

СТРАТЕГІЯ, ЗМІСТ ТА НОВІ ТЕХНОЛОГІЇ ПІДГОТОВКИ ФАХІВЦІВ У СФЕРАХ КОМП'ЮТЕРНИХ НАУК, ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ, ЕЛЕКТРОНІКИ ТА АВТОМАТИЗАЦІЇ

<https://doi.org/10.31649/1997-9266-2026-187-4-211-217>

УДК 004.8

Г. Є. Брусенцов¹

БАГАТОМОВНА КЛАСИФІКАЦІЯ ЕЛЕКТРОННИХ ЛИСТІВ В УМОВАХ ДИСБАЛАНСУ КЛАСІВ ДЛЯ МАРШРУТИЗАЦІЇ В РЕКТОРАТІ УНІВЕРСИТЕТУ

¹Національний університет «Львівська політехніка»

Розглянуто проблему ручного сортування електронних листів у ректораті місцевого університету. Багато електронних листів, що надходять до секретаріату ректора, потребують ручного перерегляду та подальшого перенаправлення до інших підрозділів університету. У такому випадку листи не оброблялися безпосередньо секретаріатом, їх потрібно переслати до семи різних підрозділів-адресатів. Проаналізовано електронні листи, призначені для відділів освіти (education), якості освіти (education quality), навчальної роботи та розвитку студентського потенціалу (education outreach), міжнародної діяльності (international office), науки (science), господарського відділу (utility) та проректора (vice rector). Початкова база даних електронних листів містила 9842 листи. Після видалення дублікатів залишилися 8153 унікальних листів. Використання алгоритму, що поєднував видалення дублікатів з пошуком подібних листів, дало 4672 унікальних записів. Лише 477 з цих потенційних записів можна вважати достатньо маркованими для навчання моделей з учителем. До трьох сімейств моделей, що досліджувалися, належать: TF-IDF з використанням логістичної регресії, TF-IDF з використанням лінійної машини опорних векторів (SVM) та XLM-RoBERTa. Найкращі результати показало поєднання TF-IDF з SVM. Поєднання TF-IDF та SVM перевершило TF-IDF з логістичною регресією за показником F-measure (макро f1: 0,6421, зважений f1: 0,7920 та точність: 0,8021) за випадкового поділу та показником F-measure (макро f1: 0,5371, зважений f1: 0,6854 та точність: 0,6947) з хронологічним розбиттям. Аналіз класів продемонстрував чудові результати для найпоширеніших класів з високою частотою, таких як "science" та "education". Проте результати для класів з нижчою частотою виявилися мінливими. Повнота (recall) для "education outreach" — нульова для обох поділів через наявність лише одного тестового запису. З 102 загальних помилок, допущених у обох комбінаціях TF-IDF, усі, крім однієї, відбулися нижче рівня достовірності 0,5. Отже, загалом ця стаття визначає автоматичне пересилання листів для ректорату університету як завдання багатомовної класифікації з екстремальним дисбалансом класів. Відсутність достатньої кількості мічених даних та обмежені лексичні моделі є найбільшою перешкодою для впровадження рішення для цієї задачі.

Ключові слова: багатомовна класифікація електронної пошти; дисбаланс класів; напівкероване навчання; TF-IDF; лінійна SVM.

Вступ

Автоматизація маршрутизації електронної пошти є актуальним і практично значущим застосуванням технології класифікації тексту, що має реальні операційні застосування. Метою цієї роботи є розроблення рішення для ректорату місцевого університету. Ректорат отримує значну кількість листів та повинен переслати їх до іншого підрозділу університету перед тим, як вживати будь-яких заходів. Існує сім можливих підрозділів, до яких пересилається пошта, зокрема відділи освіти (education), якості освіти (education quality), навчальної роботи та розвитку студентського потенціалу (education outreach), міжнародної діяльності (international office), науки (science), господарського відділу (utility) та проректора (vice rector). Тому головне оперативне питання полягає в тому, чи можна швидко класифікувати електронні листи, отримані ректоратом, щоб переслати їх до відповідних підрозділів. Також потрібно враховувати багатомовність та значний дисбаланс класів.



© Г. Є. Брусенцов, 2026. Ця стаття відкритого доступу та поширюється на умовах ліцензії [Creative Commons Attribution License 4.0](https://creativecommons.org/licenses/by/4.0/)

Електронні листи — це адміністративні та академічні повідомлення, що надходять до централізованої університетської поштової скриньки і мають бути згодом перенаправлені до відповідного(их) підрозділу(ів). У статті [1] автори довели, що електронні листи мають достатню семантичну структуру, яку можна автоматично виокремлювати. Це дозволяє ухвалювати рішення куди переслати електронний лист. До того ж, у статті [2] досліджено порівняння декількох методів класифікації тем електронних листів та сформульовано проблему маршрутизації електронних листів як задачу багатокласової класифікації тексту. Також, у статті [3] продемонстрували, що запити у службу підтримки є «шумними», «незбалансованими» та «обмеженими з операційної точки зору». Через ці фактори вони є гарною аналогією для розглянутої проблеми маршрутизації листів, які будуть сортуватися за дуже обмежених доступних ресурсів.

Окрім дослідження актуальності проблеми розглянуто доцільність вибраної методології для розв'язання поставлених задач. Так, у статті [4] автори продемонстрували, що розріджені лексичні базові моделі залишаються дуже конкурентоспроможними в категоризації тексту. Це підтверджує використання TF-IDF з лінійними моделями як серйозних базових моделей замість слабких компараторів. У статті [5] представлено модель XLM-RoBERTa для багатомовної класифікації тексту. Модель показала значно вищі показники точності, ніж інші досліджувані моделі. У статті [6] автори пояснили, як проблеми, пов'язані з незбалансованим навчанням, вимагають врахування оцінки, чутливої до класів. Таким чином, у цій роботі наголошувалося на макро-F1 та відтворенні за класами. У сукупності ці статті надають обґрунтування для порівняння лексичних та багатомовних моделей. При цьому вони підкреслюють доцільність використання метрик, що враховують проблеми, пов'язані з дисбалансом класів.

Метою дослідження є розробка та оцінка методичного підходу до автоматичної маршрутизації листів у канцелярії ректора шляхом багатомовної класифікації в умовах суттєвого дисбалансу класів і дефіциту розмічених даних.

Методи

Оригінальний набір даних складався з 9842 необроблених електронних листів. Розподіл мов у всьому наборі даних включав 3965 повідомлень українською мовою, 3442 повідомлення англійською мовою, 2403 повідомлення, що містять кілька мов, та 32 повідомлення, мова яких неідентифіковано. Підготовка набору даних відбувалася у декілька кроків:

1. Визначення міток для кожного листа.
2. Видалення дублікатів.
3. Видалення схожих листів за змістом.

Для визначення міток, якому підрозділу університету має бути переадресований лист, застосовано підхід напівкерованого навчання на основі правил. Для цього для кожного класу визначено набір ключових слів та їхні ваги. Далі для кожного класу обчислено суму ваг та визначено коефіцієнти впевненості, як співвідношення цієї суми до загальної суми.

$$S_r = \sum_{k \in K_r} w_k, \quad (1)$$

де S_r — сума ваг для класу r ; K_r — набір ключових слів класу r ; k — знайдене ключове слово в листі; w_k — вага ключового слова k .

$$C_r = \begin{cases} 100 \frac{S_r}{S} & S > 0, \\ 0, & S < 0, \end{cases} \quad (2)$$

де C_r — коефіцієнт впевненості класу r ; S_r — сума ваг для класу r ; S — загальна сума ваг усіх класів.

Якщо коефіцієнт впевненості будь-якого класу перевищував 70, то такий клас вважався міткою для цього листа.

Ключові слова та їх важливість визначено та погоджено з секретаріатом університету для отримання точніших результатів. Список таких слів продемонстровано у табл. 1 для кожного класу.

На виході отримано 8236 листів, які не підпадали під жоден клас, 742 листи з коефіцієнтом впевненості меншим за 70 та 864 листи з визначеною міткою і коефіцієнтом впевненості більшим чи рівним за 70.

Для пошуку дублікатів застосовано метод хешування SHA1. Поєднуючи тему та основний текст листів, визначався хеш для кожного листа. Після цього листи з однаковими хешами видалялися, залишаючи один екземпляр. У результаті отримано 8153 листи.

Таблиця 1

Набір ключових слів для кожного класу

Клас	Ключове слово	Вага	Клас	Ключове слово	Вага
science	проскти	2,4	edu,quality	акредитація	2,5
	конференції	2,1		кваліфікації	2,2
	наука	1,5		працевлаштування	1,3
	стаття	1,1		профорієнтація	1,1
	дослідження	1,1		якість	1,1
international.office	закордон	2,4	vice-rector	оренда	1
	мобільність	2,2		архів	1
	міжнародна	2,2		пенсійний фонд	1
	відрадження	2,0		стажування	1
	консул	2,0	edu,outreach	інформатизація	2,3
	посол	2,0		мережа	2,0
	діяльність	1,0		веб сайт	2,0
edu	студенти	2,3	utility	інформаційна сис-тема	1,8
	сесія	1,9		студмістечко	2,2
	переведення студентів	1,2		гуртожиток	2,2
	відрахування	1,1		комунальні	2,2
	поновлення	1,1		автогосподарство	2,2
	студентський відділ	1,1		послуги	1,2
	академічна довідка	1,1			

Пошук схожих листів дещо складніший і складався з декількох етапів. Спочатку, аналогічно до пошуку дублікатів, поєднано тему та основний текст листа. Після цього цей корпус тексту перетворено на вектор, використовуючи TF-IDF та n-грами розміром від 3 до 5, щоб отримати числове представлення тексту листа. Далі застосовано метод найближчих сусідів для пошуку схожих текстів. Оскільки корпуси тексту перетворено на вектори, найкращою метрикою для їх порівняння є косинус подібності [7]. Ця метрика порівнює напрям двох векторів, де -1 позначає, що два вектори напрямлені в протилежні сторони, а 1 — в одну. Останнім кроком здійснювався пошук всіх листів, у яких косинус подібності перевищував 0,9. Такі листи видалялися, залишаючи один екземпляр. У результаті отримано 4672 унікальні листи.

Після видалення дублікатів та схожих листів з 864 листів з визначеною міткою і коефіцієнтом впевненості більшим чи рівним за 70 залишилося лише 477. Розподіл класів підготовленого тренувального набору даних можна побачити на рис. 1.

Для визначення оптимальної ефективності у завданні маршрутизації порівняно три сімейства моделей: TF-IDF з використанням логістичної регресії; TF-IDF з використанням лінійної SVM; та XLM-RoBERTa (xlm-roberta-base), багатомовний кодер, описаний Conneau та ін. [5].

Вибірка з випадковим відбором складалася з 381 навчального прикладу та 96 тестових прикладів. Вибірка з хронологічним відбором складалася з 382 навчальних прикладів та 95 тестових прикладів, датованих після навчання. Параметри регуляризації для моделей сімейства TF-IDF вибрані шляхом перехресної валідації на навчальних підмножинах.

Основним показником оцінки був макро-F1 через його придатність для вимірювання збалансованої продуктивності між класами за умови значного перекошу міток [6]. До вторинних показників належали зважений F1 для оцінки загальної якості моделі з урахуванням частоти класів, точність

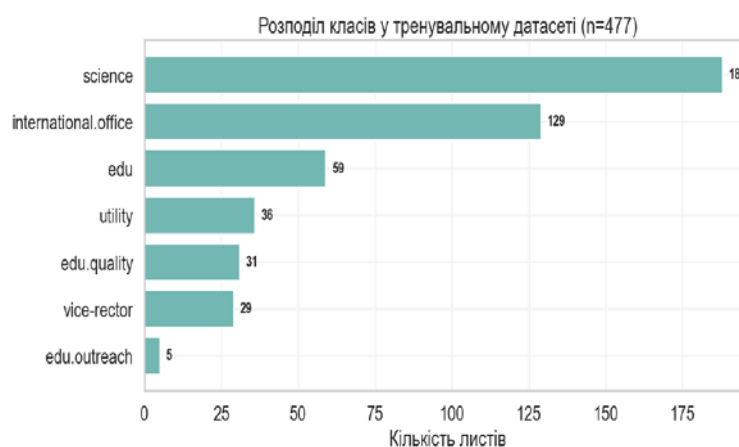


Рис. 1. Результати моделей по розбиттю набору даних та метрик

для вимірювання частки правильних позитивних передбачень, повнота (recall) за класами для оцінки здатності моделі знаходити всі об'єкти кожного класу та рівень впевненості моделі.

У цій роботі для обчислення рівня впевненості кожної моделі використовувалась нормована експоненційна функція (softmax).

Результати дослідження

У табл. 2 підсумовано основні результати оцінки.

Таблиця 2

Основні результати застосованих методів

Розбиття	Модель/метод	Кількість тренувальних рядків	Кількість тестувальних рядків	Макро-F1	Зважений F1	Точність
Випадкове	TF-IDF + логістична регресія	381	96	0,6146	0,7705	0,7813
Хронологічне	TF-IDF + логістична регресія	382	95	0,4696	0,6394	0,6526
Випадкове	TF-IDF + SVM	381	96	0,6421	0,7920	0,8021
Хронологічне	TF-IDF + SVM	382	95	0,5371	0,6854	0,6947
Випадкове	XLM-RoBERTa	381	96	0,0810	0,2245	0,3958
Хронологічне	XLM-RoBERTa	382	95	0,1066	0,2303	0,3579

Результати моделей з розбиття набору даних та їх метрики показані на рис. 2. Для обох методологій розбиття TF-IDF + SVM був членом сімейства моделей з найвищою ефективністю серед протестованих. TF-IDF + логістична регресія була конкурентоспроможною, але загалом мала нижчу ефективність, ніж TF-IDF + SVM. Тоді як протестована конфігурація XLM-RoBERTa значно поступалася за ефективністю. Отже, в контексті дослідження цього проекту найобґрунтованішим базовим варіантом є представник сімейства розріджених лексичних SVM, а не представник сімейства трансформерів.

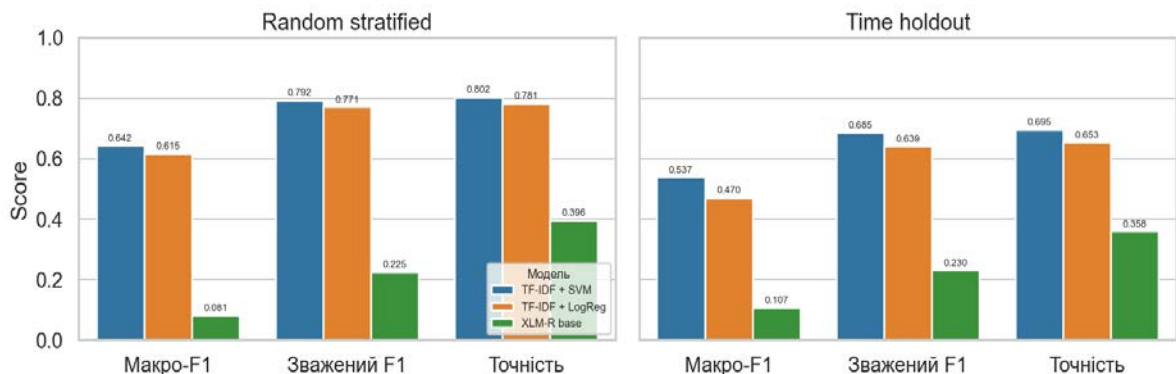


Рис. 2. Результати моделей з розбиття набору даних та їх метрики

Розбиття на основі хронології складніше, ніж випадкове розбиття даних, для обох представників сімейства TF-IDF. Різниця в макро-F1 між випадковим розподілом та хронологічним становила 0,145 для логістичної регресії та 0,105 для SVM. Ці відмінності свідчать про те, що випадково створені розподіли краще охоплюють загальний зміст листів, оскільки з часом стиль та зміст листів одної тематики може змінюватися.

XLM-RoBERTa покращився з макро-F1 0,0810 до 0,1066 між випадковим поділом та хронологічним поділом даних. Хоча ані покращення, ані погіршення в будь-якому напрямку не можна використовувати для загальніших висновків щодо стійкості представників сімейства трансформерів у цьому застосуванні.

У табл. 3 подано повноту (recall) за класами для базової моделі з найкращими результатами разом з розміром тестової вибірки для кожного розбиття.

У результатах переважали класи з вищою частотою («science», «education»). Класи з меншою частотою демонстрували нестабільну продуктивність. Клас «education outreach» показав найгірші результати в обох розбиттях через наявність лише одного тестового прикладу на розбиття, проте жоден з трьох протестованих методів не класифікував його в обох розподілах. «Education quality»

та «vice-rector» також виявилися складними, з нижчим показником повноти (recall) у хронологічних розбиттях, ніж у випадкових.

Таблиця 3

Метрики TF-IDF + SVN для кожного класу

Клас/Відділ	Розмір випадкової тестової вибірки	SVN повнота (recall) для випадкової вибірки	Розмір хронологічної тестової вибірки	SVN повнота (recall) для хронологічної вибірки
education outreach	1	0,00	1	0,00
education quality	6	0,50	8	0,375
education	12	0,75	19	0,8947
international office	26	0,7692	28	0,5714
science	38	0,9737	30	0,8334
utility	7	0,5714	3	0,6667
vice rector	6	0,6667	6	0,50

Результати, розбиті за мовами, продемонстрували, що приклади зі змішаними мовами не завжди є найскладнішими мовними групами. Для випадкових розбиттів обидва представники сімейства TF-IDF досягли макро-F1 0,8161 на прикладах зі змішаними мовами ($n = 27$), тоді як макро-F1 для українських прикладів нижчий: 0,5729 для логістичної регресії ($n = 69$), 0,6100 для SVM ($n = 69$).

Для хронологічних розподілів SVM досягла макро-F1 0,5111 на прикладах змішаних мов ($n = 29$) та 0,4933 на прикладах української мови ($n = 65$). Логістична регресія досягла макро-F1 0,4936 для прикладів зі змішаними мовами ($n = 29$) та 0,4104 для прикладів українською мовою ($n = 65$). Оскільки у хронологічних розбиттях вибірки був лише один приклад з англійською мовою, неможливо зробити висновки щодо цієї мовної групи.

Як зазначено вище, результати демонструють, що найкращою перевіреною конфігурацією є TF-IDF + SVM. Детальнішу інформацію можна побачити на рис. 3.

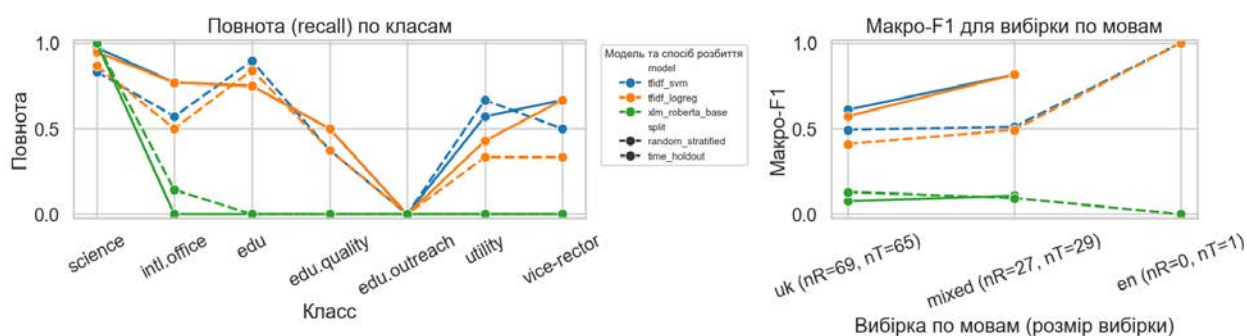


Рис. 3. Метрики по класах та мовах

У 382 прогнозах за TF-IDF + логістичною регресією та TF-IDF + SVM було 102 помилки для обох розбиттів сумарно. 21 помилка для логістичної регресії та 19 для SVM за використання випадкового поділу проти 33 та 29 відповідно за використання хронологічного поділу. Це збільшення кількості помилок відповідає зниженню макро-F1 у разі оцінювання у хронологічному розбитті.

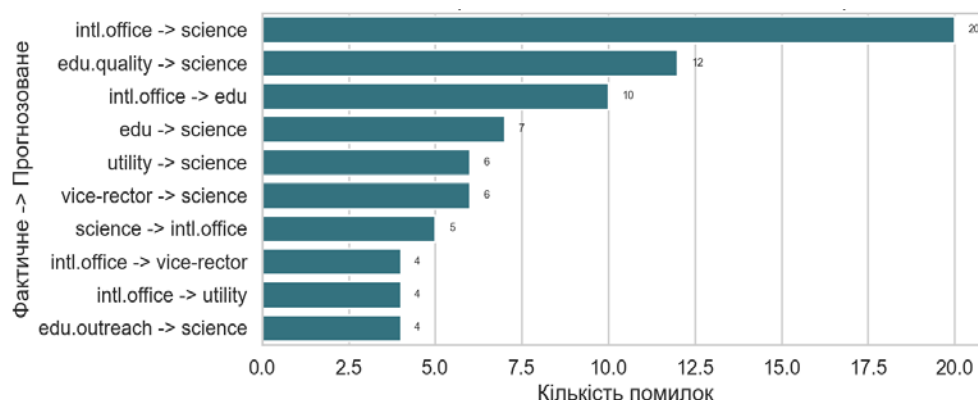


Рис. 4. Найчастіші переходи (топ 10) з помилкою, що відповідають прогнозам

Існував сильний зв'язок між рівнем впевненості та помилками базових моделей. Використовуючи 0,5 як поріг низької впевненості, 101 із 102 помилок TF-IDF опинилися нижче цього порогу.

Як впливає з рис. 4, найпоширенішими помилками базових моделей є:

- «international office», неправильно класифікований як «science».
- «education quality», неправильно класифікована як «science».
- «international office», неправильно класифікована як «education».

До того ж, у моделях TF-IDF 55 із 102 помилок припадали на категорію «science». Отже, здається, що неоднозначні електронні листи зливаються в одну широку категорію з високою частотою.

Головним емпіричним висновком цього дослідження є те, що найкращою моделлю виявилася розріджена лексична базова модель, а не багатомовний трансформер. Цей висновок узгоджується з попередніми дослідженнями, які продемонстрували, що базові моделі класифікації тексту можуть залишатися досить конкурентоспроможними в багатьох реальних застосуваннях. Особливо це актуально для задач, де навчальні дані обмежені та/або лексичні підказки, специфічні для завдання, є інформативними. Водночас, цей висновок не суперечить корисності застосування багатомовних трансформерів [5]. Натомість він вказує на те, що розмір маркованої вибірки в цьому дослідженні та її незбалансований характер не дозволили тестованій конфігурації трансформера скористатися потенційними перевагами.

Окрім точності та зваженого F-1, не менш важливими є висновки на рівні класів. Хоча ці метрики вказували на прийнятну якість класифікації для всіх категорій, макро-F1 та повнота (recall) за класами давали обережнішу оцінку. Ефективність зосереджена серед більшості категорій, тоді як низька ефективність продовжувала спостерігатися у категорій з меншою вибіркою. Це є класичною небезпекою в незбалансованому навчанні, де загальні показники ефективності можуть маскувати системні провали класів меншин [6]. Отже, базову модель SVM варто розглядати як корисний відповідний пункт, а не як комплексне рішення проблеми класифікації листів.

Аналіз рівня впевненості моделей є хорошим початком, але далеко не повним. Майже кожна помилка TF-IDF допущена у разі рівня впевненості менше 0,5. Потенційно такі листи можуть бути використаними для визначення пріоритетності для перевірки вручну.

Існує кілька обмежень щодо можливості інтерпретації цих результатів. По-перше, розмір контрольованої вибірки був відносно невеликим порівняно з розміром даних після видалення дублікатів. Після фільтрування даних для навчання залишилося лише 477 можливих повідомлень з мітками.

По-друге, оскільки мітки, використані для контрольованої оцінки, походили з напівкеруваного навчання, залишковий шум міток залишається можливою причиною помилок.

По-третє, оскільки деякі класи мали надзвичайно малу кількість екземплярів, їхні відповідні оцінки є нестабільними.

По-четверте, проблема зміни загального змісту листів з часом залишається невирішеною. Це можна побачити через існування істотних відмінностей між тим, наскільки добре кожна модель працювала з хронологічним розбиттям порівняно з випадковим.

Ці обмеження вказують на чіткий напрямок для подальших досліджень:

- розширення охоплення міток, особливо для класів меншин.
- ретельніша оцінка рівня впевненості моделі за допомогою досліджень калібрування.
- проведення додаткових послідовних часових ретроспективних тестів можуть допомогти отримати чіткіше уявлення про стабільність даних.

Висновки

У межах цього дослідження TF-IDF + SVM виявився найефективнішим базовим підходом для багатомовної класифікації електронної пошти в умовах дисбалансу класів. Він дав значення макро-F1 0,6421 на випадковому розбитті та 0,5371 на хронологічному. До того ж, логістична регресія і тестова конфігурація XLM-RoBERTa показали гірші результати, ніж TF-IDF + SVM.

Отже, одним із найважливіших позитивних результатів цього дослідження є те, що навіть у багатомовній ситуації з напівкеруванним навчанням проста та недорога розріджена модель може забезпечити достатню якість підтримки класифікації.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

- [1] S. Scerri, G. Gossen, B. Davis, and S. Handschuh, "Classifying Action Items for Semantic Email," in *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, 2010. [Online]. Available: <https://aclanthology.org/L10-1018/>.

- [2] B. Sandrih Todorovic, K. Josipovic, and J. Kodre, "Three Approaches to Client Email Topic Classification," in *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, Varna, Bulgaria, 2023, pp. 1015-1022. [Online]. Available: <https://aclanthology.org/2023.ranlp-1.109/>.
- [3] M. Marcuzzo, A. Zangari, M. Schiavinato, L. Giudice, A. Gasparetto, and A. Albarelli, "A multi-level approach for hierarchical Ticket Classification," in *Proceedings of the Eighth Workshop on Noisy User-generated Text (W-NUT 2022)*, Gyeongju, Republic of Korea, 2022, pp. 201-214. [Online]. Available: <https://aclanthology.org/2022.wnut-1.22/>.
- [4] J. Wang, and C. D. Manning, "Baselines and Bigrams: Simple, Good Sentiment and Topic Classification," in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL)*, Jeju, Republic of Korea, 2012, pp. 90-94. [Online]. Available: <https://aclanthology.org/P12-2018/>.
- [5] A. Conneau, et al., "Unsupervised Cross-lingual Representation Learning at Scale," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020, pp. 8440-8451. [Online]. Available: <https://aclanthology.org/2020.acl-main.747/>.
- [6] H. He, and E. A. Garcia, "Learning from Imbalanced Data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 9, pp. 1263-1284, Sep. 2009, <https://doi.org/10.1109/TKDE.2008.239>.
- [7] G. Salton, A. Wong, and C. S. Yang, "A Vector Space Model for Automatic Indexing," *Communications of the ACM*, vol. 18, no. 11, pp. 613-620, Nov. 1975, <https://doi.org/10.1145/361219.361220>.

Рекомендована кафедрою автоматизації та інтелектуальних інформаційних технологій ВНТУ

Дата надходження 6.04.2026

Дата прийняття після доопрацювання 23.04.2026

Дата прийняття до друку після рецензування 14.05.2026

Дата публікації 2.10.2026

Брусенцов Георгій Євгенович — аспірант кафедри систем штучного інтелекту, e-mail: heorhii.y.brusentsov@lpnu.ua . <https://orcid.org/0000-0002-3346-0164> .

Національний університет «Львівська політехніка», Львів

H. Ye. Brusentsov¹

Multilingual Email Classification Under Class Imbalance for University Rector's Office Routing

¹Lviv Polytechnic National University

This study examines the issue of manual sorting of emails in the Rector's office of a local university. Many emails received by the Rector's secretariat require manual review and subsequent forwarding to other university departments. In this case, the emails were not processed directly by the secretariat; they had to be forwarded to seven different recipient departments. Emails intended for the departments of education, education quality, education outreach, the international office, science, the utility department and the vice-rector were analyzed. The initial email database contained 9.842 emails. After removing duplicates, 8.153 unique emails remained. Using an algorithm that combined duplicate removal with a search for similar emails yielded 4.672 unique records. Only 477 of these potential records could be considered sufficiently labelled for supervised model training. The three families of models investigated were: TF-IDF using logistic regression, TF-IDF using a support vector machine (SVM), and XLM-RoBERTa. The combination of TF-IDF and SVM yielded the best results. The combination of TF-IDF and SVM outperformed TF-IDF with logistic regression in terms of the F-measure (macro F1: 0.6421, weighted F1: 0.7920 and accuracy: 0.8021) on a random split and an F-measure (macro F1: 0.5371, weighted F1: 0.6854 and accuracy: 0.6947) on a chronological split. Class analysis demonstrated excellent results for the most common high-frequency classes, such as "science" and "education". However, results for lower-frequency classes were variable. Recall for "education outreach" was zero for both splits due to the presence of only one test record. Of the 102 total errors made in both TF-IDF combinations, all but one occurred below a confidence threshold of 0.5. Thus, overall, this paper identifies the automatic forwarding of emails for the University Rector's Office as a multilingual classification task with extreme class imbalance. The lack of sufficient labelled data and limited lexical models is the greatest obstacle to implement the solution for this task.

Keywords: multilingual email classification; class imbalance; weak supervision; TF-IDF; linear SVM.

Brusentsov Heorhii Ye. — Post-Graduate Student the Chair of Artificial Intelligence Systems, e-mail: heorhii.y.brusentsov@lpnu.ua . <https://orcid.org/0000-0002-3346-0164>