Kno.e.sis Media Coverage: 2012 through September 2014

1. Lu Chen: « Twitter devient de plus en plus un moyen de vivre sa foi » A US study found that US Twitter users "atheists" would be more active on the social network that "believers" users. La Croix, France, Nov 4, 2014. [Covers Prof. Sheth’s PhD student Lu Chen’s research; mentions Kno.e.sis-Wright State University]*
2. Twitter Data Mining Reveals America’s Religious Fault Lines, MIT Technology Review, Oct 6, 2013 [Covers Prof. Sheth’s PhD student Lu Chen’s research; mentions Wright State University]
3. Atheists tweet more often than Muslims, Jews, Christians, study shows, ReligionNews, Oct 3, 2014 [Covers Prof. Sheth’s PhD student Lu Chen’s research; mentions Kno.e.sis]*
4. Digital volunteers use social media, come together during Jammu floods, OneIndia News, September 20, 2014 [Describes the role of Twitris technology resulting from Kno.e.sis’ NSF funded project for crisis coordination using social media; credited: Prof. Sheth and his PhD advisee Hemant Purohit]
5. Digital soldiers emerge heroes in Kashmir flood rescue, HindustanTimes, September 25, 2014 [Kno.e.sis researcher and Prof. Sheth's PhD student Hemant Purohit is one of 8 initiators/key coordinators of JKFloodRelief whose digital volunteers leveraged Kno.e.sis’ Twitris technology for rescue and relief coordination.]
6. India's Briefcase-Sized Voting Machines, The Atlantic, April 15, 2014 [Penultimate paragraph covers Prof. Sheth’s views/interpretation of data we saw].
7. How big data has changed India elections, CNBC, April 10, 2014. [Mentions some actionable information Kno.e.sis produced using Twitris on actual on the ground use in activities associated with India Elections 2014].
8. Analysis: See what Twitter community thinks of BJP, Congress, AAP, OneIndia News, Apr 10, 2014 [Detailed analysis of view of three main parties in 15 important states; credited: Prof. Sheth and his PhD advisee Shreyansh Bhatt]
9. India's social media election battle, BBC News, Mar 30, 2014 [Includes coverage of our analysis of India 2014 Election using Twitris]
10. Survey: Kejriwal lost popularity after Gujarat visit ahead of election, OneIndia.com, March 11, 2014 [based on our analysis of specific event in India 2014 Election using Twitirs; credited: Prof. Sheth and his PhD advisee Shreyansh Bhatt]*
11. Twitter’s top swear words revealed: Researchers find one in 13 tweets contains a curse, Daily Mail (UK)/Mail Online, February 20, 2014 [A media coverage on our CSCW2014 paper, credit: Kno.e.sis PhD student Wenbo Wang]*
12. #Cursing Study: 10 Lessons About How We Use Swear Words on Twitter, Time.com, Feb 19, 2014 [A media coverage on our CSCW2014 paper, credit: Kno.e.sis PhD student Wenbo Wang]
13. **140 Characters Of F*ck, Sh!t, And @ss: How We Swear On Twitter**, *Fast Company*, Feb 19, 2014 [A media coverage on our CSCW2014 paper; also credited: Kno.e.sis PhD student Wenbo Wang]

14. **Mapping Unstructured Clinical Notes to ICD-10 Coding**, SemanticWeb.com, Nov 26, 2013 [Covers our research on ontology/background knowledge enhanced NLP that is at the core of ezCAC and related products from ezDI; Prof. Sheth and his PhD advisee, Sujan Perera]

15. **Are we missing out on tech-aided disaster management in Uttarakhand?** *The Hindu*, July 17, 2013 [challenges in using 'tech' to support the relief and rebuilding effort; credited: Prof. Sheth and his PhD advisee Hemant Purohit]*

16. **Using crisis mapping to aid Uttarakhand**, *The Hindu*, June 27, 2013 [data resulting from coordination done by Kno.e.sis was used by Google Crisis Maps that helped crisis response planners; credited: Prof. Sheth and his PhD advisee Hemant Purohit]

17. **Twitris: Taking Crisis Mapping to the Next Level**, Tech President, June 24, 2013 [in-depth on use of Twitris in Crisis Mapping/Disaster Response, Political Movement and Campaigns; credited: Prof. Sheth and his PhD advisee Hemant Purohit]


19. **Could Twitris+ Be Used for Disaster Response?** *iRevolution*, Sept 11, 2012 [The influential blog discusses Twitris’ ability to provide insights and situational awareness for emergency response aid, reputation management etc. credits: Kno.e.sis PhD student Hemant Purohit]


22. **Web App Analyzes Tweets in Real Time for a Record of Historic Events**, Mashable.com, Feb 17, 2012 [Reviews Twitris’ monitoring/analysis of major events such as Occupy Wall Street, Egypt revolution 2011, Haiti Earthquake 2010, Iran Elections 2009]*

23. **Twitris Social Media Analysis Tackles Occupy Wall Street, 2012 Elections**, Semanticweb.com, Feb 12, 2012 [Review of Twitris’ social media analysis of social movements such as Occupy Wall Street and India Against Corruption]*

* Items not included in print version
This is a partial list of all media coverage involving Kno.e.sis and its researchers.
Twitter Data Mining Reveals America's Religious Fault Lines

Social media is playing an increasingly important role in religious activity. Computer scientists ask: How is that reflected in the Twitter stream?

Religion represents one of the most powerful forces at work in shaping today's society. So there is a great deal of interest in understanding how religion influences people's views and tolerance of each other and how this affects other aspects of society such as economic growth.

One potential tool in improving this kind of understanding is the Internet, where religious views flourish. In particular, social networks have expanded rapidly in the last few years, playing an increasingly important role in spreading opinions and information in areas as diverse as politics, sport, science, and so on. However, the role that social networks play in religion has been relatively poorly studied.

Now Lu Chen at Wright State University in Dayton, Ohio, and a couple of pals have analyzed more than 250,000 Twitter users in the U.S. who have declared an affiliation with religions such as Christianity, Hinduism, Islam, Buddhism, and Judaism. The results provide some curious insights into the nature of religious activity on Twitter and how these groups of self-declared users differ from one another.

The team began by filtering the biographies of Twitter users in the U.S. using keywords associated with various different religions. That produced a data set of more than 250,000 users from seven different groups: atheists, Buddhists, Christians, Hindus, Jews, Muslims, and undeclared. Christians were by far the largest group with over 200,000 users, while Hindus were the most poorly represented with only 200 users.

Chen and co then calculated the religiousness of each U.S. state according to their data. They then...
compared this to a "ground truth" Gallup survey of religiosity. They say the data sets demonstrate a respectable level of agreement. "They agree on 11 of the top 15 most religious states (e.g., Alabama, Mississippi, and South Carolina) and 11 of the top 15 least religious states (e.g., Vermont, New Hampshire, and Massachusetts)," they say.

However, there are some important differences. According to the Gallup poll, Utah is the second most religious state in the U.S., but one of the least religious states in the group's Twitter data set. That is almost certainly because the group did not filter the biographies for Mormonism, the dominant religion in Utah.

Chen and co go on to ask what kind of information each group tends to tweet. To find out, they collected all of the words that appear in more than 100 tweets in each group and ordered them by frequency of use. That gives them the words that are most positively associated with each religious group.

For Christians, these include words such as Jesus, God, Christ, Bible, gospel, pray, and so on. By contrast, the words most associated with atheists include science, evolution, evidence, religion, Republicans, abortion, and so on. In general, Chen and co say the words most strongly associated with different religions tend to be focused on the nature of the religion itself.

This could have important implications for society. "If our observations were to hold in a broader context, it could be seen as good for society that followers of religious groups differ most in references to religious practice and concepts, rather than in everyday aspects such as music, food, or other interests," say Chen and co.

The team also asks who people in these groups are most likely to follow. Unsurprisingly, members of one religious group are much more likely to follow members of the same group than members of a different group. "Following someone of the same religion is 646 times as likely as following someone of a different religion," say Chen and co.

It is also possible to work out the most popular accounts followed by members of each group, and this reveals some interesting fault lines. For example, the most popular account followed by Buddhists is @Dalailama, while the most popular among atheists is @RichardDawkins. All the religious groups are likely to follow @BarackObama, reflecting their American heritage.

One important question that this kind of work raises is to what extent the behavior on Twitter is indicative of the real world. One potential problem is that Twitter users themselves are not entirely representative of the general population — they tend to be young, male, and urban. What's more, all of the selected Twitter users have a self-declared interest in religion, making them more likely to be highly religious compared to most users. Just how this biases the data isn't clear.

Nevertheless, Chen and co have made an interesting start in teasing apart the behavior of religious users on Twitter. And in pursuing this line of research, there ought to be plenty more low hanging fruit.

For example, analyzing the sentiment displayed in tweets from users of different religions could be revealing, and Chen and co certainly have plans to study these kinds of ideas in more detail. "In future work, we hope to gain clues as to what makes a religion stand out, e.g., when it comes to providing emotional stability or dealing with personal setbacks," they say.

Another interesting possibility is to study tweets in regions of the world that are devastated by religious conflicts. Of course, this kind of work is complicated by all kinds of additional factors, not least of which is language.

Whatever the direction of future research, it is clear that religion and social media are likely to have an important ongoing relationship. As Chen and co point out, much religious activity focuses on the spreading and replication of religious ideas. In other words, social media and religion are ideally suited. That suggests an interesting evolution in the way religions exploit Twitter and other social media platforms — yet another dynamic to study.
In times of natural calamities, relief operations often are hampered due to lack of communication on the ground and between the organisations working engaged in relief operations. Chaos follows and also there happens to be so much of information on the cyberspace that it becomes difficult to filter the necessary information.

It is where the significance of social media comes to relevance. The new media not only connects people from across the borders but also plays a vital role at times of natural disasters. One such example is the role of social media during the recent Jammu and Kashmir floods.

Social Media volunteers come to rescue during Jammu floods
When people of Jammu and Kashmir were helpless and were completely dependent on the state government and the Centre for relief materials, a team of eight digital volunteers came forward to help. Some of them even worked during Cyclone Phailin last year in crisis mapping and response.

These volunteers, present across different time zones, have been making use of the social media and working round the clock to map all the necessary information for the volunteers who are on the ground in Jammu and Kashmir. This time they have used social media more aggressively and effectively to communicate between various organisations and volunteers. These individuals used Twitter and introduced #JKFloodRelief, now the most widely used hashtag to communicate about the relief aspect of the floods on social media, as well as the twitter account @JKFloodRelief. They also started JKFloodRelief.org which has managed to be a channel of communication for relief ops.

What is Twitris?
The team has been using one key technology called Twitris, which has helped to establish communication across various channels. Twitris came from a US National Science Foundation funded project on use of social media for emergency response headed by Prof Amit Sheth at the
Ohio Center of Excellence in Knowledge-enabled Computing (Kno.e.sis) at Wright State University.

By using Twitris, the team used social media, filtered out necessary information, which in turn helped the volunteers to keep a track where people needed to be rescued and where relief materials were supposed to reach. The curated information on the JKFloodRelief blog also managed to create awareness not just among volunteers but people all around.

Hemant Purohit, a key member of this research team and one of the eight initial coordinators of JKFloodRelief.org team, informed that within a week, relief supply comprising 15 tons of food, medicines and several other essentials were delivered due to efficient supply chain established by the team.

Twitris is a comprehensive system which analyses exact details like time, location, people and content and filters this information for further assistance.

**How do volunteers benefit in rescue operations in Jammu?**
The team then, put together some actionable tools, which could be which would be more accurate and would give exact real-time information, accessed on mobile devices. This was used by volunteers to identify those who needed to be rescued and those who were already rescued. It also enabled the volunteers to identify important people or organisations, who could respond to specific needs on ground.

For example the volunteers can mark the tweets with #SoS and indicates ADGPI as the recipient.

Even when telecom towers are down, relatives in other locations are able to tweet the exact location of the person stuck which is picked up by Twitris. Families tweet about the last known details of the relative stuck, such as the hotel name and these are then picked up by Twitris who alert the Army or other officials to rescue the person.

Prof. Amit Sheth, PI of the NSF project on leveraging social media for crisis response at Kno.e.sis, told that by using such tools and technology, JKFloodRelief team have shown how social media can play a crucial role in times of natural calamities.

JKFloodRelief have also received support from many MNCs and well known organisations like Indigo Air, Cipla, Biocon, SpiceJet, Air Asia, Goonj, Uday Foundation and more. They are supported by Twitter, Google, Facebook, Kno.e.sis Center (Wright State University) and Army's public information directorate (ADGPI) for coordination of data efforts.

The team by filtering relevant Twitter data on the Jammu floods made rescue operations easier for the volunteers who were able to locate exactly where people needed to be rescued or needed supplies.

In Kashmir's massive flood rescue operations, a group of digital volunteers - some based as far away as the United States and Singapore - has emerged as unlikely heroes, whose selfless service on social media has helped save thousands of stranded people across the devastated Himalayan region.

The Twitter account, @jkfloodrelief, set up by the group is among the foremost platforms curating and disseminating relief and rescue information from other users and putting it up on their handle and dedicated website www.jkfloodrelief.org. So useful is the information provided by the group that even the Indian army and the National Disaster Management Authority have been monitoring their updates and acting upon them.

Initially, before communications broke down, the group, working across time zones, depended on the relatives of those stranded and contacts on the ground for information. Their experiences during similar disasters in the past - Uttarakhand floods and Cyclone Phailin - encouraged them to start their efforts early and keep them more coordinated. When telecom networks crashed, the group still managed to relay information through the army.

"Our goal is to update on priority needs for relief, donation collection centers, and donation transportation logistics for public awareness on what to help via donations, where and how, and therefore, avoid mismatching of what is needed vs what is offered," said Hemant Purohit, a 27-year-old computational social scientist on crisis informatics at Kno.e.sis Center in Ohio, United States. The center has set up a dedicated 'Twitrus event' which identifies the most used hashtags and most active users, bringing their tweets together in one click.

The group's efforts supplemented those of the army which used social media extensively to coordinate its relief and rescue operations.

Meanwhile, the army has set up a dedicated WhatsApp group connecting all important stakeholders. All information received through social media channels are passed on to this group. Their efforts led to the rescue of a heavily pregnant woman on Tuesday.
"Initially we didn't have a plan in place on to address these (messages on social media requesting help). So, we put in place a WhatsApp group with all major stakeholders and started sharing these messages there to properly facilitate help," said an Army official explaining how the Indian forces are using social media.

Purohit tweets and retweets late into the night about relief and rescue efforts. His colleagues, some based out of Singapore, do the same, in between managing their regular jobs.

The group is amply supported by voluntary organisations and corporates such as Twitter India, Kno.e.sis Center, Google India, Cipla, IndiGo Airlines, DeVil On Wheels, Biocon, and Emami.

"Goonj has supplied close to 2500kg of relief materials, Emami has supplied 560kg of feminine hygiene products, Cipla has supplied 126 cartons of meds, Indigo is supplying a large amount of basic essentials like soap, toothpaste, toothpowder, biscuits, etc. from Mumbai and is on standby to provide foodgrains, Uday Foundation has sent 200kg of relief that includes kids' clothes, woolens and medicines and Biocon is on standby to provide insulin vials," explains Bhavana Upadhyaya, another core member. DeVil On Wheels has helped the organisation build connections with on-ground volunteers.

Social media organisations are not far behind as well.

"We have advised (them) on things like what kind of hashtags to use, what they should name their handle, etc." explained a Twitter official. Meanwhile, Google has launched a Onebox to give helpline information to the Army, Home Ministry, and NDRF Control Room.

© Copyright © 2013 HT Media Limited. All Rights Reserved.
Indian Democracy Runs on Briefcase-Sized Voting Machines

By Matt Ford

A polling official carries an electronic voting machine on April 9 after collecting it from a distribution center in Kerala. (Sivaram V/Reuters)

Holding India's titanic general election is no simple task. Voting is broken down into nine phases—the fifth and largest of which is scheduled for this Thursday—that are spread over six weeks. Over the six weeks, an army of 11 million election officials and security forces will staff and operate more than 935,000 polling stations in India's 543 electoral constituencies, where they will serve almost 815 million registered Indian voters. Central to this undertaking are India's 1.7 million electronic voting machines, or EVMs, the portable, affordable, and highly durable systems that help this massive exercise in democracy run smoothly.

Each EVM comes in two parts. The control unit remains with election officials at each polling place and connects by cable to the balloting unit. When a voter enters a polling booth, an official activates the balloting unit. The voter then presses one of up to 64 blue buttons next to each candidate's name and political-party symbol to cast his or her vote. India's Election Commission has produced a video explaining the process: http://www.theatlantic.com/international/archive/2014/04/indian-democracy-runs-on-briefcase-sized-voting-machines/360554/

EVMs help India overcome a number of electoral challenges. The machines are compact and portable, in contrast to bulkier booth-sized voting machines in the United States and elsewhere. They are also
built to withstand India's diverse and sometimes-harsh climate. Since they run on two 6-volt alkaline batteries, EVMs can be readily used in rural India, where two-thirds of the country's 1.2 billion citizens live, and other areas with limited or no electricity.

The symbol-oriented design also makes voting more widely accessible in a country with 287 million illiterate adults—nearly 37 percent of the worldwide total—and a multilingual electorale that speaks 22 officially-recognized languages and hundreds more unofficial ones. But perhaps the EVM's most impressive feature is its price tag: each unit costs only 10,500 Indian rupees, or about $175. By comparison, even older, used voting machines in the U.S. can cost around $6,000.

After a decade of sporadic and unsanctioned use of EVMs, India legalized the devices in 1988 alongside the existing (and often-maligned) paper-ballot system. They became standard features of elections in 1998 and the sole method for casting votes in the 2004 general election, in which almost 1.1 million EVMs were deployed in polling stations across the country. The Indian government boasts that "EVM has become the leitmotif of the world's largest democratic exercise and gets smarter with each avatar."

Official election materials cite the EVM's superiority over paper balloting by noting the reduction in environmental waste, the speediness of tabulating results, and the decrease in spoiled or improperly cast votes. Another strength, according to election officials, is the EVM's role in combating electoral fraud through "booth capturing"—an ugly tactic where a candidate’s supporters storm a polling place, sideline legitimate voters, and cast ballots—and ballot-stuffing.

But the machines have their limitations as well. EVMs can only record a maximum of 3,840 votes each (the Election Commission says each polling place should only serve about 1,500 voters) and can only list a maximum of 64 candidates at a time to vote for. Because India’s elections are staggered over a six-week period, votes are tabulated in one region and the machines are then reused in another. In March, the Election Commission estimated it would have 1.7 million ballot units and 1.8 million control units—some polling places have more than one ballot unit per control unit—for this year's election. Each Indian constituency is required to keep 10 percent more EVMs than necessary for emergency situations.
Like all electronic voting systems, EVMs also invite concerns about outside tampering. Since implementing the devices nationwide, the Election Commission has insisted that the machines are not susceptible to hacking or other forms of fraud. But a 2010 report by Indian computer-security experts challenged this claim after examining one of the machines and cited numerous vulnerabilities, especially if a malicious user had access to the EVMs in advance. "The technology’s promise was that attacks on the ballot box and dishonesty in the counting process would be more difficult," the report concluded. "Yet we find that such attacks remain possible, while being potentially more difficult to detect."

A few months after the report’s release, Mumbai police arrived at the house of Hari Prasad, one of the researchers, and arrested and interrogated him for hours about where he had obtained the device his team analyzed, before releasing him on bail. (The U.S.-based Electronic Frontier Foundation later bestowed upon Prasad its 2010 Pioneer Award for his work and his ordeal.) After repeated legal challenges by activists, the Delhi High Court ruled in January 2012 that the EVMs weren't tamper-proof and ordered the Election Commission to add a paper trail as an extra security measure against electoral fraud. Election officials pledged to upgrade 600,000 old EVMs to comply with the new guidelines and procure new ones, and voters can now file complaints if there are still problems with the devices.

Occasionally, criticism of the machines takes bizarre forms. During last year’s regional assembly elections in Chhattisgarh, for instance, the Bharatiya Janata Party (BJP) filed a formal complaint after an Indian National Congress party elder allegedly told tribal voters that the EVMs would electrocute them if they voted for non-Congress candidates. The BJP won the election, but the Election Commission’s FAQ now reassures prospective voters that there is no chance of electrocution from "short-circuitry or [any] other reason."

In India, popular sentiment toward EVMs is mixed. Amit Sheth, a professor at Wright State University in Ohio who studies social media’s influence in elections like India’s, ran a preliminary analysis of Twitter users’ feelings about the devices. He found that about 50 percent of the tweets his team analyzed were complaints, though Sheth cautions that the analysis coincided with reports of a faulty machine in an eastern Indian state that reportedly cast votes for the BJP regardless of which button was pressed. "In the last election, there were many claims about Congress tampering with EVM, usually the ruling party gets blamed more often," Sheth told me. "So far, I have not found systemic differences of one party’s view towards EVM than other party’s views."

Despite their drawbacks, EVMs help solve electoral problems that aren’t unique to India. The Indian government provided 4,130 EVMs to neighboring Bhutan last year for its legislative elections, and other developing countries, ranging from Nepal to Namibia, have also imported the Indian-manufactured machines for use in their own contests. Although it’s no panacea for poor governance or repressive regimes, this $150 device from the world’s largest democracy could soon make voting easier in burgeoning democracies worldwide.

This article available online at:


Copyright © 2014 by The Atlantic Monthly Group. All Rights Reserved.
How big data has changed India elections

Neerja Pawha Jetley | Special to CNBC.com
Thursday, 10 Apr 2014 | 6:48 PM ET

With India's 16th national election under way, some parties appear to have an edge as technology, social media and big data play a key role in connecting with voters.

"The era of big data in Indian politics has arrived," said J Ramachandra, CEO of Gramener, a data analytics startup that is working with media companies to digest the huge chunks of election information to provide voters with user-friendly maps of the state of play.

At the opposition Bharatiya Janata Party's (BJP) election backroom in New Delhi, a team of over 100 young techies and consultants have propelled prime ministerial candidate Narendra Modi to the lead position.

Through data analysis they have helped raise funds, rework advertisements and create detailed models for voter engagement in swing states as well as gender and minority voter clusters to increase the power of their micro-targeted strategy.

The result: a data-driven election campaign not very different from that of U.S. President Obama's, albeit somewhat smaller in size, scale and perhaps style.

"We have developed our own customized digital tools based on both commissioned and open source data that puts us in direct touch with voters," said Arvind Gupta, the BJP master strategist behind Modi's 3.67 million followers on Twitter, 12 million likes on Facebook and the party's 68 million page views on Google Plus.

Modi's unique digital events like 'Chai pe Charcha' (Talk over tea) are unprecedented election events that put the political leader directly in touch with people at tea stalls in villages at publicized localities through a combination
"Modi is perhaps one of the most tech-savvy politicians in the world and certainly the most active in India," says Amit Sheth, a professor at Wright State University's Knowledge Computing Center in Ohio.

Sheth, with the help of his students, is mining real-time data on the ongoing elections, which include voter sentiment, emotions and concerns in different constituencies and states. India's political parties then use this data to drive donations, enroll volunteers, and organize resources on the ground to improve the effectiveness of everything from door knocks and phone calls, to micro-messaging and social media.

Technology and data analysis have come to play a crucial role in this election which will use over 930,000 polling booths and 1.7 million voting machines, with 11 million personnel participating.

"The upset win of new incumbent Aam Aadmi Party (AAP) in Delhi polls, a few months ago, has been a catalyst for political parties to take social media data analysis seriously," said Pinstorm Founder Mahesh Murthy.

The AAP, or the Common Man's Party, has been no less constructive in using hard data to develop focused content for virally transmitted messages. Digital marketing companies like Pinstorm have been working with them on data analysis to achieve their strategic goals.

Even within the incumbent Indian National Congress (INC) – the country's oldest and largest political party – the importance of a data-driven campaign strategy has become apparent as the INC beefs up its technology component in what may be a too-little-too-late effort.

Party leader Rahul Gandhi has scrambled to build an online presence; it currently has less than 150,000 Facebook likes, around 10,000 page views on Google Plus and less than 100,000 followers on a Twitter page set up by party workers.

**Two trends underlie the change**

Two irreversible trends underlie the tech-driven fight for votes: a very large young voter base and advances in technology.

Over 100 million new young voters have been added to India's electorate this year.

"This election is about an [inspired], young, urbanizing and educated citizenry who are high on optimism, have tasted the fruits of economic buoyancy, are wired by smart phones and internet and eager to get things rolling on the ground," said R. Sukumar, Editor at Mint newspaper.

Meanwhile, technology has advanced while the field of data analytics has matured. "Data is exploding at breakneck speed all around you. The trick lies in converting it from scattered, hard-to-decipher formats and refining it for strategic goals," said BJP strategist Gupta.

A new era in "electioneering" is here that promises to change the way elections are fought from here on out, analysts say. Gray-haired political strategists who relied on hunches and intuitions to gauge complex demographics of caste, religion, community and localities are on the way out. In their place are quants, programmers and data scientists.

"Data analytics is moving out of research labs into real-time monitoring of people's reaction to politics, policy and rapid responses to crisis situations. It will play a significant role in changing some political outcomes this election – enough of them to matter," says Sheth.
Analysis: See what Twitter community thinks of BJP, Congress, AAP

by Aswathy

Bangalore, April 7: Month of March was full of election campaigns by each of the political party. Based on 900,000 tweets collected from 15 states about three major political parties (BJP, Congress and AAP), our analysis shows how people talked about and reacted to each political party.

Using Twitris, their Collective Social Intelligence platform, Professor Amit Sheth led the researchers at the Ohio Center of Excellence in Knowledge-enabled Computing (Kno.e.sis) at Wright State University processed each tweet to compute sentiment about the mentioned political party. One parameter to measure popularity is to check which political party gets most positive sentiment or least negative sentiment. Just counting negative (or positive) sentiments on a politician provides, as in this Deccan Herald story, provides little useful information about the state of electorate.

Here is our primary analysis:

The BJP is getting most positive sentiments in 8 states: WB, Delhi, Gujarat, Haryana, Kerala, Madhya Pradesh, Rajasthan, Uttar Pradesh.

The party is getting least negative sentiments in 9 states: Delhi, Gujarat, Haryana, Karnataka, Kerala, Madhya Pradesh, Rajasthan, Tamil Nadu, Uttar Pradesh.

The Aam Aadmi Party is getting most positive sentiments in 4 states: Andhra Pradesh, Karnataka, Punjab, Tamil Nadu. The AAP is getting least negative sentiments in 3 states: WB, Jharkhand, Maharashtra.

The Congress is at a loss, finds the survey. The party is getting most positive sentiments in 2 states: Himachal Pradesh, Jharkhand. It’s getting least negative sentiments in 3 states: Andhra Pradesh, Himachal Pradesh, Punjab. The party is getting positive as well as negative sentiments in one state. * See the links at the bottom for state wise data.

The above may fail to indicate true popularity, if that political party has got high number of tweets.

In other words, with high number of tweets a political party receiving most positive sentiment may also have received most negative sentiments. So we have considered a different parameter to get popularity of a political party, which is to consider most positive and (at the same time) least negative sentiments. To be specific, it measures if a party gets most positive sentiments with least negative sentiments which indicates true popularity of that party in that state. Using this, we will be able to exclude the effect of more tweets getting more positive sentiments.

It shows that BJP is being mentioned in most positive manner, at the same time, least negative compare to other political parties in 7 states while congress is mentioned in 1. AAP fails to fetch even one state even though it is mentioned in most positive way in 4 states because at the same time its negative rating is high.

Detailed data to support this analysis can be found here. More questions are analyzed here at Twitris India Election 2014 Insight page. Contact Shreyansh Bhatt, the researcher, and Prof. Amit Sheth, the executive Director of Kno.e.sis, at shreyansh@knoesis.org and amit@knoesis.org for more in-depth constituency level analysis.
INDIA

30 March 2014 Last updated at 20:02 ET

India's social media election battle

By Atish Patel

Delhi

Ahead of the general elections, political parties in India are attempting to woo voters on social media for the first time.

Politicians are taking part in Google+ Hangouts, televised interviews organised by Facebook and using the Facebook-owned smartphone messaging app WhatsApp to connect with millions of tech-savvy urban voters.

India's 16th general election - to be held in nine phases over April and May - will be closely fought, with some observers saying social media will play a vital role in deciding which party wins the most seats.

According to a report published in April 2013 by the Internet and Mobile Association of India (IAMA) and the Mumbai-based Iris Knowledge Foundation, Facebook users will "wield a tremendous influence" over the results of the polls in 160 of India's 543 constituencies.

It's a finding political parties have taken note of, with major contenders like the ruling Congress party and main opposition Bharatiya Janata Party (BJP) earmarking 2-5% of their election budgets for social media, according to an October 2013 study by IAMA and Mumbai-based market researcher IMRB International.

Big data

During the last general election in 2009, social media usage in India was minuscule.

Today, however, Facebook has 93 million users and Twitter has an estimated 33 million accounts in the country.

Many political parties have beefed up their online presence as a result.

The main opposition BJP's prime ministerial candidate, Narendra Modi, was among the first Indian politicians to set up a website and today is on Twitter, Facebook and Google+.

His main rival, Rahul Gandhi, the Congress party's undeclared candidate for PM, however, doesn't have a website and doesn't use any of the three major social networks.

Anti-corruption campaigner-turned-politician Arvind Kejriwal has amassed 1.5 million followers on Twitter since joining in November 2011, a year before he launched his Aam Aadmi Party (AAP) and over two years after Mr Modi, who has 3.6 million followers, opened his account.

"Now no serious politician is seen as being able to avoid social media altogether," said Congress government minister Shashi Tharoor, who until he was overtaken by Mr Modi last July, was the most followed Indian politician on Twitter.

"It does have a significant reach in certain segments of the population and as far as we're concerned, that's important enough to pay attention to and clearly the opposition is paying attention to it too," he added.

Taking a leaf from US President Barack Obama's presidential campaigns, India's parties are using tools to crunch the insurmountable amounts of information social media generates - what's known as big data analytics.

Pinstorm, a digital marketing agency used by some of India's biggest companies to monitor what is being discussed online, now has political parties as clients.
From its Mumbai office, the agency has been collecting, storing and analysing tens of thousands of political statements from over 100 online platforms daily for the past six months to allow parties to find supporters and tweak their political message.

The agency is able to track conversations at national and local level, making it a useful tool for both national and regional parties.

The anti-corruption AAP, taking part in its first general election after an impressive debut in local polls in Delhi last year, uses Pinstorm to "compare how we are faring against others", said Ankit Lal, the party's social media strategist.

Professor Amit Sheth and a team of researchers at the Ohio Centre of Excellence in Knowledge-enabled Computing at Wright State University have also been tracking political sentiment online since July.

He says data collected from social media could in the future replace opinion polls, which many observers say are often rigged in India.

'Dipstick of the elite'
There are some, however, who are doubtful about social media's expected effect.

Social media "is not a true dipstick. It really is only a dipstick of the elite," said Sunil Abraham, executive director of the Bangalore-based Centre for Internet and Society.

Sceptics believe with so many Indians illiterate and lacking internet access, particularly in rural swathes of the country, it is still essential for political leaders to hold rallies and spend on billboard and newspaper advertising to reach the majority of the 814 million-strong electorate.

Parties are also interacting with voters on their mobile devices and it makes sense.

There are more mobile phones in India today than toilets, according to the latest census data, and just over half of the country's 1.2 billion population owns one.

"Mobile is very integral to our strategy," said Arvind Gupta, who heads the BJP's IT and social media cell.

One of the BJP's most unique electioneering tools allows potential voters to listen in on Mr Modi's rally speeches in real time on their phones from anywhere in India. "It's our own innovation," said Mr Gupta.

The number of smartphone users is growing in India and it's how most of the country's web users go online.

That's why WhatsApp, recently purchased by Facebook, is being used by the likes of the BJP and Congress to send photos, videos and messages to potential voters.

"No other medium gives as much mass, simultaneous reach as mobile phones in India today," said Milind Pathak from One97 Communications, a Delhi-based mobile marketing firm.

Political parties like AAP have signed up tens of thousands of members by urging people to give them a missed call for free - party officers then get in touch and formally enrol them as supporters.

"Looking forward, I think the medium will continue to be a heavily-invested area for a political party," Mr Pathak said.

Atish Patel is a Delhi-based independent journalist
#Cursing Study: 10 Lessons About How We Use Swear Words on Twitter

Katy Steinmetz  @katysteinmetz  
Feb. 19, 2014

Besides “a lot”

The history of swearing dates back to at least the Anglo-Saxon days, which means that Twitter has only been around for a smidgen of the time that humans have been leveling verbal filth at one another. But Twitter—which provides a constant stream of raw data about how we communicate with each other—is also giving academics new opportunities to study that old, old habit.

At a conference about social computing held this week, three researchers from Ohio’s Wright State University presented a paper on the subject of cursing, analyzing more than 50 million tweets to find out how much Twitter users curse, when they curse and what types of users are most drawn to four-letter words. As with any research on Twitter, the results come with caveats—like the sample not being completely random and certain attributes for users being unknown. That said, here are 10 takeaways from the study led by PhD researcher Wenbo Wang, with results drawn from some 14 million users:

People curse a lot on Twitter.

Using a carefully selected list of 788 swear words—ranging from standard f*ck to Internet-y @$$—the researchers found that Twitter users curse at a rate of 1.15%, twice the normal rate found in other studies. That might sound small, but that’s more than we use first person plural pronouns, like we, our and ourselves. Researchers have estimated that people typically curse from about 0% of the time (Tipper Gore) to 3% of the time (Chris Rock on a roll).
The most common curse word used on Twitter is the f-bomb.

The f-bomb accounts for nearly 35% of all cursing on Twitter, followed by sh*t, ass, bitch, n*gga, hell and a long string of other things that make grandmothers lose their hair. This should come as little surprise given studies showing that people use 10,000 racial slurs per day on Twitter and widgets showing the many places in the world where people are dropping f-bombs on Twitter right this very moment.

People mainly curse on Twitter when they are not smiles times.

The researchers developed algorithms to assign seven emotions to tweets, practicing the flawed but developing science of “sentiment analysis.” They found that people most often cursed when they were sad (21% of cursing tweets) and angry (17%). But people did also use them to express love (7%), as in “I f*cking love you, man.”

The most popular time to curse on Twitter is after dinner.

Analyzing the time stamps on tweets, adjusted for the users’ time zones, the researchers found that cursing takes off around 5 a.m. and rises throughout the day, hitting its peak at 9 p.m. They also found that there is no hour when people aren’t fouling up the Internet. “People curse all the time throughout the day,” they write, though there are lulls during lunch time. (Bless you and your soothing ways, sandwiches.)

People curse less on Twitter after hump day.

The researchers analyzed random samples of tweets sent out over four weeks and found that there are “relatively high cursing ratios on Mondays, Tuesdays and Wednesdays,” after which ratios decrease until Saturday and then start rising again on Sunday. There can be something damnably oppressive about the work week being just on the other side of bedtime.

Retweets contain the most curse words.

Previous research, the authors note, has found that “profane comments are more popular or more widely read than non-profane comments” on social media. Retweets won this dubious popularity contest compared to tweets they categorized as “starters,” those one writes with the intent of starting a conversation; updates, messages mentioning no one with no particular expectation of response; mentions, tweets referring to someone without expecting a response; and replies.
Twitter users swear more in more “relaxed” environments.

Complementing previous research, the study found that people are more likely to speak like a sailor when they’re at places like a personal residence than a place of work. The researchers determined where people curse the most by meshing Foursquare data with the latitudes and longitudes on tweets to which users added a location. Cursing while traveling was the least popular, which may come to a shock to anyone who has engaged in air travel since Pan Am’s glory days.

Men tweeting at men are most likely to curse.

Using data from the U.S. Census Bureau listing the most popular girls’ and boys’ names in the U.S., the authors assigned a gender to the users and analyzed how much they curse in various conversations. Previous studies have also shown that people tend to curse more in “same-gender contexts” like being out with the ladies or the boys. Men were much more likely to use the f-word overall; females were much more likely to use the word bitch, perhaps while trying to reclaim it.

The only people who curse less than the most popular Twitter users are the least popular Twitter users.

Other studies have shown that the higher someone’s social ranking, the less likely they are to utter words you can’t say on television. This one found that to generally be true; the “top 1%” of users, with an average following of about 68,000 users, cursed less than every other group except the most lonely 10%, who had an average of 2.3 followers.

Slut is a term of endearment on Twitter.

Looking at that follower-based social ranking, the researchers found that cursing tweets containing the word slut were most likely to be directed at the top 1%, where celebrities reside. This, they posit, “is because some fans like to call celebrities slut regardless of their gender for fun” and present an eloquent example: “@Harry_Styles,” it begins, “slut drop on my follow button.”

This is an edition of Wednesday Words, a weekly feature on language. For the previous post, click here.
140 Characters Of F*ck, Sh!t, And @ss: How We Swear On Twitter

At best, Twitter is a place to find breaking news, thoughtful dialogue, and unexpected voices. At its worst, it’s a forum for knee-jerk reactions, trolls, and harassment. At both of these extremes, as in life, there’s bound to be some bad language along the way.

Just how much do people curse on Twitter? Who swears and what do they say?

Four mild-mannered researchers endeavored to find out, and the paper they produced—presented this week at the ACM Conference on Computer-Supported Cooperative Work & Social Computing—is a compendium of offensive language that rivals the depraved, cuss-filled brilliance of the recent film The Wolf of Wall Street.

![Cumulative distribution of curse Words](image)

**Figure 2. Cumulative distribution of curse Words: The top 7 curse words cover more than 90% of all the curse word occurrences.**

After examining a random one-month sample of 51 million English-language tweets from 14 million distinct user accounts, they came up with this conclusion: We curse a lot on Twitter, where our language is usually public, even more than we do in real life. Even more compelling, they discovered the underlying context of when and why cursing happens and who is cursing to whom.

“It’s a sizable fraction of the words we use. On average, one tweet out of 13 tweets will contain at least one cursing word,” says Wenbo Wang, a PhD researcher at Wright State University who led the study. “Because of social media, people don’t see each other. They can say things they wouldn’t say in the physical world.” Other studies have found that 0.5 to 0.7% of words we say in the physical world are curses—on Twitter, the researchers found the rate to be 1.15%. Or as the paper reads, and as Wang was too polite to repeat during our phone interview:
The most popular curse word is fuck, which covers 34.73% of all the curse word occurrences, followed by shit (15.04%), ass (14.48%), bitch (10.34%), nigga (9.68%), hell (4.46%), whore (1.82%), dick (1.67%), piss (1.53%), and pussy (1.16%).

The findings are interesting for anyone who uses Twitter, but for the team, all affiliated with Ohio Center of Excellence in Knowledge-enabled Computing, the paper will fold into work with broader societal implications related to mental health, verbal abuse, online harassment, and gender differences in online communications.

![Figure 6. Cursing Volume in Different Days of Week](image)
“Social content is extremely rich,” says the center’s director Amit P. Sheth. “The cursing issue is an expression of sentiment and emotion… it’s kind of a core issue of understanding the language.” The center is working on developing automated tools that could flag issues of worrisome harassment on social media, especially in high school and college years, or could identify depressive disorders or disposition to violence. Creating filters for kids on social media is also another potential application. Of course, timing and context is everything, and not all cursing is negative. For example, “I f*$%-ing love you” could be music to the @ recipient’s ears.

An attempt at “sentiment analysis," which has its limitations, revealed that while negative emotions won out in swearing tweets, love and playfulness (two friends saying "you whore") emerged as real signals in the data. Anger and sadness represented 22% and 17% of cursing tweets, whereas 7% seemed to express love. One in four of all tweets sampled that were classified as “angry” contained curse words.

Other insights in the paper involved timing, location, gender, and level of influence of the NSFW tweeters. People curse more and more as the day passes, reaching a peak at 12 a.m. to 1:30 a.m. before bedtime, and Mondays, Tuesdays, and Wednesdays contain the most curse words relative to tweet volume.

And as in real life, people who are in more relaxed environments like home or at a club--rather than in the office--are more likely to curse in a tweet, but the differences are smaller than in the face-to-face world (the researchers looked only at geo-located tweets for these). College and high school students aren’t shy about cursing, however, even when they are at school. Men curse more than women, but both genders are more likely to curse when directly conversing on Twitter with someone of the same gender. As for social rank, celebrities in the top 1% of follower counts on Twitter get treated better than “middle ranked” Twitter users:

The cursing ratio among tweets received by the top 1% group is the lowest across all recipient groups: these popular users receive a lot of friendly messages from their fans, e.g., “@Harry Styles follow me babe<3”, “@NiallOfficial I can’t sleep :(!”

The researchers do say there is room to improve their classification system, as the endeavor of even pinning down curse words isn’t as straightforward as it seems. After removing spam, the team had to code each tweet as “cursing” or “non-cursing.” That meant deciding what a curse is. For this, the four authors--none of whom are native English-speakers--compiled a lexicon of offensive words and asked two undergraduates (who else?) to assist in resolving ambiguities. They realized “gay” can both be used as a slur or in common descriptive speech. The lexicon also had to be modified to include all sorts of variations: “e.g., a55, @$$, $h1t, bltch, bi+ch, c0ck, f*ck, l3tch, p*ssy, and dik.”

“I think our vocabularies have increased,” says another one of the other authors, Lu Chen.
Help For HealthCare: Mapping Unstructured Clinical Notes To ICD-10 Coding Schemes

By Jennifer Zaino on November 26, 2013 6:52 AM

The health care industry – and the American citizenry at large – has been focused of late on the problems surrounding the implementation of the Affordable Care Act, the federal website’s issues foremost among them. But believe it or not, there are other things the healthcare industry needs to prepare for, among them the October 1, 2014 date for replacing the World Health Organization’s International Statistical Classification of Diseases and Related Health Problems ICD-9 code sets used to report medical diagnoses and inpatient procedures by ICD-10 code sets. ICD-9 uses 14,000 diagnosis codes which will increase to 68,000 in ICD-10, which is a HIPAA (Health Insurance Portability and Accountability Act) code set requirement.

Natural language processing has had the primary role in many solutions aimed at transforming large volumes of unstructured clinical data into information that healthcare IT application vendors and their hospital customers can leverage. But there’s an argument being made that understanding unstructured text of clinical notes that contain a huge stash of information and then mapping them to fine-grained ICD-10 coding schemes requires a combination of NLP, advanced linguistics, machine learning and semantic web technologies, and Amit Sheth, professor of computer science and engineering at Wright State University and director of the Kno.e.sis Center is making them. (See our story yesterday for a look at how the NLP market is evolving overall, including in healthcare.)

“ICD-10 has thousands of codes with millions of possible permutations and combinations. A rule-based approach is not effective to cover the huge number of ICD-10 codes.” Sheth says. Extracting the correct concepts, identifying the relationship between these concepts and mapping them to the correct code is a major challenge, with codes often formed by information from various sections of a clinical document that itself is subject to individual physicians’ style of recording information, among other factors.

Sheth gives one example of how easy it is to miss things: In the sentence: “He is having severe inflammation of appendix and peritoneum,” the typical rules-based NLP engine on its own doesn’t recognize that inflammation of appendix means appendicitis, and inflammation of peritoneum equals peritonitis (see graphic below). “Without a knowledge base, without an ontology, the NLP engine did not have enough context” to make the match to ICD-10 code K35.2, he says. When NLP is combined with semantic web technologies, however, it can help machines identify the right...
concepts, identify interconceptual relationships, understand and disambiguate abbreviations, and continuously learn and improve. He and a team of students at Kno.e.sis set about to raise the level, and their work today is the basis of the cloud-based coding product ezCAC from ezDI, which launched late last month.

Rules-based NLP solutions also require that humans input the rules – and then re-input them when exceptions occur. That takes time, which means that users have to wait as long as months for a new version of their solution to be available, he notes. That wait can be eliminated when “instead you take the NLP engine and use an ontology which has the relationships that are the forms of rules, and you get the same benefit. All you have to do is to change the knowledge base,” Sheth says, an easier proposition than recoding. Once that’s done, the next time the NLP engine with an ontology in its rack looks at a text it is equipped to have the appropriate understanding and make the right identification. Additionally, this approach avoids the costs of having to maintain staff focused on maintaining the rules.

The work at Kno.e.sis also delivered unique IP in the way of an ability to assess the richness of a knowledge base with respect to a given corpus, determine the missing domain relationships in the knowledge base, and suggest the most plausible relationships that can fill the gap created by the missing relationships. “It’s highly semi-automated in the sense that humans make the ultimate call. But it helps for keeping the ontology up with new medical knowledge at very low cost.”

Translating clinical data to the ICD-10 codes matters because, as much as any other organizations, hospitals are driven by concerns about revenue cycle management. “Hospitals don’t get paid unless they encode the bill right and send it to the insurance company,” he notes. “They need some computer-assisted coding to improve their operations.”

RELATED:
Tags: advanced linguistics, Affordable Care Act, Amit Sheth, ezCAC, ezDI, healthcare, healthcare codes, HIPAA, ICD-10, ICD-9, Kno.e.sis, Machine Learning, medical, natural language processing, NLP, Ontology, semantic, Semantic Web, World Health Organization's International Statistical Classification of Diseases and Related Health Problems, Wright State University
Technology-driven efforts are on to crowdsource information using the Internet and other communication channels and integrate it into online maps to help manage the aftermath of the natural calamity that has beset Uttarakhand.

The international experience in recent years has proved that ‘crisis mapping’ can make a difference to disaster situations.

One such Uttarakhand flood relief effort, that aims to make use of ‘digital volunteers’ in India and other parts of the world to aggregate information from diverse sources and make it more useful and actionable has been launched by Hemant Purohit, a Crisis-Response Coordination researcher at the Ohio Center of Excellence in Knowledge-enabled Computing (Kno.e.sis), Wright State University.

And there exists an International Network of Crisis Mappers, which includes members with different skill sets, and experience in the use of tools covering crowdsourcing, mapping, use of aerial and satellite imagery, geospatial platforms, advanced visualisation and computational and statistical models.
What the volunteers are doing is to monitor different channels of information on Uttarakhand, including official sources, blogs, social media, non-governmental organisations, public networks and the news media to generate ‘situation reports’ and also update with vital information an online crisis map set up by the GoogleCrisisResponse team (http://google.org/crisismap/2013-uttarakhand-floods?gl=in). The map has information on rescued people, cleared areas, people stranded, relief camps, medical centres, road networks and so on.

Information flow will gather momentum as the repair and restoration of the mobile phone network progresses in Uttarakhand.

Google had also set up an instance of its webapp, Person Finder, that makes it possible for information about missing persons to be posted online and searched, with the option of triggering alerts.

The role of crisis mapping was to bridge the gap that existed between information-seekers and providers, particularly when it came to providing insights into the situation on the ground and the action that needed to be taken, Mr. Purohit said in an e-mail interview to The Hindu.

**Haiti earthquake**

The world got a glimpse of the potential of crowdsourced mapping following the deployment of an open-source platform called Ushahidi when an earthquake hit Haiti in 2010. The information provided by the affected population and others over the Internet and the mobile phone network and the map-based capabilities of the platform helped in addressing specific disaster management requirements.

In the case of the Haiti earthquake, a volunteer team got the platform going fast and were soon tapping into social media sources like Twitter, facebook, and blogs and other media to create actionable reports. Later an international SMS number was created for people to input information relating to the quake. Soon it turned into a flood which was painstakingly processed by volunteers and turned into information that could be loaded onto an online map.

As many as 4,636 project volunteers translated 25,186 SMSs and numerous e-mails, web, and social media communications, resulting in 3,596 reports that were actionable and included enough relevant information to be mapped on Ushahidi, said a United States Institute of Peace report.

Now the trend was to try to use automation to process such large volumes of information.

“Until recently, much of our crisis response and disaster coordination relied on manual effort using web infrastructure. Ushahidi is an excellent and successful example. More recently, there is rapid progress on our ability to use automation — especially in processing large amount of information shared on social media,” said Amit P. Sheth, director, Kno.e.sis Center. “We are at a stage where we should combine human effort with machine processing,” he said.

Keywords: Uttarakhandfloods, flashfloods, landslips, Uttarakhandlandslides, Himalayanecosystem, CharDham yatra, Uttarakhandrescue, disastermanagement, IndianArmy rescue, Badrinath, Kedarnath, Rudraprayag, Pauri, Himalayanrivers, NationalDisasterRelief Force, Gaurikund, HemkundSahib, Uttarakhandpilgrimage, Googleapps, crismapping, GoogleCrisis Response, webapp, GooglePersonFinder, socialmedia

Twitris: Taking Crisis Mapping to the Next Level

Twitris – the techy marriage of “Twitter” and “Tetris” – is a platform that aims to help civil society win the “game” of big data, creating layers and layers of analysis that provides a holistic picture of an event. The idea for Twitris was born out of the chaos of the terrorist attacks in Mumbai on January 26, 2008 when Amit Sheth and his team of PhD students at the Kno.e.sis center noted that social media users played a key role in feeding information to the media.

Kno.e.sis stands for the Ohio Center of Excellence in Knowledge-enabled Computing at Wright State University in Dayton and Sheth is the founder and director of the center where he oversees the Twitris project.

The platform uses algorithm-based technology to aggregate existing information on Twitter, Wikipedia, and news sites in order to provide a fuller picture of events, disaster scenarios, as well as political movements and campaigns. The Kno.e.sis team works closely with social scientists in order to understand behavior and what types of information would be useful to aggregate. They also work with civil society groups, like humanitarian organizations, to understand how this information can be effectively applied during a disaster to speed up aid delivery.

Crisis Mapping v. Twitris

Crisis Mapping emerged as early as 2004 after the massive Indian Ocean Earthquake and Tsunami devastated Aceh, Indonesia. It was also used in 2010 during the Haiti Earthquake. Patrick Meier, one of the key figures that coordinated the Haiti mapping project, used social media and crowdsourcing to display information on food distribution, sanitation, displacement and security. It was used again in 2011 to map the Japan Earthquake and even the progression of the Gaza-Israel War in 2012.

The key difference between Twitris and crisis mapping is that it provides the context and background to understand what is happening across social media and therefore, allows for a deeper analysis of online data. Hemant Purohit, a PhD candidate at Kno.e.sis and one of the Twitris developers, describes the platform's three key components as “people, content, and network analysis.”

In other words, to gain a comprehensive understanding of an event, Twitris first mines social media for relevant live tweets. It then provides background information from sources like Wikipedia and news sites. Lastly, it analyzes interactions on social media to understand existing networks or a lack thereof.
"These three sets of information presented together give you a complete picture," says Purohit. In a way, Twitris behaves like a more comprehensive version of a magic eight ball that provides critical answers during a crisis: What if there is an informal resource center that has been set up ten blocks away from a disaster victim who needs food and fresh water - how would the two connect? What if a humanitarian aid worker lands in Uttarakhand, India and needs to better understand the region to help those on the ground?

While the platform is still being refined, this human informed, algorithm-based technology has now been applied to a number of events.

Disaster Response

Sudden surges in social media activity often follow large events and disasters, most recently after a series of Tornadoes slammed Oklahoma. “You have all these people talking and some post very important information,” explains Purohit. “How do we mine these users? How do we represent these users in a meaningful way?”

Using an algorithm, based on the number of retweets, mentions and replies, the Kno.e.sis team used Twitris to find the top 100 most influential and well connected users of social media and list them on the site with their profile information. The top 100 often consist of professionals working across a variety of sectors: academia, media, humanitarian work, politics and medicine, for example. Twitris allows users to look at communication patterns among these “influencers,” allowing a humanitarian aid worker, for example, to quickly activate the help of this network and speed up emergency response during a crisis.

Twitris also enables social media users to interact more effectively with each other through a matching program, allowing one person in need to find resources being offered by another. However, the Kno.e.sis team is still working on creating a more effective matching program and to track the number of successful matches made so far.

ABOVE: Twitris helped match tweets calling out for donations to the Oklahoma disaster to those requesting donations. (image: http://twitris.knoesis.org/)

Twitris also allows users to see what is trending on social media, a useful tool for journalists to nab their next headline. It can also help disaster victims find aid. Purohit gleaned a number of help line numbers during the Oklahoma tornadoes by looking at the trending tweets. He then posted this information at the top of the Twitris site page for easy access.

Trending topics can also point out answers to key questions. For example, why did northeast India suffer so severely from the recent floods? One of the trending topics noted: the monsoons came earlier than normal this year and the administration simply wasn’t prepared for it.
Finally, Twitris provides real time monitoring of tweets as they come in second by second. Flags appear on a map, showing the location of the tweets.

The Kno.e.sis team is hoping to work more closely with emergency responders in utilizing Twitris. They are currently working with Google Crisis Response on the floods in India and are also conducting a study with the National Center for Medical Readiness, one of the organizations that trains the air force. By the end of the year, Twitris hopes to understand how the air force's command control can use social media.

**Political Movements and Campaigns**

By analyzing social media, Twitris can provide a simple visualization of Occupy Wall Street's (OWS) performance, city by city. For example, Twitris revealed an interesting facet of the OWS movement in L.A. versus Chicago: While OWS L.A. organized and networked successfully, OWS Chicago did so on a much lesser extent.
ABOVE: OWS L.A. v. OWS Chicago networks (image: http://twitr.is.knoesis.org/)

Another method of organizing data, which Twitr is still developing, allows users to ask simple questions. For example, Twitr wanted to gauge how OWS might be viewed in light of past movements and asked how many historical figures were mentioned during OWS protests. Rosa Parks was mentioned 639 times and Howard Zinn, an American civil rights historian, 415 times. This type of human interaction generates even greater insights, Purohit notes.

During the 2012 presidential elections, Twitr tracked the way “influencers” responded to electoral events, creating a map of positive, negative, and neutral reactions to each electoral debate or event.

ABOVE: The top 100 influential social media users and how their interactions play out over the course of the 2012 presidential election. (image: http://twitr.is.knoesis.org/)

Finally, the “popular perceptions across geography” tool provides a map of sentiments and opinions across regions as specific as a city or a state. “It is extremely important for something like U.S. elections to know what are the red and blue states saying,” explains Purohit.
Verifying the Data

One gaping issue in handling public data is verifying its authenticity.

Twitris addresses this issue by applying aggregation techniques in order to find trusted information. Think of it as ox-weighing at the 1906 county fair. As the anecdote goes, over 800 attendees were asked to guess the weight of an ox. While no one person guessed correctly, the average response was quite accurate - within one pound of the weight of the ox.

Images and videos also help to verify facts and figures. Patrick Meier, who is now the Director of Social Innovation at the Qatar Foundation’s Computing Research Institute, is creating a tool called Verily that will determine whether images have been altered. It has not yet been incorporated into Twitris.

“But because information spreads so fast, if something is wrong with the videos or photos, people can immediately verify,” says Purohit. “I have seen some tweets pointing out the original source of a photo. I have seen people use a Google image search to verify photos.”

In the overwhelming world of big data, Twitris may prove a powerful tool in getting civil society to understand social media and open data in a meaningful way.

Personal Democracy Media is grateful to the Omidyar Network and the UN Foundation for their generous support of techPresident’s WeGov section.
Election 2012: The Semantic Recap
By Jennifer Zaino on November 8, 2012 2:17 PM

There’s no such thing as too much post-election coverage, is there? Alright, maybe there is. But we couldn’t let things die down without at least a nod to those in our space that have delivered the semantic industry’s own take on the topic. Here are a few you may want to review: Twitris Election Insights: “The Twitris system had an amazing night—while Nate Silver’s model might have received well deserved attention, Twitris gave better indications and insights and large majority of the polls,” wrote Dr. Amit Sheth, Kno.e.sis Ohio Center of Excellence in Knowledge-enabled Computing director and LexisNexis Ohio Eminent Scholar, in an email to us. The semantic social web application (first covered here) is a project of Kno.e.sis at Wright State University.

As no one could have missed, there’s been considerable attention paid by the media to whether President Obama’s performance post-hurricane had an impact in the election results. And, among the insights Twitris pointed out following Hurricane Sandy, and a couple of days prior to the election, was that “the structure of interaction network of top 100 influencers in the two topical communities [of each candidate] provides insights about increased positive cohesiveness for President Obama.”

During the election, Sheth also pointed to another trend Twitris spotted that also pointed to the final outcome: that it recorded a cross-over in sentiment for swing states Florida, Virginia, Ohio, Colorado and Florida. “If this holds, we are wrapping up the results by 11pm!” he wrote. The TV networks began projecting President Obama as the winner at 11.20 pm.
Parse.ly's Web Wide Trends:

Speaking of Hurricane Sandy’s possible impact on the election, a company blog published right after Hurricane Sandy by predictive content analytics platform vendor Parse.ly (which we first covered here) discussed its Web Wide Trends feature — and what it told about who got what coverage in the midst of the disaster. Parse.ly is used by publishing outfits such as Reuters, U.S. News and World Report, and The Dallas Morning News to help them understand what stories are gaining steam or flagging, and why, with the help of NLP technology, metadata and visualizations.

Says the post by Parse.ly business intern Jason Bell, the feature shows that “over the days leading up to Sandy, Obama was actually trailing Romney in terms of total media volume. Thanks to Sandy, however, Obama has gotten a quantifiable press and media coverage bump.”

Bell also mentions the launch of Parse.ly News, built by the company CTO during the storm, which monitors the latest dispatches from top online media front-pages. Users can tour what’s live at The New York Times, Chicago Tribune, The Washington Post, and many more, for a scan, and click through to the publication if they want to go deeper than the headlines.

Presidential Election News & Twitter Tracker:

Semantic text analytics vendor Saplo in conjunction with BI vendor QlikTech’s QlikView “predicted the U.S. election by analysing Twitter and News data,” wrote Saplo founder and CEO Mattias Tyrberg in a Google Plus posting.
The Presidential Election News & Twitter Tracker let users review and analyze content by candidate, subject matter, category, key words and so on. It recorded 2-million plus tweets and close to 53,000 articles on the president since March, leading up to a 39 percent positive sentiment rating to the 32 percent rating challenger Mitt Romney earned based on 1.3 million-odd tweets and just under 40,000 news articles. Romney rose to a 35 percent positive sentiment score based just on November tweets and articles to Obama’s 39 percent for that month – a big change from March when the president had just a 40 percent positive sentiment score to Romney’s 49 percent (before he was the official Republican nominee).

What may be of particularly interest, though, now that the election is over, are some of the individual topic findings and what challenges that might mean for the president going forward. Taking a look at sentiment for the president about the economy, for instance, Obama records just a six percent positive rating based on tweets and articles analysis. Social security and the debt are at 5 percent, and health care and role of government stand at four percent. Foreign policy’s a bright spot, at 20 percent positive sentiment, as is Afghanistan at 32 percent.

**Topsy Twitter Political Index**

For a state by state look at sentiment around the president and his former challenger based on what was being shared on Twitter up through the early morning hours of election day, you can still take a look at [this](#) visualization from Twindex (which we told you about [here](#)). As of that point, its analysis of the data showed that the Twitterverse that began as overwhelmingly pro-Obama saw positive sentiments gradually shifting toward Romney, and that Montana, California, New Mexico, Illinois and Kentucky were currently the most neutral states.

Back at the main Twindex hub, as of Nov. 7, not surprisingly, Obama was up 11 on positive sentiment to a score of 85, while Romney was down 2 to 57. You know what they say, to the winner go the sentiment spoils.

**RELATED:**

I recently had the pleasure of speaking with Hermant Purohit and colleagues who have been working on an interesting semantic social web application called Twitris+. A project of the the Ohio Center of Excellence in Knowledge-enabled Computing (Kno.e.sis), Twitris+ uses “real-time monitoring and multi-faceted analysis of social signals to provide insights and a framework for situational awareness, in-depth event analysis and coordination, emergency response aid, reputation management etc.”

Twitris+ packs together quite an array of social computing features, integrating spatio-temporal-thematic dimensions, people-content network analysis and sentiment-emotion subjectivity analysis. The tool also aggregates a range of social data and web resources such as twitter, online news, Wikipedia pages, other multimedia content, etc., in addition to SMS data, for which the team was recently granted a patent.

Unlike many other social media platforms I’ve reviewed over recent months, Twitris+ geo-tags content at the tweet-level rather than at the bio level. That is, many platforms simply geo-code tweets based on where a person says s/he is as per their Twitter bio. Accurately and comprehensively geo-referencing social media content is of course no trivial matter. Since many tweets do not include geographic information, colleagues at GeoIQ are seeking to infer geographic information after analyzing a given stream of tweets, for example.

I look forward to continuing my conversations with Hermant and team. Indeed, I am particularly interested to see which emergency management organizations begin to pilot the platform to enhance their situational awareness during a crisis. Their feedback will be invaluable to Twitris+ and to many of us in the humanitarian technology space.
Picking the President: Twindex, Twitris Track Social Media Electorate

By Jennifer Zaino on August 3, 2012 2:43 PM

The U.S. is a mere three months away from choosing who will be its leader for the next four years. The semantic web is a supporting player in the action.

This week, of course, saw the debut of the Twitter Political Index (Twindex), a joint effort between Twitter, Topsy, and the Mellman Group and NorthStar Opinion Research polling groups. Since the Semantic Web Blog last spoke with Topsy execs here, the company has refined its sentiment analysis to the point where it could be released for the Twindex. The sentiment analytics engine ingests hundreds of millions of English-language tweets a day and computes sentiment for all terms in Twitter, though that's not publicly available yet.

In its Twindex incarnation, Topsy aggregates the underlying sentiment score minute by minute, and then that is rolled up into an hourly and daily score for each candidate, says Rishab Aiyer Ghosh, co-founder and chief scientist at Topsy Labs. Behind the scenes, “that score is normalized so that it is on 0 to 100 scale comparing to all the other terms people talk about,” he says, which is important for keeping perspective on the candidates in context relative to whatever else may be on the mind of the collective social media conscience. It also is weighted to include the scores of the previous two days before its publication at the end of the day, and smoothed out so that it doesn't jump around in helter-skelter fashion.

The sentiment algorithm, Ghosh says, was designed specifically for dealing with short-form content, and in sample testing Topsy has found that the algorithms agree with human perceptions of positive or negative sentiment 90 percent of the time. Additionally, it has done work in the background to narrow things down based on geography, so it's pretty confident that the scores it comes up with are based on representative U.S. tweets. “We worked with the polling firms a lot to validate it, too, to see how the sentiment algorithm should be used,” adds Ghosh. “This is definitely a skeptical audience, or one you want to be skeptical, and they were pretty positive about it,” he says.

Though not to be published as part of the Twindex, Topsy itself will publish analysis of sentiment on the candidates for individual groups of states, such as swing states. “We will publish more detail based on geography,” he says. “Only 1 percent of tweets actually have geo-tagging-enabled. But we have built technologies to expand from that. We use machine learning to process billions of tweets to come up with inferences of where the location is, and that has been pretty accurate.” The Twindex site will link to that.

This approach has its advantages and disadvantages compared to traditional opinion polling. Questions and audience can’t be controlled, for example. On the other hand, you do get insight into what hundreds and hundreds of thousands of people who are just out there expressing themselves think in an instant. “You can get an opinion right now and that is valuable information for everyone trying to understand what people think.” Ghosh says.

From Twindex to Twitris (And Wolfram Alpha Too)

That’s not the only semantic way to see how we’re feeling about the 2012 election, though. Wright State University and the Kno.e.sis Ohio Center of Excellence in Knowledge-enabled Computing, headed up by director and LexisNexis
Ohio Eminent Scholar Dr. Amit P. Sheth, has as a piece of its semantic social web application project Twiris the Election 2012 360° Social Media Analysis site. Sheth, who discussed how Twiris handles social media analysis in our story here, brings similar Search & Explore, Sentiment and Network (as in influential users and connectivity by topic) analysis capabilities to the site as it did for its Occupy Wall Street and India Against Corruption Twiris-based sites. Sheth told us previously that, for example, a candidate’s team could use the Network Analysis feature to see what influential users are talking positively about their candidate, perhaps to help target donations, or the Browse by Location tag to see what issues are being discussed in a certain area, to see if that information can be parlayed into speeches the candidate will make in that locale.

Sheth, posting a Google Plus message about the Election 2012 page, said this Twiris site is the way to follow through on the comment reportedly made by Adam Sharp, head of government, news and social innovation at Twitter, about how the Twitter Political Index could be used: “When the Twitter Political Index is giving a different indication than the polls as to where the winds of the electorate are shifting, that is a signal to perhaps dig deeper and gain a better understanding for the complexities of voting behavior.” Not only can Twiris Election 2012 let you do that now, Sheth said, with a nod to its Sentiment tab, and “soon you will be able to associate correlation between real-world events with its analysis here,” he noted. Click on Sentiment to currently view weekly charts of candidate sentiment by domestic, economic, international and social issues, the candidates themselves, or their parties.

And, if you just can’t get enough of the intersection between semantic technology and elections, over at the blog of the Wolfram Alpha knowledge engine – one of the sources for Siri that can provide results based on its structured data – there was a recent discussion of how it can help “to provide some useful context and analysis, particularly when it comes to understanding past election outcomes or predicting this year’s results.”

Writing for the blog, C. Alan Joyce explains how users can leverage Wolfram Alpha to ask questions about the impact of various demographic groups on presidential election races. “You can use Wolfram|Alpha to try to understand where the senior vote might have the most impact in Florida, for example. Or you might look closer at the specific origin of the Hispanic population in Florida—increasingly split between Hispanics of Cuban origin (who have traditionally voted Republican) and other groups, such as Puerto Ricans, who tend to favor Democratic candidates,” he says.
Semantic App Helps Researchers Understand Prescription Drug Abuse

By Jennifer Zaino on June 11, 2012 10:10 AM

There’s been a lot of attention given to the issue of prescription drug abuse, in the wake of violent crimes such as one last year that left four people dead in a pharmacy shooting in Suffolk County, New York. A recent study from the Workers Compensation Research Institute also shows that prescription drug abuse is the fastest growing drug problem in the United States, with over fifteen thousand people dying last year from an overdose. And, the U.S. Senate in late May approved an amendment to reclassify drugs that contain hydrocodone, a highly-addictive substance found in Vicodin and Lortabas, among other drugs, as Schedule II substances, while giving law enforcement more tools to monitor distribution of such drugs and also decreasing access to them for non-medical purposes.

What, you may ask, does any of this have to do with semantic technologies? Dr. Amit P. Sheth, Wright State University Kno.e.sis Ohio Center of Excellence in Knowledge-enabled Computing director and LexisNexis Ohio Eminent Scholar, and Dr. Raminta Danilaityte of the school’s Center for Interventions, Treatment and Addictions Research (CITAR), have a ready answer: PREDOSE, an application for understanding pain-killer drug abuse through the semantic analysis of social media conversations. More specifically, it’s automated data collection and analysis tools to process web-based data to determine the knowledge, attitudes, and behaviors of addicts, related to buprenorphine, OxyContin and other pharmaceutical opioids. It’s a National Institutes of Health (NIH)-funded project created by a partnership between Kno.e.sis and the CITAR.

In its role providing substance-abuse related services, academic research, and services research, CITAR engages in live interview projects and population surveys with subjects. But such approaches are costly and have a time lag to coordinate information and extract results, subjects might not fully disclose information, and research questions also can unintentionally have a bias. On the other hand, on “social media [people are] uninhibited about what they want to share,” says Sheth, who also leads the Twitris social media analysis project (covered here). They represent a potentially valuable source of anonymous information on what people are doing with painkillers and other drugs that could be helpful in driving policies or other efforts for curbing abuse. Of course, on Twitter or sites like Bluelight, which provides a forum for open information and discussion board about ecstasy and other drugs, people aren’t necessarily adhering to proper language structure, and they’re using a lot of abbreviations and slang names for drugs (oxy and multiple variants for OxyContin, for instance), too.

“It’s very challenging text to analyze,” says Sheth. The combination of domain knowledge at CITAR and semantic expertise at Kno.e.sis is a strong one-two punch to start to solve the issue. Work on PREDOSE, which has a three-stage architecture, has been underway for about a year now. The first part, which has largely been tackled, is collecting the user-generated data from social media sources; the project also involves creating semantic web data to correspond to the user-generated data and providing analysis tools to researchers. The work has included the development of the PREDOSE Drug Abuse Ontology (DAO) and the PREDOSE Annotator, to help capture and identify entities and represent the semantics of colloquial expressions in user-generated commentary.

“Unless you can establish a map between a drug and all its synonyms, you may miss some important posts,” says Delroy Cameron, the key PhD student involved in the project. By establishing the mapping in content from the slang term to the standard name of the drug, the tool can facilitate analysis of the data by capturing additional relevant posts. As an example, Oxycontin also has a slang name of hillbilly heroin.
PREDOSE also already enables gauging sentiment, “to help us understand and automatically to extract information about substances people favor,” says Daniulaityte. “In the case of a new drug, for example, we can very quickly update how users are reacting to it or their opinions.”

With its annotation capabilities for entity extraction, sentiment and mood in hand, focus will be on doing more work to automatically extract relationships among entities. Already the team has made a discovery around loperamide, an over-the-counter drug used to control diarrhea, which is also known as Imodium. Having those terms mapped to each other was an important step in making it possible for the team’s analysis of extra-medical use of the term to discover that, whatever the name, it was being used to control opiate withdrawal symptoms from drugs like Oxycontin — something that had not been previously reported in the epidemiological literature, according to the project team. Other relationships to be explored could be around the methods of use of a drug, or side effects, or prices, notes Daniulaityte.

“The technology perspective is that we have done more than half the work, but ultimately our project will allow scientists to ask more complex questions,” says Sheth. For example, from all web posts with a mention of Oxycontin and also of withdrawal, what is the most common sentiment among users? Or, on the heels of the 2010 FDA-mandated Oxycontin reformulation to reduce its use for getting high, they can ask the tool to display the change in overall positive and negative mood/sentiment about that, to gauge if the policy had its intended effect and whether it might be something to extend to other prescription drugs. “That part is not done but the core technique of being able to collect all the data, annotate it and spot things for a particular purpose is there,” he says.

The demographic scale and real-time capability the tool enables potentially can be useful for helping inform policy decisions. “It adds another perspective to epidemiological data,” says Daniulaityte. “All policy decisions should be data-driven. And especially valuable about this type of research is that it provides a very rapid assessment of situations,” which is critical when drug abuse patterns change as rapidly as they do.

Down the road there also is more potential for active engagement, borrowing on some Twitris concepts, for example, to provide people with information on where to get help if they really want it, or here are some statistics about long-term abuse via a targeted message. “We would hope to make this tool public and also propose a larger follow-on that would allow other epidemiologists working in this area to also benefit from the web-based tool,” Sheth says.

RELATED:

Tags: Amit Sheth, Bluelight, Center for Interventions, CITAR, Delroy Cameron, drug abuse, hillbilly heroin, Kno.e.sis, Kno.e.sis Ohio Center of Excellence in Knowledge-enabled Computing, National Institutes of Health, newsletter, NIH, OxyContin, PREDOSE, prescription drug abuse, Raminta Daniulaityte, sentimen analytics, sentiment analysis, Social Media, Treatment and Addictions Research, Twitris, Twitter, Workers Compensation Research Institute
OHIO CENTER of EXCELLENCE
in Knowledge-Enabled Computing

www.knoesis.org

FROM INFORMATION TO MEANING