

## **1. Вступ**

У цьому проекті реалізовано невеликого, але реалістичного helpdesk-асистента для вигаданого продукту SecurePass (портал SSO/MFA).

Асистент повинен:

- відповідати на запитання користувачів на основі продуктової бази знань (KB);
- визначати настрій користувача з тексту та підлаштовувати тон відповіді.

Задача визначення настрою формулюється як 3-класна класифікація тексту (calm, annoyed, confused). Helpdesk-бот поєднує цей класифікатор з компонентом пошуку по KB (TF-IDF) і надає просте веб-демо через Streamlit.

## **2. Дані та постановка задачі**

### **2.1 Дані для класифікації настрою**

Для визначення настрою використано датасет GoEmotions. Початкові емоції агрегуються у три узагальнені стани:

- calm – нейтральні/позитивні емоції (наприклад, “neutral”, “gratitude”, “joy”);
- annoyed – злість/фрустрація (наприклад, “anger”, “annoyance”, “disappointment”);
- confused – страх/розгубленість (наприклад, “fear”, “nervousness”, “confusion”).

Отримані дані розбиваються на train/validation/test за початковим поділом GoEmotions. Класи є несбалансованими: calm домінує, annoyed та особливо confused — мінорні. Це мотивує використовувати macro F1 поряд з accuracy.

### **2.2 База знань для відповідей**

Продуктова база знань зберігається у з колонками:

- id – унікальний ідентифікатор;
- section – розділ (наприклад, MFA, Troubleshooting, Admin & Setup);
- question – FAQ-стиль запитання;
- answer – коротка відповідь;

- tags – опційні теги.

КВ покриває основні теми: SSO-логін, проблеми з MFA, браузери, сесії, безпека та конфіденційність. Вона використовується як джерело всіх конкретних відповідей бота.

## 2.3 Формулювання задач

Система розв'язує три пов'язані підзадачі:

### 1) Класифікація настрою

Вхід: текст повідомлення користувача.

Вихід: один з {calm, annoyed, confused}.

### 2) Пошук у базі знань

Вхід: текст запитання.

Вихід: top-k найбільш релевантних записів КВ.

### 3) Формування відповіді

Вхід: повідомлення користувача + настрій + знайдені записи КВ.

Вихід: текст відповіді з тоном, адаптованим до настрою, та додаткові підказки.

## 3. Методи

### 3.1 Базова модель настрою: TF-IDF + Logistic Regression

Як класичний бейзлайн використано підхід “bag-of-words”:

- Тексти перетворюються в ознаки за допомогою TF-IDF.
- Класифікатор — multinomial Logistic Regression (scikit-learn) з типовою регуляризацією і збільшеним max\_iter для збіжності.

Модель дуже швидко навчається і служить референсом, від якого можна відштовхуватись при оцінці трансформера, особливо для мінорних класів настрою.

### **3.2 Трансформерна модель настрою: DistilBERT**

Основна модель — трансформер:

- Базова модель: distilbert-base-uncased.
- Дані перетворено у HuggingFace DatasetDict з полями text та labels.
- Токенізація: AutoTokenizer, обрізання до max\_length=128.
- Модель: AutoModelForSequenceClassification з num\_labels=3 та явними id2label/label2id.

Навчання через HuggingFace Trainer:

- 3 епохи, per\_device\_train\_batch\_size=16, per\_device\_eval\_batch\_size=32;
- learning rate 5e-5, weight decay 0.01, warmup ratio 0.1;
- оцінка наприкінці кожної епохи (accuracy + macro F1);
- вибір найкращої моделі за validation macro F1.

Фінальна модель та токенайзер зберігаються й обгортаються у сервіс EmotionAnalyzer з простим інтерфейсом.

### **3.3 Пошук по KB: TF-IDF + cosine similarity**

Для пошуку по KB використовується простий класичний метод:

- Для кожного запису KB поєднується question + answer у один текст.
- Будується TF-IDF-матриця для всіх записів.
- Для запиту користувача рахується TF-IDF-вектор і cosine similarity з усіма записами KB.
- Повертаються top-k записів з найбільшим score.

Такий підхід прозорий, інтерпретований і достатній для відносно невеликої. Він інкапсульзований у класі TfidfKBRetriever з методом retrieve(query, top\_k).

### **3.4 Ядро helpdesk-бота та демо**

Клас HelpdeskBot поєднує всі компоненти:

- 1) Викликає EmotionAnalyzer.analyze(user\_message) → отримує настрій.

2) Викликає TfIdfKBRetriever.retrieve(user\_message, top\_k=3) → отримує релевантні записи KB.

3) Формує відповідь:

- top-1 KB-відповідь — основний зміст;
- 1–2 додаткові статті в форматі “You might also find these topics helpful”;
- тон залежить від настрою:
  - \* annoyed — вибачення, кроки й підтримка;
  - \* confused — спокійне пояснення, пропозиція розкласти на дрібні кроки;
  - \* calm — нейтральна інформативна подача.

Застосунок на Streamlit (streamlit\_app.py) забезпечує інтерактивний інтерфейс: користувач вводить запит, бачить відповідь бота й за бажанням відкриває блок Details з mood та KB-хітами.

## 4. Експерименти

### 4.1 Налаштування експериментів

Усі експерименти виконувались на локальній машині.

Для обох моделей використано однакові train/validation/test split'и.

Метрики:

- Accuracy — частка правильно класифікованих текстів;
- Macro F1 — середній F1 по класах calm, annoyed, confused ( кожен клас має однакову вагу, важливо для несбалансованих даних).

### 4.2 Порівняння бейзлайну та трансформера

Загальні метрики (validation та test):

TF-IDF + LogisticRegression:

- Validation: accuracy ≈ 0.7932, macro F1 ≈ 0.5640
- Test: accuracy ≈ 0.8014, macro F1 ≈ 0.5580

DistilBERT (transformer):

- Validation: accuracy  $\approx 0.8295$ , macro F1  $\approx 0.6866$
- Test: accuracy  $\approx 0.8264$ , macro F1  $\approx 0.6844$

Per-class F1 на тесті:

TF-IDF + LR (baseline):

- calm: Precision 0.82, Recall 0.95, F1 0.88
- annoyed: Precision 0.68, Recall 0.43, F1 0.53
- confused: Precision 0.60, Recall 0.17, F1 0.27

DistilBERT (transformer):

- calm: Precision 0.89, Recall 0.89, F1 0.89
- annoyed: Precision 0.64, Recall 0.68, F1 0.66
- confused: Precision 0.60, Recall 0.42, F1 0.50

Основні висновки:

- DistilBERT підвищує macro F1 з  $\approx 0.56$  до  $\approx 0.68$  (+0.12 абсолютних).
- Для calm якість приблизно однакова ( $\sim 0.88$ – $0.89$ ).
- Найбільший виграв — у мінорних класів annoyed та confused, що особливо важливо для саппорту.

#### **4.3 Приклади виконання**

# SecurePass Helpdesk Bot

Ask questions about SecurePass SSO/MFA and get answers from the knowledge base. The bot also adapts its tone to your mood.

Your question

|

Press ⌘+Enter to submit form

Send

**Your message:** How do I enable MFA for all users in SecurePass?

**Bot:** Here's the information you need:

Sign in to SecurePass, open your profile or security settings, and choose 'Set up multi-factor authentication'. Follow the wizard to register an authenticator app or another factor allowed by your organization.

You might also find these topics helpful:

- How do I see all applications available to me in SecurePass? (section: Applications & Access)
- SecurePass asks me to enable pop-ups. Is this safe? (section: Browser & Extensions)

If you need more details or a different scenario, just ask.

› Details (mood and KB hits)

# SecurePass Helpdesk Bot

Ask questions about SecurePass SSO/MFA and get answers from the knowledge base. The bot also adapts its tone to your mood.

Your question

Send

**Your message:** What the hell, SSO login keeps failing with an "invalid assertion" error. I can't log in and keep getting an MFA error, this is wasting my time.

**Bot:** I'm sorry you're running into this issue. Let's go through the most relevant solution step by step:

Provide the time of the issue, your username, the application you tried to open, the exact error message, and a screenshot if possible. This helps support reproduce and resolve the problem faster.

You might also find these topics helpful:

- What is just-in-time (JIT) provisioning in SecurePass? (section: Admin & Setup)
- I see a blank page or endless redirect after login. What can I do? (section: Troubleshooting)

If this does not resolve the problem, please let me know what you tried so we can narrow it down further.

# SecurePass Helpdesk Bot

Ask questions about SecurePass SSO/MFA and get answers from the knowledge base. The bot also adapts its tone to your mood.

Your question

**Send**

**Your message:** I'm not sure what SSO actually means. Do I still need a password or not?

**Bot:** I understand this can be confusing. Here's a clear explanation based on our documentation:

No. You sign in with your company account in SecurePass, and the connected applications trust that sign-in. You still must protect your company password and MFA carefully.

You might also find these topics helpful:

- Do I need a browser extension to use SecurePass? (section: Browser & Extensions)
- Who can see my login activity in SecurePass? (section: Security & Privacy)

If any step is unclear, tell me which part is confusing, and I can break it down into smaller steps.

## 5. Виклики, несподіванки та обмеження

Виклики та несподіванки:

- Несбалансованість і “тонкі” настрої: розрізнати annoyed та confused непросто навіть людині.
- Доменно-стилістична різниця між Reddit (GoEmotions) і helpdesk-запитами, при цьому DistilBERT показує прийнятну узагальнюваність.
- Особливості інфраструктури на Apple Silicon.

Обмеження:

- KB невелика і англомовна; TF-IDF гірше працює з синонімами та перефразуваннями, ніж dense-ембеддинги.
- Навчалася лише одна трансформерна модель (DistilBERT).
- Немає multi-turn контексту: кожне повідомлення розглядається окремо.

## 6. Висновки та подальша робота

У проекті реалізовано end-to-end helpdesk-асистента, який:

- визначає настрій користувача (calm, annoyed, confused);
- знаходить релевантні відповіді в KB за допомогою TF-IDF + cosine similarity;
- формує відповіді з тоном, адаптованим до настрою;
- має інтерактивне демо на Streamlit.

Експерименти показали, що TF-IDF + Logistic Regression дає accuracy ≈0.80, але низький macro F1 через помилки на мінорних класах. Fine-tuning DistilBERT підвищує macro F1 до ≈0.68, значно покращуючи розпізнавання annoyed та confused, що напряму важливо для якості саппорту.

Подальша робота:

- dense-векторний пошук замість TF-IDF;
- підтримка multi-turn діалогів;
- розширення KB та мультимовна підтримка.