Analyzing Listings for High-Risk Real Estate with Natural Language Processing

Anonymous Author(s)
Affiliation
Address
email

Abstract

Climate change has made some coastal real estate increasingly vulnerable to flooding. Home-sellers don't always communicate these risks, and laws in the United States are inconsistent in whether they mandate disclosures. We propose using natural language processing (NLP) to analyze real estate listing text for coastal properties in the United States. We expect this work to help policy-makers assess the effectiveness of current legislation around real estate disclosures and identify regions to target with future legislation. In particular, we propose a procedure for identifying words and phrases that are relevant to flooding risks, and a neural model to classify relevant listings by their claims around these risks.

1 Introduction

With climate change and sea level rise, many coastal properties are at increased risk for flooding. In Miami Beach, for example, Zillow Research predicts that 85.2% of homes could face yearly flooding by 2100, which accounts for more than 37 billion dollars of property [1]. Developers, however, continue to build in the Miami area, which has led the New York Times to write that they're playing a "game of hot potato" with real estate, "[trying] to hand off risky properties before getting burned" [2]. Some states have mandated that sellers communicate these risks, but legislation in the United States is patchwork. Home-sellers in Virginia, for instance, have no requirement to disclose flood risks [3] [4].

We propose using techniques from natural language processing (NLP) to analyze scraped listings from real estate websites like Zillow or Realtor, in order to help policy-makers better protect home-buyers from downstream effects of climate change. We expect our work to contribute in the following ways:

1. **Short-term, high-impact**: Assess the effectiveness of existing legislation around disclosures of flood risks
2. **Short-term, high-impact**: Identify high-risk regions where disclosures are uncommon
3. Compare listings in different regions to (1) pinpoint where buyers are anxious about flooding risks and/or (2) better understand how real estate markets respond to natural disasters

There's been significant work on high-risk markets in terms of listing *prices* [5] [6] [7]. However, we aren't aware of an existing large-scale analysis of the *text* in real estate listings, which is largely the same across the country and could therefore provide a useful method to compare how sellers behave in these markets. This gap is likely a result of the unstructured nature of the data, which makes it a natural candidate for NLP (see [8] for a related application to corporate disclosures about climate risks). In particular, as we'll discuss in the next section, we propose to first scrape the listings; then to compile a list of phrases relevant to flooding risk; and finally to train a neural model to classify the listings that contain these phrases by their claims about flooding risk. Note that we limit our scope to

Submitted to 33rd Conference on Neural Information Processing Systems (NeurIPS 2019). Do not distribute.

flooding (including hurricanes) in the coastal United States, and to the property descriptions in the listings.

We hypothesize, with respect to #3 above, that even if very few listings discuss being at-risk, a preponderance of listings that claim otherwise might be indicative of general buyer anxiety about flooding. For example, we see in Florida more than 2,500 mentions of the phrase 'high and dry' in Zillow listings, compared to only 123 in California, where flooding is less of a concern [9]. We also hypothesize that in areas with inconsistent damage (i.e. Houston after Hurricane Harvey [10]), listings for properties with damage are less likely to mention it, as opposed to areas that were more universally affected (i.e. the Florida Keys after Hurricane Irma [11]). An informal review of listings in Houston and the Florida Keys seems to support this hypothesis [9].

2 Details

2.1 Identifying relevant listings

We'll use a semi-supervised procedure, inspired by that in [8], in which we start with a short *seed list* of n-grams that we believe to indicate a risk of flooding or lack thereof. The list is as follows: 'flood' (including 'floods', 'flood zone', etc.), 'hurricane', 'high and dry', and the specific names of the four of the five costliest Atlantic hurricanes since 2015: 'Harvey', 'Irma', 'Michael', and 'Florence' [12] [13]. We exclude Hurricane Maria because most of its damage was outside the United States. These n-grams were prevalent in our informal review of Zillow listings in at-risk areas; 'Harvey', for example, returned almost 2,000 results in Texas [9].

We'll then record the relative frequencies of the most frequent n-grams in (1) the listings that contain n-grams in the *seed list* and (2) all listings in the United States. A 'human-in-the-loop' will look through the n-grams that are much more common in (1) than (2) to identify other n-grams that could mark discussion of flooding risks; we'll add these to the *seed list* to create an *expanded list*. Finally, we'll compile a list of listings that contain n-grams in the *expanded list*.

2.2 Classifying the listings

We'll use supervised learning with a neural model that contains BERT, a pre-trained language model [14], to classify the relevant listings into the following classes: (1) experienced flooding and/or at risk for flooding, (2) avoided flooding and/or claim to be safe from further flooding, and (3) neutral, other, or both of the previous two (the appendix has examples for each class). BERT achieved state-of-the-art performance on classification tasks, including sentiment analysis, and has been shown to capture a wide range of linguistic phenomena [15], which suggests it might be useful in this context.

The neural model will take as input the sentences from a listing that contain words and phrases in the *expanded list*. We'll label some of the data using Amazon Mechanical Turk to create training and validation datasets. The model will have the following structure (see [16] for an introduction to encoder-decoder models):

1. **Embedding**: BERT generates contextualized embeddings for the input words.
2. **Encoder**: The embeddings are passed through a sequence model (LSTM or GRU); the final hidden state is the *encoding* of the sequence.
3. **Decoder**: The *encoding* is passed through a feed-forward network to obtain the predicted class of the listing.

3 Extensions

The scope of this work could be expanded from flooding and sea level rise to include drought, wildfires, and other natural disasters; or from the United States to include other regions or countries. It might also be fruitful to look at the effects of variables like listing price, the home's flood zone, or if we tracked listings over time, the year of writing. Finally, NLP could be used to extract finer-grain data from the listings, like sentiment.

References

[1] Zillow. More than 386,000 homes at risk of coastal flooding by 2050, 2018.

[2] Ian Urbina. Perils of climate change could swamp coastal real estate. *The New York Times*, 2016.

[3] National Resources Defense Council. How states stack up on flood disclosure.

[4] National Association of Realtors. State flood hazard disclosures survey, 2019.

[5] Asaf Bernstein, Matthew T Gustafson, and Ryan Lewis. Disaster on the horizon: The price effect of sea level rise. *Journal of Financial Economics*, 2019.

[6] Athanasios Votsis and Adriaan Perrels. Housing prices and the public disclosure of flood risk: a difference-in-differences analysis in finland. *The Journal of Real Estate Finance and Economics*, 53(4):450–471, 2016.

[7] Di Jin, Porter Hoagland, Donna K Au, and Jun Qiu. Shoreline change, seawalls, and coastal property values. *Ocean & Coastal Management*, 114:185–193, 2015.

[8] Alexandra Luccioni and Hector Palacios. Using natural language processing to analyze financial climate disclosures. International Conference on Machine Learning, 2019.

[9] Zillow. Zillow listings, 2019.

[10] D Hunn, M Dempsey, and M Zaveri. Harvey's floods: Most homes damaged by harvey were outside flood plain, data show. *Houston Chronicle*, 2018.

[11] John P Cangialosi, Andrew S Latto, and Robbie Berg. National hurricane center tropical cyclone report: Hurricane irma. *National Oceanic and Atmospheric Administration: May*, 30, 2018.

[12] National Hurricane Center. Costliest u.s. tropical cyclones tables updated, 2018.

[13] Stacy R. Stewart and Robbie Berg. National hurricane center tropical cyclone report: Hurricane florence, 2019.

[14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[15] Ian Tenney, Dipanjan Das, and Ellie Pavlick. Bert rediscovers the classical nlp pipeline. *arXiv preprint arXiv:1905.05950*, 2019.

[16] Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *CoRR*, abs/1406.1078, 2014. URL `http://arxiv.org/abs/1406.1078`.

Appendix

Table 1: Examples from Zillow for each listing class [9]

Class	Snippet from listing
1	"...amenities that were destroyed by Hurricane Irma have been rebuilt and or replaced..."
1	"HOME NEEDS REPAIRS FROM HURRICANE IRMA."
2	"Home is very well built and Irma survivor at 170 mph winds"
2	"Home was high and dry during Harvey- Did not flood ."
2	"All of this AND FLOOD ZONE X = NO FLOOD INSURANCE REQUIRED."
3	"ACCORDIAN HURRICANE SHUTTERS."
3	"This fabulous home custom built by prominent Michael Hurd Construction..."