Predicting Amazon's Choice of HQ2 From Social Media: Evidence From the Tweets of Informed Sources

Kissan Joseph
University of Kansas

Abir Mandal
University of Mount Olive

Sumanta Singha
Indian School of Business

Social media chatter, and in particular, Twitter, is increasingly gaining popularity to generate forecasts in a wide variety of domains. We build on this body of work and set out to predict Amazon's HQ2 choice by analyzing the tweets of officials at the 20 finalist cities. Consistent with the affect infusion model (AIM) from the psychology literature, we conceptualize that the positive affect generated in successful ongoing negotiations will lead to a congruent positive spill over even in unrelated tweets. Analyzing tweet series that include a corpus of 50,238 tweets and incorporating dynamic time warping measures, our forecasting method correctly predicts Northern Virginia, favors it over two proximal cities, Washington D.C. and Baltimore, and ranks New York City 11th out of 20 cities. These forecasts match those of the betting markets. Our research thus offers an alternate and novel approach to extracting the signal from the noise in social media.

Keywords: Amazon HQ2, sentiment analysis, private information, social media chatter, dynamic time warping

INTRODUCTION

In recent times, data-driven forecasting has become very popular amongst researchers and practitioners. With increasing digital footprint, technologies today can track almost everything from user's geolocation to browsing pattern to social media impressions. Such innovative new-age data make available new insights that are credible, granular, and incredibly information-rich. One particular source, social media chatter, and in particular, Twitter, is increasingly gaining popularity to generate forecasts in a wide variety of domains (Signorini *et al.*, 2011; Ramli, 2012; Barnes, 2014; Sprenger *et al.*, 2014; Brown *et al.*, 2018; Graves *et al.*, 2018). Accordingly, in this paper, we build on this extensive literature and investigate how Twitter data can be used to extract private information and make predictions in the context of one of the most talked-about events in recent times— Amazon's search for its second headquarters in the U.S.

Background and Context

E-commerce giant Amazon recently concluded its search for a city to host its second headquarters (HQ2), choosing Northern Virginia and New York City but finally eliminating New York City. In terms of scope, HQ2 is expected to result in an investment of more than $5 billion and create in excess of 50,000 high paying jobs (Stevens *et al.*, 2017). The search process spanned a period of about 14 months and involved the following stages (Weise, 2018; Stevens *et al.*, 2019).

- **September 7, 2017:** Amazon announces it is searching for a second headquarters, one that will be co-equal to its Seattle home.
- **October 19, 2017:** Deadline for cities to submit their applications. In all, 238 proposals were received from cities across the United States and Canada.
- **January 18, 2018:** Shortlist of 20 finalists announced. The cities (areas) are: Atlanta, Austin, Boston, Chicago, Columbus (OH), Dallas, Denver, Indianapolis, Los Angeles, Miami, Montgomery County (MD), Nashville (TN), Newark (NJ), New York City, Northern Virginia (NOVA), Philadelphia, Pittsburgh, Raleigh (NC), Toronto, and Washington, D.C.
- **Spring — Summer 2018:** Amazon teams visit all 20 cities / areas, a process that the company — and the cities — keep largely under wraps.
- **November 5, 2018**: Given the clear difficulty of attracting 50,000 highly-paid technology workers from one city, Amazon pivots and decides to choose two cities for HQ2.
- **November 13, 2018:** Amazon chooses Northern Virginia and New York City as destinations for its next headquarters.
- **February 14, 2019:** Amazon announces that it is abandoning its plans to build a New York City headquarters. It further states that it will not restart the process and hunt for a second city, but will instead add jobs in other offices around the country.

Not surprisingly, the scale of the project generated substantial public fascination. Consequently, many entities offered predictions of which city Amazon would eventually pick. Wolfe (2018) summarizes some of the more prominent early predictions:

- Some analysts considered Atlanta, Georgia, the most likely city to be chosen, due to its affordability, economy, and tech-friendly environment.
- Washington, D.C. also seems to be a worthy contender, as it's the location of Amazon CEO Jeff Bezos's $23 million mansion and the home of the Bezos-owned Washington Post.
- Some analysts doubted New York City would be chosen, given high living costs.
- Wells Fargo's AI, named Aiera, predicted the top-five contenders to be Boston, Chicago, Atlanta, New York City, and Toronto, in that order.
- Some economists and housing experts chose Atlanta and Northern Virginia as the two most-likely locations.
- Bank of America's top-five cities were Atlanta, Denver, Washington D.C., Boston, and Raleigh, North Carolina.
- The analyst who predicted Amazon would buy Whole Foods, Scott Galloway, suggested that only two cities had the potential to be chosen: New York City and Washington, D. C.

Research Motivation and Question

In this research, we are motivated by the widespread interest in predicting Amazon's choice of HQ2 to similarly develop a forecast. In particular, we propose a sentiment analysis strategy that utilizes a very popular social media platform: the micro-blogging site Twitter. Towards this end, we mine the tweets from informed sources — Mayors, Governors, and the official city hall account — affiliated with each of the 20 finalist cities to forecast Amazon's city choice. Our sentiment extraction relies on established natural language processing (NLP) capabilities at multiple levels of granularity. Indeed, given the high intensity of usage on Twitter (approximately 500 million tweets are sent every day, please see: http://www.internetlivestats.com/twitter-statistics/), the sentiment contained in Twitter data has been extensively used for forecasting purposes across multiple domains as we will review subsequently. In fact,

in the marketplace today, there are several companies (e.g., www.tweetfeel.com) that offer Twitter sentiment analysis as one of their key services.

It is important to note here that the fundamental presumption behind traditional sentiment analysis is that sentiments are available and explicit. That is, individuals choose to broadcast their views and candidly speak their mind on micro blogging sites such as Twitter; consequently, tweets are a true reflection of their underlying sentiment. However, in our forecasting context, informed entities purposely refrain from tweeting details about Amazon's HQ2. This is because the involved entities are bound by non-disclosure agreements, meaning their tweets cannot legally discuss anything about Amazon's HQ2 choice or even ongoing negotiations with Amazon. Moreover, officials and committees of a particular city may wish to keep negotiations a closely guarded secret for competitive reasons. For these reasons, broadcasts pertaining to HQ2 on Twitter are limited and certainly not very explicit. It is this aspect of the forecasting problem that differentiates our work from earlier investigations.

To overcome the aforementioned limitation, we propose an alternate and novel conceptualization to forecast Amazon's choice of HQ2. Specifically, based on the affect infusion model (AIM) developed in the psychology literature (Forgas, 1995), we conceptualize that the positive affect generated by successful ongoing negotiations with Amazon among individuals and committees will lead to a congruent positive spillover effect over their non-Amazon tweets. In essence, we employ the prediction of the AIM model that mood impacts one's evaluation of new, and even unrelated, entities. Hence, if the Mayor or Governor feels positively about the Amazon deal, that positivity will be infused into evaluations or commentaries even in unrelated social dialogues. Thus, measuring positivity in social dialogues can reveal mood, which, in turn, can provide insights about the underlying and hidden status of ongoing negotiations with Amazon. Accordingly, we examine the evolution of the positive sentiments contained in the tweets of these informed sources to rank the 20 finalist cities in terms of their likelihood to land Amazon's HQ2.

The remainder of the paper is organized in the following manner. In the next section, we review the literature and formally conceptualize our forecasting mechanism. Then, we describe our data and empirical findings. Finally, we conclude by providing a summary and discussing the implications of our research.

LITERATURE REVIEW AND FORECASTING MECHANISM

In this section, we review the past work that has utilized social media for forecasting purposes across various applications. We use this review to highlight the fact that all of the extant applications forecast outcomes for which intuitive and direct measures are available. In contrast, we attempt to forecast an outcome that is private and structured to be kept secret. Consequently, there are no intuitive or direct measures available to forecast Amazon's choice of HQ2. This makes our investigation worthwhile from a practical point of view. Theoretically, it also necessitates the conceptualization of a new forecasting mechanism.

Literature Review

There is growing academic interest in employing social media to forecast outcomes. Sprenger *et al.* (2014) investigate the relationship between the level of bullishness on micro blogging forums (Twitter) and abnormal stock returns. They find that bullishness in tweets is associated with higher abnormal returns; however, they do not find a lagged relationship of bullishness with abnormal returns. Signorini *et al.* (2011) demonstrate that estimates of influenza-like illness derived from Twitter chatter accurately track reported disease levels. Brown *et al.* (2018) use the sentiment contained in tweets to forecast the outcomes of the English Premier League soccer matches as they unfold. They find that the aggregate tone of tweets contains information over and above that found in contemporaneous betting prices. In a markedly different context, Graves *et al.* (2018) analyze tweets to successfully locate geographies wherein unique opioid related topics are significantly correlated with opioid overdose death rates. In yet other applications, Hollywood utilizes data from social media to forecast demand for new films (Barnes, 2014) and the Australian Treasury department employs social media to forecast workforce participation and retail sentiment (Ramli, 2012).

In short, there is a burgeoning body of work that attempts to extract the signal from the noise in social media, and in particular, tweets. However, to date, no research has examined if social media can be used to extract private information that is structured to be kept secret. Moreover, as mentioned previously, obvious and direct measures are available to predict the outcome of interest in all of the previous studies (e.g., bullishness to predict stock performance, aggregate fan sentiment to predict match outcomes, discussion of opioid topics to predict substance abuse, etc.). However, in our context, such obvious and direct measures are unavailable because no information is forthcoming from Amazon and all of the negotiating parties are bound by non-disclosure agreements. From a practical point of view, this raises doubts as to whether it is even feasible to extract this information from social media. And, as suggested before, from a theoretical point of view, a new forecasting mechanism has to be conceptualized in order to extract this private information via social media. These twin considerations characterize our enhancements to the extant literature. We next discuss the conceptualization of our forecasting mechanism.

Forecasting Mechanism

We conceptualize our forecasting mechanism by relying on the affect infusion model developed in the psychology literature (Forgas, 1995). In brief, this model predicts that evaluative judgments about a new, and unrelated, target will generally tend to be congruent with one's current affect, meaning that positive feelings lead to more positive judgments and negative feelings lead to more negative judgments. The generalizability of this affect- congruency effect has been very well documented for targets ranging from products (Gorn *et al.*, 1993) to advertisements (Gardner and Jr, 1987), political candidates (Isbell and Jr, 1999), and evaluations of brand extensions (Barone *et al.*, 2000). In research that predates this body of work, Veitch and Griffitt (1976) demonstrate that good (bad) news induces positive (negative) affect which, in turn, leads to more positive (negative) descriptions of unrelated strangers.

Utilizing the conceptual lens of AIM and the empirical support found for it in many domains, we therefore conceptualize our forecasting mechanism as follows: the positive affect generated by ongoing successful negotiations with Amazon will manifest itself in higher levels of positivity even in unrelated, non-Amazon tweets. It then remains an empirical question as to whether this mechanism is strong enough to reveal Amazon's HQ2 choice. We next discuss our data, measures, and empirical findings.

DATA, MEASURES, & FINDINGS

Data

We use three different corpora of Twitter messages for each of the 20 finalists selected for Amazon's HQ2. These corpora include the official Twitter handles of the Mayor of the city, the Governor of the state, and the official city hall account. We periodically scrape all the tweets emanating from these sources at the 20 finalist cities / areas chosen by Amazon using the *rtweet* package in *R*, an open source programming language, and the Twitter API. Clearly, each of these sources has access to private information with respect to the Amazon deal but also has incentives to keep it secret. Since our goal is to extract private information from public tweets, we only choose these sources because they have direct familiarity with the strength and intensity of negotiations. While these entities may not have perfect information, they are certainly more "informed" than the media or professional forecasters. A complete list of the Twitter handles of the officials and city hall accounts are listed in Table 1.

Across the 60 Twitter handles (20 cities * 3 sources), we collect 50,238 tweets by archiving real-time stream for 11 months, from January 18, 2018 to November 9, 2018, just before Amazon made its formal announcement on November 13, 2018. No language, content, or any other kind of restriction or search criteria such as hashtags was imposed during the data collection process. Thus, our data set does not suffer from any kind of selection bias.

TABLE 1
LIST OF TWITTER HANDLES OF INFORMED SOURCES

City	Mayor	Mayor Twitter	Governor	Governor Twitter	Official Twitter
Atlanta	Kasim Reed	@KasimReed	Nathan Deal	@GovernorDeal	@Cityofatlanta
Austin	Steve Adler	@MayorAdler	Greg Abbott	@GregAbbott_TX	@austintexasgov
Boston	Marshy Walsh	@marty_walsh	Charlie Baker	@MassGovernor	@CityOfBoston
Chicago	Rahm Emanuel	@ChicagosMayor	Bruce Rauner	@GovRauner	@ChooseChicago
Columbus	Andrew Ginther	@MayorGinther	John Kasich	@JohnKasich	@ColumbusGov
Dallas	Mike Rawlings	@Mike_Rawlings	Greg Abbott	@GregAbbott_TX	@CityOfDallas
Denver	Mike Hancock	@MayorHancock	John Hickenlooper	@GovofCO	@CityofDenver
Indianapolis	Joe Hogsett	@IndyMayorJoe	Eric Holcomb	@GovHolcomb	@IndyCouncil
Los Angeles	Eric Garcetti	@MayorOfLA	Jerry Brown	@JerryBrownGov	@LACity
Miami	Carlos A. Gimenez	@MayorGimenez	Rick Scott	@FLGovScott	@CityofMiami
Mont. country	Isiah Legett	@CoUnTy_ExEc	Larry Logan	@LarryHogan	@MontgomeryCoMD
Nashville	Megan Barry	@MayorMeganBarry	Bill Haslam	@BillHaslam	@visitmusiccity
Newark	Raj Baraka	@rasjbaraka	Chris Christy	@GovChristie	@CityofNewarkNJ
New York City	Bill De Blasio	@NYCMayorsOffice	Andrew Cuomo	@NYGovCuomo	@nycgov
Northern Virginia	David L. Meyer	@FairfaxMayor	Ralph Northam	@GovernorVA	@CityofFairfaxVA
Philadelphia	Jim Kenney	@PhillyMayor	Tom Wolf	@GovernorTomWolf	@PhiladelphiaGov
Pittsburgh	Bill Peduto	@billpeduto	Tom Wolf	@GovernorTomWolf	@CityPGH
Raleigh	Nancy McFarlane	@NancyMcFarlane	Roy Cooper	@NC_Governor	@RaleighGov
Toronto	Norm Kelly	@norm	Liz Dowell	@LGLizDowdeswell	@ONgov
Washington DC	Muriel Bowser	@MayorBowser	Jay Inslee	@JayInslee	@DCGovWeb

Due to the nature of the micro blogging service (quick and short messages restricted to 280 characters), Twitter messages contain acronyms, targets (@), emoticons (:-)), hashtags (#), and other characters that are a part of Twitter folksonomy. Hashtags are metadata tags used for dynamic, user generated tagging. For instance, use of #Amazon or # HQ2 or both in the Twitter message are related to the Amazon HQ2 project. The following tweet by the Mayor of the City of Newark provides an example of this: @GovMurphy backs 'once in a lifetime' #amazon proposal for #NewarkNJ.' Some common hashtags used in our data set include #iloveit, #success, #bestfeeling (used to express positive sentiments), and #epicfail, #worse, #ihate (used to express negative sentiment).

Emoticons are another type of special character used for expressing sentiments in Twitter and other micro blogs and messaging platforms. Together with emojis, which express mood using facial pictographs, emoticons use letters, numbers, and punctuation to express a person's feelings or mood. For example, a symbol such as ':-)' is often used to indicate happiness and the symbol ':(' or ':-(' is typically used to express negative sentiment. In this connection, we note that Pak and Paroubek (2010) use emoticons to label a tweet as positive or negative. In addition to hashtags and emoticons, acronyms such as 'omg' for 'Oh my God,' 'lol' for 'laugh out loud,' 'rotf' for 'rolling on the floor,' and 'gr8' for 'great,' are also quite popular in Twitter to express sentiment. Lastly, Twitter users often use target '@' to refer to other users of the micro blog.

As we conjecture, there are very few explicit mentions of Amazon's choice of HQ2 or insights about negotiations. In fact, in our dataset, only 87 (0.1%) of the total 50,238 tweets collected during this period across all finalist cities mention Amazon, HQ2 or both. The vast majority of tweets are general conversations across a wide variety of miscellaneous topics. This supports our conjecture that officials and city hall accounts intend to keep details about Amazon HQ2 a closely guarded secret. A sample of verbatim tweets, and their sources, are listed in Table 2. Although they speak to a myriad of events, one can sense positivity coming through in some of them, as in the following examples: "Tennessee, it is time to lead the nation," "This week was one of the greatest weeks in Philadelphia history," and "It is a beautiful day in the Cap City."

TABLE 2
A SAMPLE OF TWEETS

Text	Twitter handle	Date
Tennessee, it is time to lead the nation. Will you join me in finishing what we began?	@BillHaslam	1/30/2018 0:38
Florida First Responder Appreciation Week is dedicated to those who protect and serve our communities every day.	@FLGovScott	1/26/2018 17:35
My statement commemorating Dr. Martin Luther King Jr. Day	@GovChristie	1/15/2018 14:50
Everyone with an interest in this space should spend some time @switchyards. Was pleased to cut the ribbon 2yrs ago	@KasimReed	1/17/2018 21:00
It' s a great day for some @HorizonLeague basketball! Proud to cheer on the @IUPUIJaguars men' s team as they take on.	@IndyMayorJoe	2/10/2018 18:14
Tuesday, we are breaking ground on the NEW Frederick Douglass Bridge - the largest @DDOTDC project in DC history!	@MayorBowser	2/8/2018 21:41
Six more weeks of winter!? Ugh, well, I hope they bring #Denver some much-needed snow!	@MayorHancock	2/2/2018 20:47
Get out of your own way.	@norm	2/10/2018 5:03
This week was one of the greatest weeks in Philadelphia history. Thank you to all the City workers who helped make	@PhillyMayor	2/9/2018 17:08

Examples shortened for citation (i.e., omission of hyperlinks)

Data Processing

In our research, we use three resources for pre-processing the Twitter data: (i) an emoticon dictionary, (ii) an acronym dictionary, and (iii) a standard lexicon of positive and negative words for polarity scoring. Choice of dictionary is critical in sentiment analysis. In this paper, we use the standard list of emoticons from Wikipedia[1] containing 172 different western emotions listed alongside their emotional states. The Wikipedia dictionary provides a rich source of emoticons mined from diverse social platforms and news articles and has less domain specific bias. As a result, Wikipedia-based emoticons have been widely used in the past (Agarwal *et al.*, 2011). We manually annotate each emoticon as positive or negative based on the associated emotional state. For example, the emotion ':-))' with label 'very happy' is considered positive and the emoticon ':-'(' with label ' crying' is considered negative. Although Wikipedia is very broad-based, we also verify our results using the emoticon data set created by Go *et al.* (2009) for a project at Stanford University as an additional robustness check. The results show that the relative ranking of cities does not qualitatively change with the choice of emoticon dictionary.

In our work, a Twitter message is considered positive if it contains positive emotions and vice-versa. However, unlike Go *et al.* (2009), which omits Twitter messages with both positive and negative emotions, we classify them under both categories.

Second, we use an acronym library from an online resource[2] that contains more than 1,000,000 human-annotated acronyms in common use across various domains. For example, 'nbd' is translated to 'no big deal,' 'fomo' is converted to 'fear of missing out,' and so on. Due to the nature of the micro blogging service, spelling mistakes in tweets are frequent. As such, for the acronyms that were either misspelt (and hence could not be manually annotated) or that were written in foreign languages, Google translators were used to decipher their meaning. In those instances where Google also failed, observations containing those acronyms were discarded from analysis.

Third, we use the lexicon of 2007 positive and 4787 negative words created by Hu and Liu (2004). These words are predominantly subjective words (adjectives or qualifiers) that express a sentiment. For the purpose of sentiment analysis, a subjective tweet differs from an objective tweet in that while the former expresses emotions, the latter is used only to represent some factual content or to ask questions. For example, 'It is raining now' is an example of an objective tweet, whereas 'I think I like Pizza' is a subjective tweet. Earlier work by Bruce and Wiebe (1999) show that subjectivity is strongly positively correlated with the presence of adjectives. Accordingly, Hu and Liu (2004) identify adjectives as *opinion words* using natural language processing methods and determine their semantic orientation, e.g., positive or negative.

After initial pre-processing and building of lexicons, the next task is tokenization and normalization. Tokenization is an important step in text analytics, which breaks a long human-readable text into machine readable components. Although tokenizing the text into words and word stems are most common, we also split the text into n-gram (bi-gram and tri-gram) and skip n-gram using *tokenizers version 0.2.0* package in R. The difference between an n-gram and skip n-gram is that in the case of the former, we choose a contiguous sequence of words of length less than n, whereas in the latter case, the words need not be contiguous. The tokenizer we use can preserve all emoticons, targets, hashtags, and other special characters as individual tokens. While tokenizing, we remove all stop-words such as 'a,' 'an,' 'the,' 'at' using a stop-words lexicon[3], which are some of the most commonly occurring words in the English language and are devoid of any sentiment value. Moreover, any word in the tweets that is found in the WordNet (Fellabaum, 1998) is considered an English word. To identify punctuation, we use the standard tag set by the Penn Treebank. Informal identifiers, acronyms, and character repetitions are reduced to their standard form wherever possible; or discarded otherwise. For example, we convert '*awesooome*' to 'awesome' and '*cooool*' to 'cool.' For normalization, all emoticons and acronyms are replaced by their actual meaning by looking up in the dictionary built during data pre- processing.

In general, sentiment analysis broadly involves counting occurrences of positive and negative words according to a reference lexicon. One of the problems of this approach is that the context of a word matters as much as its occurrence. For example, in the phrase, 'not to my liking,' the word 'liking' is preceded by the negation 'not' and is actually negative. This problem of identifying sentiment in the context of negation is known as the negation identification problem. To address this, we perform a rudimentary negation

identification check to find all instances where a positive or a negative word is preceded by a negation word within three words in the same sentence. For consistency, all negation words such as *'not,' 'no,' 'never,' 'cannot'* are replaced by the tag 'not.' We reverse the sentiment score of each word that follows a negation.

Polarity Scoring

In this section, we describe the methodology used for sentiment analysis. Polarity scoring is the task of assigning sentiment scores to individual tokens (unigram, bigram, and trigram) to find the general sentiment of the author in the opinionated text. A sentiment score of an opinionated text is the sum of the sentiment score of all constituent opinion words. Earlier work by Go *et al.* (2009) and Pak and Paroubek (2010) shows that in terms of the feature space, unigram models outperform all other models containing bigrams or trigrams with part of speech (POS) features. Agarwal *et al.* (2011) show that the accuracy of their unigram-based model is greater than the accuracy of a chance model by 20%. Accordingly, they comment that bigrams and POS features do not help in extracting sentiment. Hence, we follow a unigram-based approach. However, we do consider all bigrams in which either the following or preceding word is a negation word. For a comprehensive survey of the sentiment analysis and opinion mining research, see Pang and Lee (2008).

There exist various scoring mechanisms of an opinionated word. For example, Nielsen (2011) proposes a lexicon, AFINN, which assigns words with a score ranging between +5 to -5, with negative scores indicating negative sentiment and positive scores indicating positive sentiment. The 'nrc' lexicon by Mohammad and Turney (2010) labels words in binary fashion ("yes"/"no") across the categories of positive, negative, anger, anticipation, disgust, fear, joy, sadness, surprise, and trust. We follow the approach proposed by Hu and Liu (2004) who categorize words in a binary fashion into positive and negative. This classification is very generic and free from judgmental basis in regard to the intensity of polarity. For every positive word, we assign a score (+1) and for every negative word, we assign a score (-1). Each bigram that contains a negation word is assigned a score (-1). The sentiment score of a tweet is then simply the net score of all positive and negative words in it. Words that are not listed in the lexicon are considered neutral and assigned a zero score.

Normalization

In this section, we discuss the post-processing of the sentiment scores before they can be used for predicting the ranks of the finalist cities.

Length of Text Adjustment

Unlike some of the earlier works such as Agarwal *et al.* (2011), which normalize the score of each opinionated text by dividing each score by the range of the score, we allow the normalization to occur across the length of text. While the earlier approach takes care of the scale effect statistically, we believe that normalization by text length can more truthfully capture the intent of the tweet. The idea is that when people are strongly opinionated, they do not mince words to express their feeling, leading to fewer and stronger qualifiers. So in a way the effective score, which measures the net sentiment per word, reflects the sentiment intensity of the author. For example: 'This is a gruesome idea' has a stronger connotation than saying 'I think this is not such a great idea,' although both texts have identical net sentiment score ('gruesome' is a negative word, 'not great' is a negative bigram, both scored -1 according to our lexicon). Hence, we divide the score of each text by the number of words in the text (after removing stop words) to get the effective score.

Adjustment for Tweeting Propensity

While many people like to tweet, they are unlikely to be equal in their tweeting propensities. This leads to a systemic bias particularly when we perform sentiment analysis using *unrelated* Twitter data. For example, if an individual person tweets a lot in general, he (she) receives higher sentiment scores compared to a person who does not tweet as much. To minimize this bias and ensure parity, we consider the average score. More precisely, we sum the effective sentiment scores of all tweets made by a user in day and divide

it by the number of tweets. Thus, the average score, which measures the sentiment score per tweet, accounts for the difference in user's usage of social media.

Baseline Adjustment

Again, while many people like to tweet, they are unlikely to be equal in their baseline levels of positivity and optimism. To correct for this, we collect twelve months of data before our sample period and estimate the average sentiment per day per user in a similar manner. We subtract this baseline sentiment from the estimated sentiment score per day to arrive at the final score for a given user per day. In this way, we partially correct for individual's baseline positivity. Of course, implicit in this differencing is that the baseline period is unaffected by events that may significantly impact an individual's baseline positivity.

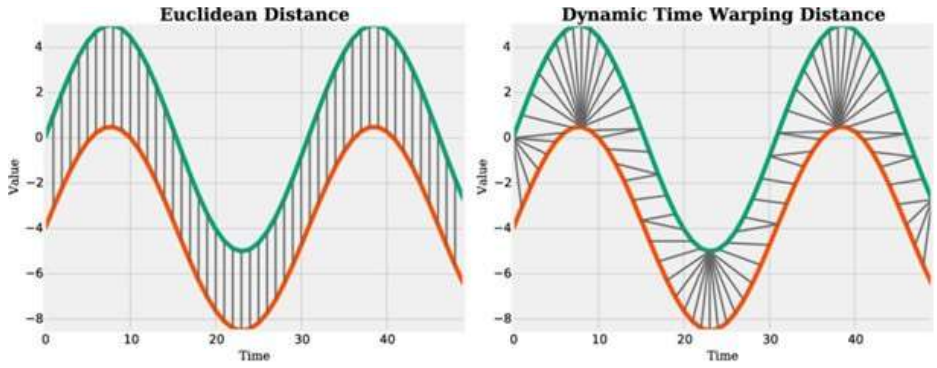
Forecasting Measure and Validation

In this section we first discuss the forecasting measure and then describe in detail the validation procedure. After initial pre-processing and tokenization, we consolidate the tweets by the Mayor, the Governor, and the official city hall account. We estimate the sentiment of each tweet based on our standard lexicon and the *score* subroutine within the *sentiment* function in R. As mentioned previously, the effective sentiment score of a tweet is simply the algebraic sum of all positive (+1) and negative words (-1), divided by the number of words in the tweet. To avoid accusations of *ex post* manipulation of the data, we simply use all the tweets that we are able to identify from a city and do not weight sources in any particular manner. Once each tweet receives an effective score, we average the score over all the tweets in a given day for every person and every city. If none of the informed sources have tweeted on any given day, a zero-sentiment score is assigned. However, such cases are rare in the data set. Since these scores contain the base effect, we subtract the base sentiment from the daily average score to get the base-adjusted daily average score. As we gather more data, we estimate the cumulative daily average score by summing the base-adjusted daily averages up to that point. We record this final cumulative sentiment score for every city, every day. A second and closely connected, score is the *ratio* of positive to negative tweets amassed cumulatively up to that point in time for each city. This measure provides another perspective on the positivity of the officials. These simple and unvarnished time series serve as our index of positive sentiment related to Amazon's choice of HQ2.

An issue that arises in final prediction is how to compare two series and conclude that one is significantly more positive than the other. The answer to this question is critical to decide the relative ranking of the cities, especially the two that visually lead the others, and in turn, the most likely winner. Unlike cross-sectional analysis, comparing time series data has some unique challenges. First, in the case of time series data, successive observations are usually auto-correlated (i.e., the i.i.d. assumption is violated), which makes the standard t-test or p-value unsuitable for comparing time series data (Nicolich and Weinstein, 1981). In our case, sentiment scores for cities have significantly high autocorrelation i.e., lag-1 autocorrelation. For example, foreshadowing our results section, the autocorrelation for both Northern Virginia and Austin are approximately equal to 0.9. This makes any inference based on independence of observations unreliable. Second, there can be temporal distortions between the two time series. To illustrate, in the case of speech signal time series, two identical messages may vary in the signal generated if one person was speaking faster than the other person or even accelerating or decelerating over the course of the observation window. This non-linear and one-to-many mapping can occur in many domains due to time shift, stretching, or bending of the time axis (Xiao, 2005). Examples include speech recognition, video indexing, signature matching, etc. This non-linear correspondence imposes additional challenges on the comparison of time series using traditional techniques. As can be well imagined, similar sentiment signals generated by two cities are likely to be shifted in time, accelerations, and deceleration as they negotiate with Amazon over the period of observation. For example, Amazon chose to visit the 20 finalists at different times thereby generating a peak at times that are temporally shifted across the cities. And, the speed of negotiations could vary due to differences in the team compositions from Amazon as well the personalities of the members of the city teams. Thus, measuring differences, or distances, between two series needs to explicitly account for these considerations.

To address this issue, a more advanced measure, dynamic time warping (DTW), proposed by Berndt and Clifford (1996), is used to determine the degree of similarity between the top two time series. The DTW algorithm optimally maps one given time series (called query) onto the whole or part of another (called reference) by expanding and contracting the time axis. Unlike Euclidean distance, which is based on one-to-one correspondence within the same time period, DTW matches two time series elements based on greatest similarity within or across the time period. Figure 1 depicts the difference between Euclidean and DTW distance (Schäfer, 2015). Once a matched pair is found, DTW measures the distance between them and then seeks to minimize the sum of total distance.

FIGURE 1
A COMPARISON OF EUCLIDEAN AND DTW DISTANCE



We use DTW package in *R* (Giorgino, 2009) to estimate the distance between two time series X and Y. The DTW algorithm works as follows: Given two time series, $X(x_1, x_2, \dots, x_n)$ and $Y(y_1, y_2, \dots, y_n)$, the minimum distance between them is given by $DTW(X,Y) = \min_{W}\{\sum_{k=1}^{K} d(x_i, y_j), W = \langle w_1, w_2, \dots w_K \rangle\}$, where w_k indicates a point (i, j) on the optimal path W between X and Y, and d is some distance measure. Put simply, the algorithm tries to find the best match $y_j \in Y$ for every $x_i \in X$ using a dynamic programming approach and computes the distance $d(x_i, y_j)$. Once the best matches are obtained for all x_i, the sum of the distances between matched pairs is called the DTW distance. If two series are identical, the DTW distance becomes zero. Since the magnitude of the DTW distance depends on the duration of the time series, we use normalized distance from DTW package, which is the computed DTW distance normalized for path length. A detailed review of the DTW algorithm can be found in Senin (2008).

Since we have only one sample (of 296 observations) for each city, only one estimate of the normalized distance can be obtained. In order to have a sense about the variability of the sample estimate, we apply the following bootstrapping technique, which is known as block bootstrap. We note here that the simple bootstrapping relies on random sampling with replacement and assumes that the data are independent and identically distributed (i.i.d). However, in time series, as mentioned earlier, successive observations are auto- correlated. Hence, any random sampling from the data will break the temporal dependencies and conventional bootstrap will fail. In contrast, block bootstrap preserves this correlation structure by resampling blocks of adjacent data instead of a single observation. Using block bootstrapping, we compare the DTW distance for different subsets of two time series — we keep one time series fixed (reference) and compare it with a subset taken from another time series (query). The length of the subset is randomized using random seeds. For every bootstrap sample, we estimate the normalized distance, called bootstrap estimates. The histogram of these bootstrap estimates provides the shape of the distribution and allows us to compute the variance and confidence interval.

Findings and Discussion

Figure 2 displays the evolution of the positive (net) sentiment among the 20 cities within the sample period — the two cities of Northern Virginia (NOVA) and Austin are at the top. For clarity, Figure 3 displays the same evolution of positive (net) sentiment but just for Northern Virginia and New York City. In Figure

4, we show the ratio of positive and negative sentiments for each of these two top two cities. Table 3 documents the final sentiment among the 20 cities as of November 9, 2018 subtracting the respective base sentiments. In Figure 2, we see a divergence in the sentiment scores of the 20 cities with about a third trending up over the sample period, a third remaining roughly level, and a third trending down. Northern Virginia and Austin seem to be separating clearly from the rest of the pack, ending at nearly identical sentiment scores (Austin = 68.62, NOVA = 72.93).

FIGURE 2
SENTIMENT TIME SERIES OF THE 20 FINALIST CITIES

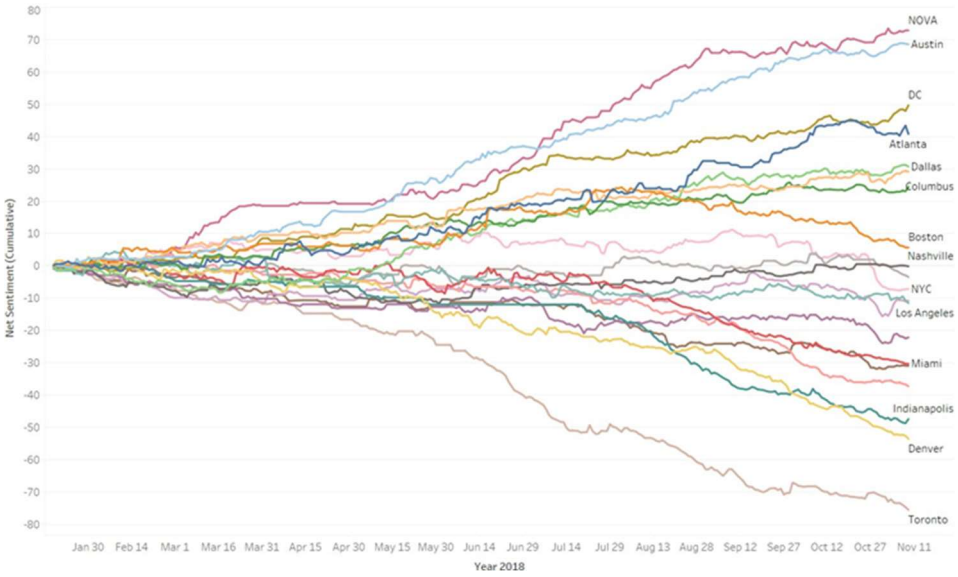


TABLE 3
SENTIMENT LEVELS OF THE AMAZON HQ2 FINALIST CITIES

City	Rank	Final Sentiment	Base sentiment
Northern Virginia	1	72.39	0.600
Austin	2	68.62	0.165
Washington DC	3	49.72	0.821
Atlanta	4	40.84	0.544
Dallas	5	30.74	0.617
Chicago	6	29.18	0.498
Columbus	7	24.47	0.949
Boston	8	5.69	1.217
Nashville	9	-0.14	0.515
Newark	10	-3.46	0.783
New York City	11	-7.16	0.552
Pittsburgh	12	-10.82	0.686
Los Angeles	13	-11.54	0.237
Philadelphia	14	-22.17	0.548
Miami	15	-30.33	0.282
Raleigh	16	-30.94	0.680
Mont. County	17	-37.27	0.582
Indianapolis	18	-47.45	0.629
Denver	19	-53.59	0.836
Toronto	20	-75.55	0.254

Comparing Northern Virginia With Austin

A closer look at Figure 3 reveals some interesting crisscross patterns between May and July of 2018, with Austin finally lagging NOVA in the net sentiment score. Figure 4 is more definitive in supporting NOVA as the preferred choice. Nevertheless, these plots only qualitatively suggest that NOVA is favored over Austin. As such, the question remains: Are the differences in scores between Austin and NOVA as portrayed statistically significant? To answer this question, we employ the statistically rigorous DTW that allows non-linear mapping.

FIGURE 3
COMPARISON OF SENTIMENT OF NOVA AND AUSTIN

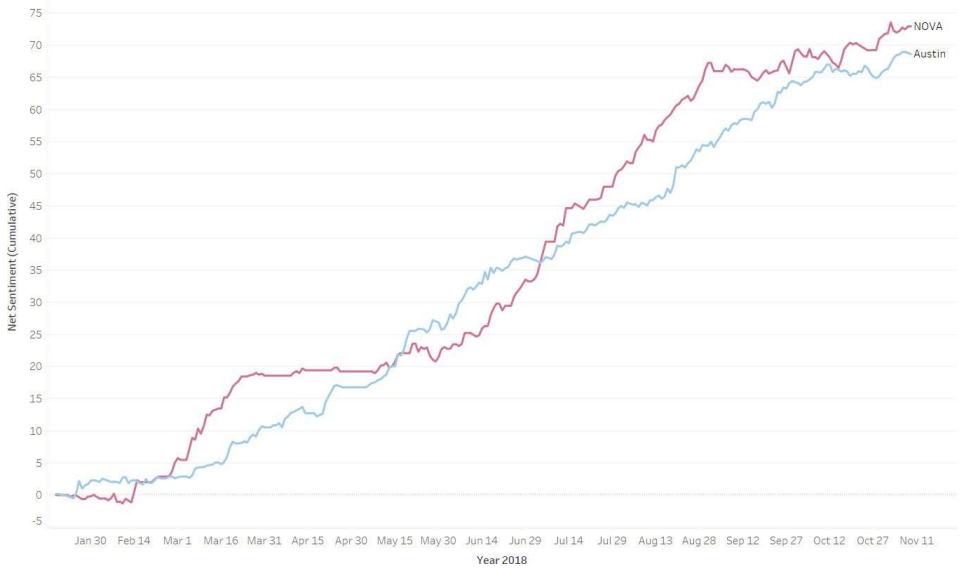


FIGURE 4
COMPARISON OF SENTIMENT RATIO OF NOVA AND AUSTIN

Figure 5 plots the optimal path W between the *net* sentiment time series of NOVA and Austin. As mentioned before, the optimal path is defined as the path of minimum total cost where the cost is measured as the sum of absolute differences of values for each match pair. Figure 6 plots the DTW distances between each matched pair based on maximum similarity. We see that DTW takes care of matching across the period as expected.

FIGURE 5
THE OPTIMAL PATH (IN BLUE) BETWEEN SENTIMENT TIME SERIES OF NOVA AND AUSTIN

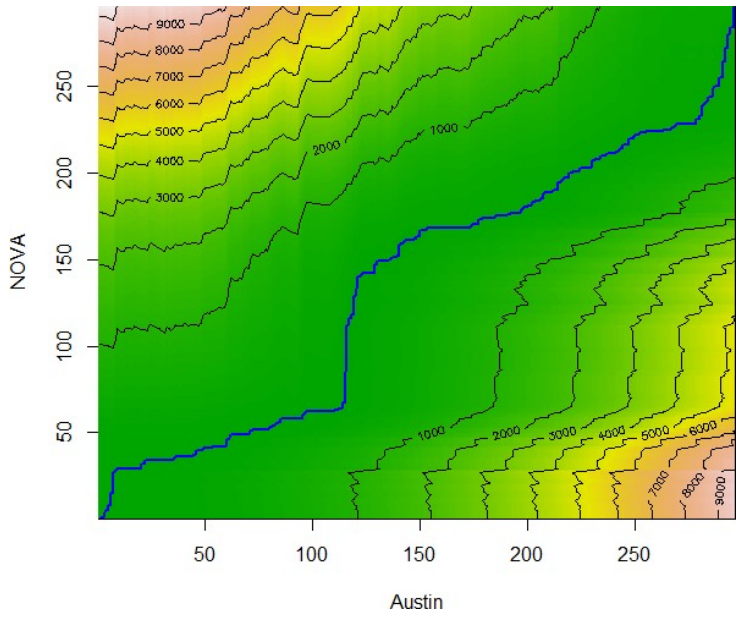
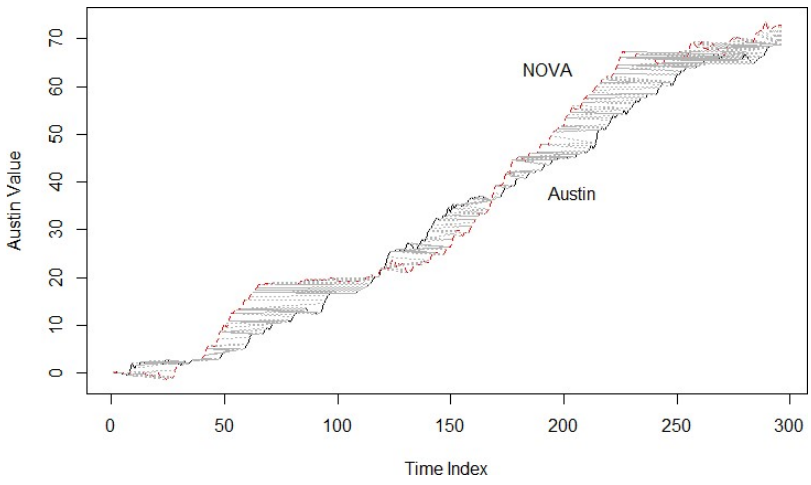


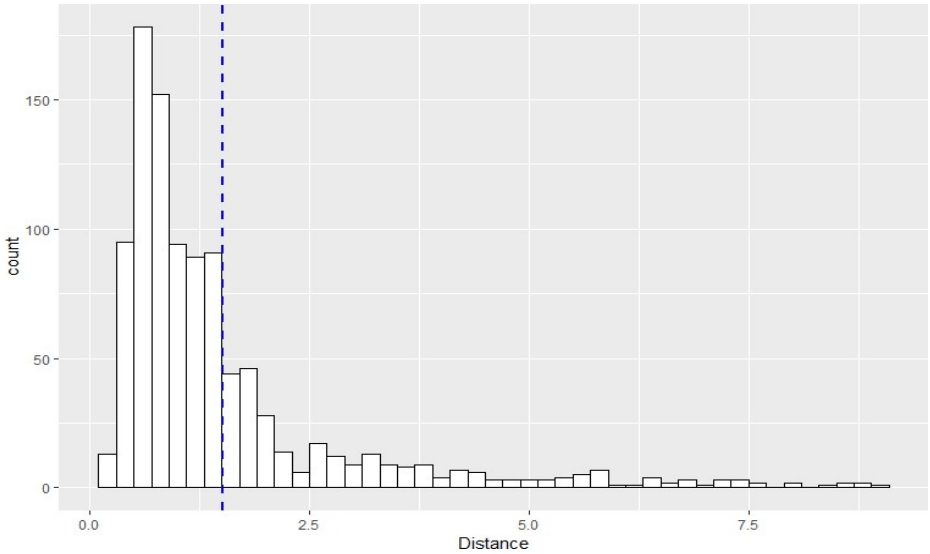
FIGURE 6
THE DTW DISTANCE MAPPING BETWEEN SENTIMENT TIME SERIES OF NOVA AND AUSTIN



Is the mean distance for net sentiment significantly different from zero? To examine that, we generate 1000 bootstrap replicates using NOVA as the reference and Austin as the query using block length anywhere between minimum 5 and maximum 296 for our example. Figure 7 presents the distribution of the bootstrap estimates and the sample mean (= 1.519) as the dotted line. The distribution is clearly not normal

and positively skewed to the right. The 95% confidence interval (CI) of the mean is found to be [1.432, 1.613]. This suggests that the mean of the distance between two sentiment time series— NOVA and Austin — is significantly different from zero, with NOVA outperforming Austin.

FIGURE 7
BOOTSTRAP DISTRIBUTION OF DTW DISTANCE BETWEEN NOVA AND AUSTIN



Similarly, Figure 8 plots the distance between matched pairs of sentiment *ratio* time series of NOVA and Austin. Is the mean distance for ratio of sentiments significantly different from zero? To examine that, we generate 1000 bootstrap replicates using NOVA as the reference and Austin as the query using block length anywhere between minimum 5 and maximum 296 for our example. Figure 9 presents the distribution of the bootstrap estimates and the sample mean (= 0.1181) as the dotted line. The distribution is clearly not normal and positively skewed to the right. The 95% confidence interval (CI) of the mean is found to be [0.1155, 0.1210]. This again suggests that the mean of the distance between two ratio sentiment time series — NOVA and Austin — is significantly different from zero, with NOVA outperforming Austin.

FIGURE 8
THE DTW DISTANCE MAPPING BETWEEN SENTIMENT RATIO TIME SERIES OF NOVA AND AUSTIN

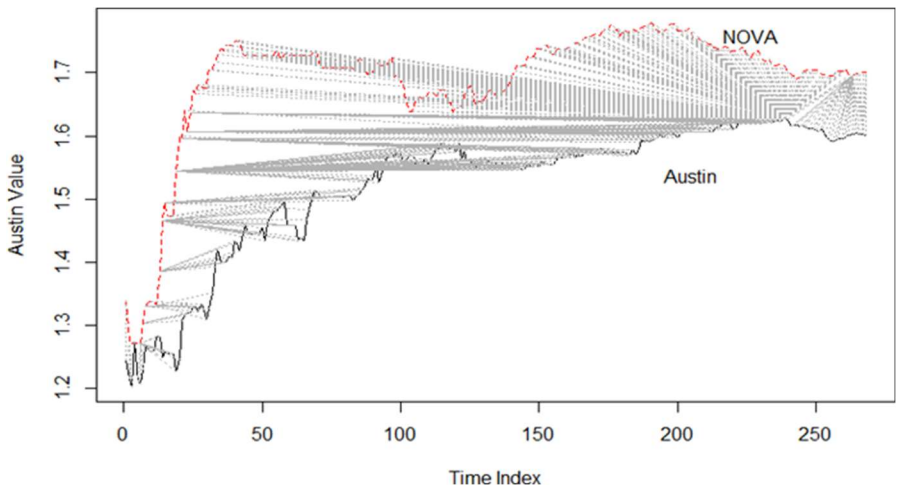
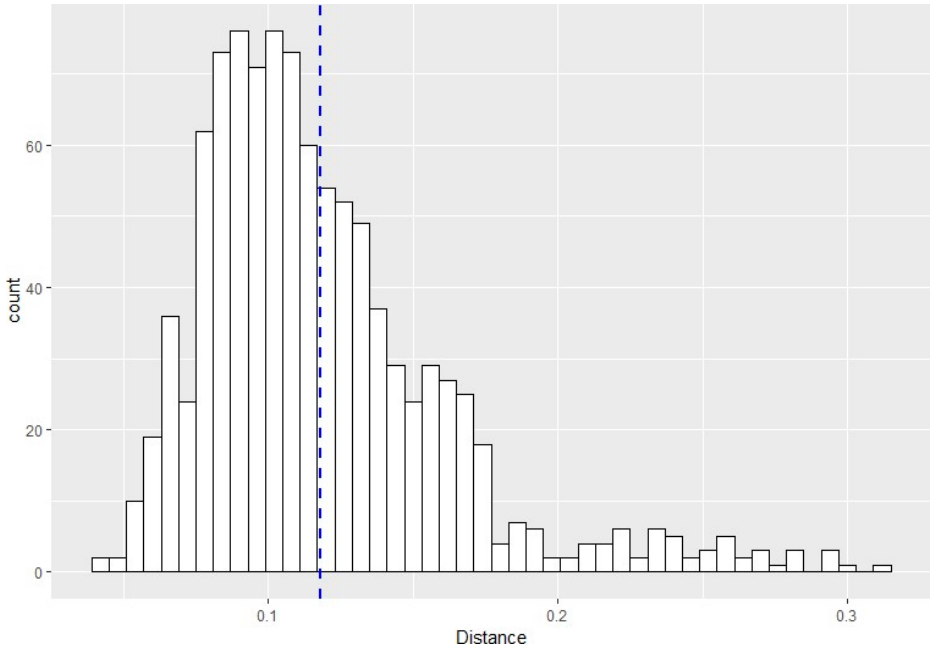


FIGURE 9
BOOTSTRAP DISTRIBUTION OF THE DWT DISTANCE FOR RATIO SENTIMENT BETWEEN NOVA AND AUSTIN



Comparing Northern Virginia With NYC, Washington D.C. and Maryland

Figure 10 reveals that sentiment at New York City stays almost level throughout the sample period but actually dips down in November, which is in marked contrast to the sentiment for NOVA. Table 3 reveals that among the three proximal choices in the Northern Virginia — D. C. — MD area, Northern Virginia is favored over Washington D.C. with Montgomery County, MD, coming in at a distant third. It further suggests only a mediocre chance for New York City (sentiment rank 11 of 20).

FIGURE 10
COMPARISON OF SENTIMENT OF NORTHERN NOVA AND NYC

Overall, what is to be made of these empirical findings? First, our forecasting model ascribes a very high likelihood to Amazon's HQ2 being located in Northern Virginia. This, by itself is a remarkable finding given Amazon's final choice. Second, our forecasting model favors Northern Virginia among the three proximal choices in the Northern Virginia — D. C. — MD area. These findings are particularly noteworthy because all these areas are in the same general vicinity. Moreover, Montgomery County, MD, actually promised one of the highest levels of monetary incentives to Amazon — an $8.5 billion incentive package (Wolfe, 2018). To put this in perspective, Northern Virginia's incentives added up to approximately $1.85 billion (Martz, 2018). Third, our forecasting model provides a mediocre score for New York City and thus excludes it from serious contention for HQ2.

Notably, all of these findings are consistent with the predictions of the betting market site Bovada as of November 12, 2018 – the day before Amazon's announcement. According to Campbell (2018), the money line bet for Northern Virginia stood at -290, Washington D.C. at +1000, and Montgomery County, MD at +2500 (in this notation, increasing numbers reflect higher payouts by the betting platform, and therefore, lower likelihood of occurrence). New York City did not even make it into the top ten.

The initial choice of a second location in New York City, and its subsequent retraction, both deserve additional commentary. In this regard, it should be noted that Amazon's initial RFP repeatedly suggested the choice of a single headquarters although the document does provide the company with some flexibility. Thus, the unexpected choice of two locations interferes with the ability of any forecasting model to correctly predict the outcome. The ability to predict Northern Virginia in the face of this wrinkle is a strength of the model.

We would even argue that our relatively low rank for New York supports the forecasting mechanism that we conceptualize in this paper. Indeed, the last few months of the decision process coincides with the political ascendancy of individuals like Alexandria Ocasio-Cortez, who were vehemently and vocally critical of Amazon HQ2 locating in New York City. In fact, immediately after Amazon's announcement of New York City, two local politicians, State Senator Michael Gianaris and City Council Member Jimmy Van Bramer categorically stated: "Too much is at stake to accept this without a fight. We will continue to stand up against what can only be described as a bad deal for New York and for Long Island City" (Lecher, 2018). Thus, it is highly likely that officials from New York negotiating with Amazon, Mayor Blasio and Governor Cuomo, were doing so under a cloud of negative political uncertainty. This negative uncertainty, in turn, is likely to have reduced their positive affect, and consequently, the level of positivity in their tweets.

SUMMARY AND IMPLICATIONS

We set out to predict Amazon's choice of HQ by analyzing the tweets of informed sources. Notably, our forecasting method ascribes a high likelihood to Amazon's final choice. Our forecasting method also favors Northern Virginia over the other proximal choices in the Northern Virginia — D. C. — MD area. In addition, the low score given for New York City by our forecasting method is consistent with the events that unfolded at the city — the cloud of negative political uncertainty is likely to have lowered the positivity emanating from officials in that city.

Our results that predict Northern Virginia are derived by explicitly accounting for the issues of non-linearity and many-to-one mapping that arise when comparing two time series. Towards this end, we employ recent advances in the literature and construct a statistical measure of difference using dynamic time warping. Our empirical results paint a consistent picture for this prediction when we use both the net positive sentiment as well as the ratio of positive to negative tweets.

We also note that our predictions of NOVA as the ultimate winner, the correct ranking among Northern Virginia, D.C., and MD area, and the relatively low score for New York city all line up very well with the predictions offered by the betting markets. As such, we believe our forecasting method compares favorably with other methods.

Overall, its predictive success notwithstanding, we believe that our research offers support for the following hypothesis: tweets capture the affective states of informed sources, which, in turn, can shed insight on events occurring in the background. Thus, more broadly, our research suggests an alternate and

novel approach to extract the signal from the noise in social media. Our essential insight is captured well by the saying of that great American poet, William Carlos Williams: *"It is not what you say that matters but the manner in which you say it: there lies the secret of the ages."*

ENDNOTES

1. https://en.wikipedia.org/wiki/List_of_emoticons
2. https://www.acronymfinder.com
3. https://github.com/quanteda/stopwords

REFERENCES

Agarwal, A., Xie, B., Vovsha, I., Rambow, O., & Passonneau, R. (2011). Sentiment Analysis of Twitter Data. In *Proceedings of the Workshop on Languages in Social Media* (pp. 30–38). Association for Computational Linguistics, Stroudsburg, PA, USA.

Barnes, B. (2014). *Hollywood Tracks Social Media Chatter to Target Hit Films*. Retrieved from https://www.nytimes.com/2014/12/08/business/media/hollywood-tracks-social- media-chatter-to-target-hit-films.html

Barone, M.J., Miniard, P.W., & Romeo, J.B. (2000). The influence of positive mood on brand extension evaluations. *Journal of Consumer Research*, *26*(4), 386–400.

Berndt, D.J., & Clifford, J. (1996). Finding patterns in time series: A dynamic programming approach. In *Advances in knowledge discovery and data mining* (pp. 229–248).

Brown, A., Rambaccussing, D., Reade, J.J., & Rossi, G. (2018). Forecasting with social media: Evidence from tweets on soccer matches. *Economic Inquiry*, *56*(3), 1748–1763.

Bruce, R.F., & Wiebe, J.M. (1999). Recognizing subjectivity: a case study in manual tagging. *Natural Language Engineering*, *5*(2), 187–205.

Campbell, S. (2018). *Amazon HQ2 Props: Northern Virginia is a Massive Favorite to Win the Bid*. Retrieved from https://www.oddsshark.com/other/amazon-hq2-betting-prop-odds

Fellabaum, C. (1998). A Semantic Network of English: The Mother of All WordNetsn. *Computers and the Humanities*, *32*, 209–220.

Forgas, J.P. (1995). Mood and judgment: the affect infusion model (AIM). *Psychological Bulletin*, *117*(1), 39.

Gardner, M.P., & Jr, F.O.W. (1987). Consumer responses to ads with positive vs. negative appeals: Some mediating effects of context-induced mood and congruency between context and ad. *Current Issues and Research in Advertising*, *10*(1-2), 81–98.

Giorgino, T. (2009). Computing and Visualizing Dynamic Time Warping Alignments in R: The Package. *Journal of Statistical Software*, *31*(7), 1 – 24.

Go, A., Bhayani, R., & Huang, L. (2009). Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford*, *1*(12), 2009.

Gorn, G.J., Goldberg, M.E., & Basu, K. (1993). Mood, awareness, and product evaluation. *Journal of Consumer Psychology*, *2*(3), 237–256.

Graves, R.L., Tufts, C., Meisel, Z.F., Polsky, D., Ungar, L., & Merchant, R.M. (2018) Opioid Discussion in the Twittersphere. *Substance Use & Misuse*, *53*(13), 2132–2139.

Hu, M., & Liu, B. (2004). Mining and Summarizing Customer Reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 168–177). ACM, New York, NY, USA.

Isbell, L.M., & Jr, R.S.W. (1999). Correcting for mood-induced bias in the evaluation of political candidates: The roles of intrinsic and extrinsic motivation. *Personality and Social Psychology Bulletin*, *25*(2), 237–249.

Lecher, C. (2018). *Politicians and protestors are looking for ways to stop Amazon's New York Hq2*. Retrieved from https://www.theverge.com/2018/11/14/18095536/amazon-hq2-protests-law-new-york

Martz, M. (2018). *How Virginia Sealed the Deal on Amazon's HQ2, the Biggest Economic Development Project in U.S. History*. Retrieved from http://schrisjones.com/news/ how-virginia-sealed-the-deal-on-amazons-hq2-the-biggest-economic-development- project-in-u-s-history/

Mohammad, S.M., & Turney, P.D. (2010). Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Genera- tion of Emotion in Text* (pp. 26–34).

Nicolich, M.J., & Weinstein, C.S. (1981). The use of time series analysis and t tests with serially correlated data tests. *The Journal of Experimental Education*, *50*(1), 25–29.

Nielsen, F.A. (2011). *AFINN. Information and Mathematical Modeling*. Technical University of Denmark.

Pak, A., & Paroubek, P. (2010). Twitter as a corpus for sentiment analysis and opinion mining. *LREc*, *10*, 1320–1326.

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, *2*(1-2), 1–135.

Ramli, D. (2012). *Treasury to Mine Twitter for Economic Forecasts*. Retrieved from https://www.afr.com/technology/enterprise-it/treasury-to-mine-twitter-for-economic- forecasts-20121030-j1k6t

Schäfer, P. (2015). *Time Series Similarity Search for Data Analytics*. Ph.D. Thesis, Mathematisch Naturwissenschaftlichen Fakultät der Humboldt-Universität zu.

Senin, P. (2008). Dynamic time warping algorithm review. *Information and Computer Science Department University of Hawaii at Manoa Honolulu, USA*, *855*(1-23), 40.

Signorini, A., Segre, A.M., & Polgreen, P.M. (2011). The use of Twitter to track levels of disease activity and public concern in the US during the influenza A H1N1 pandemic. *PloS One*, *6*(5), e19 467.

Sprenger, T.O., Tumasjan, A., Sandner, P.G., & Welpe, I.M. (2014). Tweets and trades: The information content of stock microblogs. *European Financial Management*, *20*(5), 926–957.

Stevens, L., Raice, S., & Lombardo, C. (2017). *Amazon Seeks Prime Location for Its Second Headquarters*. Retrieved from https://www.wsj.com/articles/amazon-opens-search-for-second-headquarters-city-in-north-america-1504780191

Stevens, L., Vielkind, J., & Honan, K. (2019). *Amazon Cancels HQ2 Plans in New York City*. Retrieved from https://www.wsj.com/articles/amazon-cancels-hq2-plans-in-new-york-city-11550163050

Veitch, R., & Griffitt, W. (1976) Good News-Bad News: Affective and Interpersonal Effects 1. *Journal of Applied Social Psychology*, *6*(1), 69–75.

Weise, E. (2018). *Amazon HQ2 timeline: The winners are New York City and Arlington, Virginia*. Retrieved from https://www.wcnc.com/article/news/nation-now/amazon-hq2-timeline- the-winners-are-new-york-city-and-arlington-virginia/465-1302ba8b-c506-4a6b- 917a-69236705226b

Wolfe, S. (2018). *Amazon is Now Going to Split Its HQ2 into 2 Locations After More than a Year of Intense Speculation*. Retrieved from https://www.businessinsider.com/amazon- hq2-timeline-new-york-virginia-2018-11

Xiao, H. (2005). *Similarity search and outlier detection in time series*. Ph.D. Thesis, Fudan University, Shanghai, China.