Sentiment Analysis of Big Data from Internet News

A Project Report submitted in partial fulfilment of the requirements for the degree of Master of International Business (MIB)

Submitted by:

SUVDEEP BAGUI

Roll No. 54

Department of Commerce

Delhi School of Economics

University of Delhi

Delhi – 110007

March, 2014

CERTIFICATE BY THE SUPERVISOR

This is to certify that this project titled "**Sentiment Analysis of Big Data from Internet News"** submitted in partial fulfilment of the requirements for the Degree of Master of International Business (MIB) by **Suvdeep Bagui** at Department of Commerce, Delhi School of Economics, University of Delhi is a record of original research work carried out by her under my guidance. Any material borrowed or referred to is duly acknowledged.

Suvdeep Bagui

Roll No. 54

MIB 2012-2014

Department of Commerce

Delhi School of Economics

University of Delhi

Project Guide

Ms. Rinku Wadhawan Mahindru

Assistant Professor

Department of Commerce

Delhi School of Economics

University of Delhi

TABLE OF CONTENT

DECLARATION

I, Suvdeep Bagui, student of Master of International Business, Department of Commerce, Final Semester, hereby solemnly undertake that this project is undertaken solely for the purpose of research and is in requirement for the completion of the semester course of the programme.

The complete study has been done with utmost sincerity and honesty. No information has been listed from any source without their legal consent. All sources have been mentioned wherever quoted and all facts supported by the relevant references.

Suvdeep Bagui
Roll no 54
MIB Batch of 2014

ACKNOWLEDGEMENT

First of all, I would like to take this opportunity to thank my faculty and guide, **Ms. Rinku Wadhawan Mahindru**, who guided me at every step of my study. She gave me the freedom to pursue the study as I saw fit and to choose the research problem as per my interest. Her support and mentorship in this regard was in indispensable.

Secondly, we would like to thank our friends and colleagues who too provided us with their valuable contributions to enable me complete the project.

Lastly, I express my gratitude to Harrison Kinsley of Sentdex.com for all the tutorials and support.

I must also acknowledge the presence of free open source software, programming languages and libraries which form the bedrock of my study. I also thank the thousands of people who have contributed selflessly towards them.

LIST OF FIGURES

(Populated in the database)

- **Country codes** (on www.thehindu.com)

  USA - 413
  China - 281
  Russia - 1979
  Bangladesh - 285
  Pakistan - 371

- **Word Count**: It is the total word count corresponding to the news article.

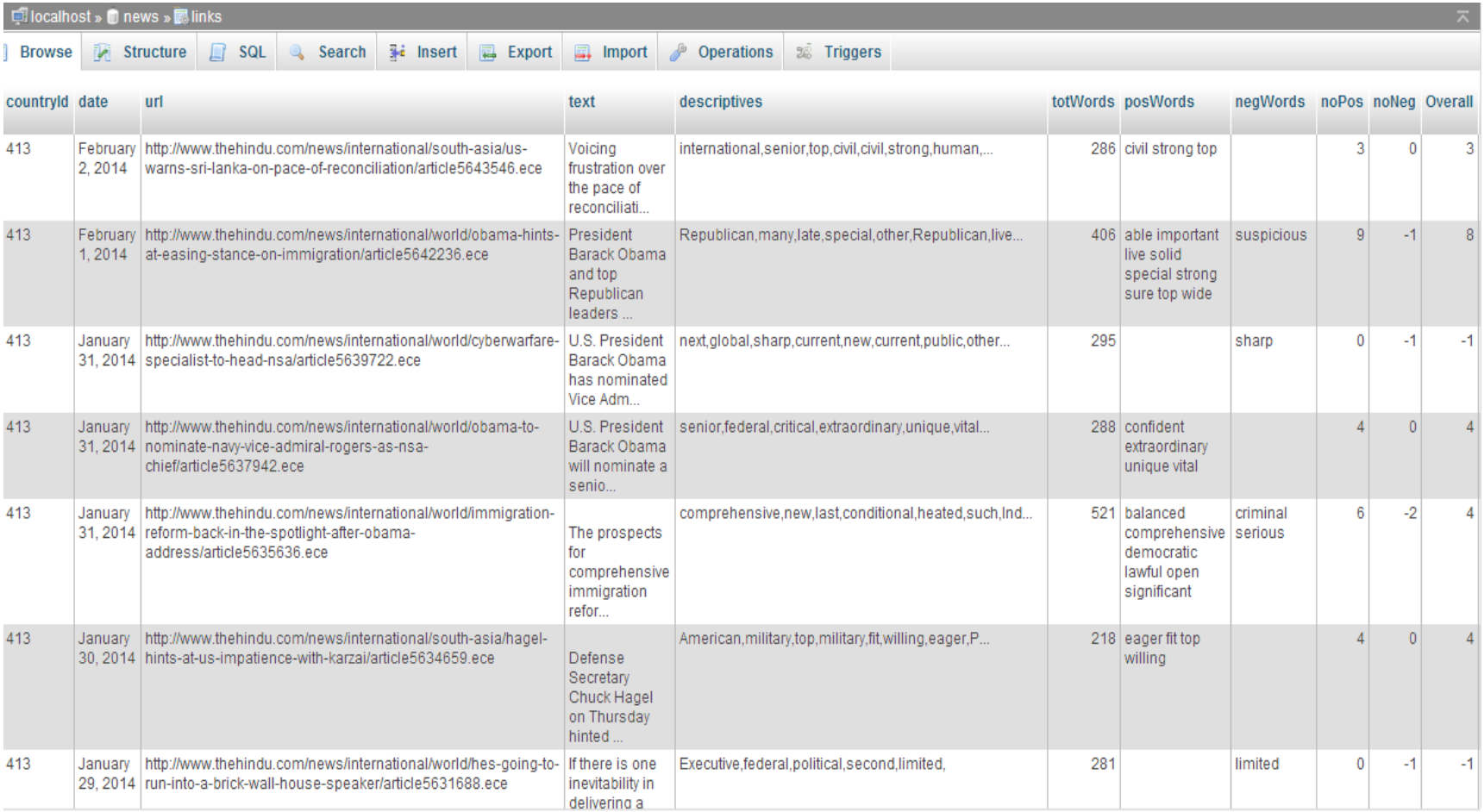
- **Descriptive words**: A list containing all the "modifier "words present in the article. Modifiers taken here are adjectives and adverbs.

- **Total Positive Words**: Is a common subset list of descriptive words and sentiment lexicon having positive connotation.

- **Total Negative Words**: Is a subset list of negative words and sentiment lexicon having negative connotation.

- **Total Sentiment Score**: Is the overall sentiment value attributed on the basis of total positive and negative word in that particular news article.

- **Positive content density**: Total Positive sentiment words/ Total Word Count of all the articles.



countryId	date	url	text	descriptives	totWords	posWords	negWords	noPos	noNeg	Overall
413	February 2, 2014	http://www.thehindu.com/news/international/south-asia/us-warns-sri-lanka-on-pace-of-reconciliation/article5643546.ece	Voicing frustration over the pace of reconciliati...	international,senior,top,civil,civil,strong,human,...	286	civil strong top		3	0	3
413	February 1, 2014	http://www.thehindu.com/news/international/world/obama-hints-at-easing-stance-on-immigration/article5642236.ece	President Barack Obama and top Republican leaders ...	Republican,many,late,special,other,Republican,live...	406	able important live solid special strong sure top wide	suspicious	9	-1	8
413	January 31, 2014	http://www.thehindu.com/news/international/world/cyberwarfare-specialist-to-head-nsa/article5639722.ece	U.S. President Barack Obama has nominated Vice Adm...	next,global,sharp,current,new,current,public,other...	295		sharp	0	-1	-1
413	January 31, 2014	http://www.thehindu.com/news/international/world/obama-to-nominate-navy-vice-admiral-rogers-as-nsa-chief/article5637942.ece	U.S. President Barack Obama will nominate a senio...	senior,federal,critical,extraordinary,unique,vital...	288	confident extraordinary unique vital		4	0	4
413	January 31, 2014	http://www.thehindu.com/news/international/world/immigration-reform-back-in-the-spotlight-after-obama-address/article5635636.ece	The prospects for comprehensive immigration refor...	comprehensive,new,last,conditional,heated,such,Ind...	521	balanced comprehensive democratic lawful open significant	criminal serious	6	-2	4
413	January 30, 2014	http://www.thehindu.com/news/international/south-asia/hagel-hints-at-us-impatience-with-karzai/article5634659.ece	Defense Secretary Chuck Hagel on Thursday hinted ...	American,military,top,military,fit,willing,eager,P...	218	eager fit top willing		4	0	4
413	January 29, 2014	http://www.thehindu.com/news/international/world/hes-going-to-run-into-a-brick-wall-house-speaker/article5631688.ece	If there is one inevitability in delivering a	Executive,federal,political,second,limited,	281		limited	0	-1	-1

Chapter 1. Introduction

Today, we live in a world that's emanating increasingly more data with every passing day. Experts quote this data amounting in Petabytes to Zeta bytes per year. This is a revolution brewing to happen or is already underway to some extent. Everything, literally, that's happening is generating some data. This presents us with new challenges and opportunities. As for opportunities, data-driven analysis, discovery and prediction hold enormous possibilities. My research endeavors to realize and experience such (some) possibilities firsthand.

The huge size, speed and complexity of big data also makes it more than tedious for humans. Thus comes the need for computational and analytical models as aid in order to make sense of the data and ultimately extract meaningful and actionable knowledge from the information. One of the method used here to analyze huge streams of textual data is natural language processing.

Main research problem

The project pertains to measurement of the various quantitative sentiment attributes of news stories. News analytics are usually derived through automated text analysis and applied to digital texts using elements from natural language processing. The data set chosen for this research is news articles from a popular Indian daily's website – The HINDU. Data (text content) was mined from 1000 articles. All these articles were analyzed for positive or negative sentiment in their articles.

Data source: News articles from Hindu Website

News subject: Country specific articles

Sample size: 200 articles per country.

Total Sample size: 1000 Articles.

Analytical Technique: Sentiment Analysis

Objective

To find hidden trends, or patterns which may be invisible to the naked eye among all these news articles.

Purpose

The main purpose of this research was explore to a niche segment of big data analytics and experience its potential first hand.

The application of Big Data and sentiment analysis to news and social media has grown from an area of research to mature product solutions since 2007. News analytics and news sentiment calculations are now routinely used by both buy-side and sell side in alpha generation, trading execution, risk management, and market surveillance and compliance. [i]

A large number of companies use news analysis to help them make better business decisions. Academic researchers have become interested in news analysis especially with regards to predicting stock price movements, volatility and traded volume.

Possible Applications in International Business and Government Intelligence

The insights from study could provide valuable inputs to Ministries of Media/ External Affairs/ Foreign Affairs. It also gives idea about a Nation's Brand image and identity. Furth more this analysis gives more understanding to public relations personnel.

Testable hypothesis

1. Does news about a particular country contain more or less opinionated /sentimental content than others?

2. Is there anything common in highly sentimental news?

Research Design

Initial efforts were to develop the data collection resources (programming scripts, server scheduling etc.) For **data mining**, the scripting language used is Python 2.7 (extensively used worldwide, also by Google.) Other libraries **beautiful soup, mechanize** might be used to automate the web scrapping. For filtering of data into meaningful information **Regular expressions** will be employed and stored in databases (**MySQL**.) The data mining exercise is expected was carried out continuously for **1 month** (till 20th March 2014.) The cumulated data finally is then statistically analyzed.

PROCESS FLOW

Data Mining → Filtering Info → Storing in Database → Final Analysis

Filters = Text content of the news articles

Database Ingredients: Country ID, News URL, News Text, Word Count, Positive| Negative Word | Score, Overall Sentiment Score

*More details on sentiment analysis and final analysis of data has been explained in the subsequent portions of this document.

Limitations of the study

- My sentiment prediction systems work just by looking at words in isolation, giving positive points for positive words and negative points for negative words and then summing up these points. That way, the order of words is ignored and important information is lost. In contrast, our new deep learning model actually builds up a representation of whole sentences based on the sentence structure. It computes the sentiment based on how words compose the meaning of longer phrases.

- Most sentiment analysis algorithms use simple terms to express sentiment about a product or service. However, cultural factors, linguistic nuances and differing contexts make it extremely difficult to turn a string of written text into a simple pro or con sentiment. The fact that humans often disagree on the sentiment of text illustrates how big a task it is for computers to get this right.[ii]

.

- Another method of generating sentiment lexicon is to make it specific to the research subject. This practice increases the accuracy.

- IT Infrastructure needed for true Big Data analysis requires considerable funds. This research conducted was magnitudes lower than what would be an actual news analytics.

- Opinion lexicon used is far from sufficient for accurate sentiment analysis.

Organization of the study

The whole has been conducted with a generic laptop running on windows 7 with 1 MBPS internet connection. Apart from report, all the data mining scripts and populated databases can be produced at request and have been appended in a CD with the report.

Chapter 2. Conceptual framework

The sentiment analysis for news model is divided into three main phases: Data mining, language processing and analysis. The goal of data mining phase is to scan the internet and collect the relevant data as much as possible with least noise. This is done by making use of website site crawlers. From this data, the most important portion of content is filtered out and stored separately for further processing.

After the mining process, the segregated news content is routed through language processing. The objective of language processing is to identify impact words in the content. These are words that impart meaning, opinion to the sentences. Very complex algorithms have been developed to achieve the same. This phase of the projects usually begins with part of speech tagging and usually selecting words that are adjectives and adverbs. These selected words are then compared with an already developed sentiment lexicon, which is sort of a database of words categories as per positive or negative connotation.

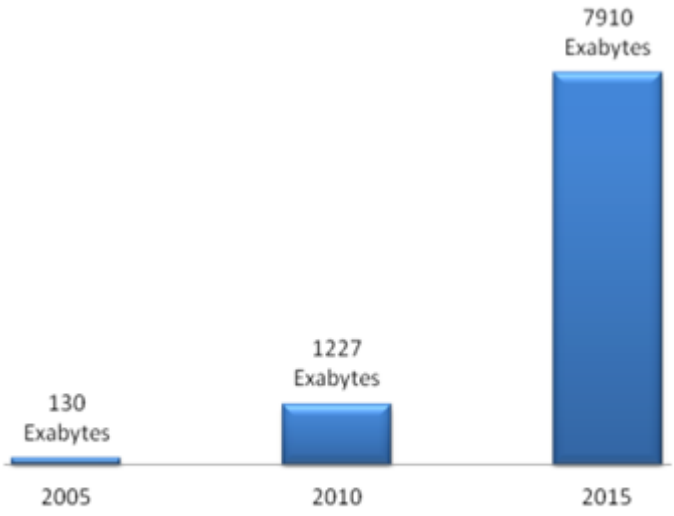
Lastly, the analysis phase works on the output of previous phase. It takes into account quantitative occurrence of the sentimental words in the news and attributes sentiment score accordingly. This concludes the analysis of a news article in isolation. However then, the whole overview picture of the research emerges when results of the all the articles are analyzed together.

List and Explanation of important terms

Introduction

The growth of data will never stop. According to the 2011 IDC Digital Universe Study, 130 Exabyte of data were created and stored in 2005. The amount grew to 1,227 Exabyte in 2010 and is projected to grow at 45.2% to 7,910 Exabyte in 2015.

The growth of data constitutes the "Big Data" phenomenon – a technological phenomenon brought about by the rapid rate of data growth and parallel advancements in technology that have given rise to an ecosystem of software and hardware products that are enabling users to analyses this data to produce new and more granular levels of insight.[iii]



A decade of Digital Universe Growth: Storage in Exabytes

There's nothing new about the notion of big data, which has been around since at least 2001. In a nutshell, Big Data is your data. It's the information owned by your company, obtained and processed through new techniques to produce value in the best way possible.[iv]

What is Big Data?

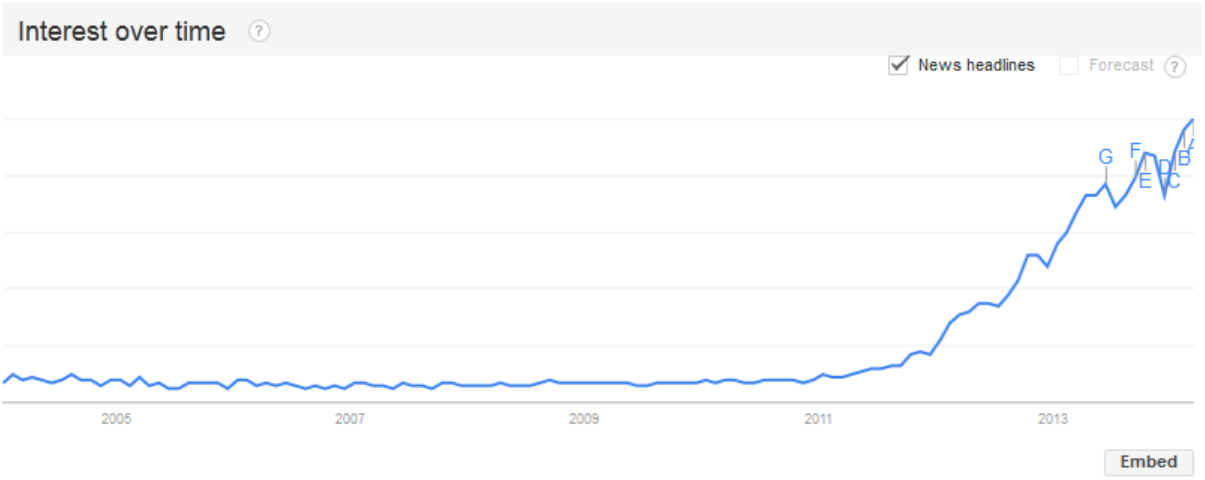
According to McKinsey, Big Data refers to datasets whose size are beyond the ability of typical database software tools to capture, store, manage and analyses. There is no explicit definition of how big a dataset should be in order to be considered Big Data. New technology has to be in place to manage this Big Data phenomenon. IDC defines Big Data technologies as a new generation of technologies and architectures designed to extract value economically from very large volumes of a wide variety of data by enabling high velocity capture, discovery and analysis. According to O'Reilly, "Big data is data that exceeds the processing capacity of conventional database systems. The data is too big, moves too fast, or does not fit the structures of existing database architectures. To gain value from these data, there must be an alternative way to process it."

Characteristics of Big Data

Big Data is not just about the size of data but also includes data variety and data velocity. Together, these three attributes form the three Vs of Big Data. The 3 Vs of Big Data

Volume is synonymous with the "big" in the term, "Big Data". Volume is a relative term – some smaller-sized organizations are likely to have mere gigabytes or terabytes of data storage as opposed to the petabytes or Exabyte's of data that big global enterprises have. Data volume will continue to grow, regardless of the organization's size. There is a natural tendency for companies to store data of all sorts: financial data, medical data, environmental data and so on. Many of these companies' datasets are within the terabytes range today but, soon they could reach petabytes or even Exabyte's. Data can come from a variety of sources (typically both internal and external to an organization) and in a variety of types. With the explosion of sensors, smart devices as well as social networking, data in an enterprise has become complex because it includes not only structured traditional relational data, but also semi-structured and unstructured data. Interpretation of Big Data can bring about insights which might not be immediately visible or which would be impossible to find using traditional methods. This process focuses on finding hidden threads, trends, or patterns which may be invisible to the naked eye. [v]

Why is Big Data so hot right now?



The pot of information (both private and public) generated by humanity has come to a recent boil. We're generating more content than ever before, but in many cases it leads to more questions and fewer answers. What is happening in the atmosphere? Which candidate do voters prefer? Which movies, books, and TV shows are going to satiate the public's appetite? Which trends are coming down the road?

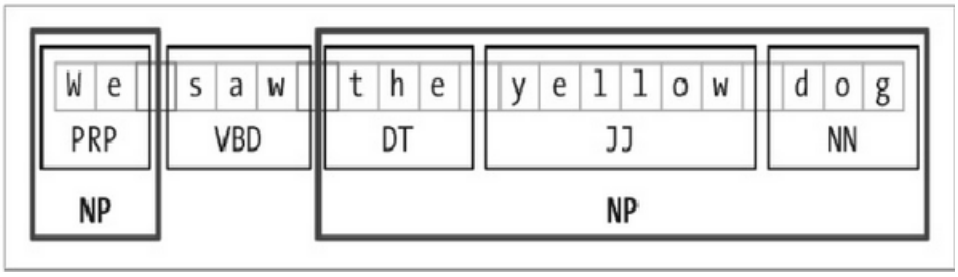
Making sense of all this content is like trying to hear what someone is whispering backstage while you're attending a booming outdoor concert. A deep need exists for the structure to parse the data to separate out the cacophony and find the useful threads to uncover opportunities. Even more potential has opened for those who can orchestrate this feat.[vi]

NLTK

The Natural Language Toolkit, or more commonly NLTK, is a leading platform for building Python programs to work with human language data. It is a suite of libraries and programs for symbolic and statistical natural language processing (NLP) for the Python programming language. NLTK includes graphical demonstrations and sample data. NLTK is intended to support research and teaching in NLP or closely related areas, including empirical linguistics, cognitive science, artificial intelligence, information retrieval, and machine learning[vii].

Basic tasks include tokenization, stemming, POS tagging, chunking, parsing, etc.

- Tokenization: getting words and punctuations out from text.
- Stemming: getting the (inflectional) root of word; e.g. root of words: plays, playing, played is 'play'
- POS (Part of Speech) tagging: tagging parts of speech into different word classes such as nouns, verbs, adjectives and adverbs. E.g. In the sentence "Ram killed Ravana", 'Ram' & 'Ravana' are both tagged as NNP (which means proper noun) and 'killed' is tagged as VBD (verb, past tense).
- Chunking: groups the similar POS tags

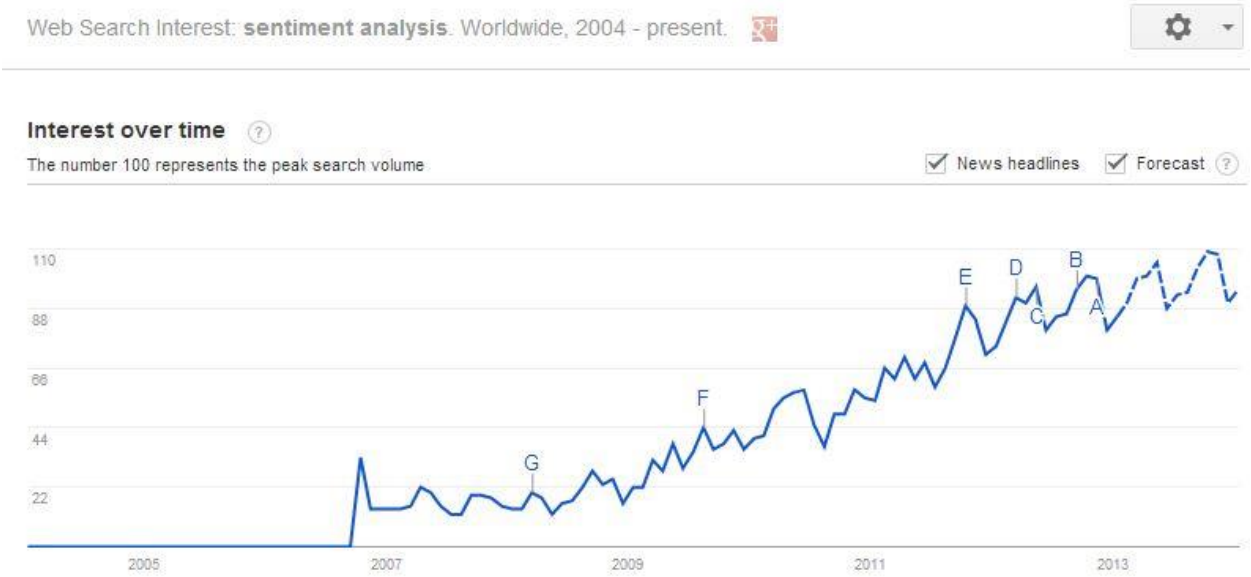


- Parsing: Identifying grammatical structures in a sentence.

Applications of NLTK:

- Machine translation (MaTra, Google translate)
- Transliteration (Xlit, Google Indic)
- Text classification (Spam filters)

Sentimental analysis



Sentiment analysis (also known as **opinion mining**) refers to the use of natural language processing, text analysis and computational linguistics to identify and extract subjective information in source materials. It normally involves the classification of text into categories such as “positive”, “negative” and in some cases “neutral”.

Generally speaking, sentiment analysis aims to determine the attitude of a speaker or a writer with respect to some topic or the overall contextual polarity of a document. A basic task in sentiment analysis is classifying the *polarity* of a given text at the document, sentence, or feature/aspect level — whether the expressed opinion in a document, a sentence or an entity feature/aspect is positive, negative, or neutral. Advanced, "beyond polarity" sentiment classification looks, for instance, at emotional states such as "angry," "sad," and "happy."

The process of sentiment analysis involves text analytics, linguistics and accepted language processing to determine and dig subjective information from source materials. It is commonly known for the term “opinion mining.” This process aims to determine how a certain person or group reacts to a topic they are being referred to. They react because they are either interested or involved. And, these reactions go to none other than their social media accounts which makes social media as one of the leading platforms in the internet where anyone can basically do

sentiment analysis. Twitter and Facebook are two of the places where one can find a lot of sentiments and they are the best considerations whenever opinion mining is done.

The applications for sentiment analysis are endless. More and more we're seeing it used in social media monitoring and VOC to track customer reviews, survey responses, competitors, etc. However, it is also practical for use in business analytics and situations in which text needs to be analyzed.

Sentiment analysis is in demand because of its efficiency. Thousands of text documents can be processed for sentiment (and other features including named entities, topics, themes, etc.) in seconds, compared to the hours it would take a team of people to manually complete.

My firm today are already developing develops applications for Sentiment Analysis with objectives like:

1. Computing customer satisfaction metrics: You can get an idea of how happy customers are with your products from the ratio of positive to negative tweets about them.

2. Identifying detractors and promoters: It can be used for customer service, by spotting dissatisfaction or problems with products. It can also be used to find people who are happy with your products or services and their experiences can be used to promote your products.

Python is a widely used general-purpose, high-level programming language. Its design philosophy emphasizes code readability, and its syntax allows programmers to express concepts in fewer lines of code than would be possible in languages such as C. The language provides constructs intended to enable clear programs on both a small and large scale.

A **regular expression** (abbreviated **regex** or **regexp**) is a sequence of characters that forms a search pattern, mainly for use in pattern matching with strings, or string matching, i.e. "find and replace"-like operations.

MySQL is (as of March 2014) the world's second most widely used open-source relational database management system (RDBMS). The SQL phrase stands for Structured Query Language.

The Hindu is an English-language Indian daily newspaper. Headquartered at Chennai, *The Hindu* was published weekly when it was launched in 1878, and started publishing daily in 1889. According to the Indian Readership Survey in 2012, it was the third most widely-read English newspaper in India (after the *Times of India* and *Hindustan Times*), with a readership of 2.2 million people. *The Hindu* has its largest base of circulation in southern India, and is the most widely-read English daily newspaper in Kerala and Tamil Nadu.

Sentiment Lexicon

- A list of positive and negative opinion words or sentiment words for English (around 6800 words). This list was compiled over many years starting from the first paper on sentiment analysis by (Hu and Liu, KDD-2004).[viii]

The sentiment lexicon is the most crucial resource for most sentiment analysis algorithms. There are three options for acquiring the sentiment lexicon: manual approaches in which people code the lexicon by hand, dictionary-based approaches in which a set of seed words is expanded by utilizing resources like WordNet 8 and corpus-based approaches in which a set of seed words is expanded by using a large corpus of documents from a single domain.[ix]

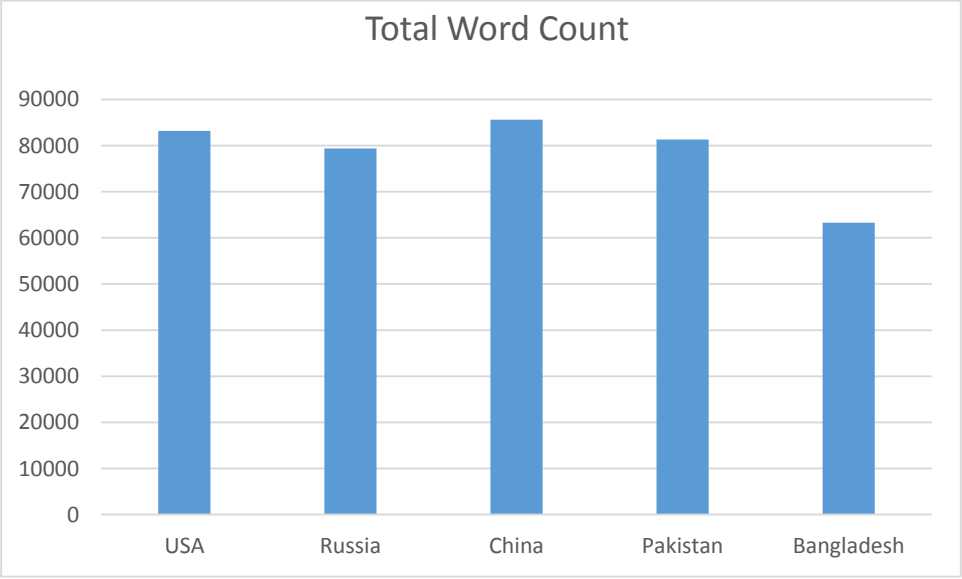
3. Analysis and Observations

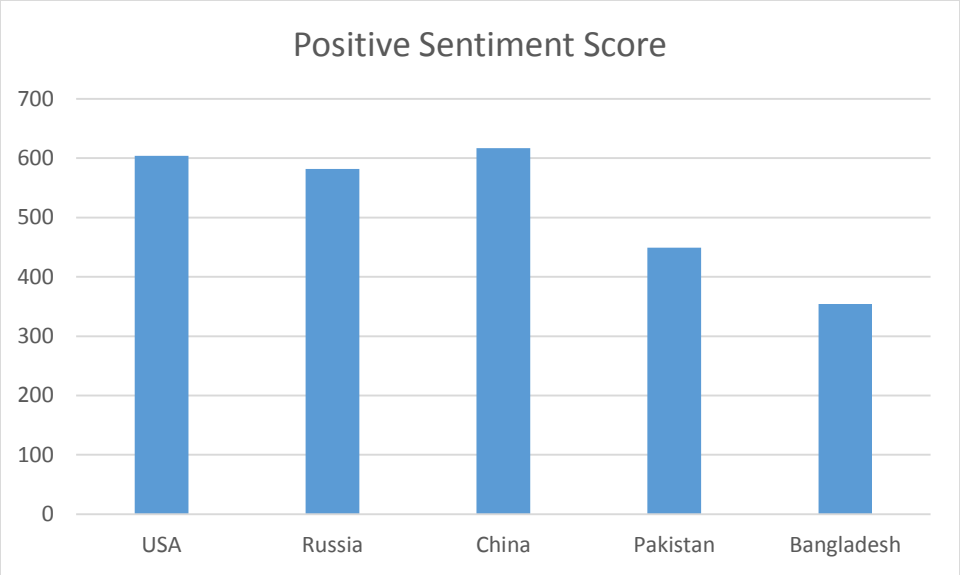
The following table shows the data aggregated from the database. Data points of interest have been highlighted and observable anomaly has been detailed subsequently.

	USA	Russia	China	Pakistan	Bangladesh
Total Word Count	83214	79356	85596	81324	63264
Positive Score	604	582	617	449	354
Negative Score	-319	-282	-314	-295	-272
Overall Score	285	300	303	154	82

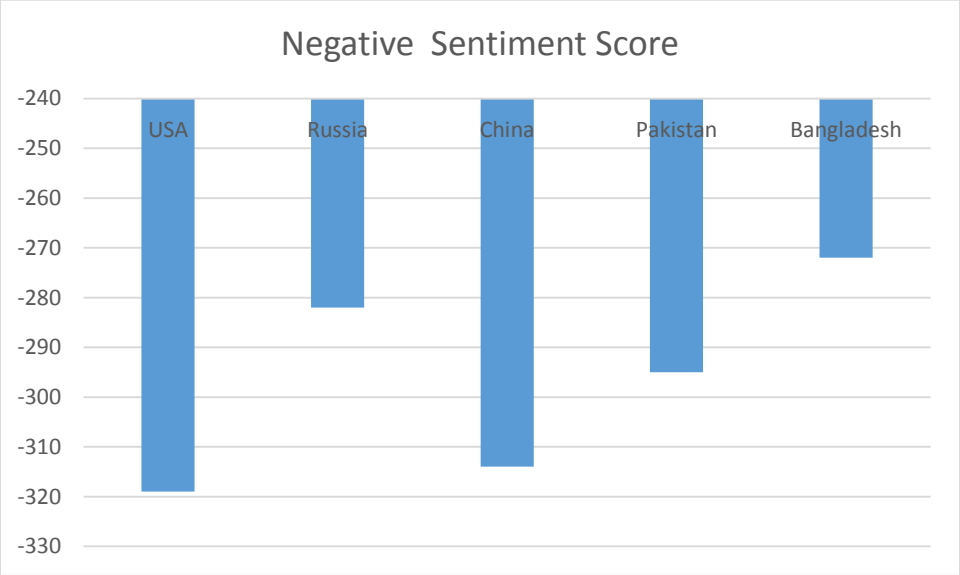
Observations

Observation 1: Although equal number of articles were compared for each county (200 countries, length of articles about Bangladesh appear to significantly shorter than others. China sees the longest articles of all the countries.)

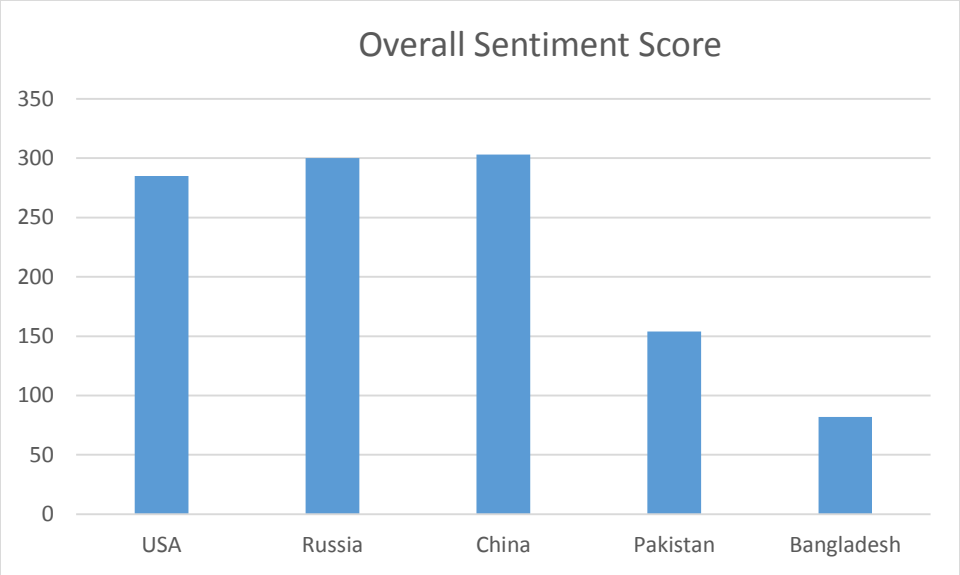




Observation 2: Our neighboring countries, Bangladesh and Pakistan have much lower positive sentiment score than others. Also surprising is that Russia has the highest positive content density of all.



Observation 3: Russia has the least negative content density of all, while surprisingly US tops in absolute terms. Our neighboring countries Pakistan and Bangladesh have a lot negative content in their articles, especially Bangladesh.



Observation 4: Overall sentiment of all the five countries is positive. However, Bangladesh and Pakistan lag considerably behind other three countries. Also, that China and Russia edge past USA, China having the highest Overall Sentiment Score.

http://www.thehindu.com/opinion/lead/for-an-unfettered-internet/article5699615.ece	27
http://www.thehindu.com/opinion/lead/the-many-shades-of-grey-in-ukraine/article5781594.ece	27
http://www.thehindu.com/opinion/lead/why-russia-needs-crimea/article5792952.ece	23
http://www.thehindu.com/opinion/op-ed/tech-industry-is-trying-out-its-election-muscle/article5661195.ece	23
http://www.thehindu.com/opinion/op-ed/bengali-hindus-in-muslimmajority-bangladesh/article5579528.ece	22
http://www.thehindu.com/news/international/world/sifting-myth-from-reality-in-a-dark-chapter-of-history/article5539021.ece	19
http://www.thehindu.com/opinion/op-ed/a-debating-group-in-search-of-politics/article5649789.ece	19
http://www.thehindu.com/opinion/lead/pakistans-terror-conundrum/article5750648.ece	19
http://www.thehindu.com/opinion/op-ed/life-on-the-razors-edge/article5674487.ece	19
http://www.thehindu.com/news/international/world/new-challenge-for-next-generation-of-party-leaders/article5570728.ece	18
http://www.thehindu.com/opinion/lead/a-tale-of-two-handshakes/article5297752.ece	18

Observation 5: Articles containing highest sentimental words are mostly from Op-ed section.

Observation 6: Articles containing highest sentimental words are mostly about politics.

Chapter 4. Summary

Every day, there is stream of data in form of news coming from thousands of sources in the world. With changing trends, the shift today is towards digital news. This also enables us to analyze using computing, *news* from a broader perspective which wasn't possible before.

This study collected 1000 articles from a popular Indian English newspaper – The Hindu. The selection is based on country tagging. 200 articles each for five countries have been collected (USA, Russia, China, Pakistan, and Bangladesh.)

The textual content extracted from the news articles has been analyzed using natural language processing for positive and negative content. Upon analyzing the whole data startling observations were made.

Significant variation in average article length for one particular country is observed. Similarly deviations have been found among countries with regard to overall sentiment, total positive and total negative sentiment. These observation reveal that there might underlying reasons for such anomalies which is a matter of further research.

This study can be applied to other fields than news as well and demonstrates the ability of Big Data analytics to reveal insights otherwise invisible.

Conclusion

Big data analytics in conjunction with sentimental analysis of NEWS can reveal valuable hidden patterns that are imperceptible to naked eye.

ENDNOTES

[i] https://en.wikipedia.org

[ii] http://www.zdnet.com/tapping-into-consumer-sentiment-on-social-networks-7000002279/

[iii] http://ovum.com/research/what-is-big-data-the-end-game/

[iv] http://www.ibm.com/developerworks/xml/library/x-datagrowth/index.html?ca=drs-

[v] http://www.emc.com/collateral/analyst-reports/idc-extracting-value-from-chaos-ar.pdf

[vi] http://www.bicoe.in/2014/03/big-data.html

[vii] http://www.nltk.org/

[viii] http://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html

[ix] http://cacm.acm.org/magazines/2013/4/162501-techniques-and-applications-for-sentiment-analysis/abstract

Bibliography:

- Mahindru, R. (2014). Big Data: It's Actually Big, The International Journal of Business & Management, 2 (1), 88.
- Godbole, N., Srinivasaiah, M., & Skiena, S. (2007). Large-Scale Sentiment Analysis for News and Blogs. *ICWSM*, *7*.
- Zhang, W., & Skiena, S. (2010, May). Trading Strategies to Exploit Blog and News Sentiment. In *ICWSM*.
- Balahur, A., Steinberger, R., Kabadjov, M., Zavarella, V., Van Der Goot, E., Halkia, M. Belyaeva, J. (2013). Sentiment analysis in the news. *ArXiv preprint arXiv: 1309.6202*.
- Nasukawa, T., & Yi, J. (2003, October). Sentiment analysis: Capturing favorability using natural language processing. In *Proceedings of the 2nd international conference on Knowledge capture* (pp. 70-77). ACM.
- Fukuhara, T., Nakagawa, H., & Nishida, T. (2007). Understanding Sentiment of People from News Articles: Temporal Sentiment Analysis of Social Events. In*ICWSM*.
- Claypool, M. (n.d.). Opinion Mining, Sentiment Analysis, Opinion Extraction. *Opinion Mining, Sentiment Analysis, Opinion Extraction*. Retrieved March 31, 2014, from http://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html
- Feldman, R. (n.d.). Communications of the ACM. *Techniques and Applications for Sentiment Analysis*. Retrieved March 31, 2014, from http://cacm.acm.org/magazines/2013/4/162501-techniques-and-applications-for-sentiment-analysis/fulltext
- Lee, L., & Pang, B. (n.d.). Opinion mining and sentiment analysis. *(Survey)*. Retrieved March 31, 2014, from https://www.cs.cornell.edu/home/llee/opinion-mining-sentiment-analysis-survey.html
- Agarwal and P. Bhattacharyya, "Sentiment analysis: A new approach for effective use of linguistic knowledge and exploiting similarities in a set of documents to be classified," in Proceedings of the International Conference on Natural Language Processing (ICON), 2005.
- R. Agrawal, S. Rajagopalan, R. Srikant, and Y. Xu, "Mining newsgroups using networks arising from social behavior," in Proceedings of WWW, pp. 529–535, 2003.