Scrapable Insights: How Twitter Can Inform on Kenyan Electrical Grid Growth and Development

Shayne T. O'Brien
SUNY Geneseo

Margaret Sy
UC Berkeley

Dr. Eric Brewer
UC Berkeley

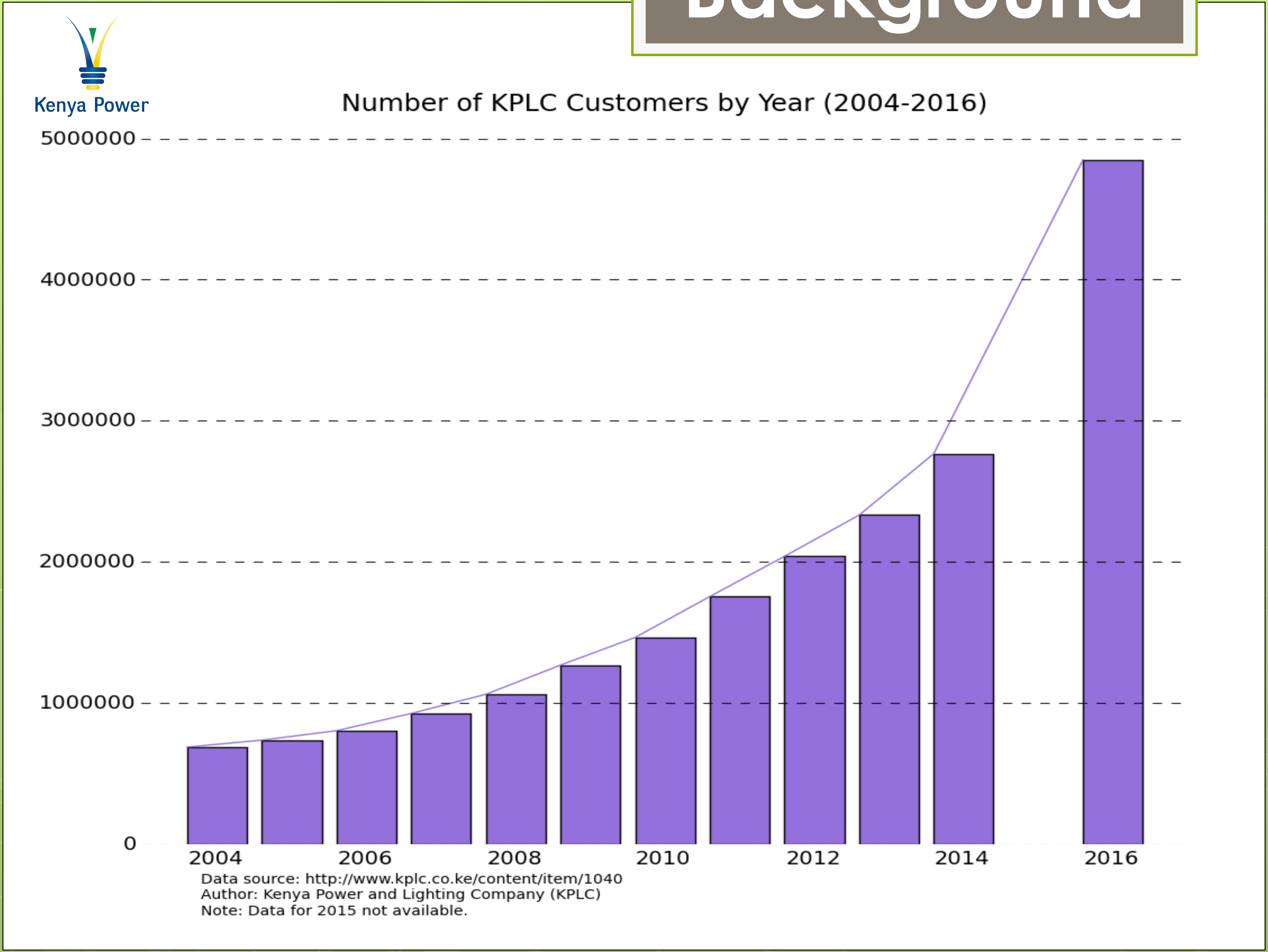
This research is funded by NSF Grant No. 1359499.

Background

- Around the world, significant portions of the world population remain “off the grid.”

- Being connected does not mean that the power supply is reliable.

Background


Kenya Power
Number of KPLC Customers by Year (2004-2016)
5000000
4000000
3000000
2000000
1000000
0
2004
2006
2008
2010
2012
2014
2016
Data source: http://www.kplc.co.ke/content/item/1040
Author: Kenya Power and Lighting Company (KPLC)
Note: Data for 2015 not available.

Research Question

What can we learn about the reliability and growth of Kenya's electrical grid over time by analyzing public customer service communications?

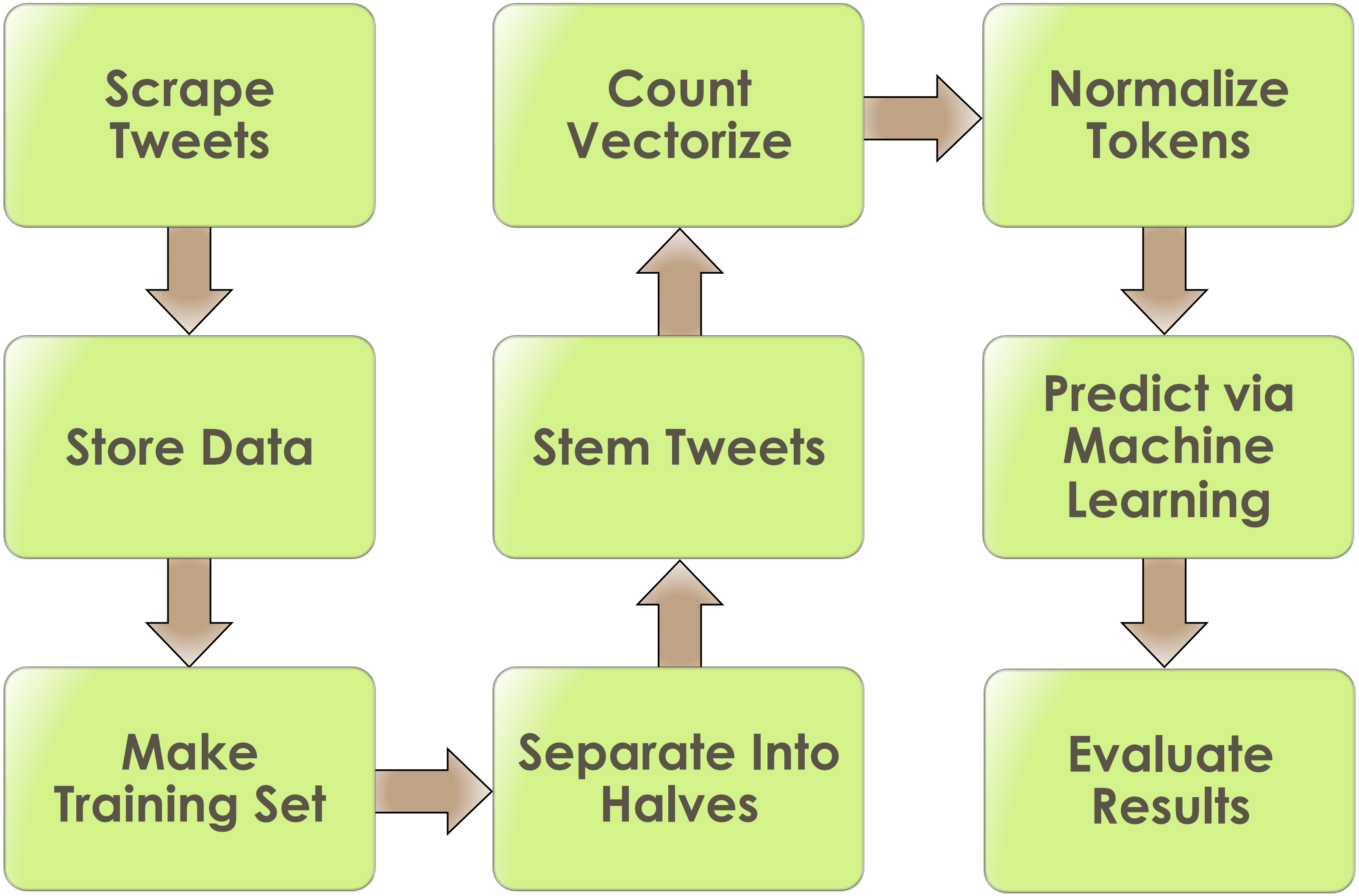
Data

Data consists of:

- Tweet ID
- Text
- Date
- Time
- Username
- Location*
- GPS Coordinates*

--> Identify outages

Workflow


Scrape Tweets
Store Data
Make Training Set
Separate Into Halves
Stem Tweets
Count Vectorize
Normalize Tokens
Predict via Machine Learning
Evaluate Results

Example Printout

Error Type	Actual	Predicted	Tweet Text
FP	[ ]	['L1']	@KenyaPower is there a reason you schedule black outs everyday in this area
FN	['L2']	[ ]	@kenyapower thank you very much lights are back kudos to your team..
FN	['L1']	[ ]	@KenyaPower Hi same story as yesterday and it's now total blackout.. Please attend
FP	[ ]	['L1']	@KenyaPower any scheduled maintenance in imara daima
FN	['L2']	[ ]	@kenyapower Thanks for the reply. Its ok though power came back. Your guys were on it quite fast. Well done.
FN	['L1']	[ ]	@KenyaPower the whole of riruta satellite.off naivasha rd.kinyanjui road.ngotho rd. plot is called jirani Centre.

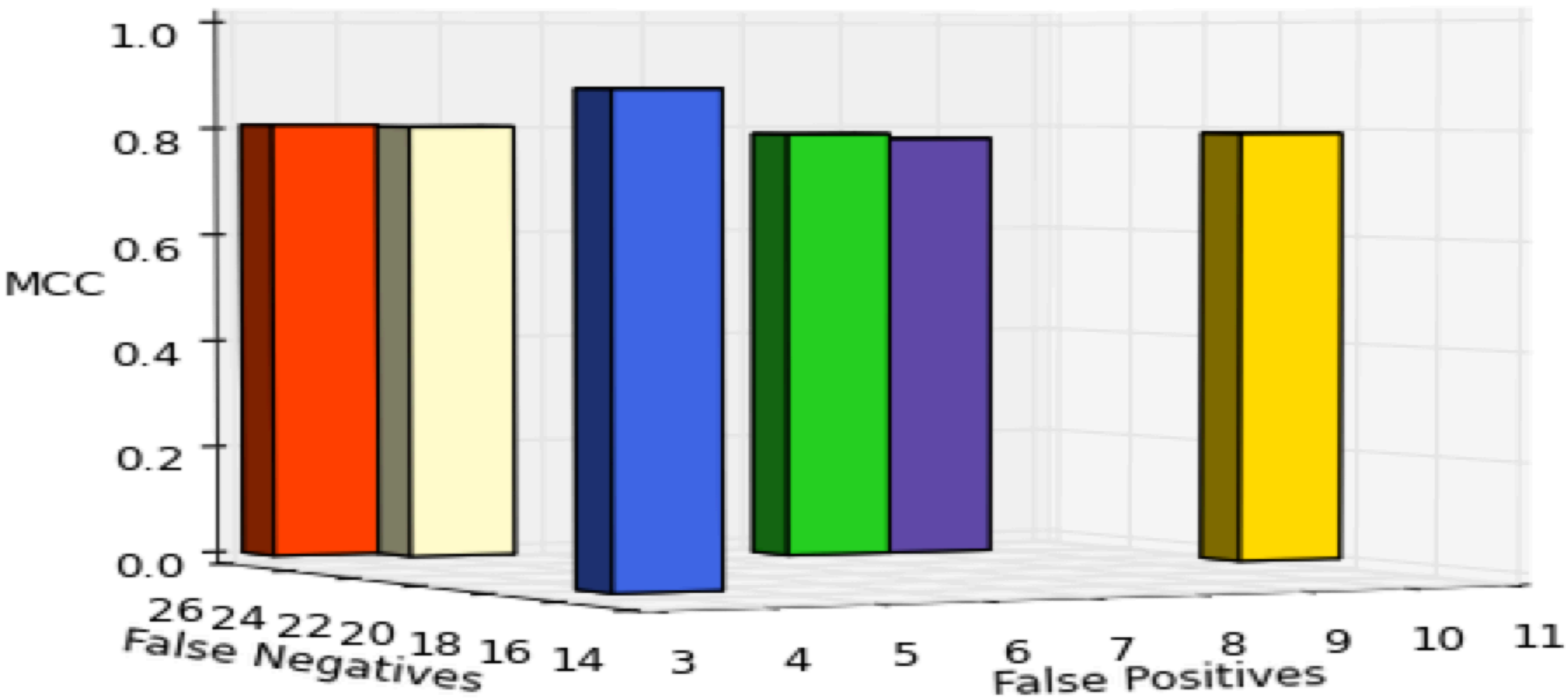
- Misclassified tweets pulled from SVM simulation.

Confusion Matrix

		Predicted condition			
	Total population	Predicted Condition positive	Predicted Condition negative	Prevalence $= \frac{\Sigma \text{ Condition positive}}{\Sigma \text{ Total population}}$	
True condition	condition positive	**True positive**	**False Negative** (Type II error)	True positive rate (TPR), Sensitivity, Recall $= \frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$	False negative rate (FNR), Miss rate $= \frac{\Sigma \text{ False negative}}{\Sigma \text{ Condition positive}}$
	condition negative	**False Positive** (Type I error)	**True negative**	False positive rate (FPR), Fall-out $= \frac{\Sigma \text{ False positive}}{\Sigma \text{ Condition negative}}$	True negative rate (TNR), Specificity (SPC) $= \frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$
	Accuracy (ACC) = $\frac{\Sigma \text{ True positive} + \Sigma \text{ True negative}}{\Sigma \text{ Total population}}$	Positive predictive value (PPV), Precision $= \frac{\Sigma \text{ True positive}}{\Sigma \text{ Test outcome positive}}$	False omission rate (FOR) $= \frac{\Sigma \text{ False negative}}{\Sigma \text{ Test outcome negative}}$	Positive likelihood ratio (LR+) $= \frac{\text{TPR}}{\text{FPR}}$	Diagnostic odds ratio (DOR) $= \frac{\text{LR+}}{\text{LR}-}$
		False discovery rate (FDR) $= \frac{\Sigma \text{ False positive}}{\Sigma \text{ Test outcome positive}}$	Negative predictive value (NPV) $= \frac{\Sigma \text{ True negative}}{\Sigma \text{ Test outcome negative}}$	Negative likelihood ratio (LR−) $= \frac{\text{FNR}}{\text{TNR}}$	

$$MCC = \frac{(TP * TN) - (FP * FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

Classifier Results

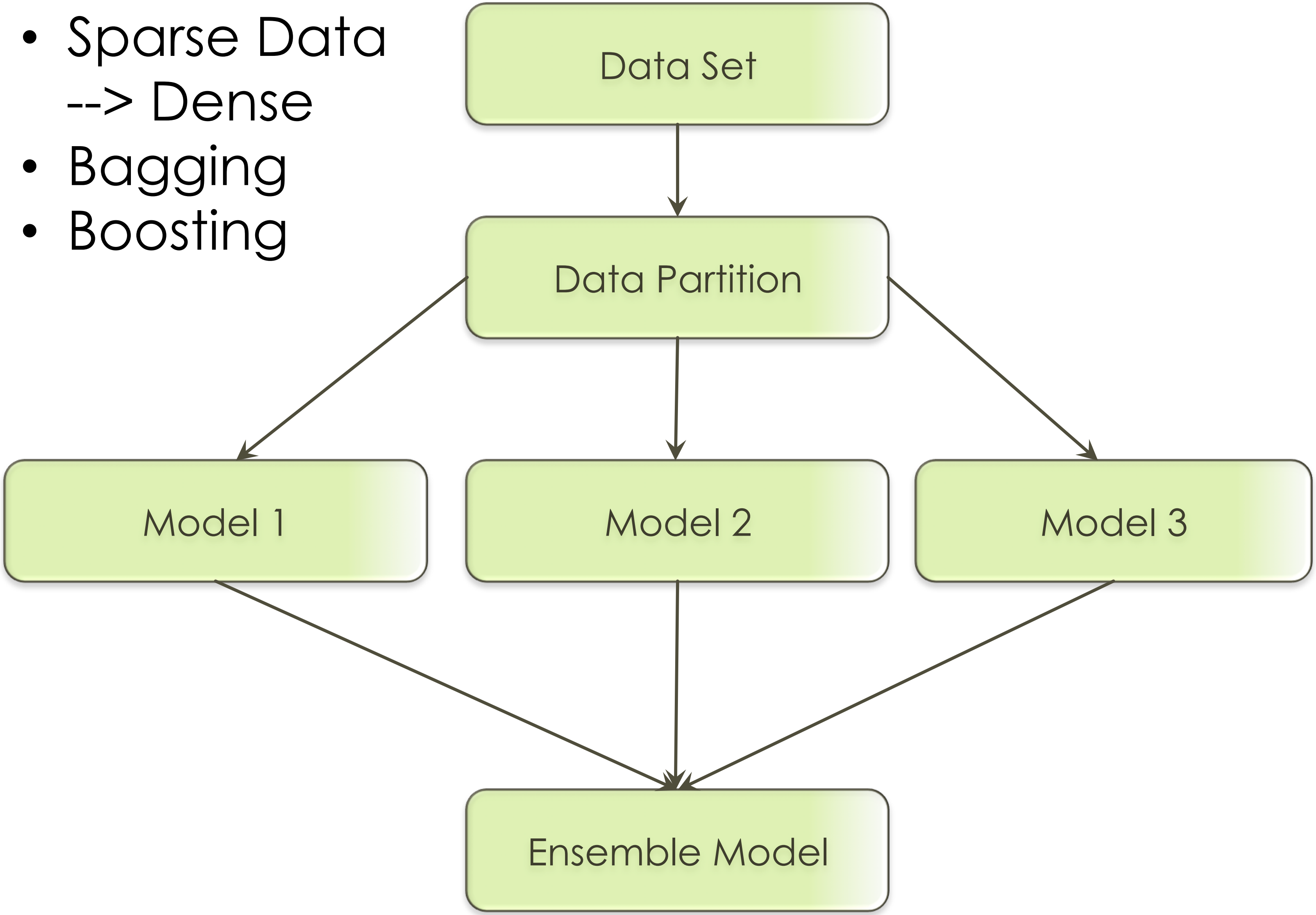


	RF	NB	SVM	KNeigh	MNB	LR
FP	3	4	3	7	8	10
FN	25	24	15	22	22	18
MCC	0.806	0.803	0.877	0.789	0.780	0.792

*** Test dataset size = 390 tweets**

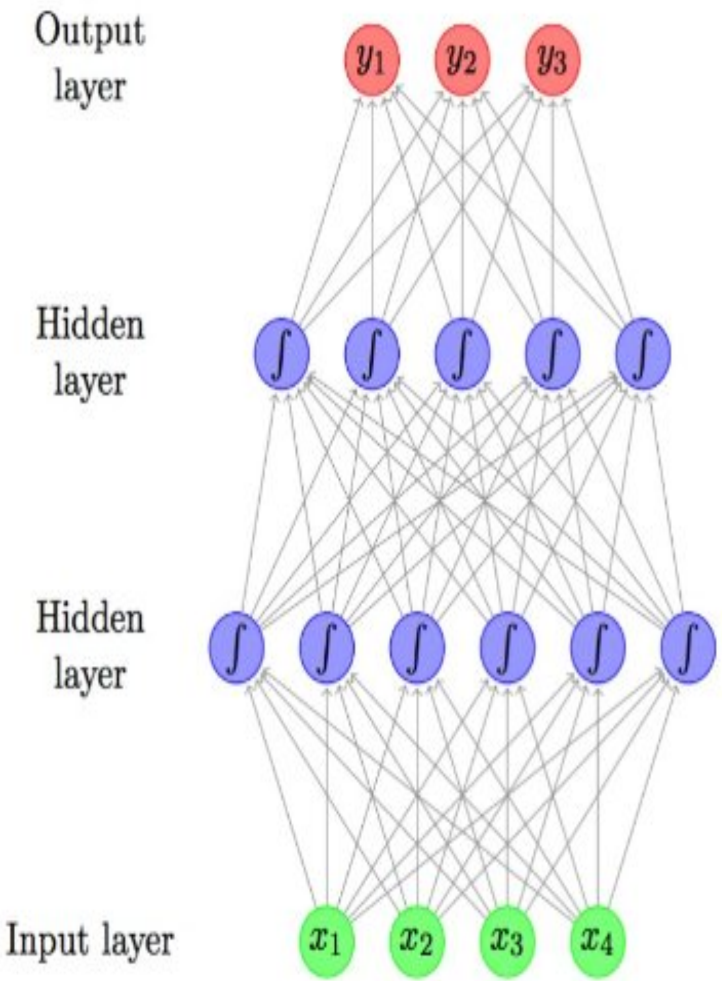
Ensemble Modeling

- Sparse Data --> Dense
- Bagging
- Boosting



Neural Network

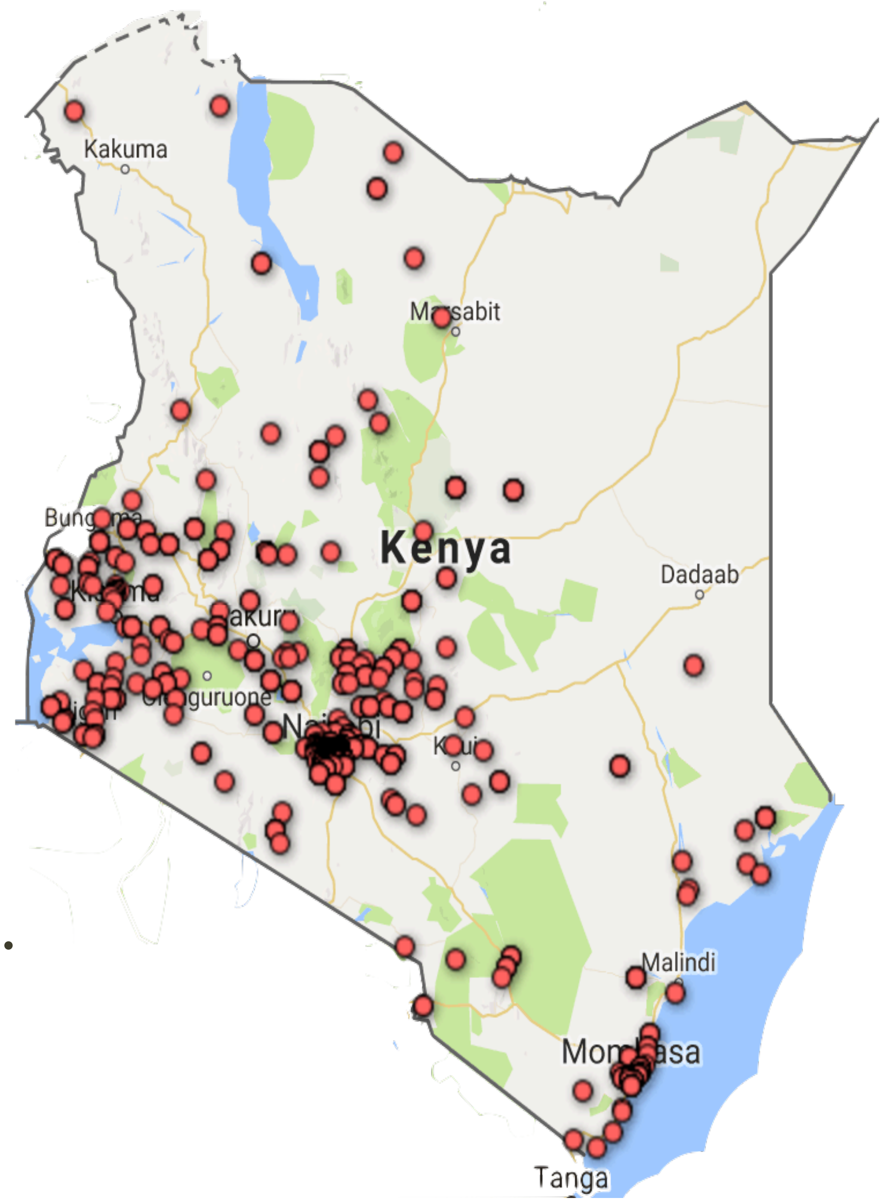
- Common technique in NLP
- Accuracy Increase
  ---> tradeoffs
- Initialize weights derived from linear model



Graphic from: http://u.cs.biu.ac.il/~yogo/nnlp.pdf

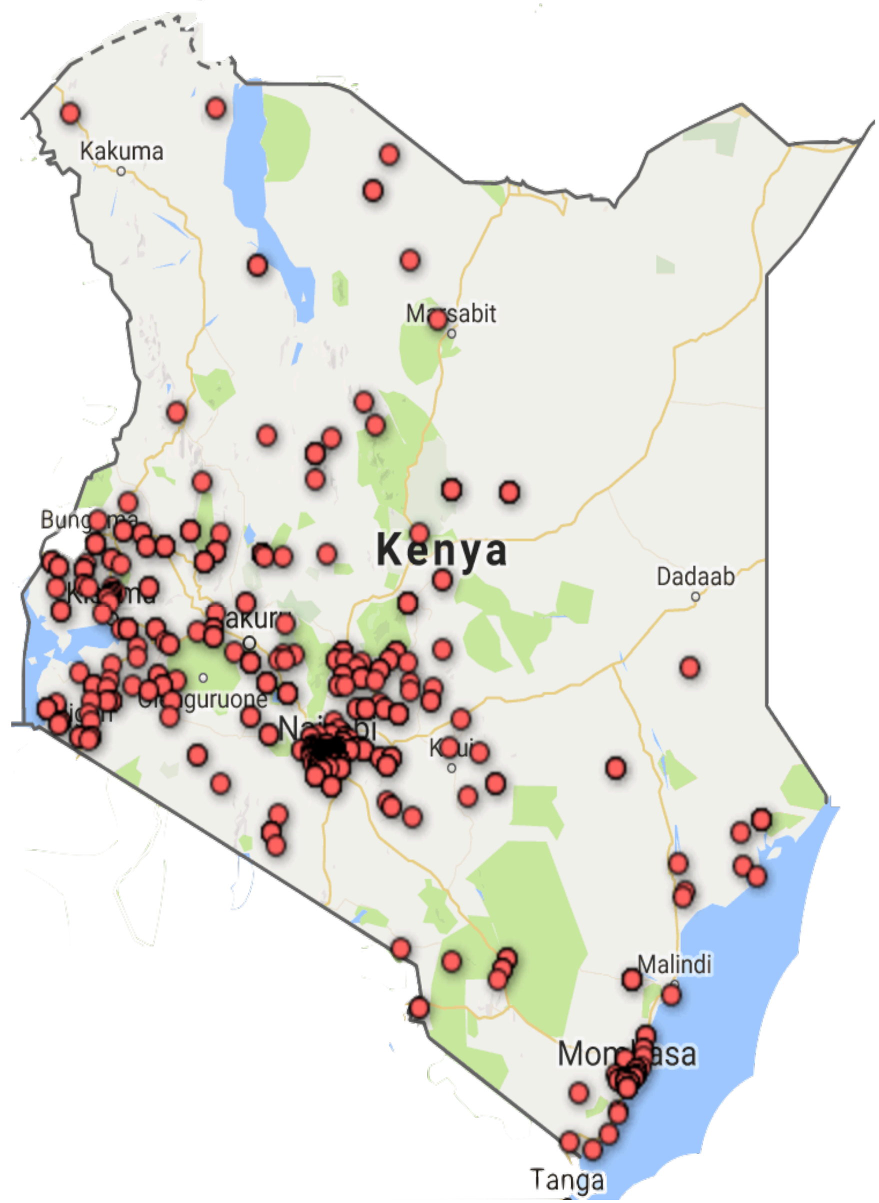
Considerations

- Not everyone has access to/uses Twitter.
- Classifiers are not perfect.
- Data are tweets.
- Not just one language.



Considerations

- Location bias.
- Time of day bias.
- Grid expansion and reliability over time.
- Data normalization is complicated.
- Privacy/security issues.



Moving Forward

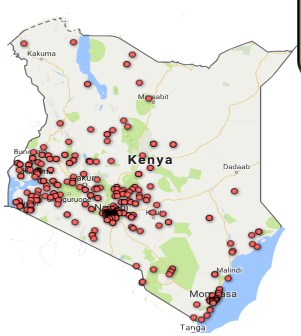
Ensemble Modeling, Neural Networks

Translation/labeling of Swahili Tweets

Improve Location Parsing

Normalize Results, Further Categorize

Classifier Implementation and Full Analysis in Google Datalab



Acknowledgements

- Dr. Eric Brewer
- Jordan and Javier
- Tiffany
- UC Berkeley

This research is funded by NSF Grant No. 1359499.

Any Questions?

Error Type	Actual	Predicted	Tweet Text
FP	[ ]	[‘L1’]	@KenyaPower is there a reason you schedule black outs everyday in this area
FN	[‘L2’]	[ ]	@kenyapower thank you very much lights are back kudos to your team..
FN	[‘L1’]	[ ]	@KenyaPower Hi same story as yesterday and it's now total blackout.. Please attend
FP	[ ]	[‘L1’]	@KenyaPower any scheduled maintenance in imara daima
FN	[‘L2’]	[ ]	@kenyapower Thanks for the reply. Its ok though power came back. Your guys were on it quite fast. Well done.
FN	[‘L1’]	[ ]	@KenyaPower the whole of riruta satellite.off naivasha rd.kinyanjui road.ngotho rd. plot is called jirani Centre.

- Misclassified tweets pulled from SVM simulation.

References

1. Brown, Peter F., Peter V. deSouza, and Robert L. Mercer. "Class-Based N-Gram Models Of Natural Language". Association for Computational Linguistics 18.4 (1992): 467-479. Web. 1 Aug. 2016.

2. Lee, Kenneth et al. "Electrification For “Under Grid” Households In Rural Kenya".Development Engineering 1 (2016): 26-35. Web.

3. "Milestones In Kenya’s Electricity Access | Kplc.Co.Ke". Kplc.co.ke. N.p., 2016. Web. 30 July 2016.

4. Mikolov, Tomas, Wen-tau Yih, and Geoffrey Zweig. "Linguistic Regularities In Continuous Space Word Representations". NAACL2013, Microsoft Research (2013): 1-6. Web. 27 July 2016.

5. Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.

6. R. Socher, Naive neural networks for very fast and accurate text classification, Private communication (2013).