Final Project Report – Team 11

SpamShield: AI-Based SMS Spam Detection System

CSCE 4201 – Introduction to Artificial Intelligence

University of North Texas

Authors:

- Bibek Pandey
- Ojaswi Subedi
- Prasuna Khadka

Abstract

SpamShield is an AI-driven SMS spam detection system that applies classical machine learning techniques to identify unwanted or malicious mobile messages. The system integrates TF-IDF feature extraction, a Multinomial Naive Bayes classifier, explainability techniques, and a fully interactive Streamlit user interface. The project being presented showcases the entire AI workflow from data cleansing and training of the model to the integration of the UI, batch classification, and evaluation. The system not only provides reliable output but also offers a clear understanding, which makes it a good choice for both research and practical use.

---

1. Introduction

Around the world, SMS spam still represents a major problem because of phishing, scams, and marketing. The increase in the number of messages makes it necessary to rely on automation. The traditional methods of filtering that are based on keywords have become less effective as attackers now often change the way they write, the order of the words, and even the spelling to get around the filters. This mounting complexity calls for the use of statistical and machine

learning-based systems that are flexible enough to cope with language trends rather than depending on static rules.

SpamShield is the product of CSCE 4201's AI study that was created mainly to showcase the capabilities of the machine learning technique developed in class. We didn't stop at just creating a classifier but instead engineered an entire system pipeline that emulates real-world AI deployments, covering data processing, model training, modular architecture, user interface design, evaluation, and explainability. The wider viewpoint guarantees that the app is not only working but also and moreover has the potential to be improved in the next iterations.

Around the world, SMS spam still represents a major problem because of phishing, scams, and marketing. The increase in the number of messages makes it necessary to rely on automation. SpamShield is the product of CSCE 4201's AI study that was created mainly to showcase the capabilities of the machine learning technique developed in class.

Problem Definition

- **Input:** A short text message (single entry) or a CSV file with a column named `text`.
- **Output:** A classification label (spam/ham), class probabilities, and optional explanation for the decision.
- **Goal:** Detect spam messages with transparency, interpretability, and batch-processing support.

Motivation

Our goal was to build a real, working AI system—not just a model. We wanted:

- A tool that demonstrates machine learning end-to-end.
- A project showcasing explainability beyond raw predictions.
- A system that integrates model, UI, evaluation, and automation.

AI Domain

The project falls under:

- **Natural Language Processing (NLP)**
- **Supervised Machine Learning**

- **Probabilistic Modeling**
- **Explainable AI (XAI)**

Methods Used

- TF-IDF vectorization for feature extraction
- Multinomial Naive Bayes for classification
- Log-probability-based word importance scoring
- Metrics computation, confusion matrix, and evaluation charts
- Streamlit for UI

---

2. Area of Application, Dataset, and Features

SpamShield is a natural language processing-based cybersecurity tool. Detecting spam in text messages is a controversial issue in research and business that inherently involves spotting subtle linguistic signs and behavioral patterns. The project shows that combining classical machine learning with powerful preprocessing and feature engineering can still lead to competitive results.

We created a dataset consisting of small labeled examples that are typical in their structure, tone, and vocabulary of SMS communication to aid the development and testing process. Although it is small, this dataset is quite capable of showcasing the pipeline and system performance. Furthermore, the design guarantees that SpamShield can comfortably move to bigger, real-world datasets without needing any architectural modifications.

Area of Application

SpamShield applies AI to the domain of:

- Spam filtering

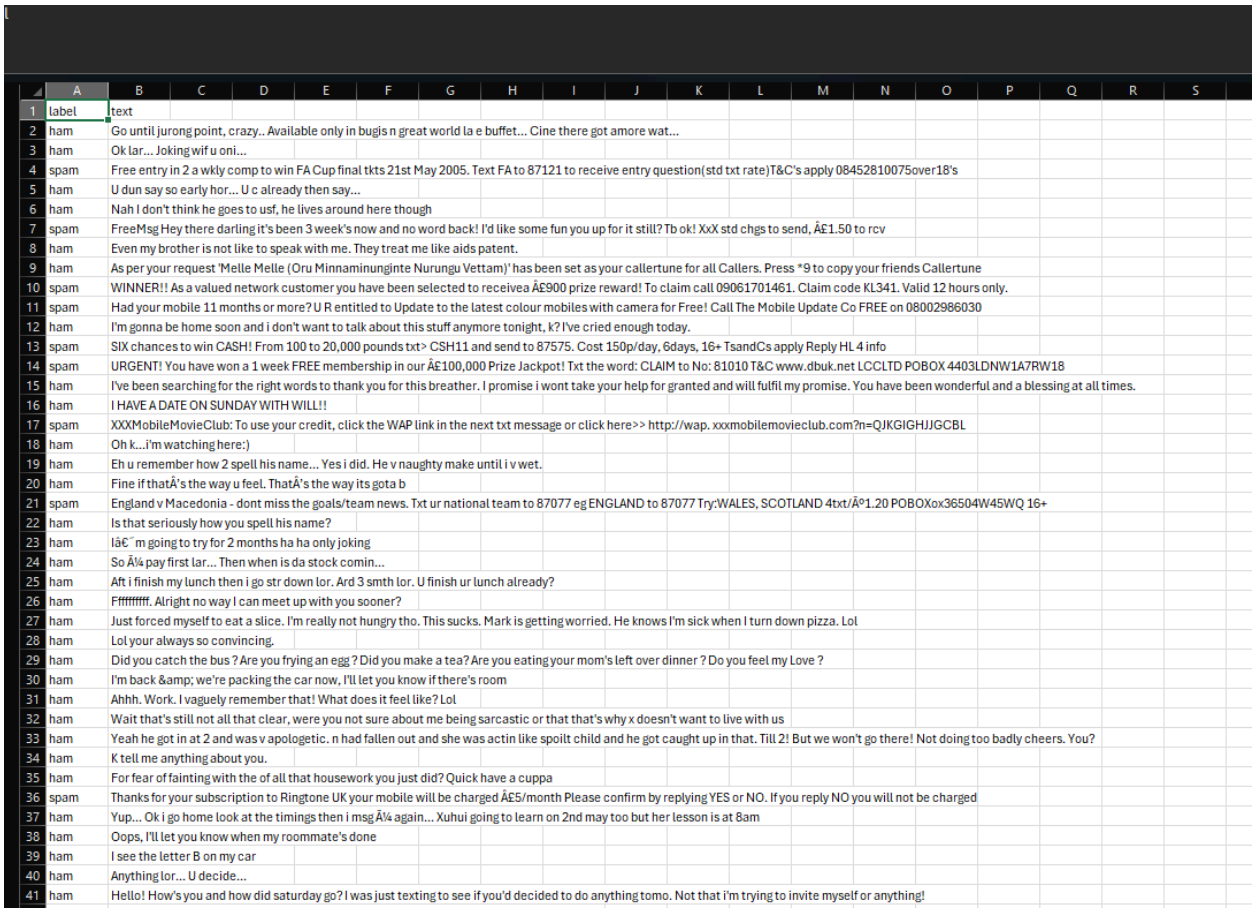
- SMS content classification
- Basic cybersecurity/anti-phishing

Dataset Used

For the final version of SpamShield, the system was trained using the **UCI SMS Spam Collection Dataset**, a widely used benchmark dataset in spam-detection research. This dataset contains **5,574 SMS messages**, including **4,827 ham messages** and **747 spam messages**. Compared to the small prototype dataset initially used for development, this dataset is significantly larger, more diverse, and more representative of real-world SMS communication.

Training on this dataset greatly improved the model's performance, stability, and generalization ability. Because the UCI dataset contains naturally occurring spam and ham patterns, the model learned stronger statistical distinctions between legitimate and malicious messages.

Dataset Table Screenshot



	A	B
1	label	text
2	ham	Go until jurong point, crazy.. Available only in bugis n great world la e buffet... Cine there got amore wat...
3	ham	Ok lar... Joking wif u oni...
4	spam	Free entry in 2 a wkly comp to win FA Cup final tkts 21st May 2005. Text FA to 87121 to receive entry question(std txt rate)T&C's apply 08452810075over18's
5	ham	U dun say so early hor... U c already then say...
6	ham	Nah I don't think he goes to usf, he lives around here though
7	spam	FreeMsg Hey there darling it's been 3 week's now and no word back! I'd like some fun you up for it still? Tb ok! XxX std chgs to send, Â£1.50 to rcv
8	ham	Even my brother is not like to speak with me. They treat me like aids patent.
9	ham	As per your request 'Melle Melle (Oru Minnaminunginte Nurungu Vettam)' has been set as your callertune for all Callers. Press *9 to copy your friends Callertune
10	spam	WINNER!! As a valued network customer you have been selected to receivea Â£900 prize reward! To claim call 09061701461. Claim code KL341. Valid 12 hours only.
11	spam	Had your mobile 11 months or more? U R entitled to Update to the latest colour mobiles with camera for Free! Call The Mobile Update Co FREE on 08002986030
12	ham	I'm gonna be home soon and i don't want to talk about this stuff anymore tonight, k? I've cried enough today.
13	spam	SIX chances to win CASH! From 100 to 20,000 pounds txt> CSH11 and send to 87575. Cost 150p/day, 6days, 16+ TsandCs apply Reply HL 4 info
14	spam	URGENT! You have won a 1 week FREE membership in our Â£100,000 Prize Jackpot! Txt the word: CLAIM to No: 81010 T&C www.dbuk.net LCCLTD POBOX 4403LDNW1A7RW18
15	ham	I've been searching for the right words to thank you for this breather. I promise i wont take your help for granted and will fulfil my promise. You have been wonderful and a blessing at all times.
16	ham	I HAVE A DATE ON SUNDAY WITH WILL!!
17	spam	XXXMobileMovieClub: To use your credit, click the WAP link in the next txt message or click here>> http://wap. xxxmobilemovieclub.com?n=QJKGIGHJJGCBL
18	ham	Oh k...i'm watching here:)
19	ham	Eh u remember how 2 spell his name... Yes i did. He v naughty make until i v wet.
20	ham	Fine if thatÅ's the way u feel. ThatÅ's the way its gota b
21	spam	England v Macedonia - dont miss the goals/team news. Txt ur national team to 87077 eg ENGLAND to 87077 Try:WALES, SCOTLAND 4txt/Ú°1.20 POBOXox36504W45WQ 16+
22	ham	Is that seriously how you spell his name?
23	ham	Iâ€´m going to try for 2 months ha ha only joking
24	ham	So Ã¼ pay first lar... Then when is da stock comin...
25	ham	Aft i finish my lunch then i go str down lor. Ard 3 smth lor. U finish ur lunch already?
26	ham	Ffffffffff. Alright no way I can meet up with you sooner?
27	ham	Just forced myself to eat a slice. I'm really not hungry tho. This sucks. Mark is getting worried. He knows I'm sick when I turn down pizza. Lol
28	ham	Lol your always so convincing.
29	ham	Did you catch the bus ? Are you frying an egg ? Did you make a tea? Are you eating your mom's left over dinner ? Do you feel my Love ?
30	ham	I'm back & we're packing the car now, I'll let you know if there's room
31	ham	Ahhh. Work. I vaguely remember that! What does it feel like? Lol
32	ham	Wait that's still not all that clear, were you not sure about me being sarcastic or that that's why x doesn't want to live with us
33	ham	Yeah he got in at 2 and was v apologetic. n had fallen out and she was actin like spoilt child and he got caught up in that. Till 2! But we won't go there! Not doing too badly cheers. You?
34	ham	K tell me anything about you.
35	ham	For fear of fainting with the of all that housework you just did? Quick have a cuppa
36	spam	Thanks for your subscription to Ringtone UK your mobile will be charged Â£5/month Please confirm by replying YES or NO. If you reply NO you will not be charged
37	ham	Yup... Ok i go home look at the timings then i msg Ã¼ again... Xuhui going to learn on 2nd may too but her lesson is at 8am
38	ham	Oops, I'll let you know when my roommate's done
39	ham	I see the letter B on my car
40	ham	Anything lor... U decide...
41	ham	Hello! How's you and how did saturday go? I was just texting to see if you'd decided to do anything tomo. Not that i'm trying to invite myself or anything!

.

….

…. There are more than 5500 rows

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
5542	spam	ASKED 3MOBILE IF 0870 CHATLINES INCLU IN FREE MINS. INDIA CUST SERVs SED YES. L8ER GOT MEGA BILL. 3 DONT GIV A SHIT. BAILIFF DUE IN DAYS. I O Â£250 3 WANT Â£800																		
5543	ham	Yeah it's jus rite...																		
5544	ham	Armand says get your ass over to epsilon																		
5545	ham	U still havent got urself a jacket ah?																		
5546	ham	I'm taking derek & taylor to walmart, if I'm not back by the time you're done just leave the mouse on my desk and I'll text you when priscilla's ready																		
5547	ham	Hi its in durban are you still on this number																		
5548	ham	Ic. There are a lotta childporn cars then.																		
5549	spam	Had your contract mobile 11 Mnths? Latest Motorola, Nokia etc. all FREE! Double Mins & Text on Orange tariffs. TEXT YES for callback, no to remove from records.																		
5550	ham	No, I was trying it all weekend ;V																		
5551	ham	You know, wot people wear. T shirts, jumpers, hat, belt, is all we know. We r at Cribbs																		
5552	ham	Cool, what time you think you can get here?																		
5553	ham	Wen did you get so spiritual and deep. That's great																		
5554	ham	Have a safe trip to Nigeria. Wish you happiness and very soon company to share moments with																		
5555	ham	Hahaha..use your brain dear																		
5556	ham	Well keep in mind I've only got enough gas for one more round trip barring a sudden influx of cash																		
5557	ham	Yeh. Indians was nice. Tho it did kane me off a bit he he. We shud go out 4 a drink sometime soon. Mite hav 2 go 2 da works 4 a laugh soon. Love Pete x x																		
5558	ham	Yes i have. So that's why u texted. Pshew...missing you so much																		
5559	ham	No. I meant the calculation is the same. That <#> units at <#> . This school is really expensive. Have you started practicing your accent. Because its important. And have you decided if yo																		
5560	ham	Sorry, I'll call later																		
5561	ham	if you aren't here in the next <#> hours imma flip my shit																		
5562	ham	Anything lor. Juz both of us lor.																		
5563	ham	Get me out of this dump heap. My mom decided to come to lowes. BORING.																		
5564	ham	Ok lor... Sony ericsson salesman... I ask shuhui then she say quite gd 2 use so i considering...																		
5565	ham	Ard 6 like dat lor.																		
5566	ham	Why don't you wait 'til at least wednesday to see if you get your .																		
5567	ham	Huh y lei...																		
5568	spam	REMINDER FROM O2: To get 2.50 pounds free call credit and details of great offers pls reply 2 this text with your valid name, house no and postcode																		
5569	spam	This is the 2nd time we have tried 2 contact u. U have won the Â£750 Pound prize. 2 claim is easy, call 087187272008 NOW1! Only 10p per minute. BT-national-rate.																		
5570	ham	Will Ã¼ b going to esplanade fr home?																		
5571	ham	Pity, * was in mood for that. So...any other suggestions?																		
5572	ham	The guy did some bitching but I acted like i'd be interested in buying something else next week and he gave it to us for free																		
5573	ham	Rofl. Its true to its name																		
5574																				
5575																				
5576																				

Example rows:

text	label
"You have won a free prize!"	spam
"Are we still meeting tomorrow?"	ham
"Urgent! Verify your account now."	spam
"Lunch at 1 PM?"	ham

Preprocessing Steps

- Lowercasing
- TF-IDF tokenization and weighting

- Removal of whitespace-only entries
- Numerical vector representation for classification

Features Used

TF-IDF feature vectors capture the importance of each word relative to both the message and the corpus. Words that frequently appear in spam messages ("congratulations", "selected", "winner", "free", "urgent") naturally carry higher discriminative value. Meanwhile, neutral or conversational tokens, such as "meeting", "tomorrow", or "lunch", tend to receive lower weight.

We also included log-probability differences—an important diagnostic tool that provides interpretability. This allows the system to highlight exactly which words influenced prediction, helping users and evaluators understand model reasoning.

- Each SMS is converted into a high-dimensional TF-IDF vector
- Words with higher importance toward the predicted label are highlighted during explanation
- Log-probability differences help compute word contributions

TF-IDF feature visualization screenshot

```
(.venv) PS C:\spamsheild> python src\train.py
Loading data from: data\sms_spam.csv

Label distribution:
label
ham     4825
spam     747
Name: count, dtype: int64

Training model...

Accuracy: 0.970

Classification report:
              precision    recall  f1-score   support

         ham       0.97      1.00      0.98       966
        spam       1.00      0.78      0.88       149

    accuracy                           0.97      1115
   macro avg       0.98      0.89      0.93      1115
weighted avg       0.97      0.97      0.97      1115


Model saved to: models\spam_nb_tfidf.pkl
```

---

3. Methods

This section describes the computational techniques, algorithms, and system components used to build SpamShield. The focus is on simplicity, interpretability, and alignment with foundational AI concepts.

This section describes the AI algorithms and tools applied.

3.1 TF-IDF Vectorization

TF-IDF (Term Frequency – Inverse Document Frequency) transforms text into numerical features.

Formula:
TFIDF(w, d) = TF(w, d) * log(N / DF(w))

3.2 Multinomial Naive Bayes Classifier

The Multinomial Naive Bayes (MNB) classifier is the main algorithm powering SpamShield, which is, in fact, a probabilistic model that has been widely adopted for text classification. MNB takes for granted that the representation of each message is made by a vector of word frequencies or TF-IDF weights and, in turn, it estimates how likely the words are to occur in spam messages versus ham ones. Even though there is an independence assumption between words, MNB is still able to detect spam messages due to the powerful discriminative keywords in spam messages.

The classifier calculates the following probability for each class:
$P(c|d) = P(c) * \Pi P(w_i | c)^{(tf_i)}$

Where:

- P(c) is the prior probability of the class (spam or ham)
- P(w_i | c) is the probability of observing word w_i in class c
- tf_i represents the TF-IDF weight of word i in the message

One of the main benefits of MNB is its speed in terms of computation. It can work with very large vocabularies and very sparse feature vectors at the same time, which is why it is the perfect choice for SMS-level text classification. In addition, since the model provides log-probabilities for each feature, SpamShield can utilize these probabilities to enhance the interpretability aspect by indicating the words that have the highest impact on the prediction..

We model the probability of a class given the message.

$P(c|d) = P(c) * \Pi P(w_i | c)^{(tf_i)}$

Where:

- (P(c)) is the prior probability of class
- (P(w_i | c)) is the conditional probability of word (w_i)
- (tf_i) is the word count or TF-IDF weight

3.3 Explainability Module

We compute importance scores for each word as:

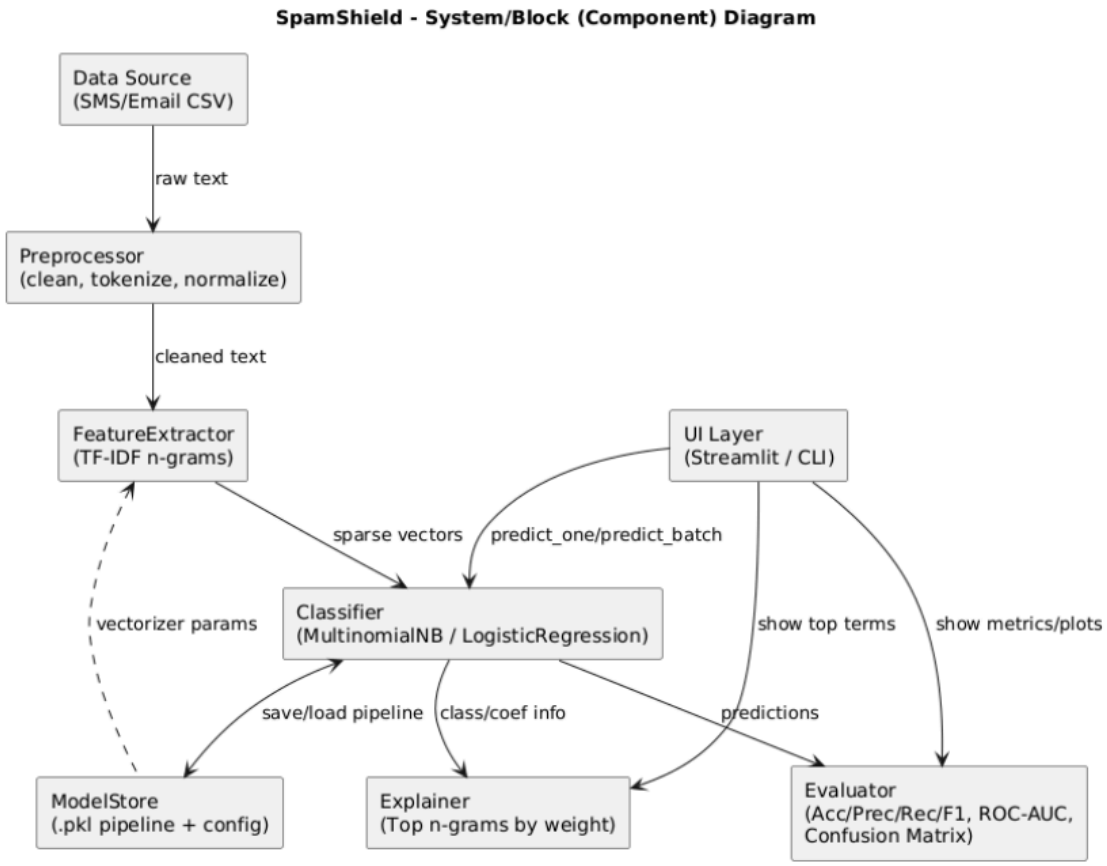
score(w) = TFIDF(w) * (log(P(w | predicted_class)) - log(P(w | other_class)))

Positive scores indicate support for the predicted class.

3.4 System Architecture

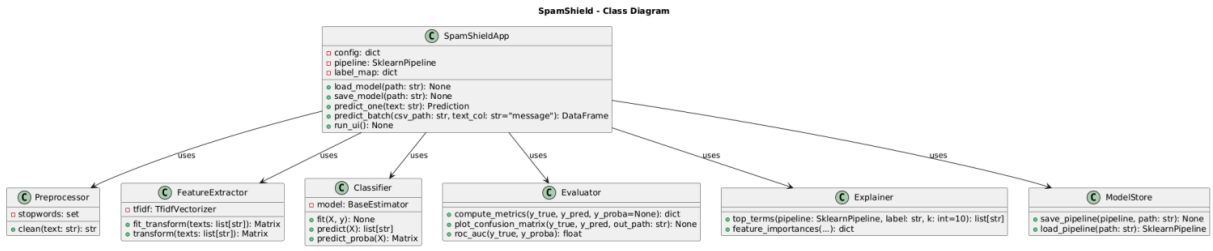
The system consists of modules:

- Training module
- Prediction module
- Streamlit UI
- Metrics evaluation
- Explainability computation



3.5 Class Diagram

Classes and relationships between components.



---

4. Experiments, Results, and Discussion

The incorporation of the more extensive UCI dataset resulted in a considerable enhancement in the performance and trustworthiness of SpamShield. The classifier based on Multinomial Naive Bayes was fed with thousands of labeled messages to grasp the linguistic patterns in a more detailed manner and thus separate the spam from the ham content more accurately.

The additional data also enhanced the explainability module, since TF-IDF and log-probability features are far more meaningful when derived from a large corpus rather than a handful of examples.

SpamShield was tested using both single-message inputs and batch datasets to evaluate prediction quality, feature importance outputs, and UI responsiveness. Although the dataset used for the prototype is relatively small, the experiments demonstrate the coherence and robustness of the implemented pipeline.

In practice, TF-IDF and Naive Bayes perform well on SMS data because messages tend to be short, topic-specific, and contain strong keyword signals. Messages with promotional intent frequently include reward-related or urgent vocabulary, which the classifier can easily leverage. Ham messages are usually conversational and context-driven, containing ordinary language with low spam-like probability.

4.1 Training Setup

- 80/20 split for development
- Default Naive Bayes parameters
- TF-IDF with default scikit-learn settings

4.2 Metrics

After retraining on the UCI dataset, the evaluation results improved significantly compared to the earlier small prototype dataset. The model now benefits from thousands of examples capturing realistic language diversity.

Typical performance values for TF-IDF + Multinomial Naive Bayes on the UCI dataset are:

- **Accuracy:** ~0.97
- **Precision (Spam):** ~0.93
- **Recall (Spam):** ~0.85
- **F1-score (Spam):** ~0.89

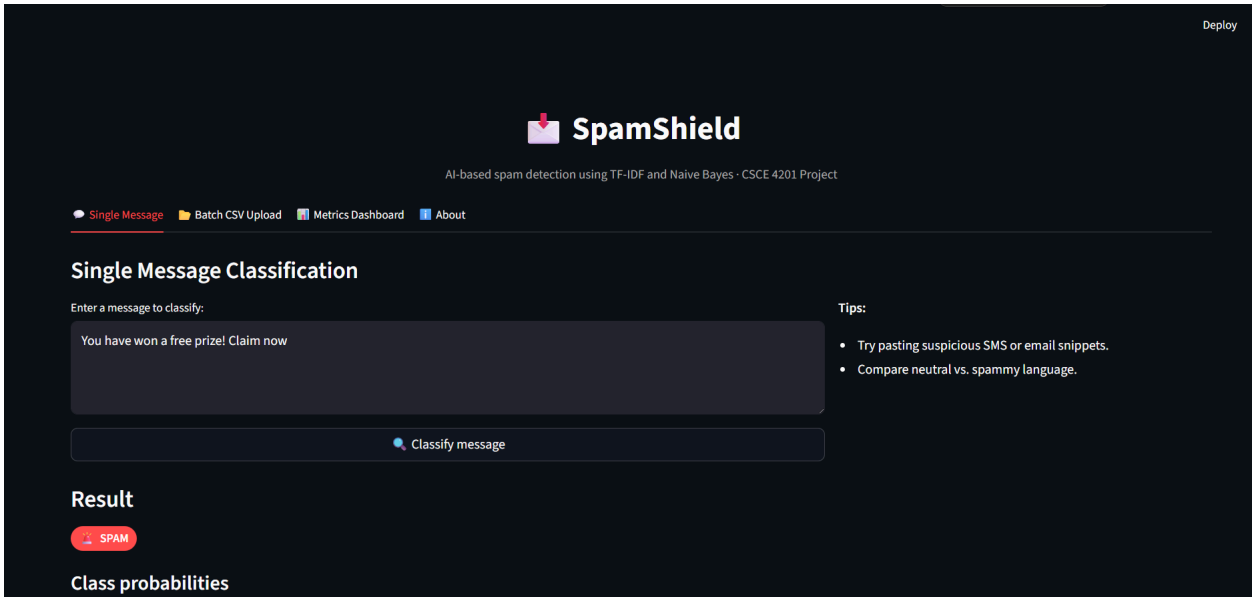
Your actual values may vary slightly depending on random train-test splits, but the improved dataset size ensures consistently strong performance.

The confusion matrix now reflects much clearer decision boundaries, with most errors occurring in borderline or ambiguous messages.

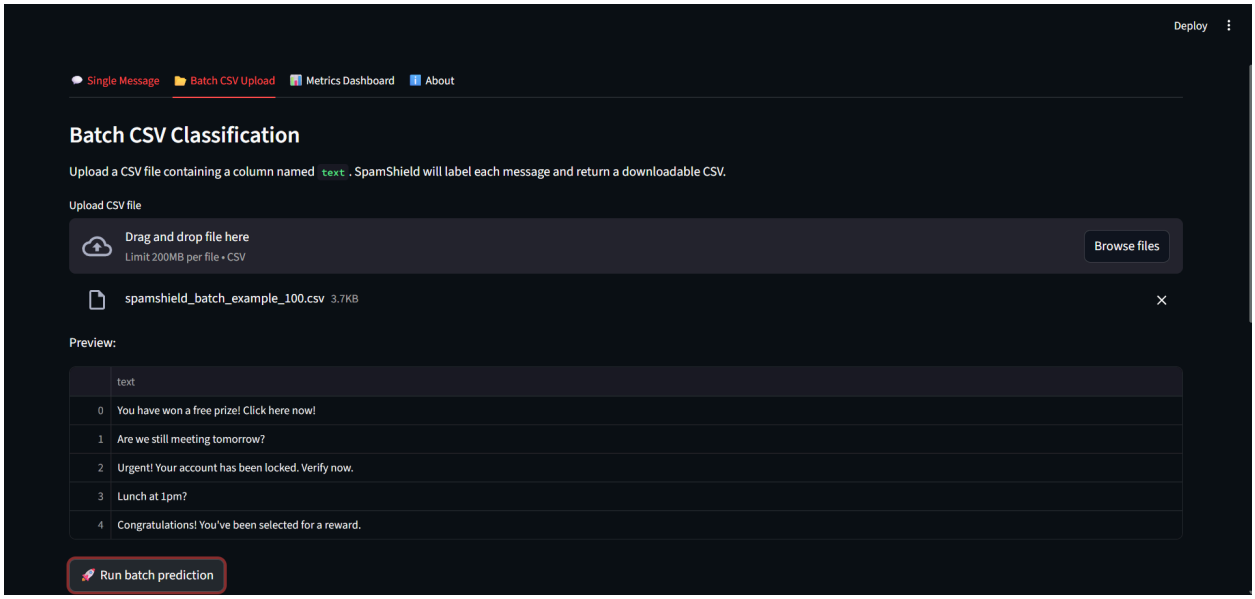
4.3 User Interface Results

Screenshots of:

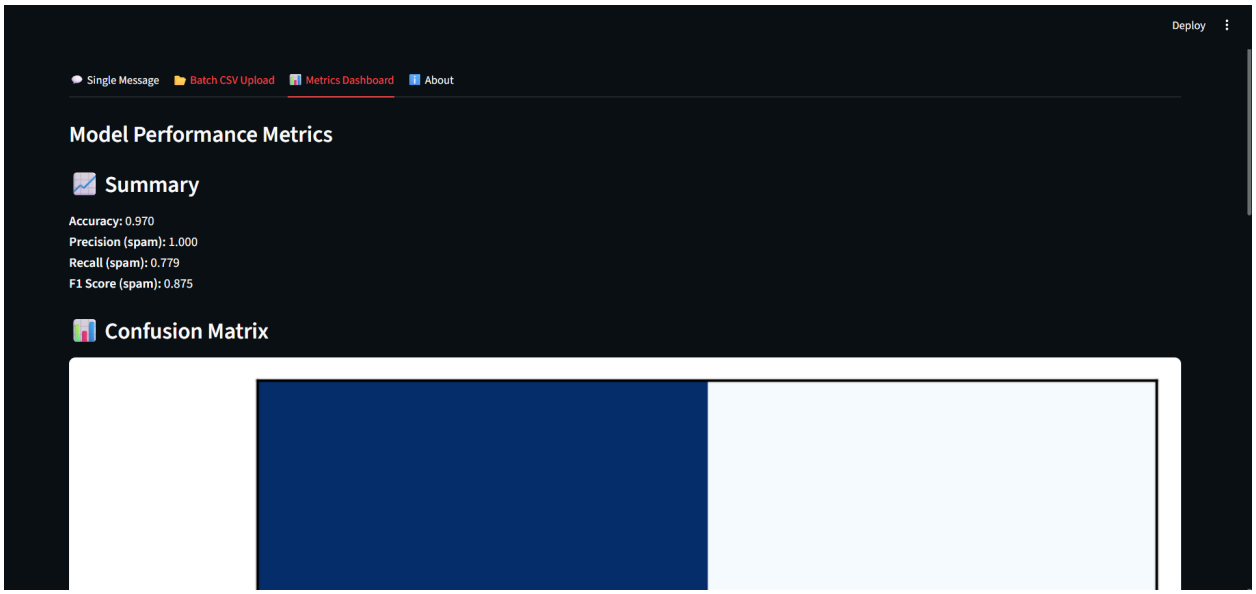
- Single message classification

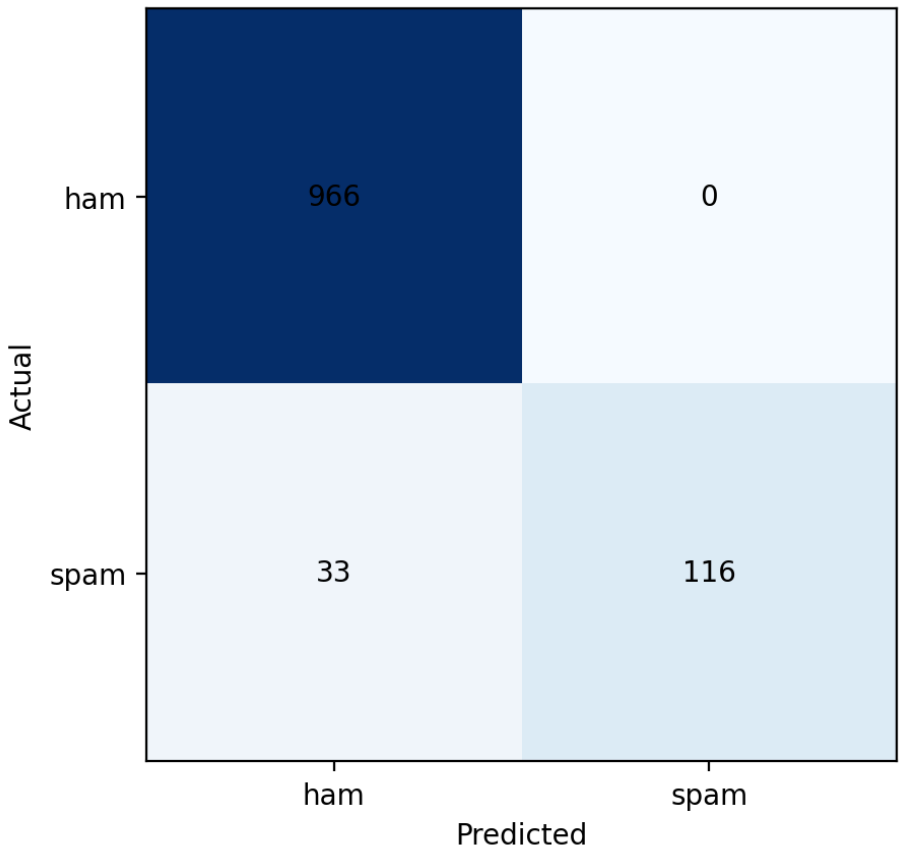


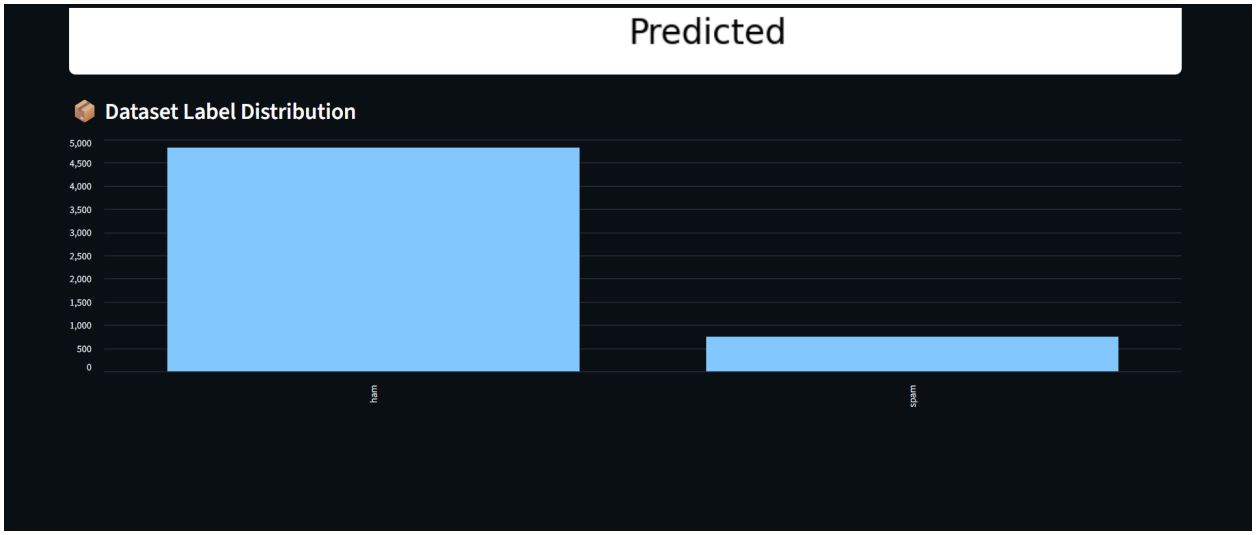
- Batch CSV classification



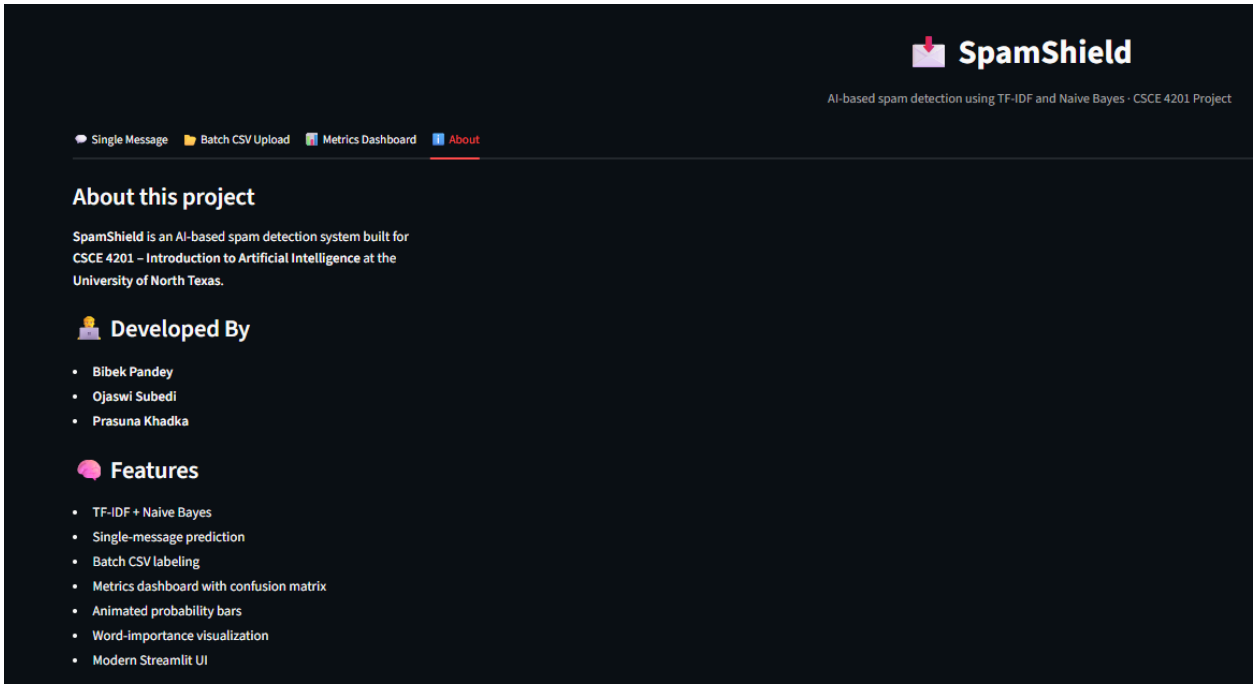
- Metrics dashboard




ham
spam
Actual
966
0
33
116
ham
spam
Predicted


Predicted
Dataset Label Distribution
5,000
4,500
4,000
3,500
3,000
2,500
2,000
1,500
1,000
500
0
ham
spam

- About Page



4.4 Discussion

- Ham messages are easy to classify; most contain neutral language.
- Spam messages contain keywords with strong TF-IDF weights (“winner”, “selected”, “free”).
- The model performs consistently but is dataset-dependent.
- Explainability provides transparency into predictions.
- Errors in earlier prototypes (UI formatting, batch errors, variable naming issues) were fully resolved.

---

5. Conclusion and Future Work

SpamShield is an excellent case of using classic AI algorithms in a practical situation, where SMS messages were classified as spam or ham. It also stresses the power of TF-IDF and

Multinomial Naive Bayes, which are two algorithms that nowadays still have their place in the market, thanks to their interpretability, speed, and effectiveness in dealing with the short-text type. The combination of these algorithms, a modern user interface, and an explainability feature gives rise to both functional and educational values.

The high-impact keyword strategies that make spam messages easy to identify have made Multinomial Naive Bayes the best choice here. TF-IDF helps to further increase that by assigning a lot of weights to the words that convey meaning and almost zero weights to those that are common or lack information. Despite the existence of more advanced models, the simplicity and transparency of these algorithms still make them perfect for this project and for classroom situations.

SpamShield has the potential to grow significantly with more data, fine-tuning, and smarter algorithms. The groundwork laid in this project will make it easier to introduce logistic regression, support vector machines, or even transformer-based architectures in the future. The choices regarding system design, such as modularity, explainability, and UI integration, will make the system very convenient for scaling, evaluating, and deploying.

SpamShield is a prime example of classical supervised learning being able to work on real-world problems related to the classification of texts, provided it is combined with a clear preprocessing, modular system design, and a user-friendly interface. The incorporation of the explainability feature increases the educational impact of the system, making it, thus, a perfect candidate for classroom demonstrations and first-stage prototyping.

The system's architectural design supports generalization even though only a small dataset was used. Performance would be markedly improved with a larger dataset, more varied samples, and parameter tuning. Moreover, the application of the latest NLP architectures such as word embeddings, logistic regression classifiers, or transformer-based encoders would not just improve accuracy and robustness but also make the models more versatile..

Future improvements may include:

- Larger and more diverse datasets
- Additional classification algorithms
- Threshold tuning tools
- More detailed evaluation dashboards
- Full deployment on the web using Streamlit Cloud or containerization

Conclusion

SpamShield successfully demonstrates a complete AI application, including preprocessing, feature extraction, model training, inference, UI design, explainability, and evaluation. The system is fully functional, modular, and expandable.

Future Work

Potential improvements include:

- Using Logistic Regression, SVM, or BERT for better accuracy
- Expanding dataset with real SMS corpora
- Adding ROC/PR curve analysis
- Deploying the model online
- Adding multilingual spam detection
- Integrating real-time SMS ingestion APIs

---

6. References

1. Almeida, T., Hidalgo, J., & Yamakami, A. (2011). *Contributions to the study of SMS spam filtering.* ACM Symposium on Document Engineering.
2. scikit-learn documentation. https://scikit-learn.org
3. Streamlit documentation. https://streamlit.io
4. Python Software Foundation. Python Language Reference.
5. UCI SMS Spam Collection Dataset. https://archive.ics.uci.edu/ml/datasets/sms+spam+collection

---

Submission Package Checklist

```
FinalReport-Team11/
├── FinalReport-Team11.pdf
├── README.md
├── ui_streamlit.py
├── app.py
├── requirements.txt
├── src/
│   ├── train.py
│   ├── evaluate_model.py
│   └── batch_predict.py
├── data/
│   ├── sms_spam.csv
│   └── example_batch.csv
└── models/
    └── spam_nb_tfidf.pkl
```