University "St. Kliment Ohridski"
Bitola
Faculty of Information and
Communication Technology - Bitola
Republic of North Macedonia

PROCEEDINGS

15th International Conference on APPLIED INTERNET AND INFORMATION TECHNOLOGIES

AIIT 2025

AIIT International Conference

Bitola, November 7, 2025

University “St. Kliment Ohridski” Bitola
Faculty of Information and Communication Technology - Bitola
Republic of North Macedonia

PROCEEDINGS
15th International Conference on
APPLIED INTERNET AND INFORMATION TECHNOLOGIES

AIIT 2025

November 7, 2025 Bitola

Proceedings publisher and organizer of the conference:
University "St. Kliment Ohridski", Bitola, Faculty of Information and Communication Technology – Bitola, Republic of North Macedonia

For publisher:
Blagoj Ristevski, PhD, Full Professor, Dean of the Faculty of Information and Communication Technologies, University "St. Kliment Ohridski", Bitola, North Macedonia

Proceedings editors:
Kostandina Veljanovska, PhD
Željko Stojanov, PhD

Conference Chairmans:
Blagoj Ristevski, University "St. Kliment Ohridski", Faculty of Information and Communication Technologies, Bitola, Republic of N. Macedonia – chair
Kostandina Veljanovska, University "St. Kliment Ohridski", Faculty of Information and Communication Technologies, Bitola, Republic of N. Macedonia – co-chair
Željko Stojanov, University of Novi Sad, Technical faculty "Mihajlo Pupin", Zrenjanin, Serbia – co-chair

Technical preparation of the proceedings:

Kostandina Veljanovska, PhD
Marija Apostoloska Kondoska, MSc
Darko Pajkovski, MSc

Cover design:
Kostandina Veljanovska, PhD
Hristina Dimova Popovska, MSc

e-Proceedings
ISBN 978-608-5003-06-8

CIP - Каталогизација во публикација

CIP - Каталогизација во публикација
Национална и универзитетска библиотека "Св. Климент Охридски", Скопје

004-049.8(062)

INTERNATIONAL conference on applied internet and information technologies AIIT 2025 (15 ; 2025 ; Bitola, Republic of North Macedonia)
Proceedings / 15th International conference on applied internet and information technologies AIIT 2025, November 7 2025, Bitola, Republic of North Macedonia ; [editors Kostandina Veljanovska, Zeljko Stojanov]. - Bitola : University "St. Kliment Ohridski", Bitola Faculty of information and communication technologies, 2025. - 477 стр. : илустр. ; 30 см
Библиографија кон трудовите
ISBN 978-608-5003-06-8

а) Информатичка технологија -- Примена -- Собири
COBISS.MK-ID 67608325

Introduction

As organizing partners of 15th International Conference on Applied Internet and Information Technologies AIIT 2025, we warmly welcome all participants, researchers, and colleagues joining us from various countries and universities, united by our shared commitment to advancing knowledge in the fields of computer science, applied Internet, and information technologies.

The AIIT conference has become a long-standing tradition of excellence and collaboration, co-organized by the Faculty of Information and Communication Technologies – Bitola, University "St. Kliment Ohridski," and the Technical Faculty "Mihajlo Pupin" – Zrenjanin, University of Novi Sad, Serbia. Over the past fifteen years, this partnership has fostered not only strong academic cooperation but also genuine friendship among our institutions and scholars.

This year's conference proudly continues that tradition, bringing together innovative research, diverse perspectives, and new insights into technologies that are shaping our digital future. The Scientific Program Committee once again faced the demanding task of selecting the highest-quality papers from more than sixty submissions spanning a wide range of topics—including Artificial Intelligence, Immersive Technologies, Mathematical Simulations, Data Science and Big Data Analytics, Knowledge and IT Management, Cybersecurity, Software Engineering, Data Mining, Digital Transformation, Behavioral Economics and Business, Social Engineering, Digital Humanities, Augmented Humanity, and Hybrid Intelligence. This ensures that the program reflects both scientific rigor and creative originality.

We would like to express our sincere gratitude to all reviewers for their dedicated work, as well as to the members of the Organizing Committee for their professionalism, commitment, and enthusiasm in preparing this event.

We are confident that these proceedings will provide an enriching and thought-provoking reading experience.

Conference chairs:

Blagoj Ristevski, University "St. Kliment Ohridski", Faculty of Information and Communication Technologies, Bitola, Republic of N. Macedonia (chair)

Kostandina Veljanovska, University "St. Kliment Ohridski", Faculty of Information and Communication Technologies, Bitola, Republic of N. Macedonia (co - chair)

Željko Stojanov, University of Novi Sad, Technical faculty "Mihajlo Pupin", Zrenjanin, Serbia (co – chair)

MAIN ORGANIZERS:

Faculty of Information and Communication Technologies - Bitola
University "St. Kliment Ohridski" University - Bitola
NORTH MACEDONIA
http://fikt.uklo.edu.mk/

Technical Faculty "Mihajlo Pupin" Zrenjanin
University of Novi Sad SERBIA
http://www.tfzr.uns.ac.rs/

ORGANIZATION PARTNERS:

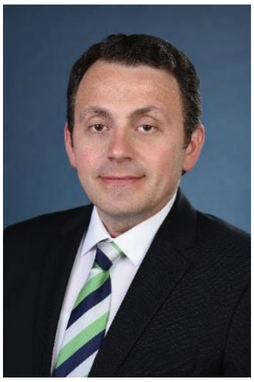
Faculty of Computer Science
Irkutsk National Research Technical University
Institute of Informational Technologies and Data Analysis
Irkutsk, RUSSIA
http://www.istu.edu/

Matrosov Institute for System Dynamics and Control Theory of Siberian Branch of Russian Academy of Sciences, Irkutsk, RUSSIA
http://idstu.irk.ru/

Irkutsk State Transport University (IrGUPS)
Irkutsk, RUSSIA
https://www.irgups.ru/

Faculty of Engineering South-west
University "Neophyte Rilsky"-Blagoevgrad
BULGARIA
http://www.swu.bg/

Conference Chairs



Blagoj Ristevski, University "St. Kliment Ohridski", Faculty of Information and Communication Technologies, Bitola, Republic of N. Macedonia (chair)

Prof. Dr. Blagoj Ristevski is a Full Professor at the Faculty of Information and Communication Technologies (FICT) at the University "St. Kliment Ohridski" - Bitola, where he currently serves as Dean. He holds a PhD in Technical Sciences from the Faculty of Electrical Engineering and Information Technologies, Institute of Computer Science and Informatics, at Ss. Cyril and Methodius University in Skopje. His research interests span Databases, Data Science, Data Mining, Big Data Analytics, Bioinformatics, Computer Graphics, and Cybersecurity. Prof. Ristevski has supervised numerous BSc, MSc, and PhD theses and has led several international research projects. He has served on the management committees of multiple COST actions, reviewed for numerous high-impact journals, and evaluated project proposals for the Horizon 2020 and Horizon Europe programs. Prof. Ristevski is also a senior member of IEEE.

Kostandina Veljanovska, University "St. Kliment Ohridski", Faculty of Information and Communication Technologies, Bitola, Republic of N. Macedonia (co – chair)

Kostandina Veljanovska, Ph.D. completed her education at the University "Sts. Kiril i Metodi", Skopje (BSc in Computer Science), at the University of Toronto, Toronto (MASc in Applied Engineering) and got her MSc and also her PhD in Technical Sciences at the University "St. Kliment Ohridski", Bitola, R. Macedonia. She has completed postdoc in Artificial Intelligence at the Laboratory of Informatics, Robotics and Microelectronics at the University of Montpellier, Montpellier, France. She worked as a Research assistant at the Faculty of Applied Science, University of Toronto, Canada. She also, worked at research team for Constraints, Learning and Agents at LIRMM, University of Montpellier. Currently, she works as a Full Professor in Artificial Intelligence and Systems, Computer Science and Computer Engineering at the Faculty of Information and Communication Technologies, University "St. Kliment Ohridski" - Bitola and serves as a Vice-dean for Science and Collaboration. Her research work is focused on artificial intelligence, machine learning techniques, intelligent systems and human - computer interaction. She participated in several international and domestic scientific projects. She has published numerous scientific papers in the area of interest, as well as several monographic items. She is a reviewing referee for well-known publishing house, journals with significant impact factor in science and also, member of editorial board of several international conferences.

Željko Stojanov, University of Novi Sad, Technical faculty "Mihajlo Pupin", Zrenjanin, Serbia (co – chair)

Željko Stojanov, Ph.D. received PhD degree in Computer science and applied informatics at University of Novi Sad, Serbia. He works as a full professor at University of Novi Sad, Technical Faculty "Mihajlo Pupin" Zrenjanin, Serbia. His research interests are in the fields of software engineering, software architecture, software life cycle, business informatics, learning and knowledge management, engineering education, and human aspects of software engineering. He is author of scientific papers published in refereed journals and in the proceedings of international conferences. He participated in several research and industrial projects at national and international levels. He has over fifteen years of experience working with small software companies as a consultant in the fields of software development, software maintenance and software process improvement.

Organizing Committee

Chairs

Kostandina Veljanovska (President), Faculty of Information and Communication Technologies - Bitola, N. Macedonia
Željko Stojanov, Technical faculty "Mihajlo Pupin", Zrenjanin, Serbia (vice-president)

Members

Blagoj Ristevski, Faculty of Information and Communication Technologies - Bitola, N. Macedonia
Ivana Berković, Technical faculty "Mihajlo Pupin", Zrenjanin, Serbia
Višnja Ognjenović, Technical Faculty "Mihajlo Pupin" Zrenjanin, Serbia
Eleonora Brtka, Technical Faculty "Mihajlo Pupin" Zrenjanin, Serbia
Dalibor Dobrilovic, Technical Faculty "Mihajlo Pupin" Zrenjanin, Serbia
Dragica Radosav, Technical faculty "Mihajlo Pupin", Zrenjanin, Serbia
Biljana Radulovic, Technical Faculty "Mihajlo Pupin" Zrenjanin, Serbia
Božidar Milenkovski, Faculty of Information and Communication Technologies - Bitola, N. Macedonia
Ljubica Kazi, Technical faculty "Mihajlo Pupin", Zrenjanin, Serbia
Snežana Savoska, Faculty of Information and Communication Technologies - Bitola,N. Macedonia
Vladimir Brtka, Technical faculty "Mihajlo Pupin", Zrenjanin, Serbia
Zoltan Kazi, Technical faculty "Mihajlo Pupin", Zrenjanin, Serbia
Siniša Mihajlović, Technical faculty "Mihajlo Pupin", Zrenjanin, Serbia
Velibor Premčevski, Technical faculty "Mihajlo Pupin", Zrenjanin, Serbia
Nikola Rendevski , Faculty of Information and Communication Technologies - Bitola, N. Macedonia
Aleksandra Stojkov, Technical faculty "Mihajlo Pupin", Zrenjanin, Serbia
Maja Gaborov, Technical faculty "Mihajlo Pupin", Zrenjanin, Serbia
Milica Mazalica, Technical faculty "Mihajlo Pupin", Zrenjanin, Serbia
Igor Vecštejn, Technical faculty "Mihajlo Pupin", Zrenjanin, Serbia
Marko Blažić, Technical faculty "Mihajlo Pupin", Zrenjanin, Serbia
Vuk Amižić, Technical faculty "Mihajlo Pupin", Zrenjanin, Serbia
Natasa Blazeska-Tabakovska, Faculty of Information and Communication Technologies - Bitola, N. Macedonia
Jovana Borovina, Technical faculty "Mihajlo Pupin", Zrenjanin, Serbia
Dalibor Šeljmeši, Technical faculty "Mihajlo Pupin", Zrenjanin, Serbia
Mimoza Bogdanoska Jovanovska, Faculty of Information and Communication Technologies - Bitola, N. Macedonia
Vladimir Šinik, Technical faculty "Mihajlo Pupin", Zrenjanin, Serbia
Nadežda Ljubojev, Technical faculty "Mihajlo Pupin", Zrenjanin, Serbia
Marina Blažeković Toshevski, Faculty of Information and Communication Technologies - Bitola, N. Macedonia
Hristina Dimova Popovska, Faculty of Information and Communication Technologies, University "St.Kliment Ohridski" - Bitola, N. Macedonia
Darko Pajkovski, Faculty of Information and Communication Technologies, University "St. Kliment Ohridski" - Bitola, N. Macedonia
Marija Apostoloska Kondoska, Faculty of Information and Communication Technologies, University "St. Kliment Ohridski" - Bitola, N. Macedonia
Milcho Prisagjanec, Faculty of Information and Communication Technologies, University "St. Kliment Ohridski" - Bitola, N. Macedonia
Ilche Dimovski, Faculty of Information and Communication Technologies, University "St. Kliment Ohridski" - Bitola, N. Macedonia
Zoran Pavlovski, Faculty of Information and Communication Technologies, University "St. Kliment Ohridski" - Bitola, N. Macedonia
Vladimir Karuović, Technical faculty "Mihajlo Pupin", Zrenjanin, Serbia
Evgeny Cherkashin, Institute of System Dynamic and Control Theory SB RAS, Russia
Anastasia Popova, Matrosov Institute for System Dynamics and Control Theory of the Siberian Branch of the Russian Academy of Sciences, Russia
Filip Tsvetanov, South-west University "Neophyte Rilsky", Faculty of Engineering, Blagoevgrad, Bulgaria

Program Committee

Bitola, North Macedonia
Andrijana Bocevska, Faculty of Information and Communication Technologies, University "St. Kliment Ohridski", Bitola, North Macedonia
Tome Dimovski, Faculty of Information and Communication Technologies, University "St. Kliment Ohridski", Bitola, North Macedonia
Marina Blažeković Toševski, Faculty of Information and Communication Technologies, University "St. Kliment Ohridski", Bitola, North Macedonia
Lela Ivanovska, Faculty of Information and Communication Technologies, University "St. Kliment Ohridski", Bitola, North Macedonia
Ilija Hristoski, Faculty of Economics - Prilep, North Macedonia
Elena Vlahu-Gjorgievska, Univerisity of Wollongong, Australia
Mimoza Mijoska, Internatonal Slavic University GAVRILA ROMANOVICH DERZHAVIN, Faculty of Technical Sciences and Informatics
Blagoj Nenovski, University "St. Kliment Ohridski", Bitola, North Macedonia
Nora Pireci Sejdiu, University "St. Kliment Ohridski", Bitola, North Macedonia
Saso Nikolovski, AUE University, Faculty of Informatics-Skopje, North Macedonia
Aybeyan Selim, International Vision University, Gostivar, North Macedonia
İlker Ali, International Vision University, Gostivar, North Macedonia
Fehmi Skender, International Vision University, Gostivar, North Macedonia
Ming Chen, Zhejiang University, China
Alexander Feoktistov, Matrosov Institute for System Dynamics and Control Theory of the Siberian Branch of the Russian Academy of Sciences, Irkutsk, Russia
Alexander Yurin, Matrosov Institute for System Dynamics and Control Theory of the Siberian Branch of the Russian Academy of Sciences, Irkutsk, Russia
Igor Bychkov, Matrosov Institute for System Dynamics and Control Theory of the Siberian Branch of the Russian Academy of Sciences, Irkutsk, Russia
Andrey Gachenko, Matrosov Institute for System Dynamics and Control Theory of the Siberian Branch of the Russian Academy of Sciences. Irkutsk, Russia
Andrey Mikhailov, Matrosov Institute for System Dynamics and Control Theory of the Siberian Branch of the Russian Academy of Sciences. Irkutsk, Russia
Anastasia Popova, Matrosov Institute for System Dynamics and Control Theory of the Siberian Branch of the Russian Academy of Sciences. Irkutsk, Russia
Alexey Daneev, Irkutsk State Transport University, Irkutsk, Russia
Denis Sidorov, Melentiev Energy Systems Institute of Siberian Branch of the Russian Academy of Sciences, Irkutsk, Russia
Viacheslav Paramonov, Matrosov Institute for System Dynamics and Control Theory of the Siberian Branch of the Russian Academy of Sciences, Irkutsk, Russia
Andrey Dorofeev, Institute of High Technologies, Irkutsk National Research Technical University, Irkutsk, Russia
Gogolák László, Subotica Tech - College of Applied Sciences, Subotica, Serbia
Zlatko Čović, Subotica Tech - College of Applied Sciences, Department of Informatics, Subotica, Serbia
Zora Konjović, University Singidunum, Centar Novi Sad, Serbia
Siniša Nešković, Faculty of organizational sciences, University of Belgrade, Serbia
Nataša Gospić, Faculty of transport and traffic engineering, Belgrade, Serbia
Branko Markoski, Faculty of technical Sciences, Novi Sad, Serbia
Željen Trpovski, Faculty of technical Sciences, Novi Sad, Serbia
Branimir Đorđević, Megatrend University, Belgrade, Serbia
Slobodan Jovanović, Faculty of Information Technology, Belgrade, Serbia
Željko Eremić, College of Technical Sciences - Zrenjanin, Serbia
Rajnai Zoltán, Obuda University, Budapest, Hungary
Tünde Anna Kovács, PhD, Óbuda University, Hungary
Zoltán Nyikes, PhD, Milton Friedman University, Hungary
Mirjana Pejic Bach, University of Zagreb, Croatia
Androklis Mavridis, Aristotel University of Thessaloniki, Greece

Madhusudan Bhatt, R.D. National College, University of Mumbai, India
Amar Kansara, Parth Systems LTD, Navsari, Gujarat, India
Narendra Chotaliya, H. & H.B. Kotak Institute of Science, Rajkot, Gujarat, India
Zeljko Jungic, ETF, University of Banja Luka, Bosnia and Herzegovina
Saso Tamazic, Univerisity of Ljubljana, Slovenia
Marijana Brtka, Centro de Matemática, Computação e Cognição, Universidade Federal do ABC, São Paulo, Brazil
Zoran Cosic, Statheros, Split, Croatia
Istvan Matijevics, Institute of Informatics, University of Szeged, Hungary
Slobodan Lubura, Faculty of electrical engineering, University of East Sarajevo, Bosnia and Herzegovina
Edit Boral, ASA College, New York, NY, USA
Dana Petcu, West University of Timisoara, Romania
Marius Marcu, "Politehnica" University of Timisoara, Romania
Aleksej Stevanov, South-west University "Neophyte Rilsky", Faculty of Engineering, Blagoevgrad, Bulgaria
Petar Apostolov, South-west University "Neophyte Rilsky", Faculty of Engineering, Blagoevgrad, Bulgaria
Filip Tsvetanov, South-west University "Neophyte Rilsky", Faculty of Engineering, Blagoevgrad, Bulgaria
Francesco Flammini, School of Innovation, Design and Engineering, Division of Product Realisation, Mälardalen University, Eskilstuna, Sweden
Deepak Chahal, Jagan Institute of Management Studies (JIMS, Rohini Sector-5), New Delhi, India
Abdel-Badeeh M. Salem, Faculty of Computer and Information Sciences, Ain Shams University, Cairo, Egypt
Dragan Peraković, University of Zagreb, Faculty of Transport and Traffic Sciences, Zagreb, Croatia
Gordana Jotanović, University of East Sarajevo, Faculty of Transport and Traffic Engineering, Doboj, Bosnia and Herzegovina
Goran Jauševac, University of East Sarajevo, Faculty of Transport and Traffic Engineering, Doboj, Bosnia and Herzegovina
Dinu Dragan, Faculty of technical Sciences, University of Novi Sad, Serbia
Gururaj Harinahalli Lokesh, Department of IT, Manipal Institute of Technology, Bengaluru, India
Ertuğrul AKBAŞ, Esenyurt University, Istanbul, Turkiye

CONTENT

Comparative Analysis of Text Mining Techniques and Tools

Marija Apostoloska-Kondoska[1], Blagoj Ristevski[2], Nikola Rendevski[3], Snezana Savoska[4]

[1, 2, 3, 4] *University "St. Kliment Ohridski" - Bitola, Faculty of Information and Communication Technologies – Bitola, Partizanska bb, Macedonia.*

[1]*marija.kondoska@uklo.edu.mk;* [2]*blagoj.ristevski@uklo.edu.mk;* [3]*nikola.rendevski@uklo.edu.mk;* [4]*snezhana.savoska@uklo.edu.mk*

Abstract:

Rapid progress in digitalisation has generated a vast volume of data, particularly unstructured and semi-structured data. Text mining is a process of extracting useful information from text documents and is a new and challenging research field that aims to address the information overload problem by leveraging techniques from machine learning (ML), natural language processing (NLP), and knowledge management. Text mining tools facilitate the analysis of unstructured textual data by extracting patterns and generating insights through techniques such as classification. In this paper, we have discussed the text mining techniques and compared two platforms: RapidMiner and KNIME.

Keywords:

Text mining, text mining tools, unstructured data, machine learning.

1. Introduction

Data mining is used for finding useful information from a large amount of data. Data mining techniques are employed to address various types of research issues. The research-related subfields in data mining are text mining, web mining, image mining, sequential pattern mining, spatial mining, medical mining, multimedia mining, structure mining, and graph mining [1].

On the other side, text mining is the process of mining useful information from textual documents. It is also called knowledge discovery in text (KDT) or knowledge of intelligent text analysis. Text mining is a technique that extracts information from both structured and unstructured data and also finds and recognizes patterns [1].

Text mining is similar to the data mining process, but they differ in the selected tool. Data mining tools are designed to handle structured data, whereas text mining can deal with unstructured or semi-structured data sets such as emails, HTML files, medical reports, full-text documents, etc. [2]. Text mining is used to find new, previously unidentified information from various written resources [1].

Structured data are data where each element is addressable and is the most organized form of data for easy storage, access, and analysis. This data type is structured into rows and columns and organized within databases and tables. Unstructured data, such as free-form text, images, and social media content, dominates information in the real world [3].

Unlike structured data, unstructured data does not follow a specific format or schema, which makes it more challenging to store and analyze. To overcome these challenges, modern storage solutions have been developed that are capable of handling large volumes and different types of data.

A data lake is a central repository that is highly scalable and can store structured, semi-structured, or unstructured raw and granular data from multiple sources [4].

Within such repositories, semi-structured data plays a particularly important role, as it strikes a balance between structured data's rigidity and unstructured data's flexibility, making it a valuable format for modern businesses looking to work with complex data sources.

The text mining process includes several stages: text pre-processing, feature extraction, text mining techniques, model training and evaluation, interpretation and visualization of results, iteration and refinement, reporting and deployment.

The rest of the paper is structured as follows. Section 2 presents the text mining process, while the subsequent section describes the fundamental techniques of text mining. Sections 4 and 5 present the obtained results of comparative analyses between RapidMiner and KNIME tools. Concluding remarks and directions for further work are given in the last section.

2. Text Mining

Text mining is a systematic sequence of steps used to extract useful information from unstructured text data, as shown in Figure 1. It consists of the following stages:

A. Text preprocessing

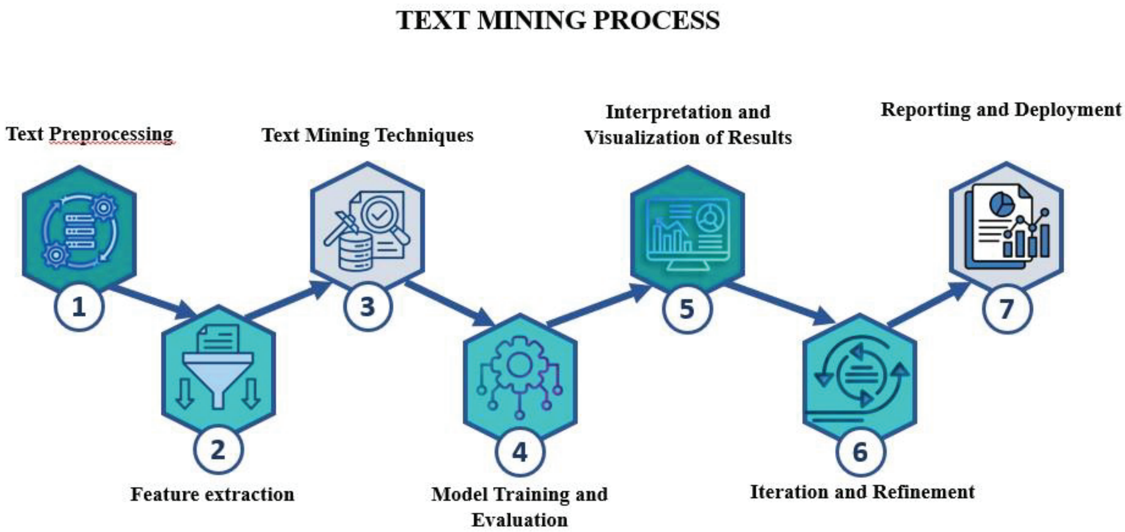


Figure 1. Text Mining Process

Pre-process the raw text data to clean and transform it into a suitable format for analysis. This step includes removing punctuation, converting text to lowercase, tokenization (splitting text into individual words or tokens), removing stop words, and stemming or lemmatization (reducing words to their root form).

B. Feature extraction

Extract attributes from the text data to represent it in a numerical or structured format that can be used for analyses. This step involves techniques such as bag-of-words representation, term frequency-inverse document frequency (TF-IDF), or word embeddings.

C. Text Mining Techniques

Used in various types of research domains like information retrieval, information extraction, natural language processing, categorization, text classification, and text clustering.

D. Model Training and Evaluation

Divide the text into training and evaluation sets using machine learning or statistical models. This step uses techniques like cross-validation or hyperparameter tuning.

E. Interpretation and Visualization of Results

Interpret and analyze results obtained from applied text mining techniques. Visualize results using graphs, charts, or diagrams.

F. Iteration and Refinement:

Analyze the result and iterate on the text mining process as necessary.

G. Reporting and Deployment

Prepare a comprehensive report summarizing the results, insights, and recommendations derived from the text mining.

3. Techniques in Text Mining

In text mining, various techniques are employed to handle unstructured textual data, enabling the discovery of patterns and the extraction of knowledge and actionable insights.

A. Information Retrieval

Information Retrieval (IR) is the process of representing, storing, and querying data collections to enable knowledge discovery based on user requests [5]. This process involves various stages, starting with representing data and ending with returning relevant information to the user. The intermediate stage includes filtering, searching, matching, and ranking operations [5].

B. Information extraction

Information extraction serves as the starting point for analysing unstructured text, where the focus is on recognizing relevant phrases and their interrelations. To accomplish this, pattern-matching methods are applied to locate predefined structures within the text. The information extraction task includes tokenization, identification of named entities, sentence segmentation, and part-of-speech assignment. Firstly, input phrases and sentences are parsed and semantically interpreted, and then the required pieces of information are added to the database. The general information extraction process is shown in Figure 2. The most accurate information extraction systems involve handcrafted language processing modules, and substantial progress has been made in applying data mining techniques to a number of these steps. This technology can be very useful when dealing with large volumes of text [6].

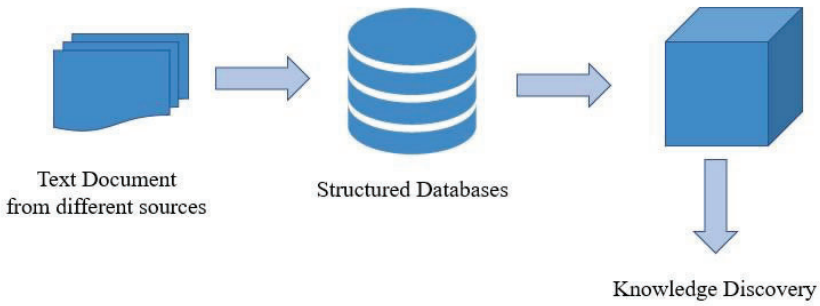


Figure 2. Information extraction [6].

C. Categorization

Categorization represents the mechanism through which entities are identified and differentiated, forming the basis for a structured understanding of ideas and objects. Categorization implies that objects are grouped into categories, usually for some specific purpose. Text categorization becomes a key technology to deal with and organize large numbers of documents. Text representation constitutes a key step in text categorization, ensuring that documents are expressed in a way that supports meaningful classification. A major problem of text representation is the high dimensionality of the feature space. The feature space with a large number of terms is not only unsuitable for neural networks but also easily causes the overfitting problem. Text categorization is the assignment of natural language documents to one or more predefined categories based on their semantic content, and is an important component in many information organization and management tasks [7].

D. Natural Language Processing

Natural language processing (NLP) refers to the automatic processing and analysis of unstructured textual information. It performs different types of analysis, such as Named Entity Recognition (NER) for abbreviation and their synonyms extraction to find the relationships among them [8]. NER identifies all the instances of specified objects from a group of documents. These entities and their instances allow the identification of relationships and other information to attain their key concept. However, this technique lacks a complete dictionary list for all named entities used for identification [8]. Natural Languages (NL) have a lot of complexities, as a text extracted from different sources does not have identical words or abbreviations. There is a need to detect such issues and make rules for their uniform identification [8].

E. Clustering

Clustering is an unsupervised approach used to group the text documents by applying different clustering algorithms. In a cluster, similar terms or patterns are grouped and extracted from various documents. Clustering is performed in a top-down and bottom-up manner. In NLP, various types of mining tools and techniques are applied for the analysis of unstructured text. Different techniques of clustering are hierarchical, distribution, density, centroid, fuzzy c-means, and k-means clustering [9].

F. Summarization

Text summarization techniques aim to generate concise summaries of long texts. It can be done in extractive or abstractive manners. Extractive summarization involves selecting important sentences or phrases from the original text, while abstractive summarization involves generating new sentences to capture the key information.

4. Text Mining Tools (RapidMiner vs KNIME)

Text mining tools can be classified into three categories, as shown in Figure 3.

1) Proprietary Text Mining Tools

These tools are commercial text mining tools owned by a company. To use these tools, a purchase is required. Although demo/trial versions are available free of charge, they have limited functionality.

2) Open-Source Text Mining Tools

These tools are available free of charge, and their source code is available; one can even contribute to their development.

3) Online Text Mining Tools

These tools can be run from the website itself. Only a web browser is required. These tools are generally simple and provide limited functionality [10].

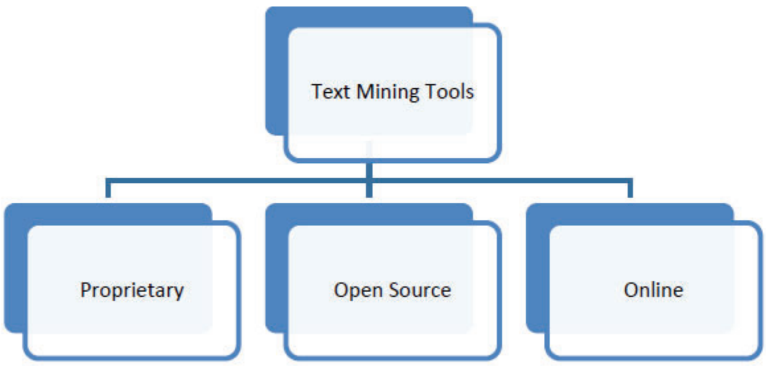


Figure 3. Types of Text Mining Tools.

Although there are many text mining tools, such as Weka, Orange, KNIME, RapidMiner, and Tanagra, we have chosen two of them that often come up in discussions: RapidMiner and KNIME. Both are powerful software programs that can help businesses and researchers in text mining and analysing data more efficiently. However, choosing between two can be a daunting task, as each has its strengths and weaknesses, as tabulated in Table 1.

RapidMiner allows users to create data workflows using a drag-and-drop interface. Like KNIME, RapidMiner offers a wide range of pre-built nodes, as well as the ability to create custom nodes using various programming languages. One of the biggest advantages of RapidMiner is its speed. RapidMiner has been designed to efficiently manage large datasets, offering faster data processing performance relative to KNIME. Additionally, RapidMiner offers a wide range of machine-learning algorithms that can be used to build predictive models, such as Naïve Bayes, Decision Tree, Random Forest, Logistic Regression, Support Vector Machine (SVM), K-Nearest Neighbours (KNN), etc.

KNIME (Konstanz Information Miner) is an open-source data analytics platform that allows users to create data workflows, which are essentially a series of interconnected nodes that represent different data processing steps. KNIME offers a wide range of pre-built nodes, as well as the ability to create custom nodes using various programming languages such as Python, R, and Java. One of the biggest advantages of KNIME is its user-friendly interface, which makes it easy for users of all skill levels to create complex data workflows without any programming knowledge. Furthermore, KNIME's community is very active, with a large number of pre-built workflows available for download, which can save users a lot of time.

Table 1: Comparison of features between RapidMiner and KNIME

Feature	RapidMiner	KNIME
Main Feature	Visual workflow	Data integration & manipulation
Strength	User-friendly, extensive pre-built components	Data integration, scripting
Weakness	Can be resource-intensive for complex workflows	Programming skills required for advanced tasks
Cost	Free and Commercial Versions, the free version supports only 10.000 rows	Free
Speed	It demonstrates faster performance when processing large datasets.	It demonstrates slower performance when processing large datasets.
Programming Language	Java	Java
Supported Data Formats	CSV, XML, JSON, others	CSV, TSV, TXT, XLSX, JSON, XML, others
Machine Learning Algorithms	Extensive library	Various, including deep learning
Data Visualization	Basic visualization components	Extensive visualization capabilities
Real-time Data Analysis	Limited support	Some support through extensions
OSs	Windows, macOS, Linux, open source	Windows, macOS, Linux, open source
Quality	Powerful and versatile, with an advantage especially in predictive analytics	The main open data mining tool that predictive analytics has made available to the general public
Techniques supported	Data mining and text analytics, text processing	Text analysis, data I/O, pre-processing and cleaning, modeling, analysis and data mining

5. Experiment

5.1 Research methodology

Besides a comparison of the features between RapidMiner and KNIME, we have conducted a comparative study of the classification algorithm Naïve Bayes on the spam emails dataset, downloaded from Kaggle [11]. The dataset includes two categories: spam and ham. The term spam refers to unsolicited or unwanted messages, typically sent in bulk for advertising or malicious purposes, whereas ham denotes legitimate, non-spam messages. The number of instances is 5572, with two attributes:

- Categories: spam, or ham (not spam).
- Message: email message.

To ensure proper evaluation of the classification models, the dataset was divided into training and test sets using an 80/20 stratified sampling approach. This technique preserved the original class distribution in both subsets, thus avoiding potential bias in the learning process. Out of the total instances, 4825 (86.6%) were labeled as ham and 747 (13.4%) as spam. After the split, the training set contained 4457 messages (3860 ham and 597 spam), while the test set included 1115 messages (965 ham and 150 spam). This distribution allowed the Naïve Bayes classifier to be trained on a representative portion of the data and tested on an independent subset that retained the original characteristics of the dataset.

In RapidMiner for classification, we used the following nodes: Read CSV, Nominal To Text, Set Role, Split Data, Process Document From Data, which contain operators for pre-processing data (Tokenize, Transform Cases, Filter Stop words), Naïve Bayes, Apply Model, and Performance, as shown in Figure 4.

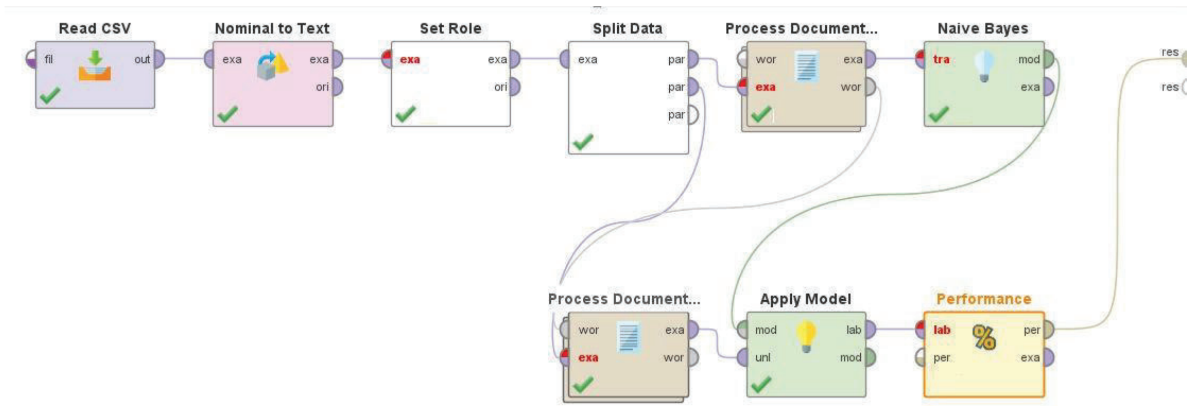


Figure 4. Classification in RapidMiner.

For the classification in KNIME, we have employed the following nodes: CSV Reader, String to Document, Nodes for Pre-processing Data (Punctuation Erasure, Number Filter, Stop Words Filter), Bags of Words Creator, Term to String, TF, Partitioning, Naïve Bayes Learner, Naïve Bayes Predictor, and Scores shown in Figure 5.

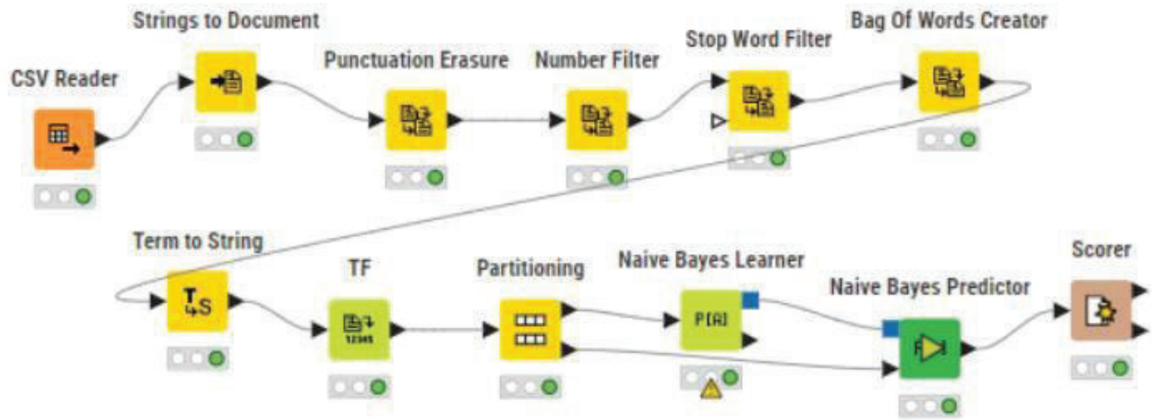
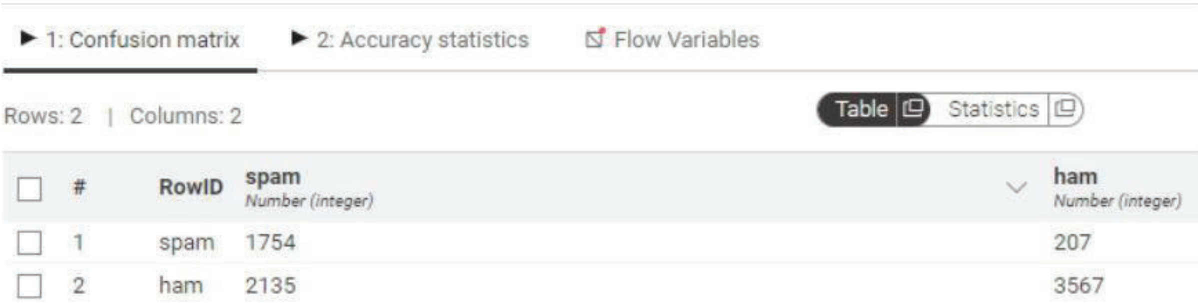


Figure 5. Classification in KNIME.

5.2 Results

In this step, we evaluate the performance of the classification algorithm using the confusion matrix, a table that reveals true versus predicted values and the performance of the classification model on the basis of accuracy, as shown in Figures 6, 7, and 8.



▶ 1: Confusion matrix ▶ 2: Accuracy statistics Flow Variables

Rows: 2 | Columns: 2 Table Statistics

#	RowID	spam Number (integer)	ham Number (integer)
1	spam	1754	207
2	ham	2135	3567

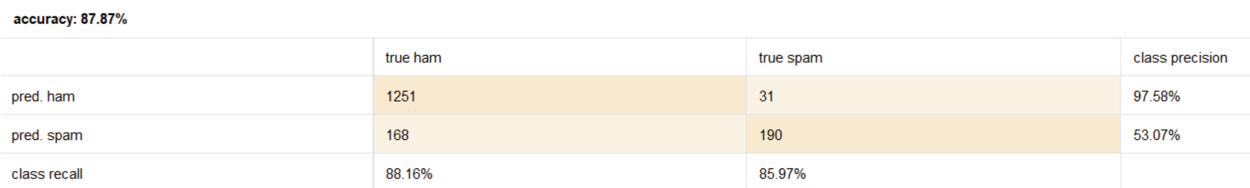
Figure 6. Confusion Matrix in KNIME.

▶ 1: Confusion matrix ▶ 2: Accuracy statistics Flow Variables

Rows: 3 | Columns: 11 Table Statistics

#	RowID	TruePositives Number (integer)	FalsePositives Number (integer)	TrueNegatives Number (integer)	FalseNegativ... Number (integer)	Recall Number (double)	Precision Number (double)	Sensitivity Number (double)	Specificity Number (double)	F-measure Number (double)	Accuracy Number (double)	Cohen's kappa Number (double)
1	spam	1754	2135	3567	207	0.894	0.451	0.894	0.626	0.6	?	?
2	ham	3567	207	1754	2135	0.626	0.945	0.626	0.894	0.753	?	?
3	Overall	?	?	?	?	?	?	?	?	?	0.694	0.393

Figure 7. Accuracy in KNIME.



accuracy: 87.87%

	true ham	true spam	class precision
pred. ham	1251	31	97.58%
pred. spam	168	190	53.07%
class recall	88.16%	85.97%	

Figure 8. Confusion Matrix and Accuracy in RapidMiner.

Table 2:
Results of classification

	Accuracy	Class Precision		Class Recall	
		predicted ham	predicted spam	ham	spam
RapidMiner	87.87 %	97.58 %	53.07 %	88.16 %	85.97 %
KNIME	69.40 %	94.50 %	45.10 %	62.60 %	89.40 %

As shown in Table 2, the accuracy value in RapidMiner is 87.87%, and in KNIME, that value is 69.4%. When we compare class precision in RapidMiner for predicted ham, which is 97.58% and for predicted spam, which is 53.07% versus class precision in KNIME for predicted ham, which is 94.5% and for predicted spam, which is 45.1%. Also, class recall in RapidMiner for ham is 88.16% and for spam is 85.97%, versus class recall in KNIME for ham is 62.6% and for spam is 89.4%.

6. Conclusion

Business decision-making relies on structured, unstructured, and semi-structured data being processed through every step of the text mining process. Text mining techniques help to improve the text mining process. Selecting the text mining technique is an important part of obtaining the most reliable results. Similarly, selecting the appropriate text mining tools is essential. Now there are a large number of tools that are available to users, which we have already classified into three categories and compared some of the features of RapidMiner and KNIME.

To conclude, RapidMiner and KNIME are powerful platforms for text mining, offering a comprehensive set of features and functionalities. When we compare them according to the obtained results, we can say that RapidMiner has better accuracy, class precision, and class recall than KNIME on the dataset that we have chosen. But choosing between the two ultimately depends on your specific needs, datasets, models, and algorithms. Regardless of which tool is selected, both RapidMiner and KNIME can help to analyze the data more efficiently and effectively.

Future work could expand the analysis by incorporating additional machine learning algorithms such as Support Vector Machines, Random Forests, Gradient Boosted Trees, or even deep learning approaches. A broader evaluation across multiple models would provide deeper insights into performance trade-offs and enhance the robustness of the findings.

References:

[1] Dr. S. Vijayarani, Ms. J. Ilamathi, Ms. Nithya, “Preprocessing Techniques for Text Mining – An Overview”, International Journal of Computer Science & Communication Networks, vol. 5(1), pp. 7-16, 2015.

[2] Vishal Gupta, Gurpreet S. Lehal, “A Survey of Text Mining Techniques and Applications”, JOURNAL OF EMERGING TECHNOLOGIES IN WEB INTELLIGENCE, VOL. 1, NO. 1, PP. 60-76, 2009.

[3] Mori, Misato. "Structured Understanding of Unstructured Data: A Cross-Domain Review of LLM-Augmented Analysis Systems." Authorea Preprints, 2025.

[4] Harby, Ahmed A., and Farhana Zulkernine. "Data lakehouse: a survey and experimental study." Information Systems 127, p: 102460, 2025.

[5] Akram Roshdi, Akram Roohparvar, “Review: Information Retrieval Techniques and Applications“, International Journal of Computer Networks and Communications Security, vol. 3, no. 9, 2015.

[6] Sonali Vijay Gaikwad, Archana Chaugule, Pramod Patil, “Text Mining Methods and Techniques”, International Journal of Computer Applications, vol.85, no. 17, pp. 42-45, 2014.

[7] B. Mahalakshmi, Dr. K. Duraiswamy, “An Overview of Categorization techniques”, International Journal of Modern Engineering Research, vol. 2, issue. 5, pp. 3133-3137, 2012.

[8] Ramzan Talib, Muhammad Kashif Hanif, Shaeela Ayesha, Fakeeha Fatima, “Text Mining: Techniques, Applications and Issues”, International Journal of Advanced Computer Science and Applications, vol. 7, no. 11, pp. 414-418, 2016.

[9] B. L. Narayana and S. P. Kumar, “A new clustering technique on text in a sentence for text mining”, International Journal of Science Engineering and Advanced Technology, vol. 3, no. 3, pp. 69–71, 2015.

[10] Arvinder Kaur, Deepti Chopra, “Comparison of Text Mining Tools”, 5th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO) (Trends and Future Directions), pp. 186-192, 2016.

[11] Spam emails dataset for Classification and Filtering [Online]. Available:
https://www.kaggle.com/datasets/abdallahwagih/spam-emails,
Accessed:18.07.2025