TopCATI Notes Sentiment Analysis

Proof of Concept and Preliminary Results

Matthew Burgess
Economist
CES Data Collection Branch
Bureau of Labor Statistics
FedCASIC 2015
3/5/2015

BLS
BUREAU OF LABOR STATISTICS
U.S. DEPARTMENT OF LABOR

www.bls.gov

Overview

- CES Overview - TopCATI Data Collection and Notes
- Sentiment Analysis Proof of Concept (R and SAS)
- Preliminary Results
- Conclusions/Going Forward

BLS

Background on the Current Employment Statistics Survey

- The BLS Current Employment Statistics (CES) survey is also known as the payroll survey or the establishment survey

- The CES Survey is the Largest multi-modal survey
  - Survey of about 143,000 businesses and government agencies, representing approximately 588,000 individual worksites

- CES publishes employment, hours, and earnings data for the nation, states, and metropolitan areas at total nonfarm and detailed industry levels

BLS

CES Data Collection

- Monthly survey of establishments to collect **employment**, **hours,** and **earnings** data

- Collect employment data for the pay period that includes the 12th of the month

BLS

Computer Assisted Telephone Interviewing

- 26% of CES reports are collected by CATI. We also use other methods like web and fax
- Interviewers at Data Collection Centers (Atlanta, Dallas, Kansas City, and Fort Walton Beach) call respondents and collect data using TopCATI software
- TopCATI software allows interviewers to take notes, schedule call times, and review reported data

BLS

Project Background

- CES is required to permanently save TopCATI interviewer notes
- TopCATI interviewer notes contain valuable, qualitative information about businesses participating in the CES survey
- Data mining techniques such as sentiment analysis may be used to quantify information contained in the notes

BLS

Sentiment Analysis Explained

- Estimating sentiment
  - Many complex algorithms have been developed – all are limited by the computer's ability to interpret language
  - Pros: facilitates analysis of millions of text notes in a short period of time
  - Cons: computers will incorrectly interpret nuanced phrases, sarcasm, etc.

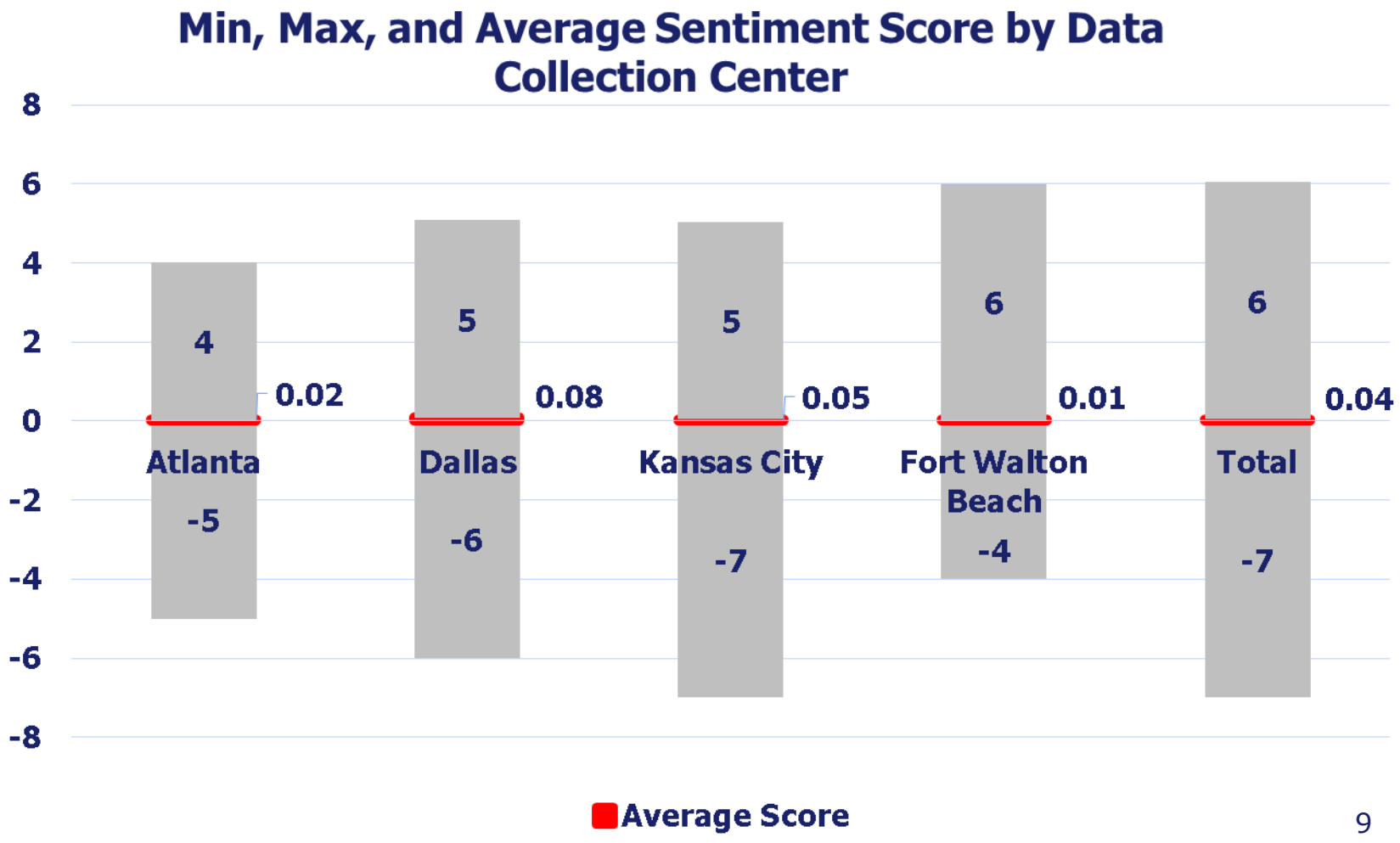
BLS

Project Methods

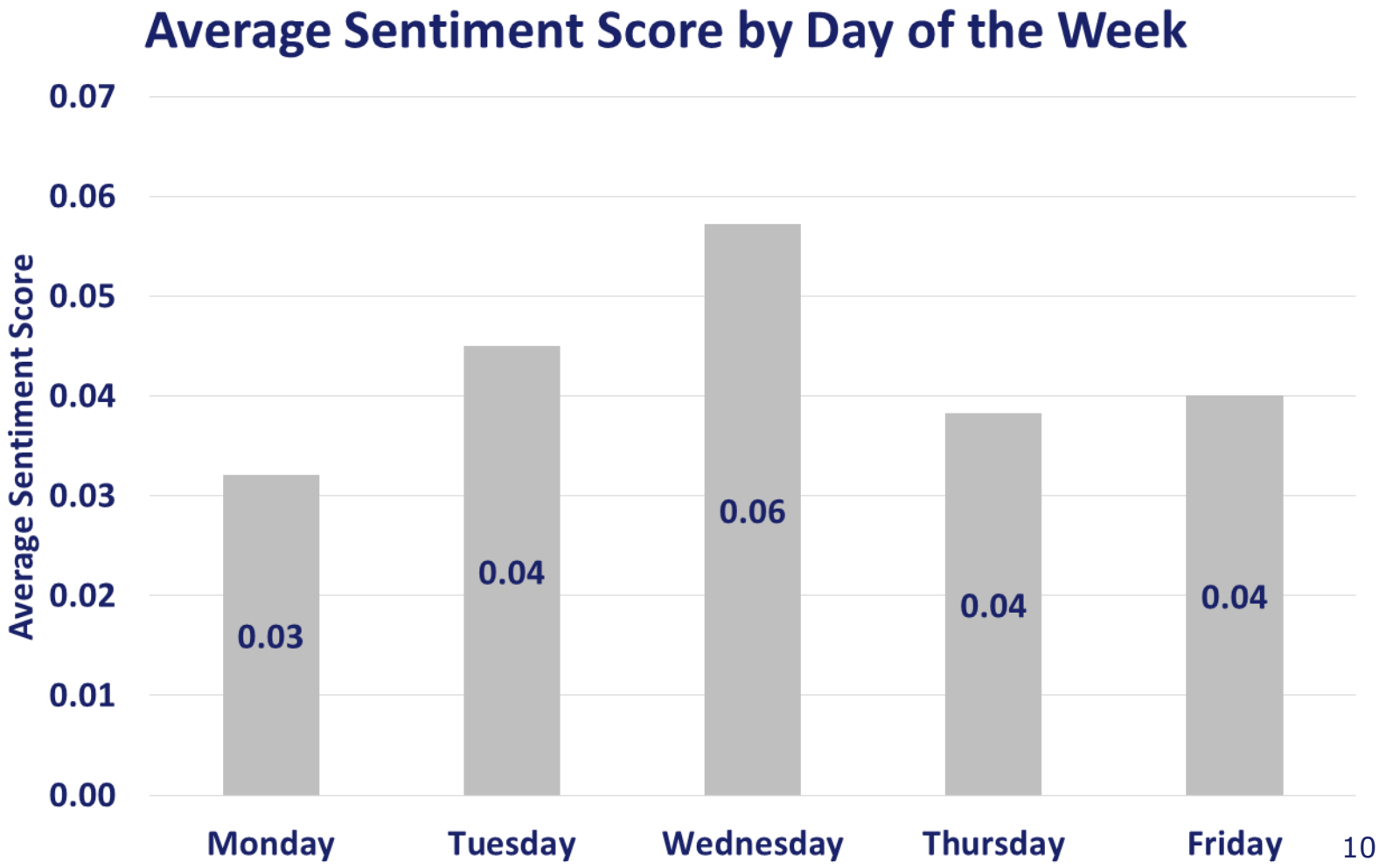
- Project methods
  - This project used an algorithm that counts the number of "positive" and "negative" words and computes an overall sentiment score for each note
  - This analysis used a list of positive and negative words categorized by researchers Hu and Liu in their "opinion lexicon" of about 6,800 words
  - The project was a proof of concept exercise and only analyzed notes from Wisconsin respondents. Sentiment Scoring code was developed by CES staff in both SAS and R; both software systems produced matching results

BLS

Preliminary Results

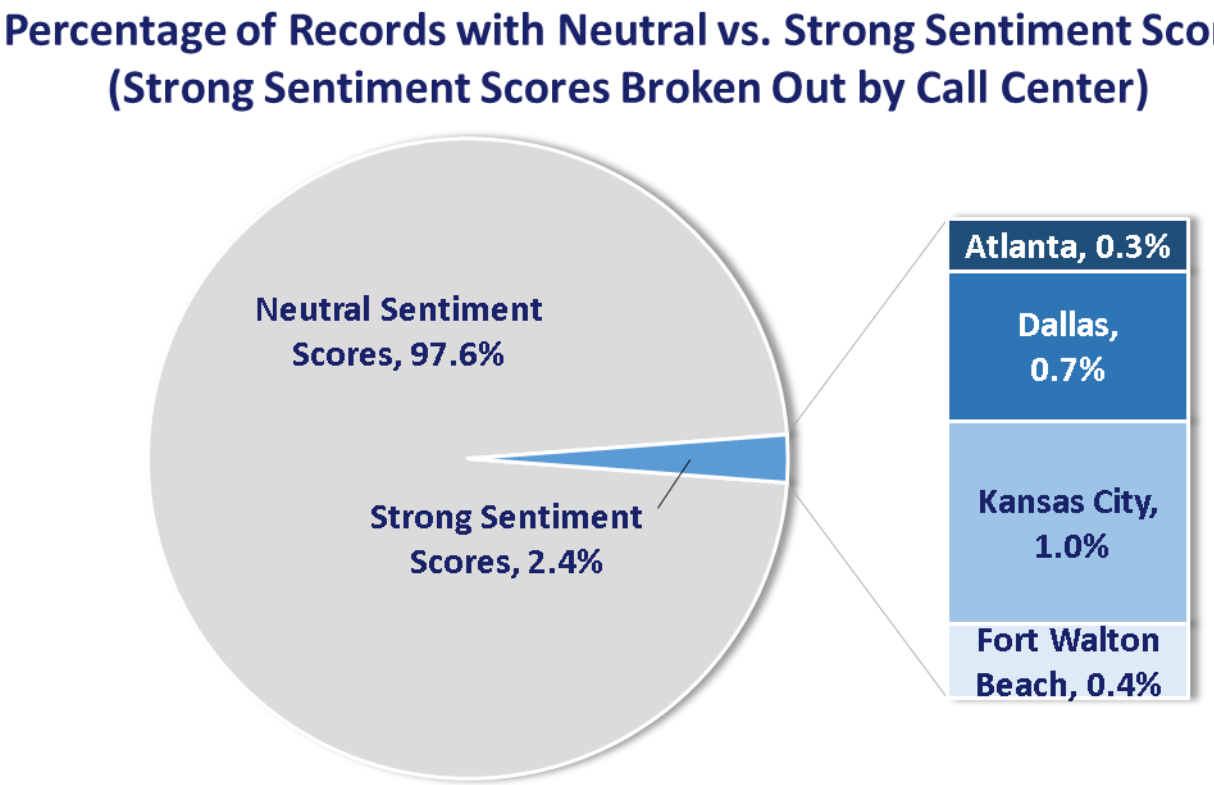


Preliminary Results

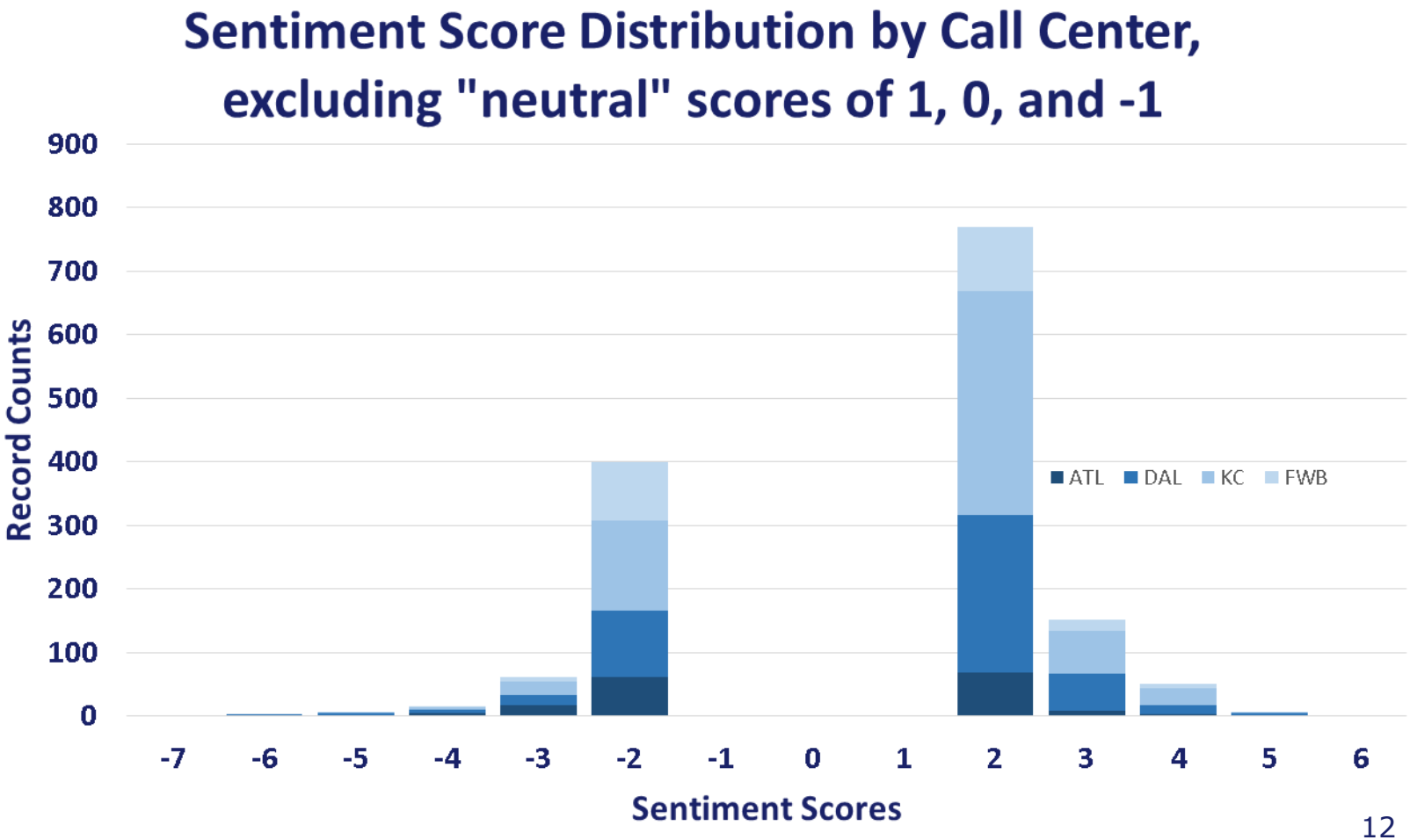


Preliminary Results

- Sentiment scores are mostly "neutral" (zero, 1, or -1)

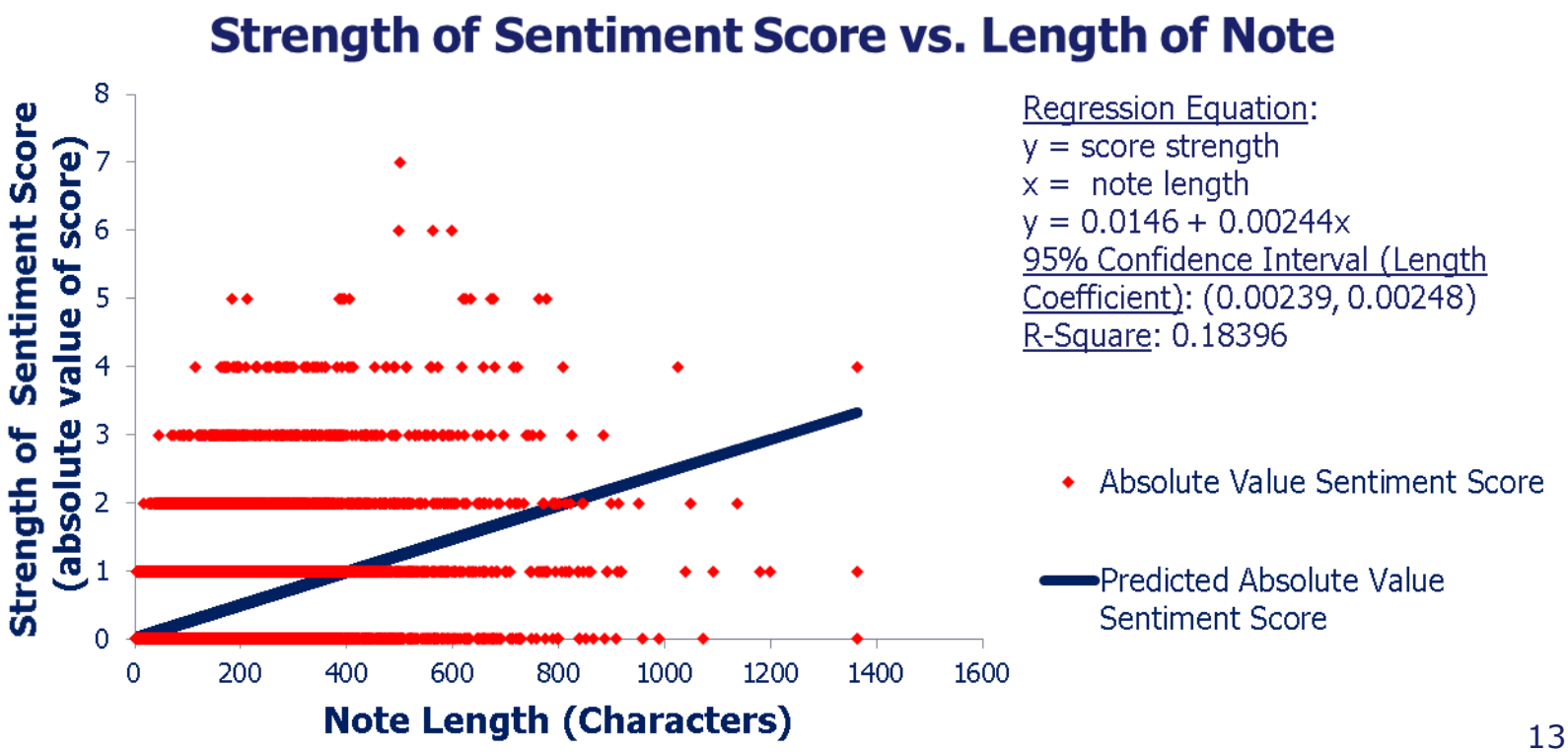


Preliminary Results



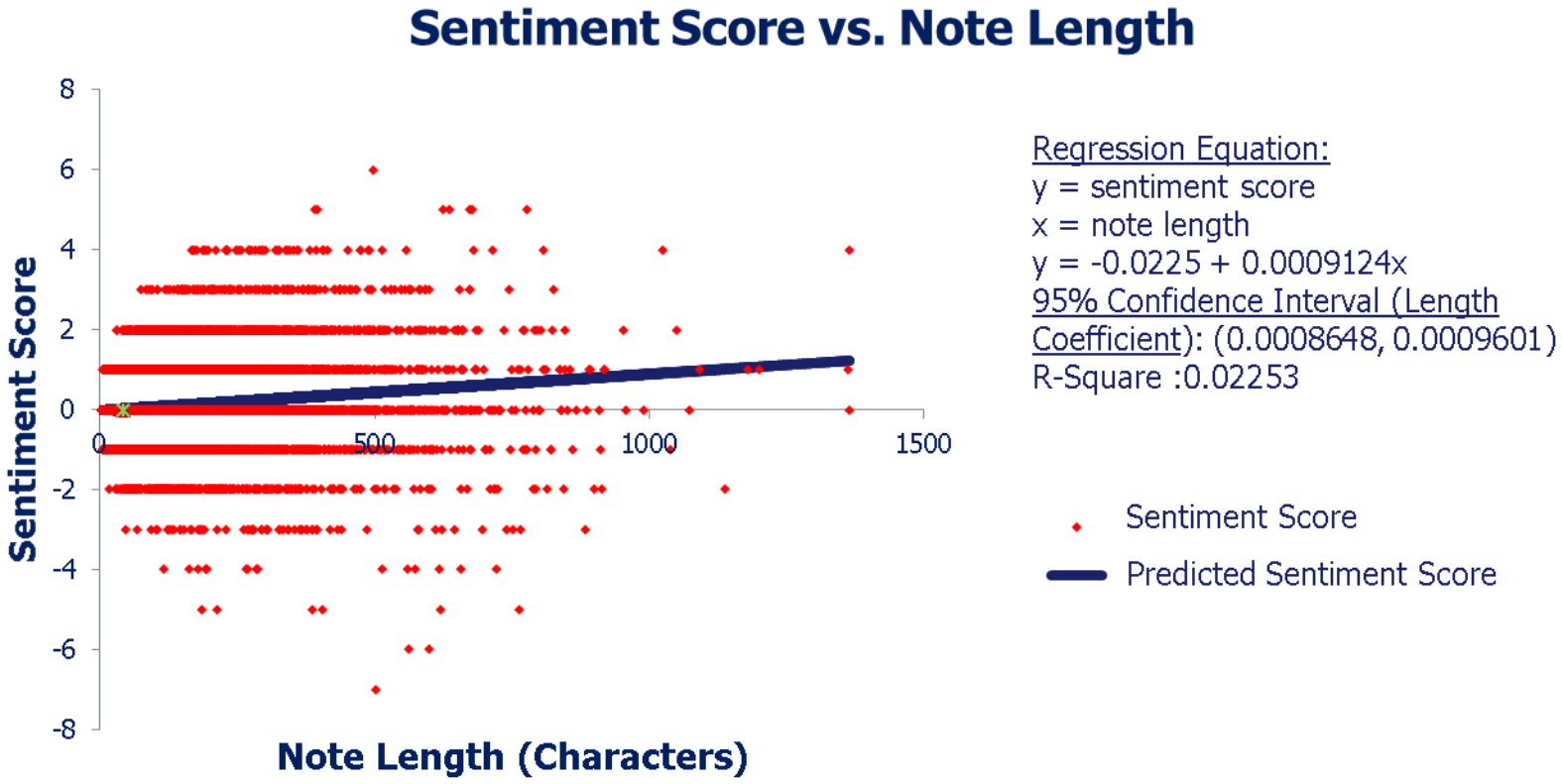
Preliminary Results

- Longer notes have better chance of getting a "strong" sentiment score



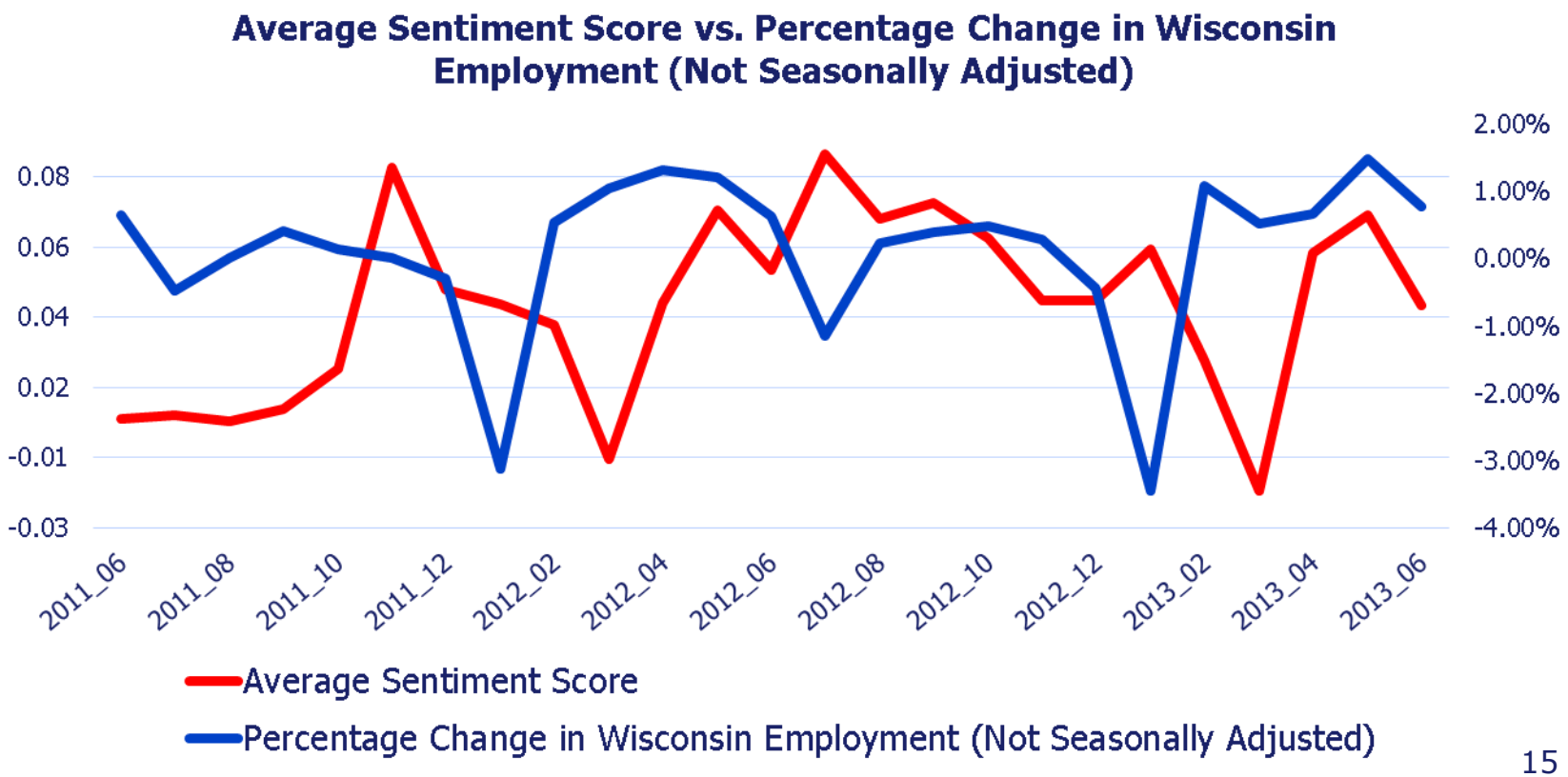
Preliminary Results

- Longer notes also have a better chance of getting a positive sentiment score



Preliminary Results

- Average score compared to change in employment over time



BLS

Conclusions

- Our preliminary sentiment score proof of concept was successful
- Sentiment analysis allowed us to quantify information from existing TopCATI notes
  - Old notes had previously been taking up storage space with unusable qualitative information
- Sentiment analysis algorithms can be applied to existing qualitative BLS data at low cost/resources

BLS

Going Forward

- Expand this research to other states and time periods for further analysis
- More targeted/complex algorithms
  - Creating a CES specific "positive" and "negative" word list
  - Using a scaling system (degree of negativity/positivity of each word)
- Seasonally adjust Sentiment Scores to compare with seasonally adjusted CES data
  - Or compare yearly average of sentiment scores and CES Data

BLS

Contact Information

Matthew Burgess

Economist

OEUS, CES Data Collection Branch

www.bls.gov/ces

202-691-6519

Burgess.Matthew@bls.gov

BLS

BUREAU OF LABOR STATISTICS

U.S. DEPARTMENT OF LABOR

Appendix

- Analysis was done on a sample of about 61,000 notes from Wisconsin
- Stats by call center:

Averages by Call Center								
Call Center	Min Score	Avg Score	Max Score	Avg Length	Max Length	Strong Sentiment	Total Record Count	% with Strong Sentiment
Atlanta	-5	0.02	4	74	958	163	7404	2.2%
Dallas	-6	0.08	5	86	1200	455	13451	3.4%
Kansas City	-7	0.05	5	69	1364	618	28581	2.2%
Fort Walton Beach	-4	0.01	6	58	1049	229	11686	2.0%
Total	-7	0.04	6	71	1364	1465	61122	**2.4%**

BLS