Using Twitter Language to Predict the Real Estate Market

Mohammadzaman Zamani[1] **and Hansen Andrew Schwartz**[1]

[1]Computer Science Department, Stony Brook University
{mzamani,has}@cs.stonybrook.edu

Abstract

We explore whether social media can provide a window into community real estate — foreclosure rates and price changes — beyond that of traditional economic and demographic variables. We find language use in Twitter not only predicts real estate outcomes as well as traditional variables across counties, but that including Twitter language in traditional models leads to a significant improvement (e.g. from Pearson $r = .50$ to $r = .59$ for price changes). We overcome the challenge of the relative sparsity and noise in Twitter language variables by showing that training on the residual error of the traditional models leads to more accurate overall assessments. Finally, we discover that it is Twitter language related to business (e.g. 'company', 'marketing') and technology (e.g. 'technology', 'internet'), among others, that yield predictive power over economics.

1 Introduction

The massive amount of text provided by users of social media like Facebook and Twitter give researchers the opportunity to investigate topics that were not previously tangible. Specifically, the study of economic outcomes has been turning to the use of social media data in order capture non-traditional factors like consumer mood. For instance, researchers have attempted to predict the stock market by measuring mood from twitter feeds (Bollen et al., 2011), used Twitter data to measure socio-economic indicators and financial markets (Mao, 2015), shown correlation of consumer confidence with sentiment word frequencies in twitter messages over time (O'Connor et al., 2010), and predicted movie revenue using social media and text mining (Asur and Huberman, 2010; Joshi et al., 2010; Yu et al., 2012).

Here, we attempt to leverage social media to understand another economic phenomena, real estate. Our goal is to determine whether language from Twitter can predict real-estate foreclosure rates and price changes, cross-sectionally across counties, beyond that of traditional economic variables. We suspect this is possible because a community's language in social media may capture economic-related community characteristics that are not otherwise easily available. However, the challenge is incorporating noisy high-dimensional language features in such a way that they can contribute beyond the robust low-dimensional traditional predictors (i.e. demographics, median income, education rates, unemployment rates).

The contributions of this paper follow. First, we show that county real estate market outcomes can be predicted from language in social media beyond traditional factors. Second, we address the challenge of effectively leveraging multi-modal feature types (i.e. socioeconomic variables, which are individually very predictive (Nguyen, 2016); and social media linguistic features, which are individually noisy) by demonstrating that a 2-step *residualized control approach* to learning a predictive model leads to more accuracy than jointly learning all feature parameters at once. This represents the first work to investigate the use of language in Twitter to predict real estate related outcomes – foreclosure and increased price rates.

2 Related Work

Much of the research on prediction of housing markets has focused on economic conditions. For instance, others have found strong relationships between housing prices and the stock market(Gyourko and Keim, 1992; Case et al., 2005), credit and income (Ortalo-Magne and Rady, 2006), past market prices (Ghysels et al.,

Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, pages 28–33,
Valencia, Spain, April 3-7, 2017. ©2017 Association for Computational Linguistics

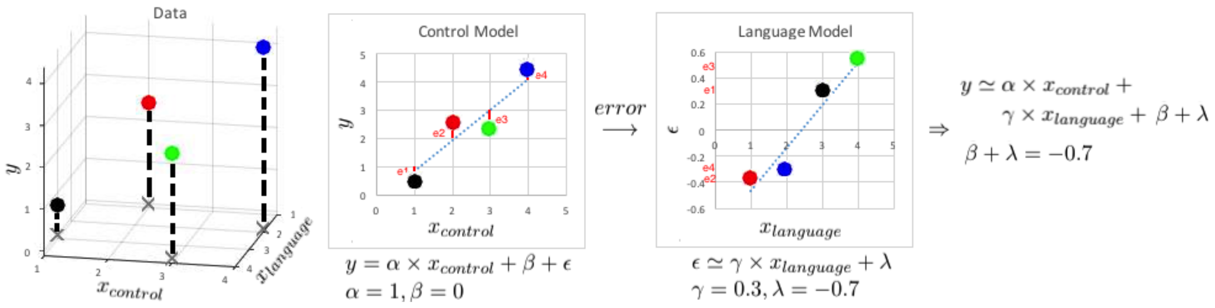

Figure 1: Procedure of building language model over the residual error of the control model.

2012; Tse, 1997), and market sentiment (i.e. from surveys) (Hui and Wang, 2014).

Except Kaplanski et al. (2012), who looked at daylight hours, few have ventured beyond direct economic factors as predictors of real estate outcomes. Our belief is that language analyses in social media can offer predictive value beyond that of economics in that they capture aspects of people's daily life that are not traditionally available to economists.

While exploiting social media language has not been studied in the real estate domain, use of language predictors has been increasing for other economic-related applications, like measuring the public health using analysis of messages in social media (Paul and Dredze, 2011; Eichstaedt et al., 2015; Culotta, 2014), in addition to predicting stock market exploiting text in social media (Bollen et al., 2011; Zhang et al., 2011; Tsolacos, 2012), and predicting political behaviour considering tweets (DiGrazia et al., 2013). Perhaps the most similar work to ours used manually selected keywords in Google searches to predict the overall US housing market (Wu and Brynjolfsson, 2013). Still, while Google has allowed researchers to tap into one aspect of the online world, search data is only available for specific scales and relying on manually-chosen keywords can restrict predictive performance (Schwartz et al., 2013). We leverage open-vocabulary features (i.e. not based on manual keyword lists) and attempt to predict real-estate at the level of US counties.

3 Language Model

We learn a model from the Twitter language of US counties to predict real estate outcomes. We extract community language features from tweets and then we learn models for the cross-county prediction task, handling both traditional predictors and linguistic predictors. We focus on two outcomes per county, foreclosure and increased price rates (zillow website, 2016), and consider a wide variety of traditional socioeconomic and demographic predictors to compare. Specifically, *socioeconomic* variables include median income, unemployment rate and percentage with bachelors degrees while *demographic* variables include median age; percentage: female, black, hispanic, foreign born, married; and population density. All variables were obtained from US Census (census bureau, 2010), and we henceforth refer to them as a whole as *controls*.

3.1 Features

We build feature vectors from the raw tweets by extracting 1, 2, and 3-grams as well as mentions of 2000 LDA topics based on posteriors we downloaded which were previously estimated from social media (Schwartz et al., 2013). Features were limited to those mentioned by at least 25% of counties, leaving us with $13,359$ 1to3-grams and all $2,000$ topics.

Since there are only $1,347$ counties, to which we plan to apply the model (data described in evaluation) but tens of thousands of predictors, we utilize feature selection and dimensional reduction to avoid overfitting. We limit ourselves to features with at least a small linear relationship to the outcome, having a family-wise error *alpha* of 200 (Efron, 2012). Then, we perform randomized principal components analysis (RPCA) , an approximate PCA based on stochastic re-sampling(Rokhlin et al., 2009), which in effect combines co-varying features and leaves a more reasonable number of parameters to estimate during learning.[1]

[1] Since the topic features are already a combination of n-grams, they are less sparse and presumably less noisy. Thus, we apply the feature selection and dimensionality reduction steps for n-grams and topics independently, keeping 90 dimensions of topics and 45 dimensions of n-grams.

	socioeconomics		demographics		socioeconomics + demographics	
	Fc	Ip	Fc	Ip	Fc	Ip
no lang	0.34	0.42	0.24	0.44	0.37	0.50
with lang (*residualized control*)	**0.41**	**0.56**	**0.39**	**0.57**	**0.42**	**0.59**

Table 1: Comparing the Pearson r of adding language model over the residual of the control model vs. control model for 'foreclosure' and 'increased price' rates. Fc stands for foreclosure rate and Ip is increased-price rate. **bold** indicates significant improvement ($p < 0.05$) over no language.

3.2 Learning

We learn four different models: (1) a *control model* using the socioeconomic & demographic variables, (2) a *language model* using only tweet-derived features, (3) a combined model using both socioeconomics & demographics and language in a single model, and (4) a language over *residualized control* model fitting language to the residual error of the control model. With the *control model* as our baseline, we investigate whether language alone (model 2) or adding language to the control model (models 3 and 4) increases accuracy. All models except the 4th are learned via $L2$ penalized ("ridge") regression (Goeman et al., 2016).[2]

Residualized Control Approach In order to effectively exploit Twitter language in our model, we suspect that we need to treat the language features (which are numerous, noisy, more biased, and non-normal) differently than the control variables (which are few, mostly unbiased, and mostly normal). In other words, simply combining the two may lead to losing the importance of the controls amongst the numerous features.[3] As depicted in Figure 1, we build a language model over the residual error of the control model, allowing independent consideration of the two sets of features and different penalties. More specifically, the training phase consists of three steps: (1) train a model using the socioeconomics & demographics, which is the control model, as in Eq.1, (2) calculate the training errors and consider this error as our new label, described in Eq.2, and (3) train a language model over this new data, which is shown in Eq.3. In the end, our model is depicted in Eq.4. In these equations α and γ are the coefficient of control features and language features, and β and λ are the interceptions. For testing purpose we feed each data to both control model and language model, and then report the summation of their predictions as the final predicted label.

$$\hat{y} = \alpha \times X_{control} + \beta \quad (1)$$

$$\epsilon = y - \hat{y} \quad (2)$$

$$\epsilon \simeq \gamma \times X_{language} + \lambda \quad (3)$$

$$\Rightarrow y \simeq \alpha \times X_{control} + \gamma \times X_{language} + (\lambda + \beta) \quad (4)$$

The resulting model, a combination of the control model and language model, is still an affine model w.r.t. the language and control features. Thus, its possible ridge-regression over all the features at once could give us the same result (i.e. hyperplane). However, since we suspect that each socioeconomic and demographic feature are more informative and less noisy than the Twitter features, we explore this two-stage learning procedure in order to bias our model toward favoring the role of socioeconomics & demographics over language features.

4 Evaluation

Here we evaluate the power of Twitter language to predict cross-county real-estate outcomes compared to demographic and socioeconomic factors.

4.1 Data Set

We are using 3 different sources of data: a *language dataset* from Twitter messages, a *control dataset* of socioeconomic and demographic variables, and an *outcome dataset* of housing related data. Our language data was derived from Twitter's 1% random stream collected from 2011 to 2013 and included 131 million tweets that are mapped to $1,347$ counties based on their self-reported location following the procedure of Eichstaedt et al. (2015). Our control data included the previously mentioned *socioeconomic* and *demographic* variables which were obtained from 2010 US Census data (census bureau, 2010). This

[2] For the control model, which has few features by comparison, the ridge penalty is essentially zero and standard multivariate linear regression produces comparable results

[3] In fact, our results show such a combined model performs only marginally better than a language alone model.

	Foreclosure	Increased-price
language	0.38	0.48
combined	0.40	0.49
residualized control	**0.42**	**0.59**

Table 2: Comparing the Pearson r of building language model over residual of control model vs. combining the language and the control features into a single model. **bold** indicates significant improvement ($p < 0.05$) over combined model.

dataset is only collected every 10 years, so the 2010 US Census is the most recent dataset for all of the *socioeconomic* and *demographic* variables at the county level.

As outcomes, our real estate data, including the foreclosure rate (the number of homes (per 10,000 homes sold) that were foreclosed) and increased-price rate (the percentage of homes with values that have increased in the past year) were downloaded from Zillow and covering 2011 to 2013 (zillow website, 2016). Considering all these data sets, we end up with 427 counties having foreclosure rate outcome data, and 717 counties having increase price rate data.[4].

4.2 Results

Table 1 reports the effect of building a language model over the residual of socioeconomics, demographics, and socioeconomics & demographics by comparing them with the control models. All of the results were produced by 10 fold cross-validation. We see a significant improvement of exploiting language ($p < 0.05$ according to paired t-test) above and beyond socioeconomic and demographic factors for both the outcomes of foreclosures (from $r = .37$ to $r = .42$) and increased price (from $r = .50$ to $r = .59$). This suggests that language on Twitter does, in fact, capture information about a community that is not captured by the traditional predictors.

We next explored whether building language model using the *residualized control approach* performs better than a model combining control and language features in a single learning step. Results are in Table 2, showing that building language model over residual performs significantly better than a combined model for both of the outcomes. In fact, the gap is .10 in Pearson r for increased price. Further, it also appears possible that the combined feature model could perform worse than the control model in some cases, presumably because the controls are lost when being fit with the language. In a sense, the *residualized control* approach utilizes a prior that each socioeconomic and demographic feature are more informative than a single word and should thus receive a different penalty parameter or be fit independently. It worth noting that this method is applicable for many different learning algorithms (e.g. SVM, deep convolutional net).

As mentioned previously, one limitation of the traditional predictors is that many are only available every 10 years as part of the US Census. We primarily focused on Twitter data that was a couple years removed from the last census, which may explain the improvement. Thus, we also ran an experiment using the Twitter data from (Schwartz et al., 2013) which spans 2009 to 2010, and found similar results: the *residualized control* approach improved the Pearson r for ‘increased price‘ from 0.36 to 0.44 and for ‘foreclosure‘ from 0.65 to 0.69. Thus, the improvements provided by the residualized control approach do not appear to be due to the fact that twitter data are newer than control data.

We have shown that Twitter language is adding predictive information about the real estate market beyond that of traditional socioeconomic predictors. So, just what exactly are tweets capturing that socioeconomics are not? Toward this, we ran a differential language analysis to identify the top 50 most predictive features (independently) of increased price, the outcome which we performed the best. Figure 2 shows the results controlled by socioeconomic and location features (US state indicator), limited to those passing a Benjamini-Hochberg False Discovery rate *alpha* of 0.01 (Benjamini and Hochberg, 1995). We see that, although each displayed n-gram was predictive beyond socioeconomics, many of them suggest a more nuanced economic characterization of a community (e.g. ‘technology’, ‘media’, ‘internet’, and ‘marketing’), suggesting avenues of future exploration for better understanding the housing market.

[4] The control and real estate datasets can be found here: `http://www3.cs.stonybrook.edu/~mzamani/datasets/eacl2017/`

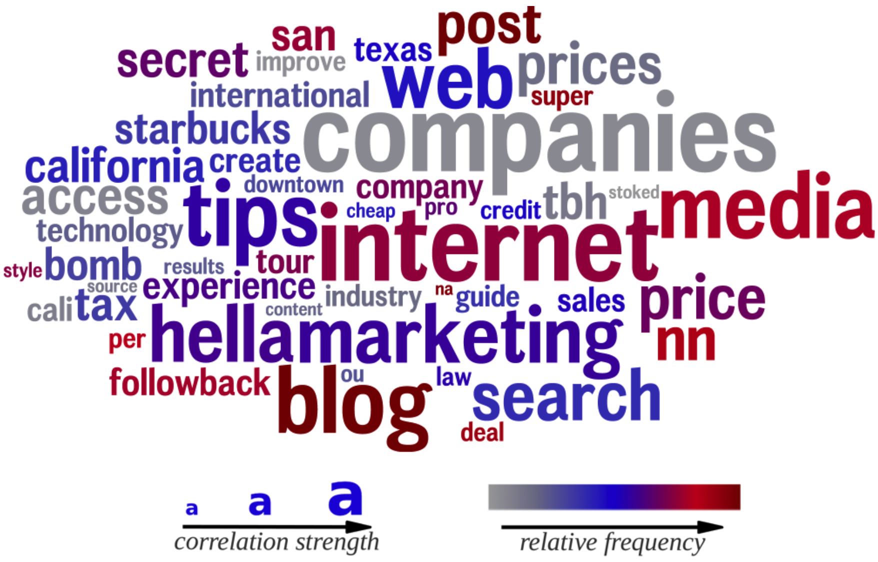

Figure 2: N-grams most predictive of 'Increased price rate' controlled by socioeconomics and location.

5 Conclusion

While the real estate market of a community is believed to be affected by many factors, traditionally only coarse economic and demographic variables have been accessible at scale to market researchers and forecasters. Here, we explored the prediction power of language in the real estate market as compared to traditional predictors, showing that language in twitter is predictive of foreclosure rates and price increases and that a *residualized control* approach to combine language features with traditional variables can lead to more accurate models. We believe this can open the door to more a nuanced and precise understanding of the real-estate market.

Acknowledgements

This work was supported, in part, by the National Science Foundation through I/UCRC CGI: Center for Dynamic Data Analytics (CDDA) (IIP 1069147)."

References

Sitaram Asur and Bernardo A. Huberman. 2010. Predicting the future with social media. In *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on*, volume 1, pages 492–499. IEEE.

Yoav Benjamini and Yosef Hochberg. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)*, pages 289–300.

Johan Bollen, Huina Mao, and Xiaojun Zeng. 2011. Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1–8.

Karl E. Case, John M. Quigley, and Robert J. Shiller. 2005. Comparing wealth effects: the stock market versus the housing market. *Advances in macroeconomics*, 5(1).

US census bureau. 2010. Profile of general population and housing characteristics: 2010 demographic profile data. `https://factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml?pid=DEC_10_DP_DPDP1&prodType=table`.

Aron Culotta. 2014. Estimating county health statistics with twitter. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems*, pages 1335–1344. ACM.

Joseph DiGrazia, Karissa McKelvey, Johan Bollen, and Fabio Rojas. 2013. More tweets, more votes: Social media as a quantitative indicator of political behavior. *PloS one*, 8(11):e79449.

Bradley Efron. 2012. *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction*, volume 1. Cambridge University Press.

Johannes C. Eichstaedt, H. Andrew Schwartz, Margaret L. Kern, Gregory Park, Darwin R. Labarthe, Raina M. Merchant, Sneha Jha, Megha Agrawal, Lukasz A. Dziurzynski, Maarten Sap, et al. 2015. Psychological language on twitter predicts county-level heart disease mortality. *Psychological science*, 26(2):159–169.

Eric Ghysels, Alberto Plazzi, Walter N. Torous, and Rossen I. Valkanov. 2012. Forecasting real estate prices. *Handbook of economic forecasting*, 2.

Jelle Goeman, Rosa Meijer, and Nimisha Chaturvedi. 2016. L1 and l2 penalized regression models.

Joseph Gyourko and Donald B. Keim. 1992. What does the stock market tell us about real estate returns? *Real Estate Economics*, 20(3):457–485.

Eddie Chi-man Hui and Ziyou Wang. 2014. Market sentiment in private housing market. *Habitat International*, 44:375–385.

Mahesh Joshi, Dipanjan Das, Kevin Gimpel, and Noah A. Smith. 2010. Movie reviews and revenues: An experiment in text regression. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 293–296. Association for Computational Linguistics.

Guy Kaplanski and Haim Levy. 2012. Real estate prices: An international study of seasonality's sentiment effect. *Journal of Empirical Finance*, 19(1):123–146.

Huina Mao. 2015. Socioeconomic indicators. *Twitter: A Digital Socioscope*, page 75.

Joseph Nguyen. 2016. 4 factors that influence real estate. "http://www.investopedia.com/articles/mortages-real-estate/11/factors-affecting-real-estate-market.asp, [Accessed: 2016-11-10]".

Brendan O'Connor, Ramnath Balasubramanyan, Bryan R. Routledge, and Noah A. Smith. 2010. From tweets to polls: Linking text sentiment to public opinion time series. *ICWSM*, 11(122-129):1–2.

Francois Ortalo-Magne and Sven Rady. 2006. Housing market dynamics: On the contribution of income shocks and credit constraints. *The Review of Economic Studies*, 73(2):459–485.

Michael J. Paul and Mark Dredze. 2011. You are what you tweet: Analyzing twitter for public health. *ICWSM*, 20:265–272.

Vladimir Rokhlin, Arthur Szlam, and Mark Tygert. 2009. A randomized algorithm for principal component analysis. *SIAM Journal on Matrix Analysis and Applications*, 31(3):1100–1124.

H. Andrew Schwartz, Johannes C. Eichstaedt, Margaret L. Kern, Lukasz Dziurzynski, Stephanie M. Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin E. P. Seligman, et al. 2013. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one*, 8(9):e73791.

Raymond Y. C. Tse. 1997. An application of the arima model to real-estate prices in hong kong. *Journal of Property Finance*, 8(2):152–163.

Sotiris Tsolacos. 2012. The role of sentiment indicators for real estate market forecasting. *Journal of European Real Estate Research*, 5(2):109–120.

Lynn Wu and Erik Brynjolfsson. 2013. The future of prediction: How google searches foreshadow housing prices and sales. *Available at SSRN 2022293*.

Xiaohui Yu, Yang Liu, Xiangji Huang, and Aijun An. 2012. Mining online reviews for predicting sales performance: A case study in the movie domain. *IEEE Transactions on Knowledge and Data engineering*, 24(4):720–734.

Xue Zhang, Hauke Fuehres, and Peter A. Gloor. 2011. Predicting stock market indicators through twitter "i hope it is not as bad as i fear". *Procedia-Social and Behavioral Sciences*, 26:55–62.

zillow website. 2016. zillow datasets. "`http://www.zillow.com/research/data/` [Accessed: 2016-11-10]".