Summer 2022

Fake Review Detection

Michael Husk
University of Virginia, michael.j.husk@gmail.com

Faryaneh Poursardar (Mentor)
Old Dominion University, fpoursar@odu.edu

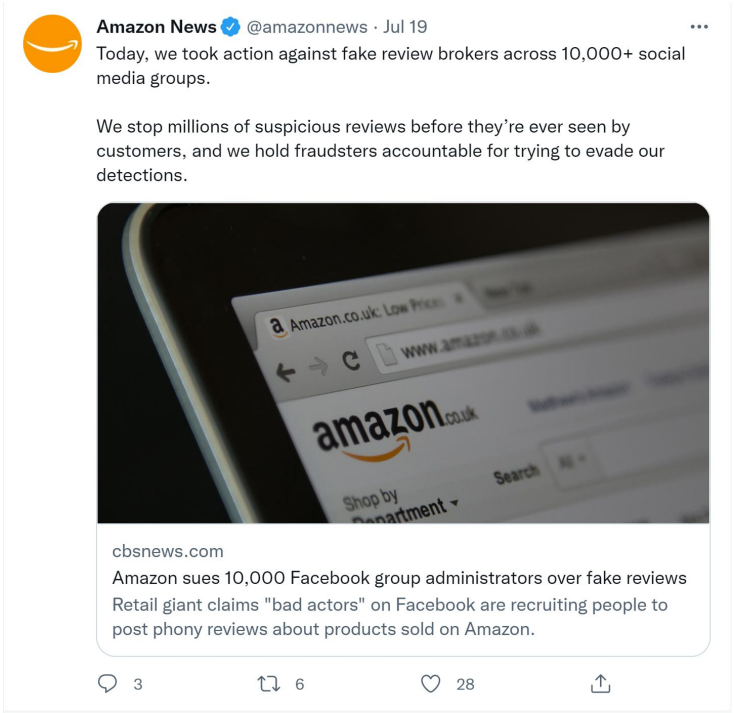
Fake Review Detection

Michael Husk
Under the mentorship of Dr. Faryaneh Poursardar

Reviews

- 93% of customers read reviews when purchasing a product online
- A single review on a product page can boost conversion rates by 354%
- Companies/sites will offer rewards for posting reviews
  - Regardless of whether you actually purchased the item

Review Manipulation in Recent News



Research

- Two major questions:

  Can we develop model(s) to detect fake reviews?

  What features are important for fake review detection?

- Use machine learning algorithms to identify deceitful reviews

Amazon Dataset

- Data collected from May 1996 - October 2018
- Over 233 million reviews
  - Separated into categories based off type
- Home and Kitchen category subset
  - About 6.8 million reviews
- Lack of gold standard data

overall	verified	reviewTime	reviewerID	asin	reviewerName	reviewText	summary
1	TRUE	08 30, 2017	APA4OJQM9I4	B00ZRD9E9I	lookatmeimmik	I received a blank c	Blank canvas???
5	TRUE	05 3, 2017	A1I7IBD13PTT	B00ZRD9E9I	Emily Wurtz	Great quality and la	Love it!
5	TRUE	02 9, 2017	A165WSYLEW	B00ZRD9E9I	Amazon Custon	Beautiful picture! It	Beautiful picture
5	TRUE	01 5, 2017	A2P8NT02T45I	B00ZRJU4QO	Lori	My daughter loves	Great tapestries!
5	TRUE	11 7, 2016	A305X87NGIO	B00ZRJU4QO	lisa Tanner	Very happy with thi	Very happy with
5	TRUE	08 13, 2016	A2OX9LRH978	B00ZRJU4QO	CC July	I love this	Exactly as pictur
5	TRUE	05 5, 2016	A1BHJ2TT7ZD	B00ZRJU4QO	april	my daughter loved	Five Stars
1	TRUE	12 31, 2017	A23PVGJNWBI	B00ZRJIOAM	Richard Dean C	no place on the info	vary vary used a
5	TRUE	03 28, 2017	A1GYRV1IPDB	B00ZRJIOAM	James L Busch	I have 2 of these &	Love them
5	TRUE	07 30, 2018	A2NQEIVS5KS	B00ZRJOKQ4	AMA	Very nice and very	Five Stars
5	TRUE	03 12, 2017	A29D7Q0JTRX	B00ZRJOKQ4	Saundra M.	Absolutely love this	Comfortable
5	TRUE	11 12, 2016	A3JFMGDNQE	B00ZRJOKQ4	Amazon Custon	What an attractive	Favorably adds t
5	TRUE	03 3, 2016	A2T2K48DJTV	B00ZRJOKQ4	Olearfam	Beautiful fabric! Wo	Five Stars
3	TRUE	02 9, 2016	A3PDXPCW7U	B00ZRJOKQ4	Michelle	I was disappointed	Not great

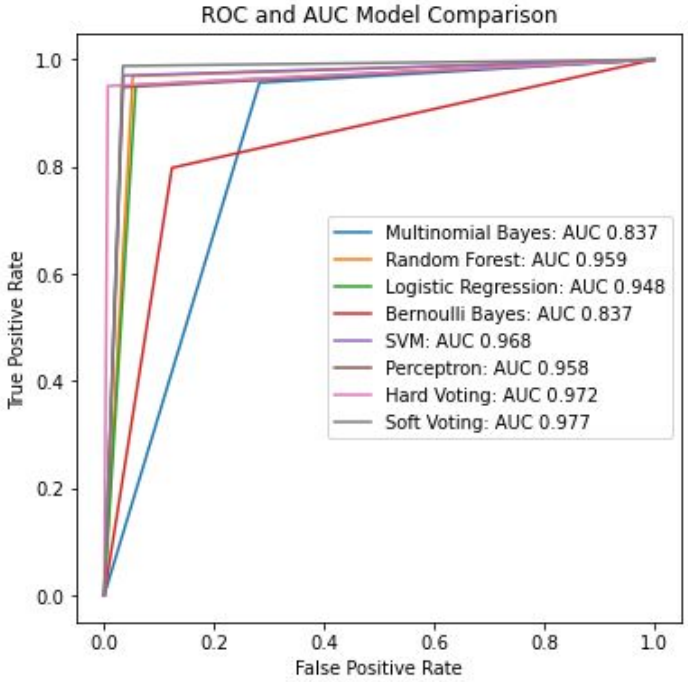
Building the Dataset

- Extracted reviews based off of specific review terms used, to form two datasets that were “fake” or “authentic”
- Fake data
  - Deceptive keywords
- Authentic data
  - Reviews that contained phrases “I bought “ or “I purchased”

disclaimer	discount for review
discount to review	for the purpose of a review
free for my review	free for review
free reviewer’s sample	free sample
free to review	Freebie
in exchange for a review	in exchange for my honest
in exchange of a review	in return for a review
in return of a review	product for review
product for test	review for product
review sample	review unit
reviewing purposes	sample for an honest review
sample for review	sent this for review
testing and review purposes	product sent for review

Models - only review text

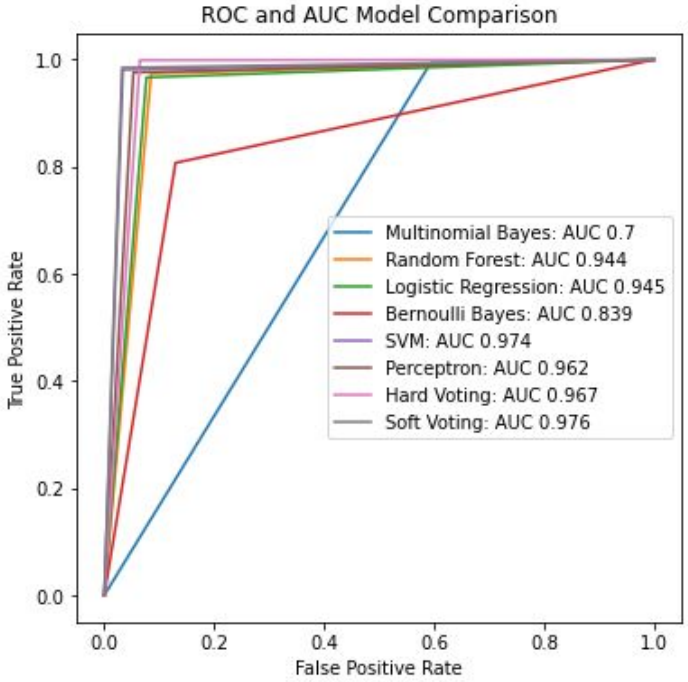
	Accuracy	Precision	Recall	F1
Multinomial Bayes	0.838	0.775	0.957	0.856
Logistic Regression	0.948	0.944	0.952	0.948
Bernoulli Bayes	0.837	0.868	0.798	0.831
Random Forest	0.959	0.95	0.969	0.96
SVM	0.968	0.966	0.97	0.968
Perceptron	0.958	0.967	0.949	0.957
Hard Voting	0.972	0.993	0.951	0.972
Soft Voting	0.977	0.967	0.988	0.978



Models - more features

- Review text
- Summary
- Whether or not the review was verified

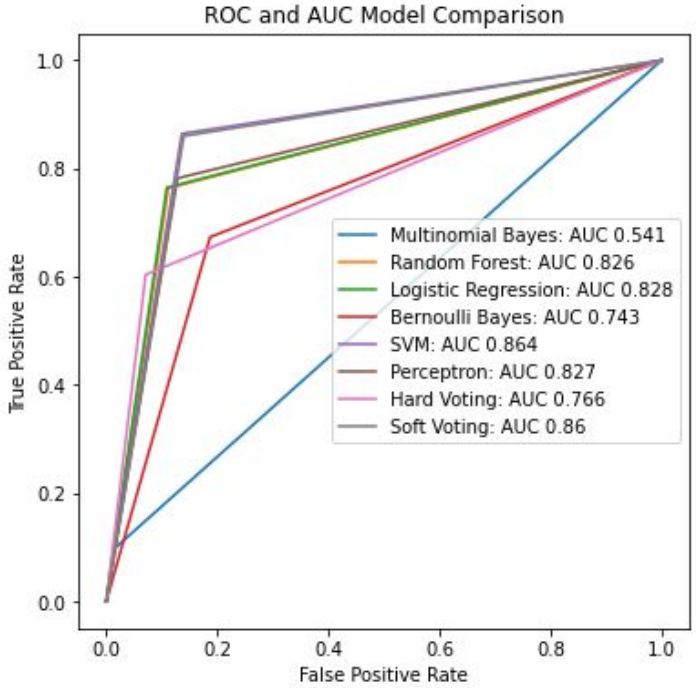
	Accuracy	Precision	Recall	F1
Multinomial Bayes	0.723	0.661	0.994	0.794
Logistic Regression	0.947	0.936	0.967	0.951
Bernoulli Bayes	0.836	0.879	0.807	0.842
Random Forest	0.947	0.93	0.974	0.952
SVM	0.974	0.971	0.981	0.976
Perceptron	0.963	0.956	0.976	0.966
Hard Voting	0.97	0.947	0.999	0.972
Soft Voting	0.977	0.973	0.985	0.979



Models- weighting features

- Review text
- Summary
- Verification
- 5 stars
- Product had multiple reviews on the same day
- Max weight 4, weight >= 3 labeled fake

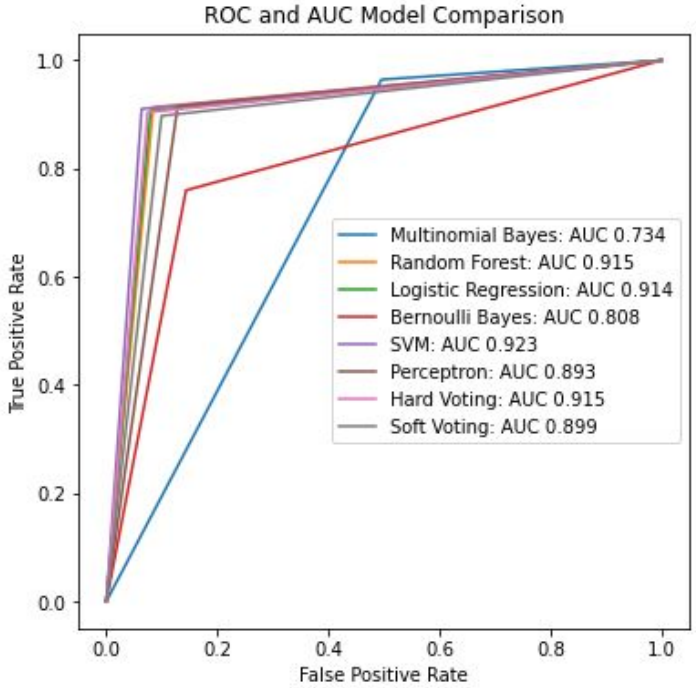
	Accuracy	Precision	Recall	F1
Multinomial Bayes	0.692	0.793	0.095	0.17
Logistic Regression	0.849	0.777	0.765	0.771
Bernoulli Bayes	0.767	0.641	0.673	0.657
Random Forest	0.847	0.772	0.764	0.768
SVM	0.864	0.759	0.864	0.808
Perceptron	0.842	0.751	0.782	0.766
Hard Voting	0.821	0.809	0.604	0.691
Soft Voting	0.86	0.754	0.86	0.803



Models- weighting features

- Review text
- Summary
- Verification
- 5 stars
- Product had multiple reviews on the same day
- Max weight 4, weight >= 2 labeled fake

	Accuracy	Precision	Recall	F1
Multinomial Bayes	0.732	0.656	0.964	0.781
Logistic Regression	0.914	0.918	0.907	0.912
Bernoulli Bayes	0.809	0.839	0.76	0.797
Random Forest	0.915	0.914	0.913	0.914
SVM	0.923	0.934	0.91	0.922
Perceptron	0.893	0.875	0.915	0.894
Hard Voting	0.915	0.923	0.904	0.913
Soft Voting	0.899	0.899	0.897	0.898



Conclusion

- Lack of gold standard data
- Started with just review text, added more features for prediction
- Achieved models with ~92% accuracy and F1
- With more time:
  - Deep learning models
  - More features e.g. user behavior, length of review