Appendix A: Corpus Creation and Preprocessing

We constructed our corpora using the Lexis-Nexis and ProQuest databases, searching the full text (including headlines) of all articles published by the 17 newspapers in our sample for terms referencing Latinos or Hispanics. Our search terms were Latino!, Latina!, Latinx, and Hispanic!. [1] We then sorted articles in chronological order and eliminated any duplicates, as determined by (near) identical titles or opening lines. In addition, we remove any articles erroneously captured by our search criteria, such as the very few articles referring using the word "latinate."

To prepare these articles for sentiment analysis, we spell out common contractions in full (he's -> he is), and convert all words to lower-case so that variations in capitalization will not affect our valence calculations. Finally, we strip accents (umlauts etc.) and we separate punctuation and special characters (#, @, etc.) from regular letters, so that words are always separated by spaces, not punctuation.

Representative Corpus Construction

Given the search constraints of LexisNexis and ProQuest, we cannot obtain a random selection of articles published by a given newspaper. To approximate a random sample, we search for articles 1) containing terms unlikely to be associated with either positive or negative valence, and 2) appearing in our 17 newspapers on a randomly selected set of days evenly distributed across our 20-year period of interest.

---

[1] Three of our newspapers were not available for the entire 21-year period: the *New York Post* became available only starting December 5, 1997, and the *Arizona Republic* beginning January 1, 1999. Availability of the *Las Vegas Review-Journal* ended December 31, 2012.

In identifying terms unlikely to be associated with positive or negative sentiment, we turn to the labMT lexicon also used in our sentiment analysis. This lexicon assigns valences to all common English words. We select words with a valence of 5.00 exactly (the midpoint of the scale) that are also among the 4,000 most common words in the English language. This produces a list of 18 words: because, per, standard, situation, carbon, assess, throw, liver, plain, supervisor, something, throat, whereas, boot, fourth, stir, price, and odds. LexisNexis does not permit a search for the word 'because'; we therefore search for any one of the remaining 17 words in articles published by each paper.

Since an open-date search would generate an enormous number of articles, we randomly select 3 days in every calendar year from 1996 to 2016 (in order to avoid overweighting one part of the year, we make sure at least one of these dates is in each half of the calendar year). We use the same dates for each of our 17 sources, in order to maintain constant treatment across them. Once selected, the representative corpus articles undergo the same preprocessing steps as articles from our main corpus. This procedure nets just over 48,000 articles that are not likely to be biased in any way by content, trends over time, or other factors that might skew their average valence. Although the search words are selected because they are neither positive nor negative, they are embedded in articles about "random" topics that contain sets of words ranging from highly positive to highly negative and that are representative of the wide variety of articles published by US newspapers.

Appendix B: Lexica Used

As described in the text, we use eight different sentiment analysis lexica. Each is widely used by scholars interested in the sentiment analysis of various types of texts. Although a number of them are constructed from the same sources, the actual overlap between them is surprisingly small: although the smallest of the lexica contains 3,731 words, only 331 words are captured by all eight lexica with the same polarity (positive or negative).

In addition to containing different sets of words, the lexica also vary in how they assign valence: Four identify words simply as positive or negative, while the others assign words a range of values indicating how strongly positive or negative they are. In addition, two of the lexica specify word stems, indicating they will accept any endings to a word (such 'wildcard' specifications increase the effective size of these lexica considerably). Finally, the ratio of negative to positive words included varies considerably, from 0.40 (labMT, the only lexicon with more positive than negative terms) to 2.39 (HuLiu), with an average ratio of close to 1.5. Table B1 offers a brief overview of the different lexica.

Table B1: Sentiment Analysis Lexica

Name & citation	Construction of original lexicon	Additional processing here / comments	Positive terms	Negative terms
HuLiu (Hu and Liu, 2004)	Manually constructed by scholars at the University of Illinois in Chicago, based on WordNet (Miller, 1995).	Developed for social media; contains terms such as "f*ck".	2003 (+1)	4783 (-1)
labMT (Dodds et al., 2011)	Used Mechanical Turk coders to code the 'happiness level' of the most frequent 5,000 words from four separate sources: Twitter, Google Books (English), music lyrics (1960 to 2007), and the *New York Times* (1987 to 2007). Full lexicon has 10,222 entries	Filtered out words with low valence scores (absolute value < 1), as recommended by the creators of the lexicon.	2668 (range from 1 to 3.5)	1063 (range from -1 to -3.5)
LexicoderSD (Young and Soroka, 2012)	Manually constructed; starting point was all words from the General Inquirer (GI) (Stone and Hunt, 1963), the Regressive Imagery Dictionary (RID) (Martindale, 1975), and Roget's Thesaurus with the same valence in all 3 dictionaries or with the same valence in 2 and omitted from the third. Targeted at political and economic news.	Includes wildcards to specify any ending acceptable for a given stem.	1615 (+1), of which 1043 stems	2768 (-1), of which 1971 stems
MPQA (Wilson et al., 2005)	Used words from GI, from Hatzivassiloglou and McKeown (1997), and from their own prior work (Riloff and Wiebe, 2003)	Used only single words (no multi-word phrases) Averaged valence for words with multiple entries. 'strong' polarity is given a value of 1, 'weak' polarity gets ½.	2299 (range from 0.175 to 1.00)	4150 (range from -0.175 to -1)

NRC (Mohammad and Yang, 2011; Mohammad and Turney, 2011)	Coded all words from Roget's thesaurus that occur at least 120,000 times in Google's n-gram corpus, using 5 different MT coders for each word.		2312 (+1)	3243 (-1)
SentiWordNet (Baccianella et al., 2010)	Assigns valences to the synonym sets (synsets) in the online semantic dictionary WordNet. Starting from 'paradigmatically' positive or negative words, propagated valence across the entire dictionary using the network structure implied by synsets sharing words. Full lexicon has 29,436 entries.	For words with multiple valences (e.g. in multiple synsets), averaged the values. Filtered out words with low aggregate valence scores (absolute value < 0.1)	11116 (range from 0.1 to 1)	13106 (range from -0.1 to -1)
SOCAL (Taboada et al., 2011)	"Sentiment Orientation CALculator", manually constructed from all words in a 400-text corpus of Epinions reviews, movie reviews from the Polarity Dataset (Pang et al., 2002), and GI.		3716 (range from 0.5 to 5.0)	6341 (range from -0.5 to -5.0)
WordStat	Constructed by Provalis (makers of WordStat), by combining word lists from GI, RID, and Pennebaker's Linguistic and Word Count dictionary (LIWC) (Tausczik and Pennebaker, 2010) and searching WordStat's internal dictionary for potential synonyms.	Includes wildcards to specify any ending acceptable for a given stem.	5539 (+1), of which 337 stems	9539 (-1), of which 578 stems

References

Baccianella S, Esuli A and Sebastiani F. (2010) SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. *LREC*.

Dodds PS, Harris KD, Kloumann IM, et al. (2011) Temporal patterns of happiness and information in a global social network: Hedonometrics and Twitter. *PLoS one* 6: 1-25.

Hatzivassiloglou V and McKeown KR. (1997) Predicting the semantic orientation of adjectives. *Proceedings of the 35th annual conference of the Association for Computational Linguistics*: 174-181.

Hu M and Liu B. (2004) Mining and summarizing customer reviews. *The ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD-2004).* Seattle, WA: ACM.

Martindale C. (1975) *Romantic progression: The psychology of literary history,* Washington, DC: Hemisphere.

Miller GA. (1995) WordNet: A Lexical Database for English. *Communications of the ACM* 38: 39-41.

Mohammad SM and Turney PD. (2011) Crowdsourcing a word-emotion association lexicon. *Computational Intelligence.* Wiley-Blackwell.

Mohammad SM and Yang T. (2011) Tracking sentiment in mail: HOw genders differ on emotional axes. *Workshop on computational approaches to subjectivity and sentiment analysis.* Portland, OR: ACL.

Pang B, Lee L and Vaithyanathan S. (2002) Thumbs up? Sentiment classification using machine learning techniques. *Conference on Empirical Methods in NLP*. Philadelphia, PA.

Riloff E and Wiebe J. (2003) Learning extraction patterns for subjective expressions. *Proceedings of the 2003 conference on Empirical Methods in Natural Language Processing*.

Stone PJ and Hunt EB. (1963) A computer approach to content analysis: Studies using the General Inquirer. *Proceedings of the Spring Joint Computer Conference*: 241-256.

Taboada M, Brooke J, Tofiloski M, et al. (2011) Lexicon-based methods for sentiment analysis. *Computational Linguistics* 37: 267-307.

Tausczik YR and Pennebaker JW. (2010) The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology* 29: 24-54.

Wilson T, Wiebe J and Hoffmann P. (2005) Recognizing contextual polarity in phrase-level sentiment analysis. *Proceedings of the Human Language Technologies Conference / Conference on Empirical Methods in Natural Language Processing*.

Young L and Soroka S. (2012) Affective News: The Automated Coding of Sentiment in Political Texts. *Political Communication* 29: 205-231.

Appendix C: Valence Calculation

The basic valence calculation process for a particular sentiment lexicon is straightforward: we add up the valence values for each word we encounter that is captured by the lexicon (either directly or through a wildcard), and divide the result by the total number of words in the text. Dividing by word length provides an adjustment for the intensity of sentiment: five negative words in one sentence will have a much stronger impact than five negative words spread across 50 sentences.

This process ignores two important ways in which a word's sentiment may be modified: through intensification and through negation. Following Taboada et al., we use a list of 216 different intensifiers and apply their associated multiplying factor to the subsequent valence word (Taboada et al., 2011). Some of these intensify the strength of a word, some weaken it, and some change the polarity. The default multiplier for a valence word is 1. Intensifier values are added to this default, and the result is multiplied by the valence. For example, "slightly" has a multiplier of -0.5, which means that a subsequent word's valence is multiplied by (1 + -0.5) = 0.5. We handle negation in a parallel fashion by identifying words that shift polarity, such as not, no, nor, nothing, never, and nowhere, and add them to the list of polarity-shifting words in Taboada et al. (2011), such as "hardly." To get a polarity shift, we need a multiplier below -1. We assign our negation and polarity-shifting words a multiplier of -1.5, so that the valence is multiplied by (1 + -

1.5) = -0.5.[2] While this approach is of course not perfect, tests show that it does improve the accuracy of sentiment measurements (Taboada et al. 2011).

Calibration and aggregation

The process just described generates eight different valence measures. For any given corpus of texts, one of these may perform better than the others; however, we have no way of knowing *ex ante* which. Moreover, given that each of the lexica we use has been carefully constructed, it is likely that their differences reflect different strengths, not simply superiority or inferiority. Accordingly, we are interested in combining them to take advantage of these different strengths. A straightforward way to do so is to average them; indeed, doing so produces a valence measure that better identifies sentiment than any of the individual measures, on a large corpus of texts whose sentiment is known.

To aggregate and calibrate the individual valence measures, we generate scaling information from the representative corpus. We apply each of the eight lexica to all the articles in that corpus, and calculate the scaling parameters necessary to produce a mean of 0 and a standard deviation of 1. Since this corpus is made up of articles we have no reason to believe are biased in a positive or negative direction away from the average

---

2 Multiple consecutive intensifiers (including negation) are handled by simply multiplying them by one another. The result is applied to the next word if it is a valence word; otherwise the intensification factor resets to 1. The only exception is if the next word is one of a very small list of stopwords (a, an, the, and, to, as), in which case we skip over that word and look at the one that follows.

newspaper article, we can be reasonably confident that a valence of 0 (after the rescaling) indicates an article of average valence within the context of the US print media. We average the eight rescaled measures thus produced to get one overall valence measure. This average will have a standard deviation less than 1, since the valence measures produced by the different lexica are correlated. As a final step, therefore, we divide the average valence measure by the standard deviation of this measure across the representative corpus.

The same calibration adjustments are then applied to the valence calculations for our main corpus. This produces an overall valence measure with a straightforward interpretation: the sign of the valence of an article indicates whether it is positive or negative, and the size of the valence is expressed in standard deviations relative to our representative corpus. Another strength of this approach is that it allows us to directly compare the tone of coverage across different bodies of texts even when they are neither generated nor processed at the same time: all we need is to use the same scaling parameters.

Appendix D: Themes and Root Words

In order to test the effect of negative themes previously identified in existing Latinx scholarship on the valence of articles in our Latinx corpus, we code each article for the presence of a root word related to one of four themes: criminality, immigration, illegal immigration, and economic threat. Root word lists for each theme were chosen through author team discussion and were supplemented by results from a collocation analysis performed on the entire dataset.

Negative Themes	**Related Root Words**
Criminality	'crime*', 'criminal*', 'violen*', 'gang*', 'devian*', 'misdemeanor*', 'larcen*', 'burglar*', 'law-breaker*', 'lawbreaker*', 'law breaker*', 'felon*', 'knife', 'knives', 'murder*', 'rape', 'rapes', 'rapist*', 'robber*', 'theft*', 'homicide*', 'assault*', 'abuse*', 'cartel*', 'drug*', 'addict*', 'trafficking*', 'thug*','parole', 'probation', 'imprison*', 'stole*', 'steal*', 'fraud*'
Immigration	'immigra*', 'migrat*'
Illegal Immigration	'illegal immigra*', 'unlawful immigra*', 'unauthorized immigra*', 'alien*', 'undocumented', 'deport*'
Economic Threat	'unemploy*', 'welfare', 'poverty', 'homeless*', 'poor*'

We identified potentially positive themes by conducting sets of collocation analyses on the subset of our articles that had a positive valence, and on the entire dataset by searching for words in close proximity to both Latinx words and a list of 73 positive adjectives (see Appendix E and Appendix F). We then inductively assembled the results

into potentially positive themes (see Appendix G) and used the root words of our collocation analyses to construct thematic variables.

Potentially Positive Themes	**Related Root Words**
Achievement	'achievement*', 'contribution*', 'award*', 'talent*', 'stride*', 'thriv*', 'attainment', 'outperform*', 'surpass*'
Culture	'cultur*', 'flair', 'pride', 'fest*', 'tradition*', 'cuisine*', 'heritage', 'vibran*', 'richness', 'torta*'
Size	predominantly', 'disproportionally', 'predominately', 'overrepresentation', 'outnumber*', 'heavily', 'sizable'
Origin	'descen*', 'origin*', 'surname*', 'ancestry'
Groupness	'communit*', 'population*', 'societ*', 'enclave*'
Leaders	'player*', 'performer*', 'writer*', 'athlete*', 'actress*', 'musician*', 'ministr*', 'businessman', 'comic*', 'heartthrob*', 'bombshell*', 'vicar*', 'heroine*', 'fighter*'
Growth	'influx*', 'surg*', 'grow*', 'boom*', 'swell*'

Appendix E: Positive Adjectives for Collocation Analysis

In order to identify words within the Latinx dataset that are likely to be positive, we undertook collocation analyses for words proximate both to our Latinx root words and to the 73 most positive adjectives identified by our sentiment lexica. The results yield root words that are systematically associated with both Latinx and positive adjectives.

The most positive adjectives from our lexica are:

'admirable'
'amazing
'angelic'
'awesome'
'beautiful'
'better'
'best'
'blessed'
'blissful'
'breath-taking'
'breathtaking'
'brilliant'
'celebrated'
'commendable'
'dazzling'
'delightful'
'divine'
'ecstatic'
'euphoric'
'excellent'
'exceptional'
'exquisite',
'extraordinary',
'fabulous'
'fantastic'
'fascinating'
'flawless'
'friendly'
'fun'
'generous'
'glorious'
'gorgeous'
'great'
'happy'
'harmonious'
'heavenly'
'high-quality'
'ideal'
'immaculate'
'impeccable'
'impressive'
'incredible'
'joyful'
'magnificent'
'marvellous'
'marvelous'
'outstanding'
'peerless'
'perfect'
'phenomenal'
'prestigious'
'priceless'
'rapturous'
'romantic'
'scrumptious'
'selfless'
'sensational'
'spectacular'
'splendid'
'stupendous'
'sublime'
'successful'
'sweetest'
'terrific'
'top-flight'
'top-notch'
'top-rate'
'unmatched'
'unsurpassable'
'wonderful'
'wonderous'
'wondrous'
'worthwhile'

Appendix F: Collocation Analysis to Identify Potentially Positive Words

We identified potentially positive themes by conducting two types of collocation analysis.

First, we conducted three collocation analyses on articles in our corpus that had a positive valence, in order to understand which words were closely associated with Latinx words in positive articles. We conducted analyses examining the one word preceding a Latinx word (the L1 collocate), a two-word window around Latinx words (L1R1 collocates), and a ten-word window around Latinx words (L5R5 collocates).

L1 collocates of Latinx words, positive articles only:

predominately	among	diluted
predominantly	mostly	thriving
wise	swelling	young
woo	non	mainly
wooing	galvanized	lure
luscious	attract	educate
heavily	appease	lone
sizable	empower	distinctly
surging	energize	fellow
galvanize	large	celebrate
prominent	motivate	premier
alienate	affecting	vibrant
booming	primarily	uninsured
mobilize	outnumber	eligible
largely	recruit	registered
recruiting	first	growing
influential	stereotypical	

L1R1 collocates of Latinx words, positive articles only:

origin
descent
population
predominately
predominantly
wise
woo
wooing
comprise
luscious
surname
heavily
heartthrob
community
bombshell
surging
sizable
clientele
advocacy
galvanize
prominent
alienate
vote
growing
culture
booming
heritage
whites
mobilize
largely
electorate
ministry
heavyweight
mostly
constitute
among
influential
outreach
ancestry
swelling
outnumber
liaison
heroine
market
comprised
enclave
representation
cuisine
compose
non

L5R5 collocates of Latinx words, positive articles only:

predominately
interchangeably
biannual
comprise
predominantly
origin
underrepresentation
population
descent
richness
woo
monolithic
torta
cater
wooing
influx
whites
marital
constitute
bloc
comprised
growing
wise
outnumber
liaison
surging
upwardly
clout
advocacy
rapidly
sizable
swelling
booming
surname
heavily
dropout
vicar
assiduously
clientele
disproportionately
galvanize
geared
lagged
heartthrob
outreach
surpass
attainment
outperform
diluted
proportionately

Second, we conducted collocation analyses that searched for words in our entire corpus that were proximate to Latinx words *and* to a list of 73 positive adjectives (see Appendix E). These words are associated with Latinx and with positivity in a greater concentration than they are present elsewhere in our corpus. We conducted analyses examining a two-word window around Latinx words and positive adjectives (L1R1 collocates), and examining a ten-word window around Latinx words and positive adjectives (L5R5 collocates). We omitted an analysis with the one word preceding a Latinx word and positive adjectives (the L1 collocate), as single words preceding a positive adjective were unlikely to yield relevant results.

L1R1 collocates of positive adjectives and Latinx words:

representation
actors
man
players
represent
female
politician
writers
decisions
are
particularly
society
performers
athletes
actress
actor
qualified
customer
migration
candidate
pride
resource
fare
musicians
businessman
talent
tradition
cast
represented
kid
film
promote
american
comic
contributions

L5R5 collocates of positive adjectives and Latinx words:

wise
attracting
fared
educating
fighters
players
serve
flair
awards
achievement
strides
pride
models
recognizes

Appendix G: Collocation Root Words Grouped by Theme

We drew on the results of the collocation analyses outlined in Appendix E and Appendix F to inductively identify potentially positive themes associated with Latinx in news coverage. We used the root words associated with each theme to construct thematic variables in our corpus (see Appendix D).

The thematic groups created from the collocation process are:

Potentially Positive Themes	**Related Collocation Words**
Achievement	achievement, contributions, awards, talent, strides, thriving, attainment, outperform, surpass
Culture	flair, pride, fest, tradition, culture, cuisine, heritage, vibrant, richness, torta
Groupness	community, population, society, enclave
Leaders	players, fighters, performers, athletes, writers, actress, musicians, businessman, comic, heartthrob, bombshell, vicar, ministry, heroine
Growth	influx, surging, growing, booming, swelling
Origin	descent, origin, surname, ancestry
Size	predominantly, disproportionately, predominately, predominantly, overrepresentation, outnumber, heavily, sizable