ClaimBuster:
The Quest to Automate Fact-Checking

Chengkai Li
Associate Professor, Department of Computer Science and Engineering
Director, Innovative Database and Information Systems Research (IDIR) Laboratory
University of Texas at Arlington

Arlington Technology Association, Dec. 2nd, 2015
The Innovative Database and Information Systems Research (IDIR) Laboratory

Research areas

- Big Data and Data Science (Database, Data Mining, Web Data Management, Information Retrieval)

Theme of current research

- building large-scale human-assisting and human-assisted data and information systems with high usability, low cost and applications for social good

Research directions

- computational journalism
- crowdsourcing and human computation
- data exploration by ranking/skyline/preference queries
- database testing
- entity search and entity query
- graph database usability

©2015 The University of Texas at Arlington. All Rights Reserved.
Our Computational Journalism Project


- **Story finding**: finding and monitoring number-based facts pertinent to real-world events. The facts are leads to news stories.
- **Fact checking**: discovering and checking factual claims in political discourses, social media, and news.
Publications

- Detecting Check-worthy Factual Claims in Presidential Debates. Naeemul Hassan, Chengkai Li, Mark Tremayne. CIKM 2015, pages 1835-1838.


- Online Frequent Episode Mining. Xiang Ao, Ping Luo, Chengkai Li, Fuzhen Zhuang, and Qing He. ICDE 2015, pages 891-902.


Publications (cont’d)


- iCheck: computationally combating "lies, d-ned lies, and statistics". You Wu, Brett Walenz, Peggy Li, Andrew Shim, Emre Sonmez, Pankaj K. Agarwal, Chengkai Li, Jun Yang, Cong Yu. SIGMOD 2014, pages 1063-1066.

- Incremental Discovery of Prominent Situational Facts. Afroza Sultana, Naeemul Hassan, Chengkai Li, Jun Yang, Cong Yu. ICDE 2014, pages 112-123.

- Discovering General Prominent Streaks in Sequence Data. Gensheng Zhang, Xiao Jiang, Ping Luo, Min Wang, Chengkai Li. ACM TKDD, 8(2):article 9, June 2014.

Publications (cont’d)

- **On "One of the Few" Objects.** You Wu, Pankaj K. Agarwal, Chengkai Li, Jun Yang, Cong Yu. KDD 2012, pages 1487-1495.

- **Prominent Streak Discovery in Sequence Data.** Xiao Jiang, Chengkai Li, Ping Luo, Min Wang, Yong Yu. KDD 2011, pages 1280-1288.

- **Computational Journalism: A Call to Arms to Database Researchers.** Sarah Cohen, Chengkai Li, Jun Yang, Cong Yu. CIDR 2011, pages 148-151. *(3rd place in best Outrageous Ideas and Vision (OIV) Track paper competition)*
CLAIMBUSTER

The Quest to Automate Fact-Checking
Disclaimer

- I am not politically savvy.
- No interest in discussing politics.
People Make Claims All The Time

“… our Navy is smaller than it's been since 1917”, said Republican candidate Mitt Romney in third presidential debate in 2012.

http://en.wikipedia.org/wiki/Mitt_Romney
http://www.thebrainchildgroup.com/
Fact Checking is not Easy

“… our Navy is smaller than it's been since 1917”, said Republican candidate Mitt Romney in third presidential debate in 2012.


http://en.wikipedia.org/wiki/Mitt_Romney

©2015 The University of Texas at Arlington. All Rights Reserved.
Fact Checking is not Easy

“… our Navy is smaller than it's been since 1917", said Republican candidate Mitt Romney in third presidential debate in 2012.

http://en.wikipedia.org/wiki/Mitt_Romney
http://en.wikipedia.org/wiki/United_States_Navy

©2015 The University of Texas at Arlington. All Rights Reserved.
Existing Fact Checking Projects

Journalists and reporters spend good amount of time on fact checking

The U.S. military is at risk of losing its "military superiority" because "our Navy is smaller than it's been since 1917. Our Air Force is smaller and older than any time since 1947."

— Mitt Romney on Monday, January 16th, 2012 in a Republican presidential debate in Myrtle Beach, S.C.

PolitiFact.com, FactCheckEU.org, FullFact.or, Factcheck.org

Numerous Claims to Check. Rise of Fact-Checkers

Republican candidate debate, August 6, 2015.¹
  9 facts checked by factcheck.org
  8 facts checked by CNN
  24 facts checked by PolitiFact

64 active fact-checking sites in 2015, 44 in 2014.²

Limitations of Current Fact-Checking Practices

- Journalists spend hours going through documents to identify claims.
- Research on factual claims can take days (finding data sources, finding experts).
- Significant time gap between speech and reporting times.
The Holy Grail: Automated, Live Fact-Checking
The Holy Grail

Source: Bill Adair
ClaimBuster

- political discourses (debates, interviews), advertisements, live events on TV and online video streams
- social media (e.g., twitter)
- web pages
- news articles

assist data analysis; solicit analyses from professionals

repository of already-checked claims

detector

important factual claims

no match

matched

display existing fact-checkers, delivered via browser extensions, mobile and smart-TV apps

©2015 The University of Texas at Arlington. All Rights Reserved.
ClaimBuster

- political discourses (debates, interviews), advertisements, live events on TV and online video streams
- social media (e.g., twitter)
- web pages
- news articles

assist data analysis; solicit analyses from professionals

repository of already-checked claims

detector

important factual claims

no match

matched
display existing fact-checkers, delivered via browser extensions, mobile and smart-TV apps
Kay Hagan is overly partisan.

Republicans suggest her achievements are thin soup. "The only thing Kay Hagan has accomplished in Washington is becoming an automatic 'yes' vote for whatever new tax or regulation President Obama wants," North Carolina GOP Chairman Claude Pope said.

Original claim made by: 1

Supporting Arguments

- Frank R. Lautenberg(D) and Kay Hagan(D) agreed on 90.02% of the votes they cast between 2011-01-05 (start of session 2011) and 2013-01-01 (end of session 2012)

- Thomas Harkin(D) and Kay Hagan(D) agreed on 92.6% of the votes they cast between 2013-01-03 (start of session 2013) and 2014-05-09 (end of session 2014)

Counter Arguments

Generated Counter Arguments

- Bernard Sanders(Independent) and Kay Hagan(D) agreed on 85.77% of the votes they cast between 2011-01-05 (start of session 2011) and 2013-01-01 (end of session 2012)

- Kay Hagan(D) voted 89.83% of the time with the Democrat party majority vote between 2008-01-02 and 2012-01-02.
In 2011, Miguel Cabrera had 197 in hits, 30 in homeruns, 0.34 in batting average; only 6 other players have ever beaten this record;

Vladimir Guerrero: 197 in hits, 44 in homeruns. 0.35 in batting average in 2000
Todd Helton: 216 in hits, 42 in homeruns. 0.37 in batting average in 2000; 209 in hits, 33 in homeruns, 0.36 in batting average in 2003
Mike Piazza: 201 in hits, 40 in homeruns, 0.36 in batting average in 1997
Albert Pujols: 212 in hits. 43 in homeruns. 0.36 in batting average in 2003
Alex Rodriguez: 215 in hits, 36 in homeruns, 0.36 in batting average in 1996
Larry Walker: 208 in hits, 49 in homeruns, 0.37 in batting average in 1997

Responses

The same claim (i.e. "no more than 6 other players have ever beaten this player's record in some year in 'hits', 'homeruns', 'batting average'") can be made for 41 other players.

The other player are: Albert Belle in 1995 (4), in 1998 (1); Adrian Beltre in 2004 (1); Dante Bichette in 1995 (6), in 1998 (3); Barry Bonds in 2001 (0), in 2002 (0), in 2003 (1), in 2004 (1); Bret Boone in 2001 (6); Ellis Burks in 1996 (2); Vinny Castilla in 1998 (1); Carlos Delgado in 2000 (4); Jacoby Ellsbury in 2011 (4); Darin Erstad in 2000 (0); Nomar Garciaparra in 2000 (1); Adrian Gonzalez in 2011 (4); Luis Gonzalez in 2001 (0); Ken Griffey in 1997 (3), in 1998 (5); Vladimir Guerrero in 2000 (1), in 2002 (4), in 2004 (4); Tony Gwynn in 1995 (2), in 1997 (0); Josh Hamilton in 2010 (3); Todd Helton in 2000 (0), in 2001 (1), in 2003 (1), in 2004 (4); Matt Holliday in 2007 (1); Ryan Howard in 2006 (1); Derek Jeter in 1999
ClaimBuster to be 2016-Ready

2016 Presidential Debates
(Speeches, debates, interviews, social media, news)

Factual claims recommended to be checked

Journalists investigate the claims
(or checked by algorithms, citizens, crowd)

©2015 The University of Texas at Arlington. All Rights Reserved.
Finding Important Factual Claims: A Classification Problem

Human Annotation → Ground Truth → Feature Extraction → Feature Vectors → Learning Algorithm → CLAIMBUSTER → 2016 Presidential Debates → Important factual claims

©2015 The University of Texas at Arlington. All Rights Reserved.
Dataset: Presidential Debate Transcripts

- 20k sentences by presidential candidates: removed very short (< 5 words) sentences
3 Classes of Sentences

Important factual claims

“We spend less on the military today than at any time in our history.” “The President’s position on gay marriage has changed.” “More people are unemployed today than four years ago.”

Unimportant factual claims

“I was in Iowa yesterday.” “My mother enjoys cooking.” “I ran for President once before.”

Sentences with no factual claims (just opinions, questions & declarations)

“Iran must not get nuclear weapons.” “7% unemployment is too high.” “My opponent is wishy-washy.” “I will be tough on crime.” “Why should we do that?” “Hello, New Hampshire!” “Our plan is to reduce tax rate by 10%.”

Goal: Given a future sentence, find the class it belongs to.
Ground Truth Collection

- In 3 months, we accumulated 226 participants.
- Used 600 screening sentences to detect spammers & low-quality participants.
- Admitted sentences which are agreed by at least 2 top-quality participants.
- 8015 such sentences.

<table>
<thead>
<tr>
<th>Class</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>CFS</td>
<td>1673</td>
</tr>
<tr>
<td>UFS</td>
<td>482</td>
</tr>
<tr>
<td>NFS</td>
<td>5860</td>
</tr>
</tbody>
</table>

©2015 The University of Texas at Arlington. All Rights Reserved.
OI: Wages are going up for the first time in a decade.

Will the general public be interested in knowing whether (part of) this sentence is true or false?

- There is **no** factual claim in this sentence.
- There is a factual claim but it is **unimportant**
- There is an **important** factual claim.

Submit  Skip this sentence  Modify My Previous Responses
I was in a state where my legislature was 87 percent Democrat.
Feature Selection

- 6201 features in total
- Used a Random Forest Classifier to calculate importance of each feature.
- Most Important Feature: POS tag ‘Cardinal Number’
Important Features

<table>
<thead>
<tr>
<th>Word</th>
<th>percent; people; jobs</th>
</tr>
</thead>
<tbody>
<tr>
<td>POS tag</td>
<td>noun; cardinal number; past tense; preposition</td>
</tr>
<tr>
<td>Entity Type</td>
<td>Quantity; Country; FieldTerminology; Person</td>
</tr>
<tr>
<td>Concept</td>
<td>United States Senate; Barack Obama</td>
</tr>
</tbody>
</table>
Implementation: Python NLP/ML Tools

Data wrangling
- Use NLTK (Natural Language Toolkit) to transform debate files into structured data format
- Use mysql-python-connector to store extracted features into a MySQL database
- Use matplotlib to plot classifiers’ performance.

Feature extraction
- Use AlchemyAPI (Python wrapper) to extract rich features of sentences

Classification
- Use scikit-learn to build classification models
Evaluation: Classification

- 4-fold cross validation
- Algorithms: Naive Bayes, Random Forest & Support Vector Machine
- Support Vector Machine performed better than others in general.

<table>
<thead>
<tr>
<th></th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>NFS</td>
<td>0.90</td>
<td>0.96</td>
<td>0.93</td>
</tr>
<tr>
<td>UFS</td>
<td>0.65</td>
<td>0.26</td>
<td>0.37</td>
</tr>
<tr>
<td>CFS</td>
<td>0.79</td>
<td>0.74</td>
<td>0.77</td>
</tr>
</tbody>
</table>
Evaluation: Ranking

- Measured accuracy of top-K sentences.
- ClaimBuster has a strong agreement with high-quality human coders on the check-worthiness of sentences.

<table>
<thead>
<tr>
<th>K</th>
<th>P@K</th>
<th>NDCG@K</th>
</tr>
</thead>
<tbody>
<tr>
<td>25</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>50</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>100</td>
<td>0.960</td>
<td>0.970</td>
</tr>
<tr>
<td>200</td>
<td>0.940</td>
<td>0.951</td>
</tr>
<tr>
<td>300</td>
<td>0.853</td>
<td>0.881</td>
</tr>
<tr>
<td>500</td>
<td>0.690</td>
<td>0.840</td>
</tr>
</tbody>
</table>
Case Study: Republican/Democratic Primary Debates in 2015

- Near real-time experiment with 2015 Aug. 6th Republican primary debate
- Transcript grabbed from closed captions of the Fox News channel using TextGrabber
- 1393 sentences
- 71% of the fact-checks from CNN, factcheck.org & PolitiFact were ranked by ClaimBuster within top 18%.
## Case Study: Republican/Democratic Primary Debates in 2015

<table>
<thead>
<tr>
<th>CNN Claim</th>
<th>Associated sentence(s)[From TextGrabber]</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Part of this iranian deal was lifting the international sanctions on general sulemani.</td>
<td>0.415</td>
</tr>
<tr>
<td>2</td>
<td>I would go on to add – &gt;&gt; you don’t favor – &gt;&gt; i have never said that.</td>
<td>0.511</td>
</tr>
<tr>
<td>3</td>
<td>A majority of the candidates on this stage supported amnesty.</td>
<td>0.295</td>
</tr>
<tr>
<td>4</td>
<td>Timely the medicaid is growing at one of the lowest rates in the country.</td>
<td>0.534</td>
</tr>
<tr>
<td>4</td>
<td>We went from $8 billion in the hole to $5 million in the black.</td>
<td>0.773</td>
</tr>
<tr>
<td>5</td>
<td>And the mexican government is much smarter, much sharper, much more cunning and they send the bad ones over because they don't want to pay for them.</td>
<td>0.215</td>
</tr>
<tr>
<td>6</td>
<td>[Not found in the transcript]</td>
<td>N/A</td>
</tr>
</tbody>
</table>
Case Study: Republican/Democratic Primary Debates in 2015

- Real-time experiment with 3 more Republic debates and 2 Democratic debates so far
- Tweeted important factual claims to https://twitter.com/ClaimBusterTM, live!
claims from debates

(1/2) #PolitiFactThis [BERNIE SANDERS?] We have the highest rate of childhood poverty, and we're the only country in the world --

(2/2) #PolitiFactThis virtually the only country that doesn't guarantee paid family and medical leave.

(1/2) #PolitiFactThis It wasn't the bill that I wanted, but yet it turned out to be one of the more significant pieces of

claims in Tweets

@realDonaldTrump If you make $25,001, should you pay $2,500 when $25,000 you pay nothing? Pay only on amt over $25K. #Trump2016 #DonaldTrump

1.3 million bicycles recalled for crash hazard: (KMSP) - Nearly 1.3 million bicycles in the United States are ... bit.ly/1MGt2HO

@richluchette $272 million project, more than 3,000 jobs
Demo

http://idir.uta.edu/claimbuster

©2015 The University of Texas at Arlington. All Rights Reserved.
2016 Republican Party Presidential Debate. Sept. 16, 2015, 7 p.m.
Venue: Ronald Reagan Presidential Library, Simi Valley, California. Broadcasted by: CNN.
Speakers: Dana Bash, Jeb Bush, Ben Carson, Chris Christie, Ted Cruz, Carly Fiorina, Hugh Hewitt, Mike Huckabee, John Kasich, Rand Paul, Marco Rubio, Jake Tapper, Donald Trump, Scott Walker
Transcript Source: http://time.com/4037239/second-republican-debate-transcript-cnn/

Chronological Order | Order by Score

Most Check-worthy

Least Check-worthy

0.18 The eleven leading Republican candidates for president are at their podiums.
0.17 They are ready to face off, and if you've been watching this race, you know anything could happen over the next few hours.
0.16 To viewers who are just joining us, welcome to the Air Force One Pavilion of the Ronald Reagan Presidential Library.
0.29 Our thanks to the staff here and especially to former first lady Nancy Reagan for this impressive setting with Ronald Reagan's presidential plane as our backdrop.
0.29 This debate is airing on CNN networks in the United States and around the world.
0.08 It's also being broadcast on the Salem Radio Network.
0.17 I know everyone is very eager to get started.
0.18 But first, I want to explain the ground rules tonight.
0.18 My name is Jake Tapper.
0.18 I'll be the moderator.
0.24 I will be joined in the questioning by Salem Radio Network talk show host Hugh Hewitt.
0.18 He worked in the Reagan administration for six years.
0.22 And by CNN's chief political correspondent Dana Bash.
0.16 I will ask follow-up questions, I will attempt to guide the discussion.

©2015 The University of Texas at Arlington. All Rights Reserved.
2016 Republican Party Presidential Debate. Sept. 16, 2015, 7 p.m.

Venue: Ronald Reagan Presidential Library, Simi Valley, California. Broadcasted by: CNN.

Speakers: Dana Bash, Jeb Bush, Ben Carson, Chris Christie, Ted Cruz, Carly Fiorina, Hugh Hewitt, Mike Huckabee, John Kasich, Rand Paul, Marco Rubio, Jake Tapper, Donald Trump, Scott Walker

Transcript Source: http://time.com/4037239/second-republican-debate-transcript-cnn/

Most Check-worthy <=1.0<=0.9<=0.8<=0.7<=0.6<=0.5<=0.4<=0.3<=0.2<=0.1 Least Check-worthy

0.24 That’s a fact.
0.33 And when the people of Iowa found that out, I went to No.
0.46 1 and you went down the tubes.
0.29 Governor Walker?
0.13 Jake, yeah, absolutely, I’ll take this on, because this is an issue that’s important in this race.
0.31 Just because he says it doesn’t make it true.
0.19 The facts are the facts.
0.78 We balanced a $3.6 billion budget deficit, we did it by cutting taxes - $4.7 billion to help working families, family farmers, small business owners and senior citizens.
0.23 And it’s about time people in America stand up and take note of this.
0.30 If you want someone that can actually take on the special interest of Washington, which you yourself said you were part of, using the system, we need somebody that will stand up and fight for average Americans to put them back in charge of their government.
0.16 I’m the one who is taking that on.
0.23 I’ll do that as your next president.
0.17 Let’s move on.
0.23 Jake, Jake.
0.13 A phenomenon going on in the race right now is the political...
0.25 OK, Governor Kasich, go ahead.
In fact, today, on the front page of the Wall Street Journal, they fired another 25 or 30,000 people saying we still haven't recovered from the catastrophe.
20 ideas receive support from Knight Prototype Fund for media and information projects

Media Coverage
You are Invited

Acknowledgment

UTA Students

- Naeemul Hassan
- Gensheng Zhang
- Fatma Dogan
- Joseph Minumol

Collaborators

- Bill Adair (Duke)
- Pankaj Agarwal (Duke)
- James Hamilton (Stanford)
- Ping Luo (Chinese Academy of Science)
- Mark Tremayne (UTA)
- Min Wang (Visa)
- Jun Yang (Duke)
- Cong Yu (Google Research)
Acknowledgment

Funding sponsors

Disclaimer: This material is based upon work partially supported by the National Science Foundation Grants 1018865, 1117369, 1408928 and 1565699, a Knight Prototype Fund from the Knight Foundation, 2011 and 2012 HP Labs Innovation Research Awards, and the National Natural Science Foundation of China Grant 61370019. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the funding agencies.
Thank You! Questions?

○ http://ranger.uta.edu/~cli  http://idir.uta.edu  cli@uta.edu

○ Demos  ClaimBuster  idir.uta.edu/claimbuster

○ Please help us label the data  http://bit.ly/claimbusters

©2015 The University of Texas at Arlington. All Rights Reserved.