



FAKE GOODS DETECTION

From review

Laksamee Pichitpajongkit
Chontira Mahatamavadee



Agenda

01

Problems & Objectives

02

Scope of data

03

Data Analysis

04

Conclusion

Problems

Fake goods detection from customer's review on marketplace platform



Cannot see and try the real product

May be a risk of receiving counterfeit product or no quality.



Spend a long time

To decide which stores should buy.



Money loss

No repeat purchase in the next time

Objectives



Recommend where store you should buy

That will receive quality products or low risk of receiving counterfeit product.



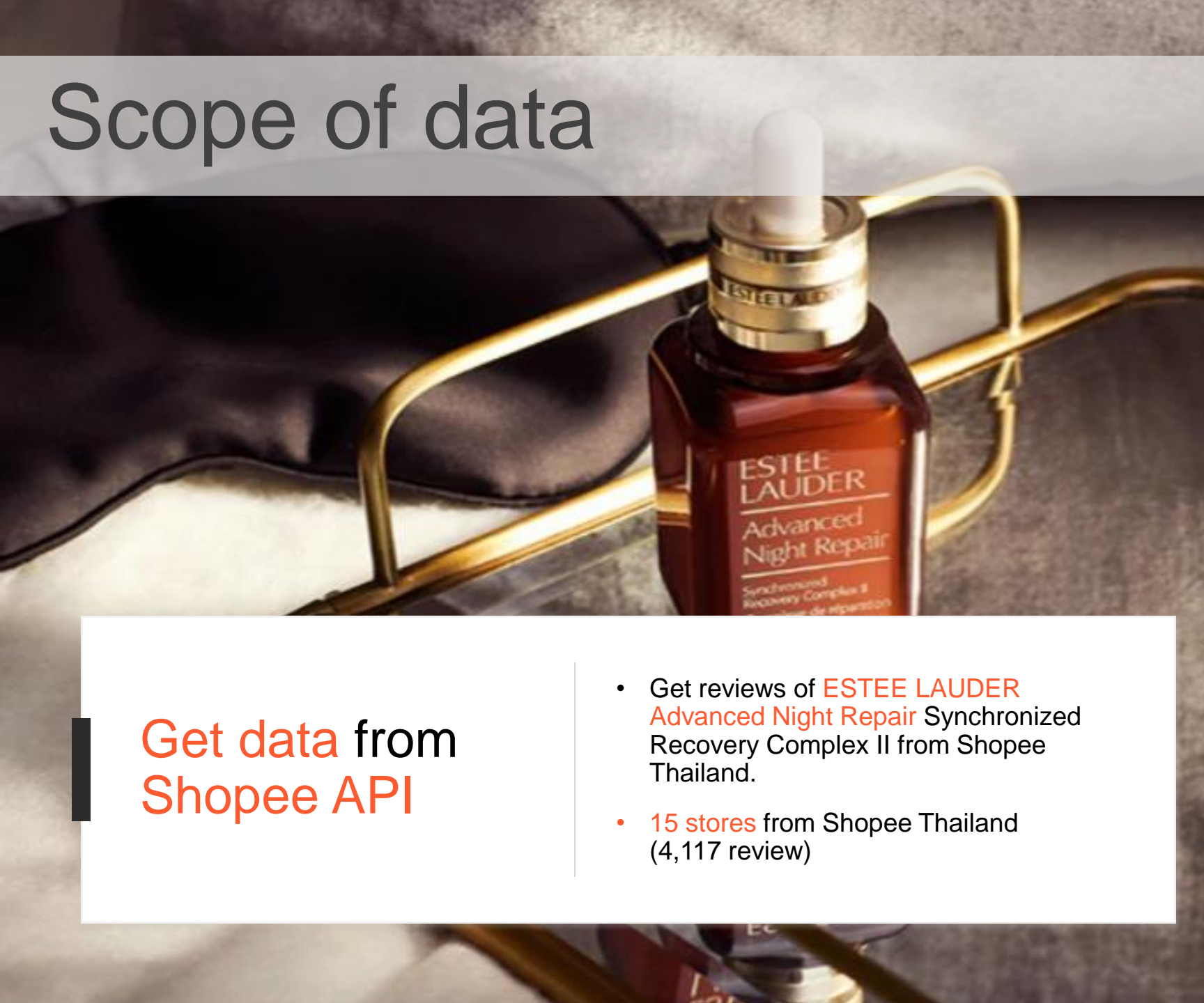
Reduce the time

To make a purchase decision of the buyer.

Scope of data

Get data from Shopee API

- Get reviews of **ESTEE LAUDER Advanced Night Repair Synchronized Recovery Complex II** from Shopee Thailand.
- **15 stores** from Shopee Thailand (4,117 review)



4.8 เต็ม 5
★★★★★

ทั้งหมด 5 ดาว (4.1พัน) 4 ดาว (555) 3 ดาว (87) 2 ดาว (26) 1 ดาว (30)

ความคิดเห็น (1.2พัน) รูป (596)

★★★★★
เนื้อครีมเหลวติดปกดี เหมือนผสมมา
2019-09-19 15:41
มีประโยชน์กับคุณ?

★★★★★
เนื้อและกลิ่นแปลกๆคะ
2019-08-15 21:38
มีประโยชน์กับคุณ?

★★★★★
คิดว่าเจอปอมค่ะ เพราะใช้ทุกวันคะ เนื้อและกลิ่นไม่โอเคเลยคะ เสียใจง่า สิ่ง 15 ml มาด้วย
2019-07-25 22:42
2

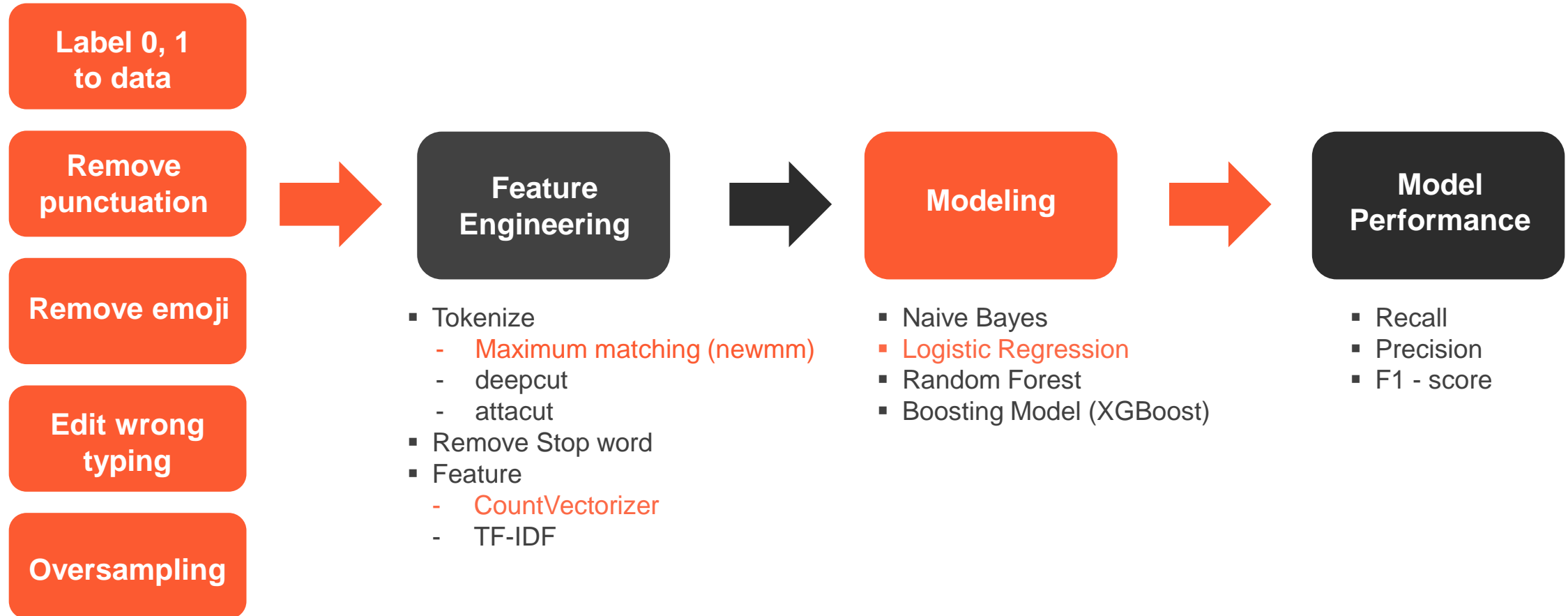
★★★★★
เคยใช้ของแท้แล้วเนื้อสัมผัสไม่เหลวขนาดนี้ ตัวหนังสือที่เขียนบนขวดดูหนา คิดว่าไม่น่าแท้คะ เสียความรู้สึกมาก
2019-06-27 23:34
2

★★★★★
ใสและฉีกส้นและสีแปลก เนื้อบางไหลง่ายเกินไป ไม่เหมือนที่เคยใช้จากเคาเตอร์
2019-06-11 14:39
1

★★★★★
ซื้อร้านนี้ประจำ ซื้อมาหลายครั้ง และหลายๆตัวแล้ว ส่งไวตลอด ของแท้แน่นอน ไม่เคยคิดหวัง การบรรจุภัณฑ์หมดแล้ว ซื้อร้านนี้ประจำ ซื้อมาหลายครั้ง และหลายๆตัวแล้ว ส่งไวตลอด ของแท้แน่นอน ไม่เคยคิดหวัง การบรรจุภัณฑ์หมดแล้ว ซื้อร้านนี้ประจำ ซื้อมาหลายครั้ง และหลายๆตัวแล้ว
2019-11-12 14:58
มีประโยชน์กับคุณ?

★★★★★
หอมดีคะ ไม่มีชาหรือหรือเสียวๆ ขอบคุ้คะ หอมดีคะ ไม่มีชาหรือหรือเสียวๆ ขอบคุ้คะ หอมดีคะ ไม่มีชาหรือหรือเสียวๆ ขอบคุ้คะ หอมดีคะ ไม่มีชาหรือหรือเสียวๆ ขอบคุ้คะ หอมดีคะ ไม่มีชาหรือหรือเสียวๆ ขอบคุ้คะ
2019-11-12 16:10






Data Analysis



Model Performance

Model	Feature	Raw data			Edit wrong typing			Edit wrong typing Oversampling		
		Precision	Recall	F1-score	Precision	Recall	F1-score	Precision	Recall	F1-score
Naïve Bayes	newmm	0.26	0.5	0.35	0.44	0.73	0.55	0.26	0.95	0.41
	deepcut	0.32	0.39	0.35	0.43	0.68	0.53	0.22	0.91	0.36
	attacut	0.26	0.5	0.35	0.44	0.73	0.55	0.26	0.95	0.41
Logistic Regression	newmm	0.18	0.11	0.14	0.67	0.55	0.6	0.47	0.77	0.59
	deepcut	0.33	0.17	0.22	0.53	0.36	0.43	0.47	0.68	0.56
	attacut	0.18	0.11	0.14	0.67	0.55	0.6	0.47	0.77	0.59
Random Forest	newmm	0.67	0.11	0.19	0.71	0.23	0.34	0.75	0.41	0.53
	deepcut	0.67	0.11	0.19	0.88	0.32	0.47	0.64	0.32	0.42
	attacut	0.33	0.06	0.1	0.57	0.18	0.28	0.45	0.23	0.3
Gradient Boosting - XGBoosting	newmm	0.12	0.06	0.08	0.6	0.27	0.37	0.27	0.77	0.4
	deepcut	0	0	0	0.57	0.18	0.28	0.34	0.73	0.46
	attacut	0.12	0.06	0.08	0.6	0.27	0.37	0.27	0.77	0.4

Result

			Predicted probability	Actual probability
1	 <p>(Shop ID: 1514953) Beautymaniashop</p>		1%	1%
	 <p>(Shop ID: 4343306) Pinaza</p>		1%	2%
3	 <p>(Shop ID: 773245) nanana_th</p>		2%	1%
4	 <p>(Shop ID: 50796065) Beautystuffss</p>		3%	3%
5	 <p>(Shop ID: 83258020) ssuuuh</p>		38%	46%

counter serum code
 สินค้า
 เปรียบเทียบ
 lot
 test review
 pack
 ของแท้
 ขวด
 check
 brand
 เปรียบเทียบ
 เหมือนกัน
 เหมือนกับ

Business idea

The screenshot shows a Shopee search results page for 'estee lauder advanced night repair serum'. The top navigation bar is orange with the Shopee logo, a search bar containing the product name, and a shopping cart icon. Below the navigation bar, there are filters for 'ประเภทสินค้า & เซอร์วิส' (Product Type & Service), 'การรับประกัน' (Warranty), and 'การส่งของ' (Shipping). The main content area displays a list of products from various sellers. The first seller is 'Hiso_Shop', which has a 4.8-star rating and 79 items. The products are displayed in a grid, each with a product image, name, price, and a 'Buy' button. The products are: '50ML/100ML Estee Lauder Advanced Night Repair...', 'ANR Estee Lauder Advanced Night Repair...', 'ESTEE LAUDER "NEW..."', 'Estee Lauder Advanced Night Repair...', and '#2#Estee Lauder Advanced Night Repair...'. The page also shows a sidebar with filters for 'ค้นหาแบบละเอียด' (Advanced Search), 'ค้นหาตามหมวดหมู่' (Search by Category), and 'ส่งจาก' (Ship from).

Key points



Quick win closing sell
Spend less time shopping



Trustworthy on platform
Get more product quantity



Positive Customer Satisfaction
Be happy



Thank You



Appendix

Data preparation

1. Data preprocessing

- **Remove punctuation**

Punctuation that remove are `!"#$%&'\()*+,-./:;<=>?@[\\]^_`{|}~` and `"'!`

- **Remove emoji**



- **Edit wrong typing**

‘แพค’, ‘แพ็ค’, ‘แพ็ก’ → **pack**

‘เซรัม’, ‘ซีรัม’, ‘เซลัม’, ‘ซีลัม’ → **serum**

‘เคาน์เตอร์’, ‘เคาเตอร์’, ‘เคาท์เตอร์’ → **counter**

‘แฟลตเซล’, ‘แฟลสเซล’, ‘แฟลชเซลล์’ → **flash sale**

‘เวป’, ‘เว็บ’, ‘เวปไซส์’, ‘เวฟไซต์’ → **website**

‘โค้ด’, ‘ไค้ด’, ‘ไค้ช’, ‘ไค้ต’, ‘โคด’ → **code**

Data preparation

2. Feature Engineering

- **Tokenize**

- newmm

- Maximum Matching algorithm for Thai word segmentation (dictionary-based)

- deepcut

- Deep Learning based Thai word segmentation (learning-based approach)

- attacut

- wrapper for AttaCut. (learning-based approach)

- **Remove stop word**

A stop word is a commonly used word (such as “the”, “a”, “an”, “in”) that a search engine has been programmed to ignore, both when indexing entries for searching and when retrieving them as the result of a search query.

Good to know

Dictionary-based:

Algorithms in this category rely on the use of dictionaries with a mechanism to decide whether to tokenize a particular sequence of characters. Some of algorithms are Chrome’s v8BreakIterator and PyThaiNLP’s newmm.

Learning-based:

Unlike dictionary-based, algorithms in this group learn to split words based on labelled data. The learning problem is typically formulated as binary classification on sequence of characters.

Data preparation

2. Feature Engineering

- **CountVectorizer**

CountVectorizer converts text documents to vectors which give information of token counts.

- **TF-IDF**

It is a method to convert documents into vector such that vector reflects the importance of a term to a document in the corpus.

Modeling

1. Naïve Bayes

			Edit wrong typing				Edit wrong typing Oversampling				Raw data		
model	tokenize	feature	precision	recall	f1		precision	recall	f1		precision	recall	f1
Naive Bayes	newmm	CountVectorizer	0.44	0.73	0.55		0.26	0.95	0.41		0.26	0.5	0.35
		TF-IDF Vectors	0	0	0		0.24	0.86	0.38				
	deepcut	CountVectorizer	0.43	0.68	0.53		0.22	0.91	0.36		0.32	0.39	0.35
		TF-IDF Vectors	0	0	0		0.24	0.91	0.38				
	attacut	CountVectorizer	0.44	0.73	0.55		0.26	0.95	0.41		0.26	0.5	0.35
		TF-IDF Vectors	0	0	0		0.24	0.86	0.38				

Modeling

2. Logistic regression

model	tokenize	feature	Edit wrong typing				Edit wrong typing Oversampling				Raw data		
			precision	recall	f1		precision	recall	f1		precision	recall	f1
Logistic Regression	newmm	CountVectorizer	0.67	0.55	0.6		0.47	0.77	0.59		0.18	0.11	0.14
		TF-IDF Vectors	1	0.09	0.17		0.41	0.82	0.55				
	deepcut	CountVectorizer	0.53	0.36	0.43		0.47	0.68	0.56		0.33	0.17	0.22
		TF-IDF Vectors	0.5	0.05	0.08		0.33	0.77	0.47				
	attacut	CountVectorizer	0.67	0.55	0.6		0.47	0.77	0.59		0.18	0.11	0.14
		TF-IDF Vectors	1	0.09	0.17		0.41	0.82	0.55				

Modeling

3. Random forest

model	tokenize	feature	Edit wrong typing				Edit wrong typing Oversampling				Raw data		
			precision	recall	f1		precision	recall	f1		precision	recall	f1
Random Forest	newmm	CountVectorizer	0.71	0.23	0.34		0.55	0.27	0.36		0.67	0.11	0.19
		TF-IDF Vectors	0.75	0.14	0.28		0.83	0.23	0.36				
	deepcut	CountVectorizer	0.88	0.32	0.47		0.58	0.32	0.41		0.67	0.11	0.19
		TF-IDF Vectors	0.44	0.18	0.34		0.86	0.27	0.41				
	attacut	CountVectorizer	0.57	0.18	0.28		0.8	0.36	0.5		0.33	0.06	0.1
		TF-IDF Vectors	0.67	0.18	0.29		0.71	0.23	0.34				

Modeling

4. Boosting Model (XGBoost)

			Edit wrong typing				Edit wrong typing Oversampling				Raw data		
model	tokenize	feature	precision	recall	f1		precision	recall	f1		precision	recall	f1
Boosting Model	newmm	CountVectorizer	0.6	0.27	0.37		0.27	0.77	0.4		0.12	0.06	0.08
		TF-IDF Vectors	0.56	0.23	0.32		0.26	0.64	0.37				
	deepcut	CountVectorizer	0.57	0.18	0.28		0.34	0.73	0.46		0	0	0
		TF-IDF Vectors	0.62	0.23	0.33		0.36	0.64	0.46				
	attacut	CountVectorizer	0.6	0.27	0.37		0.27	0.77	0.4		0.12	0.06	0.08
		TF-IDF Vectors	0.56	0.23	0.32		0.26	0.64	0.37				