the semantic separability between classes.

This limitation becomes more important because reliable academic-correspondence classification requires not only an appropriate learning architecture but also a preprocessing pipeline capable of producing semantically meaningful text from heterogeneous institutional documents. Deep-learning text-classification research consistently shows that model performance is shaped by the quality of textual representations and by the extent to which preprocessing preserves information relevant to the target class (Minaee et al., 2021). Indonesian multilabel classification using IndoBERT embeddings likewise demonstrates that contextual representation quality remains central when categories share overlapping linguistic features (Nabiilah et al., 2024). The challenge becomes sharper when official correspondence contains letterheads, institutional identifiers, OCR-induced typographical noise, irregular formatting, and highly recurrent administrative phrases that can function as shortcut features instead of genuine semantic evidence. Research on zero-shot document categorization has already emphasized the importance of transferable semantic representations when labeled task-specific examples are unavailable or severely restricted (Song et al., 2019). Performance variability reported in limited-data academic classification also cautions against assuming that a model effective in one document domain will exhibit equivalent behavior under different class distributions and linguistic structures (Molina & Sanchez, 2022). A scientifically defensible classification pipeline must consequently evaluate document extraction, noise removal, semantic representation, training-data allocation, and final classification performance as interacting components rather than attributing predictive quality to the classifier alone.

The practical urgency of this problem is evident in the UNISSULA case examined in this study, where only 145 official correspondence documents were available from BAAk, LPPM, LPKA, LKPI,

and BAU and had to be classified into Application Letters, Decision Letters, Invitation Letters, and Assignment Letters. Such a dataset is too limited to justify assumptions associated with data-intensive supervised training, while its imbalanced and institution-specific structure also complicates conventional performance interpretation. Accuracy alone may obscure class-level weaknesses in imbalanced classification, making precision, recall, and F1-score important complementary indicators of predictive behavior (Mursyidah et al., 2024). Precision, recall, and F1-score provide distinct perspectives on prediction reliability, retrieval completeness, and their harmonic balance, which becomes particularly relevant when minority classes contain relatively few observations (Aminah & Aminah, 2024). The proposed research consequently positions SetFit not as a generic replacement for larger transformer models, but as an efficient semantic-classification mechanism specifically matched to a low-resource institutional environment. Its engineering relevance lies in combining heterogeneous PDF extraction, correspondence-specific text cleaning, dynamic N-shot sampling, contrastive representation learning, and lightweight classification within one operational workflow. This position addresses a practical document-management problem while simultaneously testing whether few-shot semantic learning can remain reliable when the available training evidence is both scarce and structurally constrained.

The objective of this study is to develop and evaluate an efficient few-shot classification framework for Indonesian academic correspondence using SetFit under a severely limited labeled-data setting. The study progressively examines training configurations from one to six examples per category and selects the configuration that provides the most effective classification performance while maintaining strict separation between training and testing documents. Its theoretical contribution lies in extending the empirical understanding of few-shot semantic classification from general benchmark tasks toward a highly formal, repetitive, and institution-specific Indonesian document domain. Its methodological contribution is the integration of heterogeneous PDF text extraction, targeted removal of correspondence-specific noise, dynamic N-shot data allocation, contrastive sentence-representation learning, and multi-metric performance evaluation within a single reproducible workflow. The research also contributes an engineering-oriented implementation perspective by connecting the selected classification model with a lightweight Streamlit prototype rather than treating model evaluation as an isolated computational exercise. More broadly, the study provides evidence on the extent to which an efficient few-shot architecture can support intelligent document management when institutional data scarcity prevents the use of conventional large-scale supervised training.

RESEARCH METHODS

This study was designed as an empirical quantitative experiment to develop and evaluate an automatic classification system for Indonesian academic correspondence under severely limited labeled-data conditions. The dataset comprised 145 official PDF documents collected from Biro Administrasi Akademik (BAAk), Lembaga Penelitian dan Pengabdian kepada Masyarakat (LPPM), Lembaga Pengembangan Kemahasiswaan dan Alumni (LPKA), Lembaga Kajian dan Penerapan Nilai-Nilai Islam (LKPI), and Biro Administrasi Umum (BAU) at Universitas Islam Sultan Agung, with each document assigned to one of four functional classes: Application Letter, Decision Letter, Invitation Letter, or Assignment Letter. Because text classification requires a consistent machine-readable representation before category learning can be performed, the document-processing pipeline standardized heterogeneous PDF inputs prior to model training (Minaee et al., 2021). Native digital PDFs were extracted directly using PyMuPDF, whereas scanned PDFs were processed using Doctr OCR; the resulting raw text was cleaned by removing institutional letterheads, correcting identified typographical errors, deleting irrelevant characters such as `_`, `|`, `!`, `$`, `{}`, and `[]`, and normalizing excessive spaces and paragraph formatting. Each cleaned document was manually assigned a ground-truth label and stored in a CSV dataframe containing the document name, cleaned text, and validated class before being transferred to the SetFit pipeline, whose architecture combines sentence-level representation adaptation with a lightweight classification stage (Tunstall et al., 2022). The pretrained backbone used in the experiment was `intfloat/multilingual-e5-base`, and no architectural component or hyperparameter absent from the original final-project implementation was added to the journal version. The selection of a few-shot strategy was consistent with the study objective of exploiting prior model knowledge when only a small number of supervised examples are available for a domain-specific classification task (Wang et al., 2021).

Model evaluation followed an incremental N-shot experiment from 1-shot to 6-shot, allowing the number of training examples per category to increase systematically while documents not selected for training remained outside the training subset. This experimental structure reflects the central principle of few-shot learning, in which model behavior is examined under progressively limited numbers of labeled examples rather than through a conventional fixed percentage split (Parnami & Lee, 2022). In the final 6-shot configuration, six documents from each of the four correspondence categories were allocated to training, yielding 24 training documents, while the remaining documents designated for evaluation were kept separate to prevent data leakage. The implementation converted the sampled dataframe into a Hugging Face Dataset, initialized the multilingual-e5-base model through SetFit, applied CosineSimilarityLoss for contrastive representation learning, trained the classifier using SetFitTrainer, generated predictions through model.predict, and preserved the selected model using save_pretrained. Classification performance was assessed with a confusion matrix and accuracy, where accuracy represented the proportion of correctly classified documents across the evaluated observations (Mursyidah et al., 2024). Precision, recall, and F1-score were additionally retained to evaluate class-level prediction reliability, retrieval completeness, and the balance between both dimensions (Aminah & Aminah, 2024). Because the final project reported a weighted F1-score for the imbalanced test distribution, the article retains that reported evaluation measure without introducing an unreported recalculation, and the selected SetFit model was subsequently connected to the documented Streamlit prototype using the original database-lookup mechanism rather than an unreported real-time OCR inference procedure.

RESULTS AND DISCUSSION

Dataset Preparation and Text Preprocessing

The empirical dataset comprised 145 official correspondence documents collected from five authorized administrative and academic units at Universitas Islam Sultan Agung, namely BAAk, LPPM, LPKA, LKPI, and BAU. These documents represented different institutional contexts but were organized for classification according to functional correspondence categories rather than the organizational identity of the issuing unit. This distinction is important because digitalization in higher education increasingly requires information systems that organize administrative content according to operational meaning instead of merely preserving its original storage structure (Nugroho et al., 2025). The dataset contained both digitally generated PDF files and scanned documents, creating heterogeneity at the document-acquisition stage before any semantic modeling could be performed. Such heterogeneity is methodologically relevant because text-classification systems rely on standardized textual representations before discriminative linguistic patterns can be learned reliably (Minaee et al., 2021). The preparation process consequently served as an engineering interface between heterogeneous institutional files and the structured text representation expected by the subsequent SetFit classifier. This initial standardization also reduced the risk that variation in PDF format, rather than correspondence semantics, would become an uncontrolled source of model behavior.

A	B	C	D	E	F	G	H	I	J	K	L	M
1	nama_file	teks_hasil										
2		YAYASAN BADAN WAKAF SULTAN AGUNG UNIVERSITAS ISLAM SULTAN AGUNG (UNISSULA) Jl. Raya Kaligawe Km.4 Semarang 50112 Telp.(024) 6563564 (B Sel) Fax.(024) 6562455 N email: informas@unissula.ac.id web: www.unissula.ac.id BIRO REKTOR Bismillah Membangun Generasi Khaira Ummah SURAT EDARAN MBKM Nomor: 1329/A.1/SA/024 Kepada: Yth. Dekan, Wakil Dekan I/Sekretaris Fakultas dan Kakua Program Studi di lingkungan Universitas Islam Sultan Agung Assalamu'alaikum wr. wb Puji syukur kehadirat Allah SWT atas limpahan nikmat dan karunia- Nya, sholawat serta salam senantiasa tercurah kepada Nabi Muhammad SAW berserta keluarga, sahabat dan para pengikutnya hingga akhir jaman. Dengan ini kami menghimbau kepada Bapak/Ibu Dekan, bagi mahasiswa yang akan melakukan perkuliahan mandiri belajar magang flagship atau mandiri, studi independen, kampus mengajar, kegiatan wirausaha, proyek kemanusiaan, pertukaran pelajar, penelitian atau riset, KKN tematik wajib diinputkan ke dalam Aplikasi MKBM (app mbkm). Berikut link untuk pengisian ke MKBM https://appmbkm.unissula.ac.id/ Demikian atas perhatian dan kerjasama yang baik kami ucapkan banyak terimakasih. Waassalamu'alaikum wr. wb Semarang: 21 Rajab 1445H 02 Februari 2024 M AM SUMAA Wakil Rektor I, Andre Sugiyono, S.T./M.M., Ph.D. Tembusan 10.603.028 1. Yth. Rektor UNISSULA. UNSSNK edaran pengisian 2. Yth. Wakil Rektor II & III UNISSULA YAYASAN BADAN WAKAF SULTAN AGUNG										

Figure 1. Raw BAAk Letter Extraction Result

A1	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	nama_file	teks_hasil													
2		YAYASAN BADAN WAKAF SULTAN AGUNG UNIVERSITAS ISLAM SULTAN AGUNG (UNISSULA) Jl Raya Kaligawe Km 4 Semarang 50112 Telp (024)6583584 (8 Sal) Fax:(024) 6582455 email:untor.usicunsula.s.id web:wVV unissula.ac.id BIRO/REKTOR Bismilont Membangun Generash KharaUmmah Nomor 7967/D. 1/SA/XII/2019 Lamp. Hal PEMBERITAHUAN Kepada : Yth. Bapak/Ibu, Dr. H. Heru Sulisty, M.Si Pimpian Universitas, Dekan Fakultas, Ka. Prodi, Ka. Lembaga, Ka. Biro, SPI Universitas Islam Sultan Agung di- Semarang Assalamu'alaikum Wr. Wb. Puji syukur kita panjatkan kehadiran Allah SWT. yang telah melimpahkan rahmat dan hidayah-Nya kepada kita. Shalawat dan salam semoga tercurah kepada Nabi Agung Muhammad SAW. beserta keluarga dan para sahabatnya. Kami menginformasikan bahwa akan diadakan Pelaksanaan Audit Mutu Internal (AMI) Periode Genap 2018 - 2019 pada tanggal 2 - 12 Desember 2019, adapun ruang lingkup dan area serta jadwal pelaksanaan. Audit terlampir. Kami mohon Bapak/Ibu pimpinan dapat mengupload dokumen terkait standar yang di audit ke sim.unissula.ac.id paling lambat tanggal 2 Desember 2019. Atas perhatiannya disampaikan terima kasih. Wassalamu'alaikum Wr. Wb. Semarang, 01 Rabi'ul Akhir 1441H 28 November 2019 M EKIOTUNISSULA, - :: a - UNI LEL IrHPrabowo Setiawan, MT, Ph.D vaxs NIK/210293017 Tembusan: 1. Yth. Wakil Rektor I, II dan III UNISSULA. 2. Arsip. Dipindai dengan 1CamScanner 0 CamScanner YAYASAN BADAN WAKAF SULTAN AGUNG UNIVERSITAS ISLAM SULTAN AGUNG (UNISSULA)													

Figure 2. Raw LPPM Letter Extraction Result

A1	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	nama_file	teks_hasil												
2		YAYASAN BADAN WAKAF SULTAN AGUNG UNIVERSITAS ISLAM SULTAN AGUNG (UNISSULA) UNISSU Jl. Raya Kaligawe Km.4 Semarang 50112Telp. (024) 6583584 (8 Sal) Fax. (024)6582455 email: informasi@unissula.ac.id web : www.unissula.ac.id LKPI Bismillah Membangun Generasi Khaira Ummah Nomor 037/D. 1/LKPI-SA/VI/2025 Lamp Hal Permohonan Tutor Pengajar Bahasa Inggris Kepada : Yth. Kepala CILAD Universitas Islam Sultan Agung (UNISSULA) Di - Semarang Assalamu'alaikum Wr. Wb. Puji syukur kami panjatkan kehadiran Allah SWT. Atas segala limpahan rahmat dan karunia-Nya. Sholawat serta salam semoga tetap tercurah kepada Nabi Muhammad SAW. beserta para keluarga, sahabat, dan pengikut-pengikutnya sampai akhir zaman. Amin. Diberitahukan dengan hormat bahwa, berdasarkan surat Rektor Universitas Islam Sultan Agung nomor 5694/J/SA/VI/2025 tentang Peningkatan Kompetensi Bahasa Internasional kami mohon kepada Bapak untuk berkenan mengirimkan 1 (satu) orang tutor untuk mengajar bahasa inggris di Lembaga Kajian dan Penerapan Nilai- nilai Islam (LKPI) UNISSULA. Demikian surat permohonan ini, atas perhatian Bapak kami sampaikan terima kasih. Wassalamu'alaikum Wr. Wb. Semarang, 06 Muharram 1447H 02 Juli 2025 M - Pl, LKPI UN Farlian, S.Pd.I., S.Hum., M.Pd.LA 1Permohonan T NIK. 2115 16 025 YAYASAN BADAN WAKAF SULTAN AGUNG UNIVERSITAS ISLAM SULTAN AGUNG (UNISSULA) UNI Jl. Raya Kaligawe Km.4 Semarang 50112Telp. (024) 6583584 (8 Sal) Fax. (024)6582455 email: informasi@unissula.ac.id web : www.unissula.ac.id LKPI												

Figure 3. Raw LKPI Letter Extraction Result

A1	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	nama_file	teks_hasil												
2		YAYASAN BADAN WAKAF SULTAN AGUNG UNIVERSITAS ISLAM SULTAN AGUNG (UNISSULA) Jl. Raya Kaligawe Km.4 Semarang 50112 Telp.(024) 8583584 (8 Sal) Fax.(024) 8582455 email: informasi@unissula.ac.id web: www.unissula.ac.id Biro Rektor Bismillahi Membangun Generasi Khairal Ummah Nomor 7461/D.1/SA/VIII/2025 Lampiran : 1 (satu) lembar Perihal : PEMBERITAHUAN PEKAN TA'ARUF MAHASISWA BARU UNIVERSITAS ISLAM SULTAN AGUNG (UNISSULA) TAHUN AKADEMIK 2025/2026 Kepada : Yth. Dekan Fakultas di lingkungan Universitas Islam Sultan Agung Assalamu'alaikum Wr. Wb. Puji syukur kami panjatkan kehadirat Allah SWT yang telah melimpahkan rahmat, hidayah serta inayah-Nya kepada kita semua. Sholawat serta salam senantiasa tercurahkan kepada Nabi Muhammad SAW. Bahwa berdasarkan Keputusan Rektor UNISSULA Nomor 7403/A.1/SA/VIII/2025 tentang Perubahan Kalender Akademik Universitas Islam Sultan Agung (UNISSULA) Tahun Akademik 2025/2026, Pelaksanaan Pekan Ta'aruf Mahasiswa Baru Universitas Islam Sultan Agung (UNISSULA) Tahun Akademik 2025/2026, dilaksanakan sebagai berikut: TANGGAL AGENDA KEGIATAN TEMPAT 1-4 4 September 2025 Universitasir Auditorium UNISSULA 5- 7 September 2025 Fakultas Fakultas Demikian surat pemberitahuan ini disampaikan. Atas perhatian dan kerjasama yang baik, diucapkan terima kasih. Wassalamu'alaikum Wr. Wb. Semarang, 14 Shafar 1447 H 08 Agustus 2025 M a.n Rektor, Wakil Reator III EQ												

Figure 4. Raw BAU Letter Extraction Result

hasil_ekstraksi_LPKA

File Edit Tampilan Sisipkan Format Data Alat Gemini Ekstensi Bantuan

Menu 100% 123 Default... 10 B I A

A1 nama_file

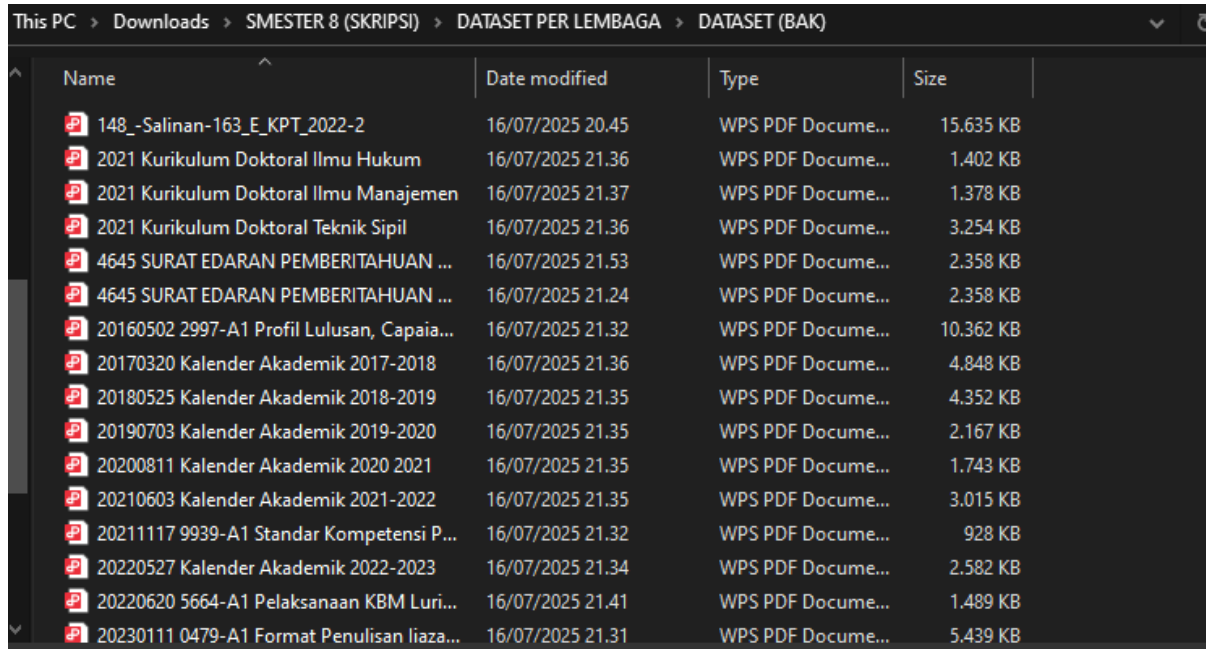
A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	nama_file	teks_hasil												
2		YAYASAN BADAN WAKAF SULTAN AGUNG UNIVERSITAS ISLAM SULTAN AGUNG (UNISSULA) Jl. Raya Kaligawe Km.4 Semarang 50112 Telp. (024) 8583584 (8 Saly Fax 1024) 8582455 email : informasi@unissula.ac.id web: www.unissula.ac.id BAGAFENGEMANGAN KEMAHASISWAAN DAN ALUMNI Bismillahi Membangun! Generasi Khataummany Nomor : 0539/LPKA/SA/IX/2025 Lampiran : 1 Lembar Perihal : Permohonan Pinjam Tempat Kepada : Yth. Roktor Universitas Islam Sultan Agung (UNISSULA) di - Semarang Assalamu'alaikum Wr. Wb. Puji syukur kita panjatkan kehadiran Allah SWT yang telah melimpahkan rahmat dan hidayahNya kepada kita. Shalawat dan salam semoga tercurah kepada Nabi Agung Muhammad SAW beserta keluarga dan para sahabatnya. Schubungan akan dilaksanakannya kegiatan "Latihan Rutin BSO Futsal Fakultas Ilmu Keperawatan", maka LPKA mengajukan permohonan pinjaman tempat guna mensupport kegiatan tersebut yang lnsya Allah dilaksanakan pada: Hari : Setiap Hari Rabu Tanggal : 30 September - 31 Oktober 2025 Waktu : 19.00 WIB - 22.00 WIB Tempat : Sportcenter UNISSULA Demikian permohonan ini kami sampaikan, atas perhatiannya diucapkan terimakasih. Wassalamu'alaikum Wr. Wb. Semarang, 8 Rabiul Akhir 1447H 30 September 2025 M Ka LPKAUnissula, Ady Fajar Putranto,SE NIK. 11001 424 Tembusan :												

+ hasil ekstraksi LPKA

Figure 5. Raw LPKA Letter Extraction Result

Text extraction was implemented through two processing routes determined by the physical structure of each PDF file. Native digital PDFs were processed directly using PyMuPDF because their embedded text layer could be accessed without optical character recognition, whereas scanned documents required an OCR stage before their contents could be transformed into machine-readable text. This dual-path design avoided excluding image-based institutional records and allowed both document types to enter a common downstream processing workflow. The use of transformer-based and pretrained representations in text classification presupposes that input text has first been converted

into a computationally consistent form from which contextual relationships can be modeled (Amatriain et al., 2023). Indonesian NLP research similarly demonstrates that the quality of downstream linguistic modeling depends on how textual input is represented before classification or generation tasks are executed (Cahyawijaya et al., 2021). The resulting extracted texts were consolidated before cleaning so that subsequent processing rules could be applied consistently regardless of whether the original document was digitally generated or scanned.



Name	Date modified	Type	Size
148_Salinan-163_E_KPT_2022-2	16/07/2025 20.45	WPS PDF Docume...	15.635 KB
2021 Kurikulum Doktoral Ilmu Hukum	16/07/2025 21.36	WPS PDF Docume...	1.402 KB
2021 Kurikulum Doktoral Ilmu Manajemen	16/07/2025 21.37	WPS PDF Docume...	1.378 KB
2021 Kurikulum Doktoral Teknik Sipil	16/07/2025 21.36	WPS PDF Docume...	3.254 KB
4645 SURAT EDARAN PEMBERITAHUAN ...	16/07/2025 21.53	WPS PDF Docume...	2.358 KB
4645 SURAT EDARAN PEMBERITAHUAN ...	16/07/2025 21.24	WPS PDF Docume...	2.358 KB
20160502 2997-A1 Profil Lulusan, Capaia...	16/07/2025 21.32	WPS PDF Docume...	10.362 KB
20170320 Kalender Akademik 2017-2018	16/07/2025 21.36	WPS PDF Docume...	4.848 KB
20180525 Kalender Akademik 2018-2019	16/07/2025 21.35	WPS PDF Docume...	4.352 KB
20190703 Kalender Akademik 2019-2020	16/07/2025 21.35	WPS PDF Docume...	2.167 KB
20200811 Kalender Akademik 2020 2021	16/07/2025 21.35	WPS PDF Docume...	1.743 KB
20210603 Kalender Akademik 2021-2022	16/07/2025 21.35	WPS PDF Docume...	3.015 KB
20211117 9939-A1 Standar Kompetensi P...	16/07/2025 21.32	WPS PDF Docume...	928 KB
20220527 Kalender Akademik 2022-2023	16/07/2025 21.34	WPS PDF Docume...	2.582 KB
20220620 5664-A1 Pelaksanaan KBM Luri...	16/07/2025 21.41	WPS PDF Docume...	1.489 KB
20230111 0479-A1 Format Penulisan liaza...	16/07/2025 21.31	WPS PDF Docume...	5.439 KB

Figure 6. Letter Data Before Text Extraction

The raw extracted text was not immediately suitable for classification because it contained letterheads, institutional identifiers, OCR-induced typographical distortions, irregular spacing, and characters that did not contribute directly to the functional meaning of the correspondence. These elements were treated as potential sources of representational noise because a high-capacity language model can exploit recurrent surface features even when those features are not conceptually related to the intended class. Transformer-based models are particularly effective at learning discriminative textual patterns, which makes careful control of irrelevant recurrent cues important when the target dataset is small and institution-specific (Anwar et al., 2021). The preprocessing routine removed institutional letterheads, normalized spaces and paragraph structure, corrected identified typographical errors, and eliminated irrelevant characters such as `_`, `|`, `!`, `$`, `{}`, and `[]`. Table 1 reproduces the preprocessing examples recorded in the original experiment and illustrates how source-specific noise was converted into cleaner textual input without changing the intended substantive content. The procedure was not designed as aggressive linguistic normalization because excessive transformation could remove administrative expressions that differentiate the correspondence categories. This balance between noise reduction and semantic preservation is important for Indonesian contextual classification, where pretrained embeddings must retain locally meaningful lexical and syntactic distinctions (Nabiilah et al., 2024).

Table 1. Preprocessing Data

No.	Before	After
1	YAYASAN BADAN WAKAF SULTAN AGUNG UNIVERSITAS ISLAM SULTAN AGUNG (UNISSULA) Jl. Raya Kaligawe Km.4 Semarang 50112 Telp.(024) 6583584 (8 Sal) Fax.(024) 6582455 N email: informasi@unissula.ac.id web www.unissula.ac.id	_ (Letterhead removed)

BIRO REKTOR Bismillah Membangun Generasi Khaira		
Ummah		
2	Btsmlll,ahit'tu.hmanlrrahlm RRTKTOR ITNTERSITAS IsLlrM SITLT4II AGUNG	Bismillahirrahmanirrahim REKTOR UNIVERSITAS ISLAM SULTAN AGUNG 14 August 2023 M 2012 KEPUTUSAN PERUBAHAN KALENDER
3	L4 Agustus 2023 M	
4	2}12	
5	KEPUTUSA}I	
6	PERUBAIAN	
7	TTALET{DER	

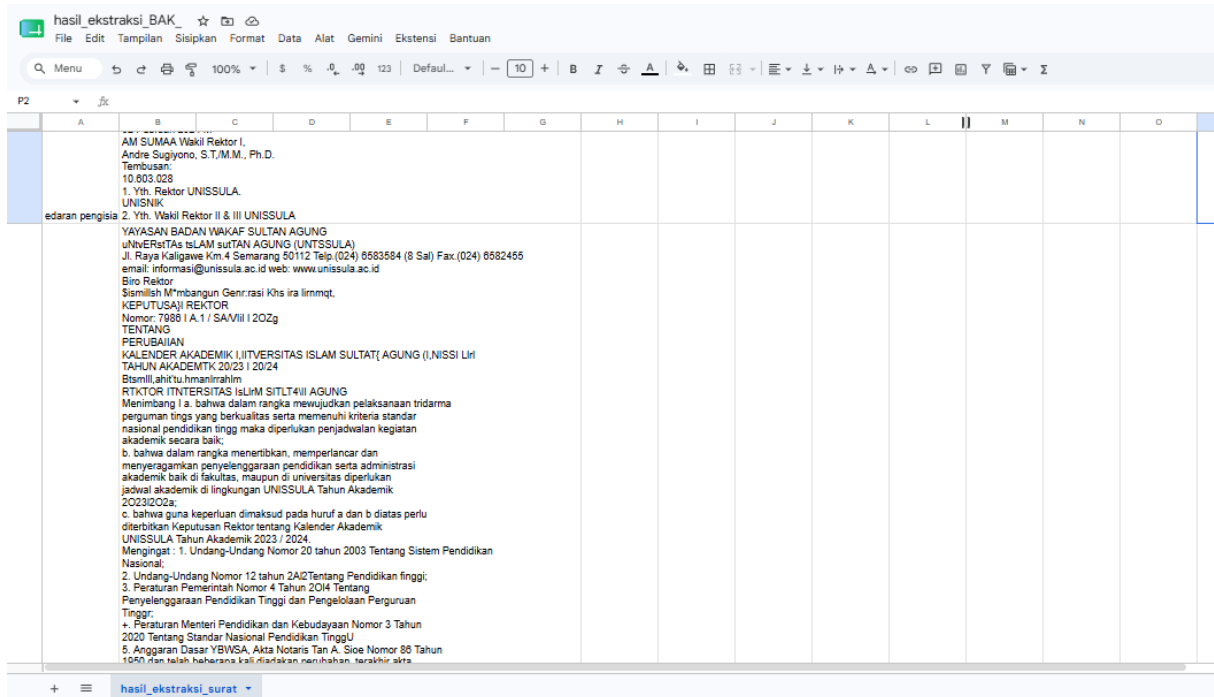


Figure 7. Data Before Preprocessing

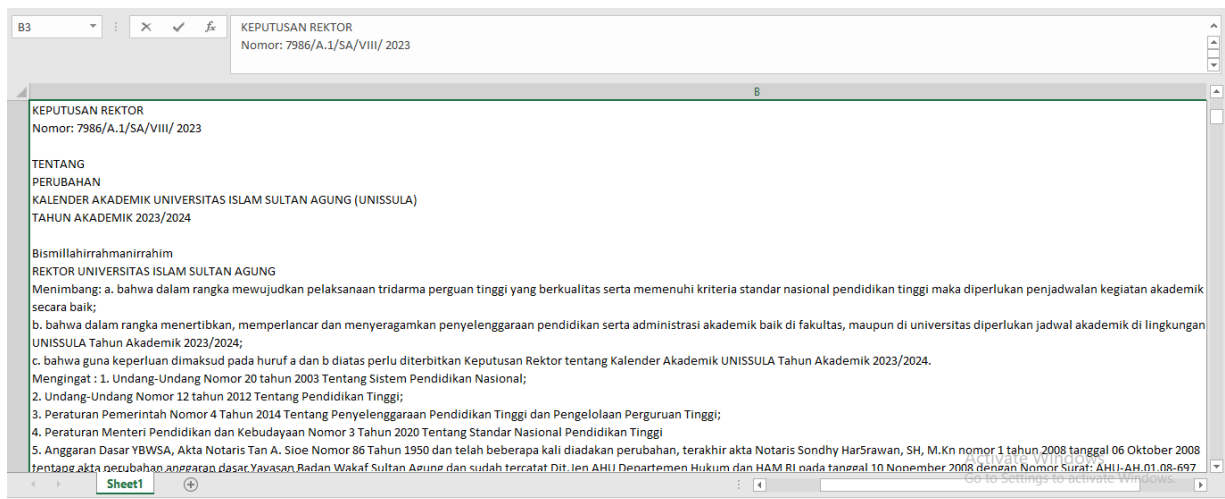


Figure 8. Data After Preprocessing

The before-and-after representations indicate that the cleaning stage addressed both predictable document formatting and errors introduced during text extraction. Removing the letterhead is especially important because institutional names and organizational identifiers occur repeatedly and could become shortcut features that permit category prediction without requiring the classifier to learn the substantive

intent of the correspondence. Deep-learning text classifiers are capable of exploiting highly informative lexical correlations even when those correlations are incidental to the conceptual classification objective (Minaee et al., 2021). The source experiment consequently attempted to direct learning toward the informational body of each document by eliminating features judged to be unrelated to correspondence function. This choice is compatible with Indonesian pretrained-language-model research showing that contextual representations can capture semantic differences more effectively when input sequences preserve meaningful linguistic content rather than noisy formatting artefacts (Koto et al., 2020). An independent ablation experiment comparing classification with and without letterhead removal was not reported, so the magnitude of the cleaning effect cannot be isolated quantitatively. The contribution of preprocessing should therefore be interpreted as the establishment of a controlled input condition for the few-shot experiment rather than as a separately validated performance improvement.

After preprocessing, each cleaned document was manually assigned a ground-truth class corresponding to Application Letter, Decision Letter, Invitation Letter, or Assignment Letter. Manual labeling was required because the subsequent supervised few-shot procedure needed an explicit reference category against which model predictions could be evaluated. Clear target definitions are particularly important in Indonesian document classification because contextual embeddings can only be interpreted as task-specific predictors after the semantic representation has been associated with stable class labels (Juarto & Yulianto, 2023). The cleaned text, document identifier, and validated label were subsequently stored in a CSV dataframe that served as the direct input to the SetFit training workflow. This structured representation separated document acquisition and extraction from the learning stage, preventing the classifier from operating directly on heterogeneous PDF files. Similar Indonesian classification research has demonstrated that pretrained semantic models can produce strong results when textual observations are represented consistently before task-specific adaptation (Syazali & Yulianti, 2025). The resulting dataframe also enabled the controlled selection of training examples for the incremental N-shot experiments conducted in the next stage.

The significance of this preparation strategy becomes greater in a few-shot configuration because every training document contributes a relatively large share of the evidence available to the classifier. Few-shot learning is designed to exploit prior model knowledge when only a small number of labeled task-specific examples can be supplied, which makes sample quality particularly consequential (Wang et al., 2021). Reviews of few-shot methodologies likewise emphasize that performance under scarce supervision can be highly sensitive to the representativeness and consistency of the selected examples (Parnami & Lee, 2022). A document containing extraction errors, residual institutional identifiers, or an incorrect ground-truth label could consequently exert greater influence than the same error would in a large supervised corpus. The present workflow mitigated that risk by completing text extraction, noise removal, normalization, and manual labeling before the documents were subjected to dynamic N-shot sampling. The standardized dataset provided a consistent empirical basis for SetFit adaptation without introducing synthetic augmentation, additional correspondence documents, or preprocessing operations that were absent from the original experiment. This design preserves the limited-data character of the study while ensuring that the subsequent evaluation reflects the behavior of the few-shot classifier on prepared institutional text rather than uncontrolled variation in raw document format.

SetFit Training and Few-Shot Performance

The SetFit experiment was designed around a controlled separation between documents used for model adaptation and documents retained for independent evaluation, reflecting the limited-data character of the research rather than a conventional percentage-based train-test split. In the selected 6-shot configuration, six documents from each of the four correspondence categories were assigned to the training set, resulting in 24 training documents, while the remaining eligible documents were placed in the testing pool. This design is consistent with few-shot learning, where model adaptation is intentionally performed with a small number of labeled examples so that sample efficiency becomes part of the experimental objective rather than merely a dataset constraint (Wang et al., 2021). The source implementation also applied a two-layer validation procedure to prevent overlapping documents from appearing simultaneously in the training and testing partitions, thereby reducing the risk of data leakage. The training distribution was manually balanced across the four classes, whereas the test distribution retained the original imbalance of the available correspondence data. SetFit was selected because its

architecture was specifically developed to perform text classification efficiently under restricted-label conditions without depending on large prompt-based models (Tunstall et al., 2022). This experimental arrangement allowed the subsequent performance comparison to reflect the classifier's ability to generalize from scarce institutional examples rather than memorization arising from duplicated records.

Distribusi label latih:	
label	
Surat Keputusan	6
Surat Permohonan	6
Surat Tugas	6
Surat Undangan	6
Distribusi label uji (manual):	
label	
Surat Permohonan	58
Surat Keputusan	26
Surat Undangan	13
Surat Tugas	2

Figure 9. SetFit Training and Testing Dataset Distribution

The semantic backbone used in the training pipeline was intfloat/multilingual-e5-base, which provided the pretrained sentence-level representation required by the SetFit architecture. The sampled training dataframe was converted into a Hugging Face Dataset before the pretrained model and its final classification component were initialized through SetFitModel.from_pretrained. Transformer-based representations are particularly suitable for this stage because they encode contextual dependencies across textual sequences rather than relying only on isolated lexical frequencies (Lin et al., 2022). Contrastive learning was configured through CosineSimilarityLoss, which encouraged semantically related correspondence examples to occupy closer positions in the representation space while separating documents associated with different classes. This mechanism follows the conceptual logic of contrastive SetFit training, which has shown advantages over conventional fine-tuning when labeled examples are scarce (Kilic et al., 2023). The implementation subsequently used SetFitTrainer to control the training procedure, trainer.train() to execute the adaptation process, model.predict() to generate predictions on the testing data, and model.save_pretrained() to preserve the selected model for later integration.

Table 2. Main Components of the SetFit Training Process

No.	Code	Function	Description
1	df.sample(min	Dynamic Sample Selection (N-Shot)	Randomly samples documents from each class according to the number of shots, with an automatic limit to prevent excess-data errors.
2	Dataset.from_pandas	Dataset Format Conversion	Converts the tabular data structure into the Hugging Face Dataset format required by SetFit.
3	SetFitModel.from_pretrained	Base Model Creation	Loads the base model architecture (multilingual-e5-base) and integrates it with the final classifier (classification head).
4	CosineSimilarityLoss	Contrastive Learning Configuration	Sets the loss function to bring vector representations of similar documents closer and move representations of different documents farther apart.

5	SetFitTrainer	Training Parameter Configuration	Configures the machine-learning rules, including the evaluation metric (accuracy), batch size, learning epochs, and data iterations.
6	trainer.train	Execute Training	Runs the full training process, from sentence-meaning adjustment (fine-tuning) to category-prediction training.
7	model.predict	Model Prediction Testing	Generates predictions for the testing set using the trained model to evaluate its classification capability.
8	model.save_pretrained	Final Model Export	Saves the artificial-intelligence model from the best iteration to Google Drive so that it can be integrated with the Streamlit website application.

Model development was evaluated incrementally from 1-shot to 6-shot rather than relying on a single arbitrarily selected training size. At each experimental stage, the number of training examples per category increased by one, while the remaining eligible correspondence documents continued to function as testing data. This arrangement made it possible to observe how relatively small additions of labeled information affected the model's ability to classify new documents. Such incremental analysis is methodologically appropriate in few-shot learning because the central question concerns how much supervision is required before a pretrained model begins to produce useful task-specific generalization (Parnami & Lee, 2022). The original learning-curve output shows training accuracy remaining at 1.00 throughout the displayed N-shot configurations, whereas testing accuracy fluctuated as the number of examples increased. The final 6-shot configuration was reported as the best-performing setting, with testing accuracy reaching approximately 85.9%, consistent with the final reported accuracy of 85.87%. This pattern indicates that additional examples improved generalization up to the selected configuration, although the variability across intermediate shot levels also shows that performance did not improve in a strictly linear manner.

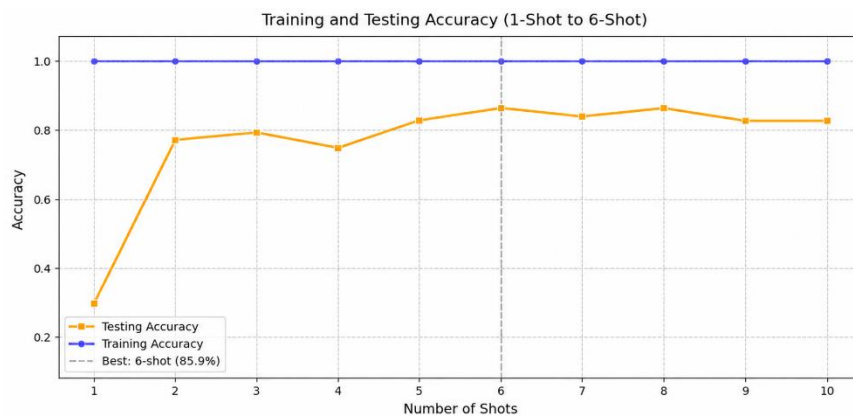


Figure 10. Training and Testing Accuracy from 1-Shot to 6-Shot

The training-loss trajectory was monitored to assess whether representation adaptation progressed toward a stable solution during SetFit fine-tuning. The original graph shows a pronounced decline in loss during the early stage of training, followed by values that remained close to the lower part of the plotted scale. This declining pattern was interpreted in the source study as evidence that the model progressively reduced its internal fitting error while learning distinctions among the correspondence categories. Contrastive representation learning is expected to produce such behavior when the optimization process increasingly aligns examples from related categories and separates semantically different examples in embedding space (Kilic et al., 2023). The reported trajectory is also

compatible with the general purpose of pretrained transformer adaptation, where a broadly learned semantic representation is adjusted toward a narrower downstream classification objective (Amatriain et al., 2023). No additional numerical loss values, confidence bounds, or alternative convergence statistics were reported in the final project, so the graphical output remains the primary evidence for this training behavior.

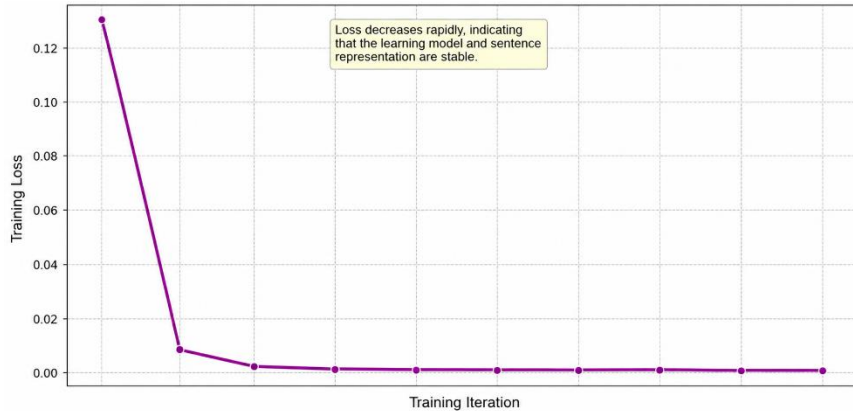


Figure 11. Training Loss

The selected 6-shot model was also examined through a diagnostic comparison between training and testing error to assess whether the observed learning behavior was accompanied by an obvious overfitting pattern. In the original figure, the training-error line remained near the baseline, while the testing-error line stayed above it and changed within a relatively limited range across the displayed observations. The testing curve did not exhibit the pronounced upward divergence that the source study associated with severe deterioration in generalization. This observation is relevant because few-shot classifiers can easily memorize a very small training subset when the semantic representation does not transfer adequately to unseen examples. SetFit was developed precisely to improve label efficiency by adapting sentence representations before training the lightweight classifier, which can reduce dependence on large supervised datasets (Tunstall et al., 2022). Related SetFit research in another formal document domain has also reported that useful classification can be achieved when only a limited number of labeled examples are available (Khaled & Manayer, 2024). The graph nevertheless provides descriptive rather than statistical evidence because repeated-run variance, confidence intervals, or an independent formal overfitting test were not reported in the source study.

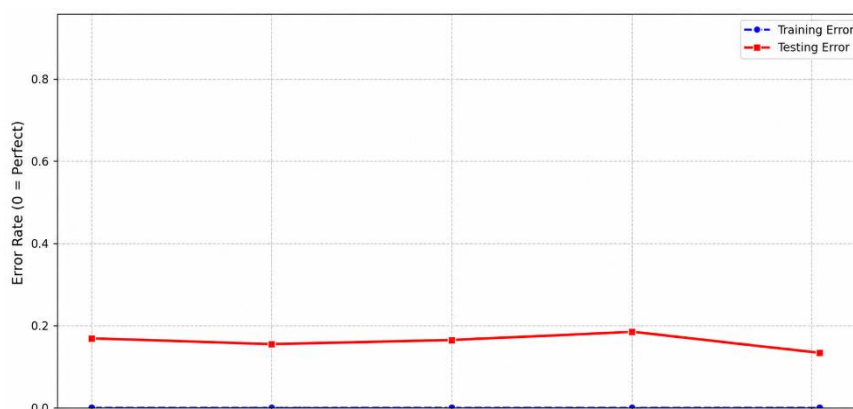


Figure 12. 6-Shot Error Rate

The combined evidence from the learning curve, declining training loss, and error-rate diagnostics led the original experiment to retain the 6-shot configuration as the final model for evaluation and prototype integration. Its principal computational advantage is that the selected classifier required only six labeled examples per class during the adaptation stage, which is consistent with the sample-efficiency objective of few-shot learning. The achieved performance is also conceptually

aligned with previous work showing that pretrained language representations can support strong classification behavior even when task-specific labeled resources are limited (Koto et al., 2020). The non-monotonic variation across the N-shot experiments should still be treated as an important empirical feature rather than ignored, because limited-data models can remain sensitive to which examples are selected for adaptation. Similar concerns have been raised in research on classification under restricted academic datasets, where performance may vary according to sampling conditions and class representation (Molina & Sanchez, 2022). The retained model was exported through `save_pretrained` and subsequently used as the classification engine for the Streamlit prototype without modifying the training architecture documented in the final project. This stage establishes the empirical basis for the next evaluation, where performance must be examined not only through aggregate accuracy but also through the confusion matrix and class-level precision, recall, and F1-score.

Classification Performance and Prototype Implementation

After the selected 6-shot SetFit model had been trained and preserved, it was integrated into a lightweight Streamlit interface to demonstrate the practical use of the classification pipeline. The interface was designed around a direct PDF-upload interaction so that users could submit an academic correspondence document and obtain a predicted category without navigating through a complex menu structure. In the source implementation, inference at the prototype stage did not perform real-time OCR; instead, the uploaded filename was matched against preprocessed text stored in `DATASET_SKRIPSI_FIX.CSV`, which allowed the experiment to isolate classification behavior from additional recognition errors. This implementation choice is consistent with the broader engineering principle that modular system design can separate data acquisition, representation, and predictive inference into independently controllable components (Zhou et al., 2020). The prototype should consequently be interpreted as a demonstration of the trained classifier rather than as a complete production system capable of processing arbitrary unseen scanned documents end-to-end. Its practical value lies in showing that a few-shot semantic classifier trained from a small institutional dataset can be exposed through a simple user-facing application without altering the trained model architecture.



Figure 13. Academic Correspondence Classification System Interface

The interaction sequence remained intentionally compact in order to keep the interface focused on the classification function itself. Once a PDF was uploaded, the application displayed the filename as confirmation that the document had been received and activated the command used to initiate the classification process. After execution, the interface returned a predicted correspondence category and provided an expandable panel containing the cleaned text that had been supplied to the trained model. This type of transparency is useful because contextual language models derive predictions from distributed textual representations that are less directly interpretable than conventional rule-based classifiers (Amatriain et al., 2023). The displayed cleaned text also provides a practical means of confirming that letterheads and other preprocessing noise had been removed before classification was performed. Similar applications of pretrained Indonesian language representations demonstrate that reliable downstream classification depends on the quality of the linguistic information passed to the

model rather than on superficial document formatting alone (Nabiilah et al., 2024). The prototype interface consequently links the earlier preprocessing decisions with the final prediction stage while retaining the lightweight implementation documented in the original study.

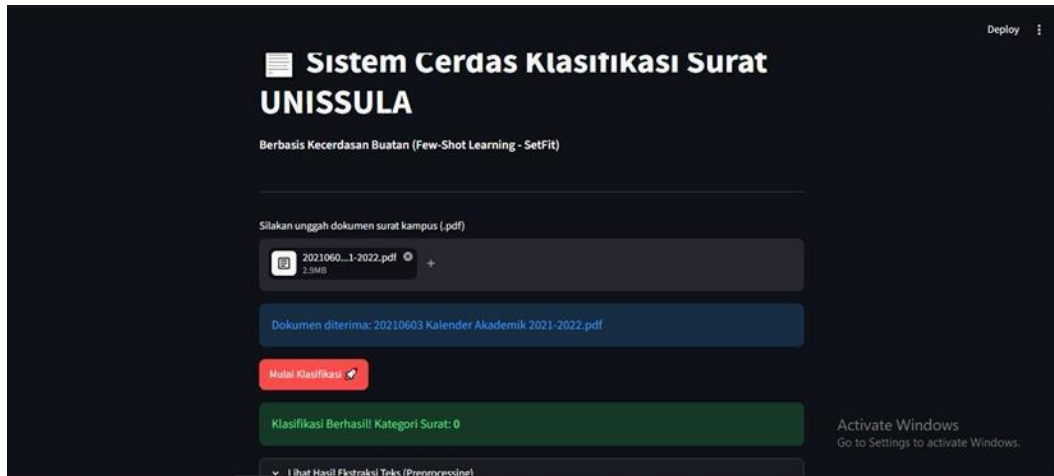


Figure 14. Starting the Correspondence Classification Process

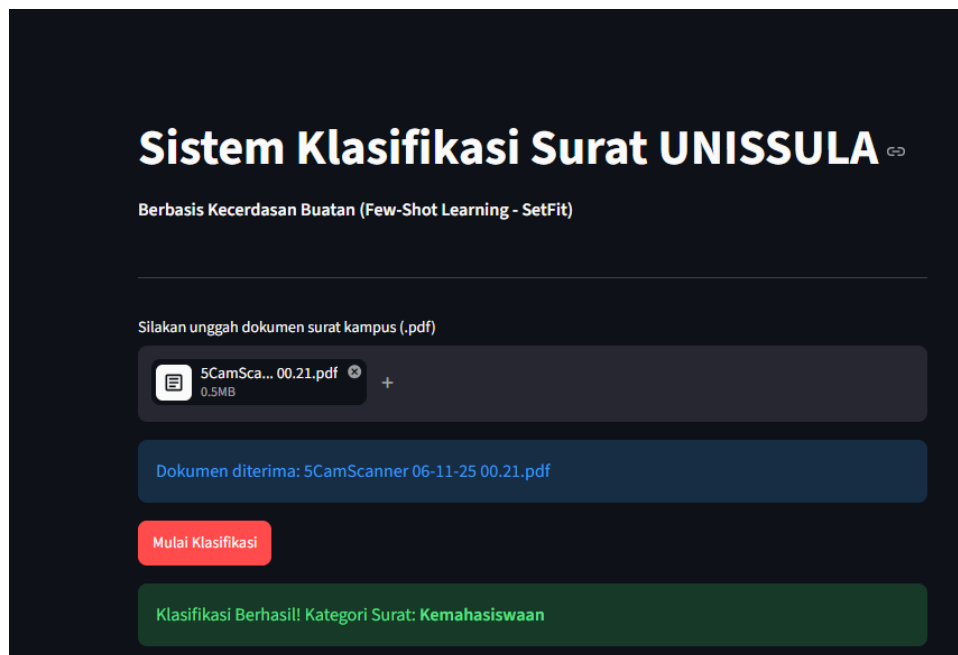


Figure 15. Classification Result Display

The final model was evaluated by comparing predicted labels with the ground-truth labels assigned during dataset preparation, using a confusion matrix and complementary classification metrics. The thesis initially states that 95 documents were supplied as testing observations, whereas the final classification report shown later in the same source contains a total support of 92 documents. Because no validated explanation is provided for this numerical difference, neither value is recalculated or silently reconciled in the present article. Accuracy remains useful for summarizing the proportion of correct classifications, but its interpretation is incomplete when class frequencies are unequal because high-frequency categories can dominate the aggregate score (Mursyidah et al., 2024). Precision, recall, and F1-score provide additional information concerning prediction reliability, retrieval sensitivity, and the balance between those two dimensions at class level (Aminah & Aminah, 2024). The confusion matrix reproduced in the original experiment shows most observations concentrated along its principal diagonal, indicating that a substantial proportion of the evaluated documents were mapped to their

reported target labels. Such a pattern provides direct empirical evidence of useful discrimination while still permitting inspection of misclassification behavior that aggregate accuracy alone cannot reveal.

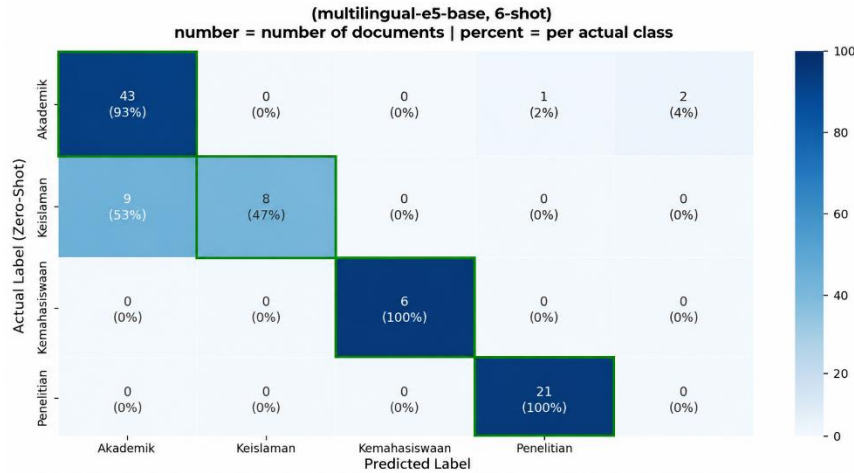


Figure 16. SetFit Confusion Matrix

Table 3 presents the final 6-shot classification report exactly as documented in the source study, preserving the reported class labels, metric values, and support counts. The report shows an exact stated accuracy of 85.87%, while the rounded value displayed in the tabular classification output is 0.86. A weighted-average F1-score of 0.85 indicates that the model achieved comparatively strong aggregate performance after accounting for the unequal frequency of observations across the reported classes. Weighted aggregation is particularly informative when the class distribution is imbalanced because each class contributes according to its observed support rather than receiving identical weight regardless of sample frequency. Few-shot document classification research has similarly shown that useful performance can be achieved with SetFit when only restricted labeled examples are available, although the resulting class-level behavior may remain uneven (Khaled & Manayer, 2024). The reported performance also supports the central rationale of efficient few-shot classification, in which pretrained semantic representations are adapted with substantially fewer labeled examples than conventional supervised training would ordinarily require (Tunstall et al., 2022). The metrics should nevertheless be interpreted jointly rather than using accuracy as the sole indication of model quality.

Table 3. Final 6-Shot Classification Metrics

Class	Precision	Recall	F1-score	Support
Akademik	0.83	0.93	0.88	46
Keislaman	1.00	0.47	0.64	17
Kemahasiswaan	1.00	1.00	1.00	6
Penelitian	0.91	1.00	0.95	21
Umum	0.33	0.50	0.40	2
Accuracy	-	-	0.86	92
Macro average	0.81	0.78	0.77	92
Weighted average	0.88	0.86	0.85	92

Source: Final 6-shot classification report reproduced from the original study; reported exact accuracy = 85.87%

```
=====
LAPORAN KLASIFIKASI FINAL (6-SHOT)
=====
```

	precision	recall	f1-score	support
Akademik	0.83	0.93	0.88	46
Keislaman	1.00	0.47	0.64	17
Kemahasiswaan	1.00	1.00	1.00	6
Penelitian	0.91	1.00	0.95	21
Umum	0.33	0.50	0.40	2
accuracy			0.86	92
macro avg	0.81	0.78	0.77	92
weighted avg	0.88	0.86	0.85	92
Accuracy	: 85.87%			

Figure 17. SetFit Evaluation Metrics

The class-level values in Table 3 reveal substantial heterogeneity that is obscured by the aggregate accuracy figure. Akademik, the largest reported class with 46 observations, obtained precision of 0.83, recall of 0.93, and an F1-score of 0.88, indicating that most relevant cases were retrieved although some predictions assigned to this category were not correct. Keislaman produced perfect precision of 1.00 but recall of only 0.47, which means that the predictions made for this class were highly precise while a considerable proportion of its actual observations were not retrieved. Kemahasiswaan recorded precision, recall, and F1-score of 1.00 across six observations, whereas Penelitian achieved an F1-score of 0.95 across 21 observations. The Umum class exhibited the weakest result, with precision of 0.33, recall of 0.50, and F1-score of 0.40, but its support consisted of only two observations, making the metric highly sensitive to a single error. Variability of this type is characteristic of limited-data classification, where minority-class estimates can change sharply when only a small number of test instances are available (Molina & Sanchez, 2022). The difference between the macro F1-score of 0.77 and weighted F1-score of 0.85 further demonstrates that strong performance in higher-support classes contributes materially to the aggregate evaluation.

The final results also require careful interpretation because the source study contains an unresolved discrepancy between the four functional correspondence categories defined in the research design and the five labels displayed in the final classification report. The methodological sections define Application Letter, Decision Letter, Invitation Letter, and Assignment Letter as the intended target categories, whereas the final output shown in Table 3 uses Akademik, Keislaman, Kemahasiswaan, Penelitian, and Umum. No validated mapping is provided in the thesis that would allow these naming schemes to be converted into one another without introducing assumptions, so the final metric labels are preserved exactly as they appear in the original output. This issue does not erase the reported 85.87% accuracy or weighted F1-score of 0.85, but it limits how confidently class-specific findings can be interpreted against the originally stated four-category objective. Research on low-resource and cross-domain classification likewise emphasizes that transfer performance depends not only on model architecture but also on the consistency of task definitions and label semantics (Song et al., 2019). The observed aggregate performance remains compatible with the broader capacity of pretrained representations to support Indonesian-language classification under limited supervision (Koto et al., 2020). From an engineering perspective, the experiment demonstrates a functioning SetFit-based classification pipeline and lightweight prototype, while label consistency and production-level real-time OCR remain methodological issues that should be resolved before broader institutional deployment.

CONCLUSION

This study demonstrated the feasibility of using SetFit for efficient few-shot classification of Indonesian academic correspondence under severely limited labeled-data conditions. The experimental pipeline combined heterogeneous PDF text extraction, correspondence-specific preprocessing, manual ground-truth labeling, dynamic N-shot sampling, and semantic classification based on the multilingual-e5-base model. Among the evaluated configurations, the 6-shot setting was selected as the best-

performing model, using six training documents for each class and achieving a reported accuracy of 85.87%. The final evaluation also produced a weighted-average F1-score of 0.85, indicating that the classifier maintained relatively strong predictive performance despite the restricted training set and imbalanced class distribution. The preprocessing stage was important for reducing the influence of letterheads, typographical distortions, irregular formatting, and other document-specific noise before semantic representation learning. The training-loss and error-rate patterns further indicated stable adaptation within the experimental configuration reported in the study. These findings support the use of efficient few-shot learning as a practical alternative when institutional document collections are too limited to support conventional data-intensive supervised training.

The study also shows that an intelligent document-classification workflow can be connected to a lightweight Streamlit prototype without requiring modifications to the trained SetFit architecture. Its methodological contribution lies in integrating document extraction, controlled text cleaning, progressive 1-shot to 6-shot experimentation, contrastive representation learning, and multi-metric evaluation within a single empirical workflow. The results should nevertheless be interpreted within the scope of the available institutional dataset because the final project contains inconsistencies between the correspondence categories defined in the methodological design and the labels displayed in the final evaluation output. The prototype also relies on a database-lookup mechanism during inference rather than performing real-time OCR on arbitrary newly uploaded scanned documents. These limitations do not negate the reported classification performance but restrict the extent to which the current system can be generalized to production-scale university document management. Future development should prioritize consistent label definitions, larger representation of minority classes, and direct integration of real-time OCR into the inference pipeline. Within the validated experimental scope, SetFit provides a computationally efficient basis for developing intelligent classification systems for Indonesian academic correspondence when labeled data are scarce.

REFERENCES

- Amatriain, X., Sankar, A., Bing, J., Bodigutla, P. K., Hazen, T. J., & Kazi, M. (2023). Transformer models: an introduction and catalog. *arXiv preprint arXiv:2302.07730*.
<https://arxiv.org/abs/2302.07730>
- Aminah, S., & Aminah, S. (2024). Perbandingan Keakuratan Sistem Rekomendasi Produk Berbasis Content-Based Filtering Dan Collaborative Filtering Pada E-Commerce Shopee Menggunakan Matrik Precision, Recall Dan F1-Score. *Online Repository of Universitas NU Kalimantan Selatan*, 1-18. <https://repository.unukase.ac.id/id/eprint/268/>
- Anwar, M. T., Permana, A. K., Ambarwati, L., & Agustin, D. (2021). Analyzing Public Opinion Based on Emotion Labeling Using Transformers. *2021 2nd International Conference on Innovative and Creative Information Technology, ICITech 2021*, 74-78.
<https://doi.org/10.1109/ICITech50181.2021.9590110>
- Cahyawijaya, S., Winata, G. I., Wilie, B., Vincentio, K., Li, X., Kuncoro, A., Ruder, S., Lim, Z. Y., Bahar, S., Khodra, M. L., Purwarianti, A., & Fung, P. (2021). IndoNLG: Benchmark and Resources for Evaluating Indonesian Natural Language Generation. *EMNLP 2021 - 2021 Conference on Empirical Methods in Natural Language Processing, Proceedings*, 8875-8898.
<https://doi.org/10.18653/v1/2021.emnlp-main.699>
- Juarto, B. ., & Yulianto. (2023). Indonesian News Classification Using IndoBert . *International Journal of Intelligent Systems and Applications in Engineering*, 11(2), 454–460. Retrieved from <https://ijisae.org/index.php/IJISAE/article/view/2654>
- Khaled, A., & Manayer, A. (2024). Few-shot Learning Approach for Arabic Scholarly Paper Classification using SetFit Framework. *Wseas Transactions on Communications*, 23, 89-95.
<https://doi.org/10.37394/23204.2024.23.12>
- Kilic, B., Bex, F., & Gatt, A. (2023). Contrast is all you need. *arXiv preprint arXiv:2307.02882*.
<https://doi.org/10.48550/arXiv.2307.02882>
- Koto, F., Rahimi, A., Lau, J. H., & Baldwin, T. (2020). IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP. *COLING 2020 - 28th International Conference on Computational Linguistics, Proceedings of the Conference*, 757-770.
<https://doi.org/10.18653/v1/2020.coling-main.66>

- Lin, T., Wang, Y., Liu, X., & Qiu, X. (2022). A survey of transformers. *AI Open*, 3(1), 111-132. <https://doi.org/10.1016/j.aiopen.2022.10.001>
- Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., & Gao, J. (2021). Deep learning--based text classification: a comprehensive review. *ACM computing surveys (CSUR)*, 54(3), 1-40. <https://doi.org/10.1145/3439726>
- Molina, M., & Sanchez, J. (2022). Performance Variability in Akademik Classification. *NeurIPS*, 7-9. <https://doi.org/10.52591/ixai202012123>
- Mursyidah, M., Davi, M., & Novitri, S. D. (2024). Klasifikasi Sentimen Review Pengguna terhadap Aplikasi Instagram menggunakan Algoritma Random Forest. *Jurnal Infomedia: Teknik Informatika, Multimedia, dan Jaringan*, 9(2), 106-115. <http://dx.doi.org/10.30811/jim.v9i2.6069>
- Nabiilah, G. Z., Alam, I. N., Purwanto, E. S., & Hidayat, M. F. (2024). Indonesian multilabel classification using IndoBERT embedding and MBERT classification. *International Journal of Electrical and Computer Engineering*, 14(1), 1071-1078. <https://doi.org/10.11591/ijece.v14i1.pp1071-1078>
- Nugroho, Y. A., Widodo, A., Pebrina, E. T., Iskandar, J., & Nadeak, M. (2025). Digitalization in Higher Education: How Information Systems Improve Operational and Strategic Performance. *Indonesian Journal of Management and Economic Research (IJOMER)*, 2(01), 90-98. <https://doi.org/10.70508/dyrdm592>
- Parnami, A., & Lee, M. (2022). Learning from few examples: A summary of approaches to few-shot learning. *arXiv preprint arXiv:2203.04291*. <https://doi.org/10.48550/arXiv.2203.04291>
- Song, Y., Upadhyay, S., Peng, H., Mayhew, S., & Roth, D. (2019). Toward any-language zero-shot topic classification of textual documents. *Artificial Intelligence*, 274, 133-150. <https://doi.org/10.1016/j.artint.2019.02.002>
- Syazali, M. R., & Yulianti, E. (2025). Classification of Economic Activities in Indonesia Using IndoBERT Language Model. *Jurnal Ilmu Komputer dan Informasi*, 18(2), 155-165. <https://doi.org/10.21609/jiki.v18i2.1446>
- Tunstall, L., Reimers, N., Jo, U. E. S., Bates, L., Korat, D., Wasserblat, M., & Pereg, O. (2022). Efficient few-shot learning without prompts. *arXiv preprint arXiv:2209.11055*. <http://arxiv.org/abs/2209.11055>
- Vergni, L. (2021). Evaluasi kesamaan antara indeks kekeringan melalui analisis korelasi dan uji Kappa Cohen di wilayah Mediterania. *Nat Hazards* 108, 2187–2209 (2021). <https://doi.org/10.1007/s11069-021-04775-w>
- Wang, Y., Yao, Q., Kwok, J. T., & Ni, L. M. (2021). Generalizing from a Few Examples: A Survey on Few-shot Learning. *ACM Computing Surveys*, 53(3), 1-34. <https://doi.org/10.1145/3386252>
- Zaplana, I., Hadfield, H., & Lasenby, J. (2022). Closed-form solutions for the inverse kinematics of serial robots using conformal geometric algebra. *Mechanism and Machine Theory*, 173. <https://doi.org/10.1016/j.mechmachtheory.2022.104835>
- Zhou, M., Duan, N., Liu, S., & Shum, H. Y. (2020). Progress in Neural NLP: Modeling, Learning, and Reasoning. *Engineering*, 6(3), 275-290. <https://doi.org/10.1016/j.eng.2019.12.014>