Toronto's Evolution of Attitudes Towards Real Estate

David Veitch
University of Toronto
Department of Statistical Sciences
david.veitch@mail.utoronto.ca

December 19 2018

1 Introduction

For many Toronto residents, buying a house represents the best financial decision they have ever made. Real house prices in Toronto roughly tripled from 1985 to 2018, and did not suffer a meaningful correction during the 2008 financial crisis [6]. Given residential real estate is an asset class widely held by the general public, it is frequently written about in the press, with many newspapers devoting entire sections to it on certain days of the week. This report looks to measure the evolution of Toronto's attitude towards real estate via applying natural language processing techniques to newspaper articles.

2 Data

The Toronto Star is one of the most widely circulated newspapers in Toronto and across Canada [1], and has been in print since 1892. To obtain the data used in this report, a search was conducted on Proquest's Canadian Newsstream database for articles from the Toronto Star with the following keywords: housing affordability, home affordability, home prices, real estate prices, housing costs, house prices, housing bubble, real estate bubble, real estate market, housing market, home sales, housing sales, homebuyers, and home-buying. The metadata associated with 8942 of the most relevant articles (as determined by Proquest's search engine) was obtained, spanning a period from 1985-2017. Fields in the metadata of particular interest were article title and article abstract.

Sample Article	
Title	House prices up moderately, survey finds
Date	Jul 20, 1985
Publication	Toronto Star
Abstract	For this type of home, prices in April ranged from an average of $86,000 in Burlington and Newmarket to $145,000 in the Islington-Kingsway area of Etobicoke. Prices for the two-storey homes ranged from $116,000 in Newmarket to $245,000 in central Toronto. In Brantford, for example, a detached two-storey house jumped 21 per cent to $115,000 from $95,000. In Sudbury, a bungalow went to $61,500 from $51,000, an increase of 20.5 per cent.

3 Analysis

3.1 Overview

Two lines of analysis were conducted to measure the evolution of attitudes towards real estate found in these articles. The first line of analysis applied probabilistic topic models [2] to extract topics from the corpus (the collection of documents), and measure their evolution over time. This is a form of unsupervised learning which only necessitates specifying the number of topics in advance. The second line of analysis applied sentiment analysis tools to measure the sentiment embedded in the corpus over time.

3.2 Modelling Topics

A probabilistic topic model was used to determine the distribution of topics in each article's abstract. Specifically a latent Dirichlet allocation (LDA) was fit to the data. A LDA model is generative model where the documents observed are assumed to arise from a process with both observed (words) and hidden (the topic structure) random variables. The generation of the observed texts can be described by the following equation:

$$P(\beta_{1:K}, \theta_{1:D}, \boldsymbol{z_{1:D}}, \boldsymbol{w_{1:D}}; \eta, \alpha) = \prod_{i=1}^{K} P(\beta_i; \eta) \prod_{d=1}^{D} P(\theta_d; \alpha) \prod_{n=1}^{N} P(z_{d,n} \mid \theta_j) P(w_{d,n} \mid z_{d,n}, \beta_{1:K}) \tag{1}$$

$$\begin{aligned}
\beta_{1:K} &\text{ where } \beta_k \sim \text{Dirichlet}(\eta),\ \text{ a given topic k's distribution over words} \\
\theta_{1:D} &\text{ where } \theta_d \sim \text{Dirichlet}(\alpha),\ \text{ a document d's distribution over topics} \\
\boldsymbol{z_{1:D}} &\text{ where } \boldsymbol{z_d} \text{ is a vector of length } N \text{ (the number of words in a document)} \\
z_{d,n} &\text{ represents the topic assigned to word n in document d} \\
w_{d,n} &\text{ represents word n in document d}
\end{aligned}$$

Using the Gensim topic modelling package [7], which relies on a modification of a variational Bayes algorithm to learn $\theta_{1:D}, \beta_{1:K}$, and $\boldsymbol{z_{1:D}}$ [3], topics were extracted from the articles' abstracts. Five topic models were fit for values of K (the number of topics) of 5, 7, 10, 15, and 20. Seven topics appeared to give a unique set of ideas with minimal overlap. From the words which appear frequently in each topic, the topics were named as follows:

Extracted Topics				
Generic	**Home Sales**	**Renting**	**Corporate**	**Builders**
metro	sales	metro	lots	builders
million	metro	room	company	mortgage
mortgage	price	controls	office	ontario
province	prices	rent	room	builder
pay	increase	million	development	program
value	million	living	lot	warranty
government	mortgage	ontario	community	association

Economic	**Government Policy**
rates	tax
prices	government
bank	ontario
interest	land
inflation	provincial
rate	social
economy	million

3.2.1 Example

To see how LDA assigns topics to articles we can observe topics assigned to two example articles:

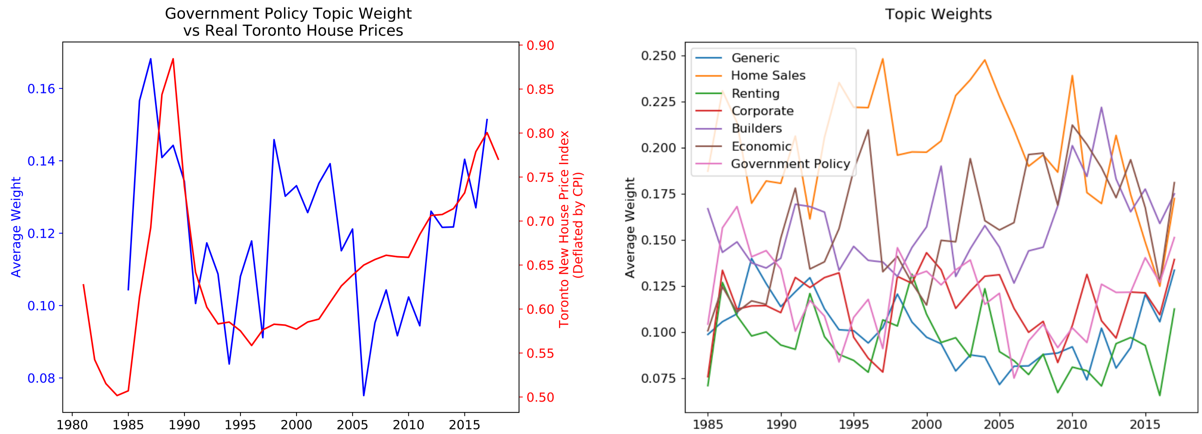
Example 1	
Title	Property tax debates are political reality
Date	Jan 5, 2017
Abstract	"By adopting these measures, city council would avoid significant property tax hikes, and, as we all know, property tax is regressive and has a significant impact on seniors," then-mayor David Miller said in September 2007, launching a campaign to sell Toronto residents on the idea of a "fair tax plan for Toronto" that included "revenue tools," including a vehicle registration tax and a land transfer tax.
Key Topics	Government Policy 97.6%

Example 2	
Title	Toronto home sales off to a strong start; Board eyes 6,000 sales for month Average price up 7% to $332,000
Date	Feb 19, 2005
Abstract	Up to Feb. 15, 2,924 homes were sold, a 14 per cent increase over the same period last year, said Toronto Real Estate Board president Ron Abraham. Areas that have experienced a surge in activity include Davisville, Willowdale and York Mills. Many parts of central Toronto including the downtown, Rosedale and Lawrence Manor have had many sales.
Key Topics	Home Sales 90.3%, Government Policy 6.7%

3.3 Topic Trends

Using the topic weights assigned to each article (for each article the topic weights sum to one) we can observe how certain topics have trended over time. This is done by averaging over the topic weights of a given topic for articles in a given year (note that each year y has n_y articles).

$$TopicWeight(Topic = t, Year = y) = \frac{1}{n_y} \sum_{i=1}^{n_y} TopicWeight(Topic = t, Year = y, Article = i) \tag{2}$$



(a) Government Policy Topic vs. House Prices

(b) All Topics

Figure 1

The most interesting thing to observe in Figure 1 regarding topic weights is the way in which the average weight of the government policy topic has increased alongside housing prices. This makes intuitive sense since as house prices increase one would expect a demand from the public for the government to provide affordable housing, or as in Example 1, an increased possibility that governments increase property taxes.

3.4 Sentiment Analysis

A basic sentiment analysis algorithm was run on the articles' abstracts to model the overall sentiment of different years. The Python package NLTK [5] was used to perform the analysis.

An article's sentiment is evaluated based on the number of positive words, negative words, and neutral words in the abstract. To assess whether a given word is positive or negative the Hiu Lu Opinion Lexicon [4], a widely used lexicon for sentiment analysis, was used. Each word in an abstract was checked against a list of approximately 7,000 words in the lexicon to determine its sentiment. Also, any words following a negation were assigned the opposite sentiment of the underlying word.

3.4.1 Example Sentence

Example Sentence												
Sentence	The	market	is	great,	fantastic	even.	Not	a	bad	time	to	buy!
With Negation	The	market	is	great,	fantastic	even.	Not	a_NEG	bad_NEG	time_NEG	to_NEG	buy!_NEG
Cleaned	the	market	is	great	fantastic	even	not	a_NEG	bad_NEG	time_NEG	to_NEG	buy_NEG
Sentiment				Positive	Positive				Positive			

3.4.2 Illustrative Results

To demonstrate the ability for the sentiment analysis algorithm to pickup the underlying sentiment of articles we can observe how articles from 1988 (the middle of the 1980s boom in house prices) are classified compared to articles from 1995 (after a house price bust), and 2015 (the middle of another boom).

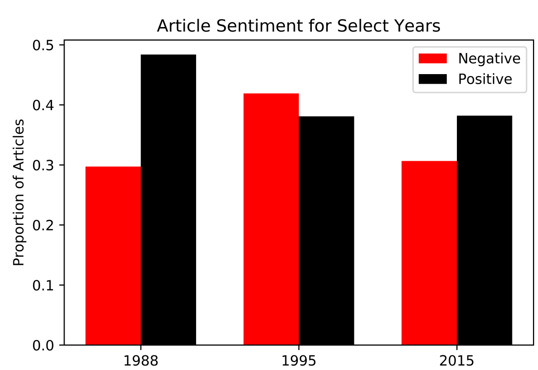


Figure 2

Figure 2 shows what percentage of articles in these years were classified as positive or negative (an article is classified as positive if its abstract contains more positive than negative words and vice-versa). Unsurprisingly sentiment in 1988 and 2015 was on balance positive, whereas 1995 was skewed negative. What is notable is the extent of the enthusiasm in 1988; this is somewhat unsurprising given 1988 was a banner year for real estate. In this year the Toronto New Housing Price Index rose 22% in real terms as opposed to a more modest gain of 2% in 2015.

4 Conclusion & Future Directions

From the above analysis it is evident that topic modelling and sentiment analysis tools can use newspaper articles to evaluate changing attitudes towards real estate. This analysis is a good first step, but there are many ways this analysis can be extended, particularly with respect to sentiment analysis.

Notably, a future avenue for further research is around a real-estate specific sentiment lexicon. Most publicly available sentiment lexicons are based upon text corpora from online reviews (e.g. movies, or online shopping). However, many words relevant to real estate do not register on these lexicons. The most obvious example would be the words 'bull' or 'bear', common terms for good and bad markets respectively. These words do not register as either positive or negative in most lexicons. Building a more robust lexicon for real-estate would improve the accuracy of these sentiment analysis algorithms. This task could potentially be done via supervised machine learning techniques where an individual could classify certain articles within a real-estate related corpus as positive or negative and a machine learning algorithm could learn the sentiment of words appearing in these, and other similar articles, but not in common lexicons.

References

[1] Canada's Top 20 Daily Newspapers, 2015. http://www.cision.ca/trends/canadas-top-20-daily-newspapers/.

[2] David Blei and John Lafferty. Topic Models. http://www.cs.columbia.edu/~blei/papers/BleiLafferty2009.pdf, 2009.

[3] Matthew Hoffman, Francis R. Bach, and David M. Blei. Online learning for latent dirichlet allocation. In J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems 23*, pages 856–864. Curran Associates, Inc., 2010.

[4] Minqing Hu and Bing Liu. Mining and summarizing customer reviews. In Won Kim, Ron Kohavi, Johannes Gehrke, and William DuMouchel, editors, *KDD*, pages 168–177. ACM, 2004.

[5] Edward Loper and Steven Bird. Nltk: The natural language toolkit. In *In Proceedings of the ACL Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics. Philadelphia: Association for Computational Linguistics*, 2002.

[6] Canada Mortgage and Housing Corporation. Are ontario home prices in a corrective or price bust phase?, June 2018. http://publications.gc.ca/collections/collection_2018/schl-cmhc/nh12-295/NH12-295-2018-6-eng.pdf.

[7] Radim Řehůřek and Petr Sojka. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May 2010. ELRA. http://is.muni.cz/publication/884893/en.