Fake Reviews Detection

SHOBHIT JAIN, PAWAN VAIDYA, ARIHANT CHHAJED, ANISH HEGDE,

Introduction and why is it needed?

- Amazon is facing a major problem to get the fake reviews out of their platforms.
- They have applied various techniques but still they have not been able to fully eradicate these fake reviews out of their platform.
- It is necessary to verify the genuineness of the reviews before posting them. Removing fake reviews would increase the faith of buyers while buying the products and at the same time help increase the trust of buyers on the online marketplace platform. Hence it would be a win-win situation for buyers, sellers(as they will get more orders and more profit) and the online platform(as they will get more traffic and orders through their platform).

PROBLEM STATEMENT

We are addressing the most common problem in today's ecommerce business, to determine the fakeness or the truthfulness of any entity. Here we are considering the Amazon reviews dataset to separate fake and genuine reviews for the product.

1	marketpla	customer	review_id	product_id	product_parent	product_title	star_ratin	helpful_vc	total_vote	vine	verified_p	review_he	review_bc	review_da	year
2	US	33631101	RXVC5R2C	571233724	977049896	Qi Advanced Bante	5	22	24	N	N	"I don't ev	I must hav	########	
3	US	12547443	R2HHJYS2'	1453772863	517108048	LAX California: An	5	0	0	N	Y	Great boo	Great boo	########	
4	US	27717622	R8B169XA	345482476	62929266	Shadow Divers: Th	5	1	1	N	N	An excelle	I read Sha	########	
5	US	24112715	R2KSHOG\	1631061240	383464621	From Frank Everyd	5	1	1	N	N	Love these	I have bee	########	
6	US	12289876	R9XKF91V.	1580170234	531575098	Let it Rot!: The Gar	3	3	25	N	Y	All about c	This little l	########	
7	US	15943263	RWPNZQV	380791714	534499833	Wayside School Bc	5	0	0	N	Y	The packa	I rememb	########	

Dataset

- https://s3.amazonaws.com/amazon-reviews-pds/readme.html

- **Attributes Used out of Dataset:-**
- Prod_id
- Customer_id
- Helpful_votes
- Review_id
- Review_body
- Star rating
- Review date
- Sentiment

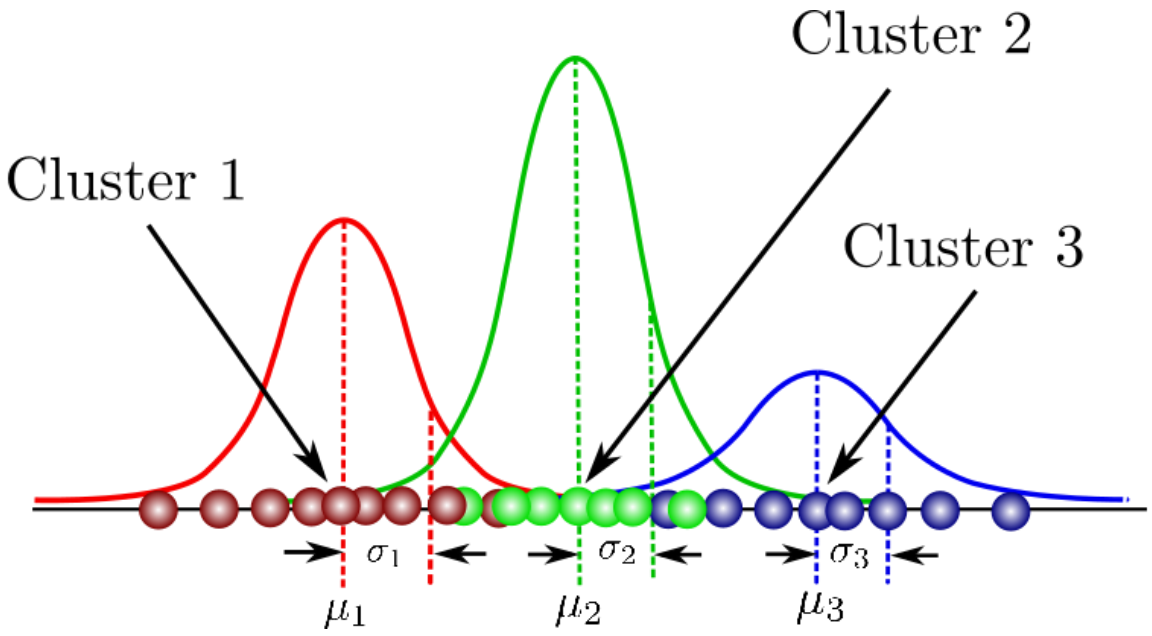
Theory and Conceptual Study

1) Gaussian Mixture Model

- It's a function comprised of several Gaussians, each identified by k belongs to {1, ……, K}, where K is the number of clusters of our dataset. Each Gaussian k is the mixture of the following parameters:

2) Silhouette Width

- Silhouette is a way of validating and interpreting consistency within clusters of data. It gives an adequate graphical representation of how well each object has been classified. It tells how similar an object is to its own cluster.



Preprocessing of Dataset

- We took the following steps to pre-process the dataset:

1) We converted the dataset which was in TSV Format(originally), then we used AWS Glue to convert data into CSV Format.

2) Then we analyzed the dataset & decided to pick only few useful attributes from the dataset.

3) We used Stanford Core NLP to assign a sentiment value for summary feature of each product in the dataset.

PROPOSED SOLUTION AND MODEL

1) Initially, we calculated the average of all ratings based on overall features considering Product ID.

2) Since we have to classify each review based on its sentiment value, as we are considering the clustering Model. We are calculating the mean for all sentiment values.

3) Then we took the non negative difference of 2 values - average rating and overall rating of product ID. This is known as Delta Value.

4) Then we took the non negative difference of mean sentiment value & sentiment of a summary feature of product ID.

5) We had the need to create a custom feature that was an amalgamation(combined mixture) of 3 features:- delta values of sentiment, rating & helpful feature. Hence we created this new custom feature.

6) In order to detect the similarities, we implemented the Gaussian mixture Model also known as GMM. This helped to create clusters which would help us differentiate between genuine and fake reviews. There is a strong possibility that there may be a possible fake reviewer. We can identify this easily after doing an outlier analysis. We have taken this as the basis of our technique to classify reviewer as fake or genuine.

Conclusion and future work

- Currently we are identifying the product as fake based on average sentiment value. So, if the product has thousands of positive reviews but only 1 or 2 negative reviews, then it would classify those specific reviewers as fake. It might be a case that the reviewer was genuine, and he/she must have received a product which was damaged in delivery. Hence, we need to figure out another effective way to avoid such wrong classification in such rare cases.

Sno.	Team Member	Contribution	Contribution
1	Shobhit Jain	Applied the Gaussian Mixture Model for the clustering of reviewers into fake and genuine category.	25%
		Detected the sentiment values using the stanford core NLP library for the available features.	
		Coded the implementation for computing silhouette width(wrt different cluster Size)	
		Prepared the Report.	
2	Arihant Chhajed	Did Data Pre-processing.	25%
		Applied the Gaussian Mixture Model for the clustering of reviewers into fake and genuine category.	
		Created own handpicked identifier (Delta sentiment, delta ratings) based on sentiment & overall(ratings) columns.	
		Coded the implementation for computing silhouette width(wrt different cluster Size)	
3	Pawan Vaidya	Performed cluster analysis and labeled each reviewer.	25%
		Created the data visualization and plots for the output result.	
		Created own handpicked identifier (Delta sentiment, delta ratings) based on sentiment & overall(ratings) columns.	
		Prepared the Report.	
4	Anish Hegde	Did Data Pre-processing, converted data to CSV Format, analysed important features.	25%
		Performed cluster analysis and labeled each reviewer.	
		Created the data visualization and plots for the output result.	
		Applied the Gaussian Mixture Model for the clustering of reviewers into fake and genuine category.	
		Came up with new features	


THANK
YOU!