SMS Spam Detection Model

MIDS W266 - Summer 2023
Sunny Shin, Raymond Tang, Karsyn Lee

Agenda

01 Introduction, Data and Challenges

02 Models and Performance

03 Conclusion

Introduction

OUR MODEL	Spam Detection Model for SMS • Spam -> identity theft, malware, and unwarranted charges on your phone bill
RELATED MODELS	Spam detection most commonly modelled for email - has been a profitable business
WHY IS THIS IMPORTANT	March 2023, (FCC) is cracking down on spam text messages with new rules for telecom companies - can cost telecom company various fines for each reported SMS • New market for SMS Spam detection from telecom companies that has never existed before

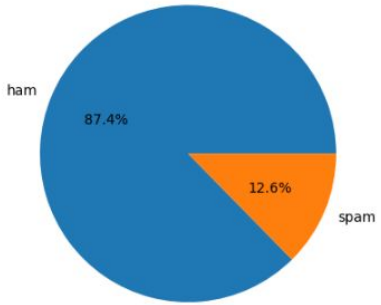
Challenges

When comparing to **email** spam detection, which is widely practiced in industry using NLP, **SMS** spam detection raises these new challenges

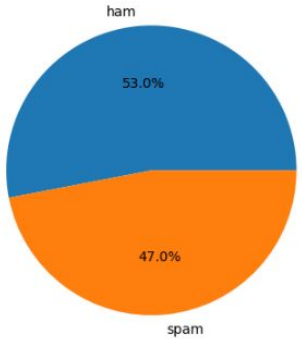
- Data source
    - Lack of available & recent data due to privacy
- Data characteristics
    - message ambiguity due to short length
    - limited header information
    - presence of emojis and abbreviations

Data

- University of California, Irvine Machine Learning Repository (donated June 2012)
  - 5,574 total instances - legitimate(ham) vs. spam
  - User receives ratio 10/90 spam to ham SMS messages - data is representative
  - **Imbalanced dataset** - required resampling technique



Original dataset



Dataset after resampling

Models

Features	Baseline	LSTM	CNN	BERT*	BERT-CNN*	BERTweet**	BERTweet-CNN**
Vectorization	X	X	X	X	X	X	X
Logistic Regression	X						
Convolutional Filters			X		X		X
Pretrained tokenizer	X SpaCy-eng-sm	X Word2Vec	X Word2Vec	X BERT built-in	X BERT built-in	X BERTweet built-in	X BERTweet built-in
Pretrained model				X bert-base-cased	X bert-base-cased	X bertweet-base	X bertweet-base

* The BERT model was pretrained on the Toronto BookCorpus (800M words) and English Wikipedia (2,500M words)
** The BERTweet model however was pretrained on corpus consists of 850M English Tweets (16B word tokens ~ 80GB), containing 845M Tweets streamed from 01/2012 to 08/2019 and 5M Tweets related to the COVID-19 pandemic

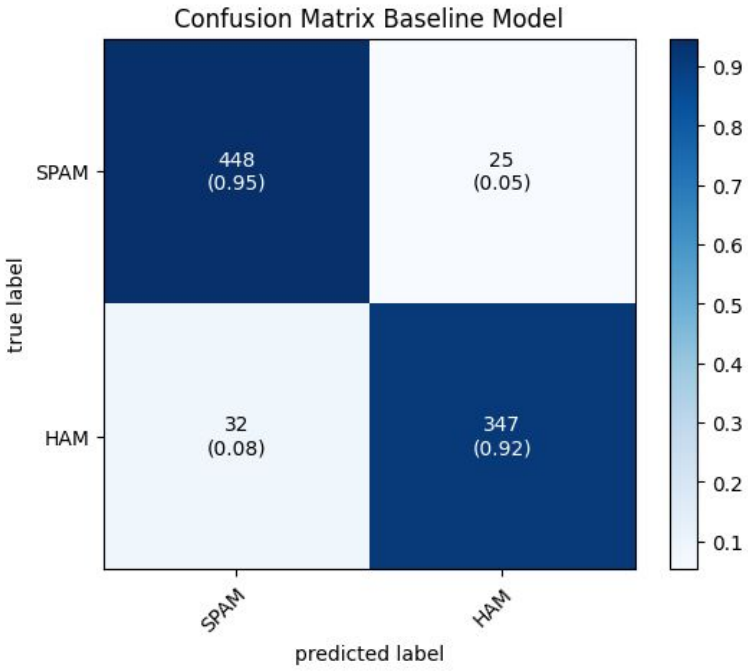
Performance

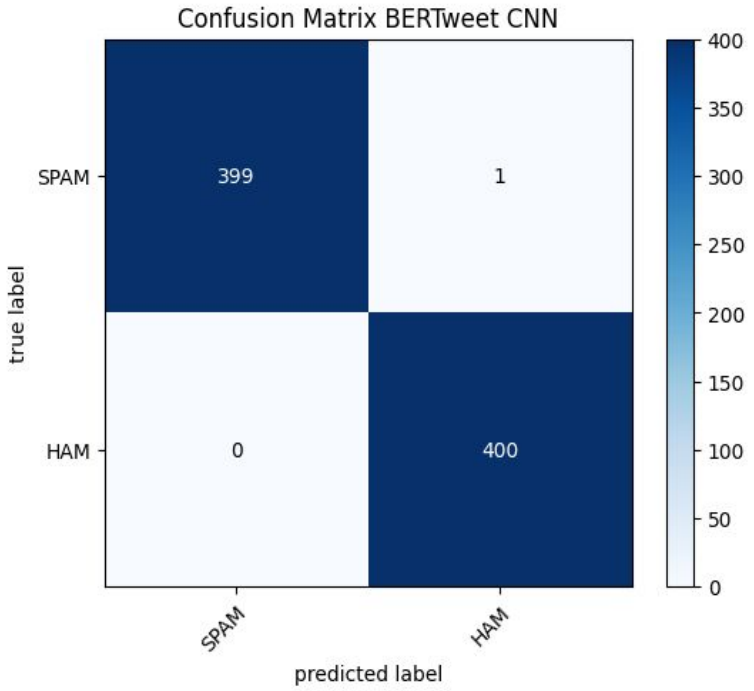
Metrics	Baseline*	LSTM	CNN	BERT**		BERT-CNN	BERTweet**		BERTweet-CNN
				Pooler	CLS		Pooler	CLS	
Accuracy	0.9425	0.9542	0.9918	0.9061	0.9660	0.9988	0.9624	0.9589	0.9988
Precision	0.9444	0.9191	0.9951	0.8980	0.9620	0.9974	0.9852	0.9763	0.9978
Recall	0.9327	0.9927	0.9879	0.8967	0.9633	1.0000	0.9318	0.9238	1.0000
F1 Score* **	**0.9385**	**0.9545**	**0.9915**	**0.8936**	**0.9610**	**0.9986**	**0.9566**	**0.9478**	**0.9989**

* Logistic regression was used as a baseline model for the classification task
** BERT and BERTweet models were trained with both the pooler token and CLS token approaches
*** F1 score was a primary performance measurement since it is a more harmonic mean of precision and recall. It's calculated as:

$$F1\ score = 2 * \frac{Precision\ * Recall}{Precision + Recall}$$

Model Improvement





Conclusion

- **BERTweet-CNN** model performs the best at the F1 score of 0.9989
- Societal impact:
  - Improvement in user experience improvement
  - Increased trust in SMS communication and importance of communication security
- Limitations:
  - Small number of observations and not recent data
  - Ever-changing environment of texting languages and culture
  - Increased risk of blocking ham SMS messages
- Future state:
  - Further training our model with a larger dataset
  - Data augmentation to fix the imbalance issue
  - Creating *spam inbox* for SMS messages

Thank You!

Questions?