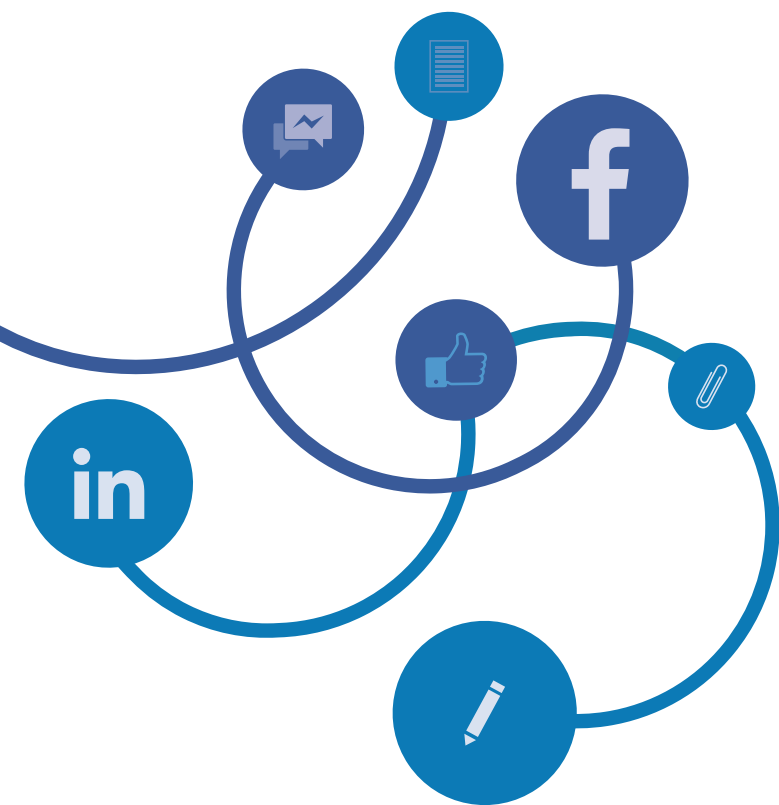


KEY CHALLENGES OF **MINING DATA** FROM **FACEBOOK & LINKEDIN**

Social Media users are generating astonishing amounts of data on a day-to-day basis. Making sense of the data helps in developing actionable patterns from which businesses and other organizations can gauge user behavior and preferences. However, the challenges manifest themselves at the data mining phase, in spite of the availability of robust analytical tools and algorithms.



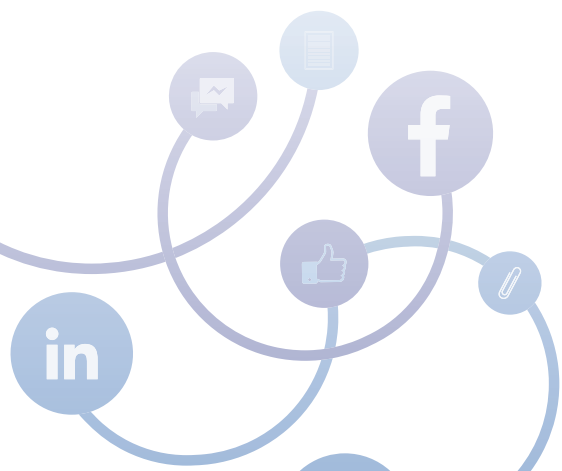
<i>INTRODUCTION</i>	<i>2</i>
<i>EXTRACTING RELEVANT DATA</i>	<i>3</i>
<i>THE FALLACY OF BENCHMARKING</i>	<i>4</i>
<i>BIG DATA PARADOX</i>	<i>5</i>
<i>NOISE REMOVAL FALLACY.....</i>	<i>6</i>
<i>MISSING VALUES</i>	<i>7</i>
<i>DUPLICATE DATA</i>	<i>8</i>
<i>OUTDATED INFORMATION</i>	<i>9</i>
<i>LIES AND EXAGGERATIONS</i>	<i>10</i>
<i>THE MENACE OF REPUTATION MANAGERS</i>	<i>11</i>
<i>DISTORTIONS</i>	<i>12</i>
<i>ISSUE OF SERIOUSNESS.....</i>	<i>13</i>
<i>BIAS AND PRESUMPTION</i>	<i>14</i>
<i>PRIVACY.....</i>	<i>15</i>

INTRODUCTION

Social media mining entails delving into the mass of data that comes from interactions through the social media and applying a variety of tools and disciplines including algorithms, statistics, sociology, ethnography, optimization and others, to analyze and extract meaningful patterns and trends.

Data is growing at an exponential pace and social media data is a big contributor to this growth. Millions of people spend countless hours on various social media channels such as Facebook and LinkedIn to connect, communicate, interact, create and share user-generated data at an unprecedented rate. Such “social big data” offers unparalleled opportunities for marketers and others, who make use of such data to understand how their targeted customers behave and what they prefer, to target them better in a customized way and deliver better products and services.

However, for all the possibilities, the field of social media mining is rife with issues and challenges.



01

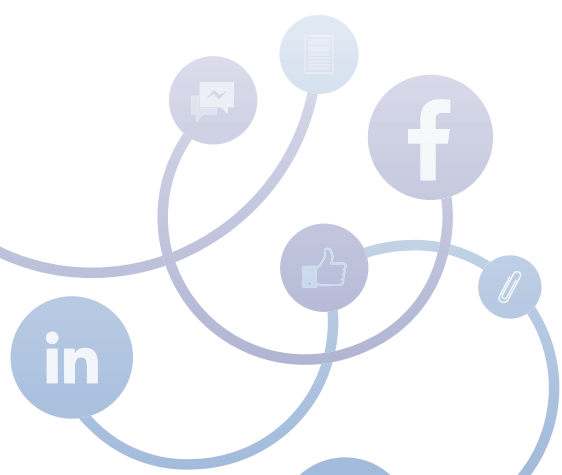
EXTRACTING RELEVANT DATA

There is no dearth of data in the social media space for anyone interested to mine it. Facebook alone collects more than **500 terabytes of data a day**, and that figure is growing. It is almost impossible for anyone to collect and analyze all this data, and make sense of it.

Data miners use application programming interfaces (API) from social media websites to collect data. Such APIs can deliver only a limited amount of data daily, and also do not reveal the population's distribution. Without such insight, data miners have no clue whether the samples they extract are reliable representatives of the full data. In any case, more of data does not necessarily beget more good or relevant data.

Data in social media sites such as Facebook and LinkedIn is inherently unstructured and “noisy”, which requires new approaches and new computational methods, different from the way data is traditionally extracted and mined.

Furthermore, such data is inseparable from social networks, and as such any meaningful extraction of data requires application of social theories and research, over and above statistical and data mining methods.

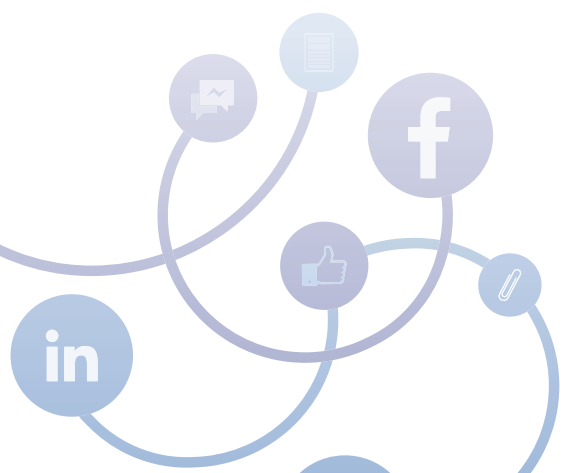


02

THE FALLACY OF BENCHMARKING

Having applied the correct techniques, the biggest challenge facing social data miners is determining what is useful to extract. There is no one-size fits all approach, and many data miners fall into the trap of benchmarking not just the method, but also the content and structure from others. Data miners would ape best practices when it comes to techniques and methods; but when it comes to the content, they would do well to consider business strategy, or specifically the purpose for which the data is required.

For instance, when it comes to data mining to promote a book, a subset of “likes” would be of critical importance, whereas a subset of “customer names” would be totally irrelevant to predict whether an individual would buy a book or not. Putting the API to innovative use is critical here, and benchmarking may actually become counterproductive.



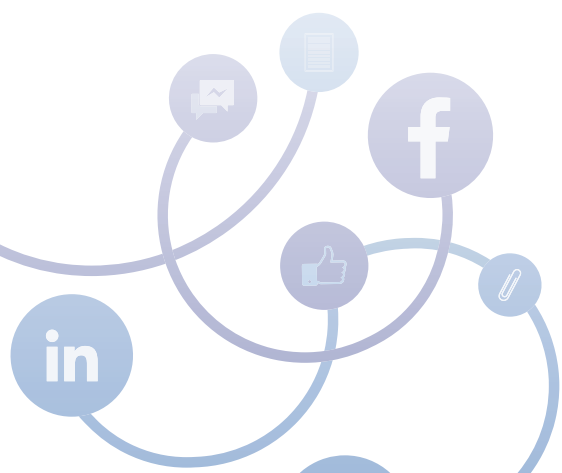
03

BIG DATA PARADOX

Even as data grows at an exponential pace, and user generated content in the social media is a big contributor to such growth, there is still very little relevant data available when it comes to individuals. Very often the information available would not tell the complete picture.

The challenge before data miners is to aggregate data at the individual's level from multiple sources, and at a multi-dimensional level.

Extracting data from one channel or one source would rarely suffice. Data miners would, for instance, need to build up a candidate's professional profile from LinkedIn and look for their hobbies and values from Facebook, all the while keeping an eye open for contradictions.




04

NOISE REMOVAL FALLACY

Noise is the distortion of data, and it needs to be removed before running algorithms, to prevent distorted data from making the algorithm churn out results that are wide off the mark. Many filtering algorithms effectively combat noise; but in social media, removing noise raises more issues than it solves.

Classic data mining has an extensive stage of preprocessing data to remove “garbage data” and thereby eliminate distortions. However, a big chunk of Facebook and LinkedIn data is such “garbage” data, and worse, such data lies tangled with other data. Any attempt to remove useless data would invariably result in the loss of valuable information as well. In any case, removing “garbage” or “noise” may actually result in the removal of a big chunk of data, accentuating the problem of insufficient data at the individual level.

The various types of API—bulk APIs for data loading, advanced streaming APIs that assist in push notification integrations, social APIs that facilitate collaboration, and metadata APIs, which define permissions, field types, data access guidelines, and user experience—all have the power to achieve more than 1.3 billion transactions per day. These APIs make it very easy to connect the mobile device or wearable smart devices to the customer information residing in servers, to develop apps for personalized experiences.



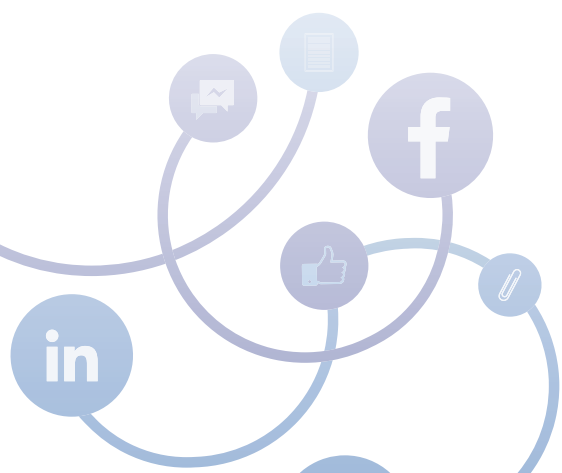
Salesforce has revamped Heroku and ExactTarget along with the launch of Salesforce1. The new Heroku1 allows companies to build and deploy communities for custom facing apps, to connect stakeholders. ExactTarget Fuel allows companies to engage with customers on a one-to-one basis, at scale.

05

MISSING VALUES

Many individuals avoid mentioning their personal information such as age or location on social media websites, for privacy and safety reasons. For data miners, this leads to critical gaps and it requires a painstaking search using smart algorithms to collect the required data from multiple sources; and even then, success would be invariably limited.

Another dimension to this issue is people being inactive intermittently. Not all people spend a consistent amount of time on social media all year round. They may spend more time on Facebook when they are free, and when work pressure increases, their presence may be minimal. Similarly, many members may log in and brush up their LinkedIn profile only when they consider a job switch. These sporadic activities can lead to big gaps in timelines and distort the efficacy of the data in a big away.



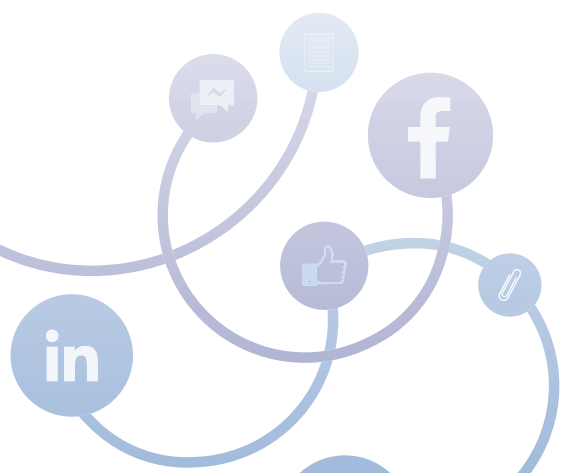
06

DUPLICATE DATA

If critical gaps inconvenience the data miner, duplicate data gives them nightmares.

Duplicate posts, duplicate blogs, and even duplicate profiles are all too common occurrences in social media channels such as Facebook.

Data miners have their task cut out trying to identify and weed out duplicates, before they distort algorithms.



07

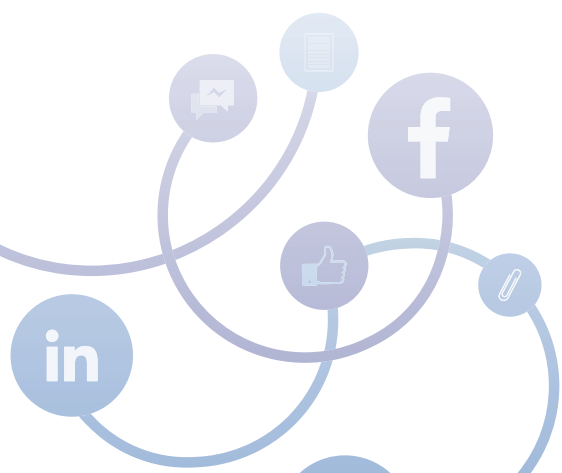
OUTDATED INFORMATION

The world is not static. People change jobs, acquire new competencies, views and outlooks change, relationships evolve and break, and people move on with their lives. Unfortunately for the data miner, social media accounts such as Facebook and LinkedIn need not necessarily capture all such changes.

Many social media profiles are full of outdated and obsolete information that present a completely different picture than the actual situation on the ground.

Marketers engaging candidates based on obsolete information may rake up demons from their past, making such engagements meaningless, or even counterproductive. For instance, a company may refer to an outdated LinkedIn profile to make a job offer to a candidate already hired and working for them since the previous month, and the gaffe would have far-reaching implications if the offered package is more than what the employee is presently getting.

Data miners need to be wary of accounts that do not show much recent activity, have not been updated for a while, do not have many followers, and display other tell-tale signs of obsolescence.

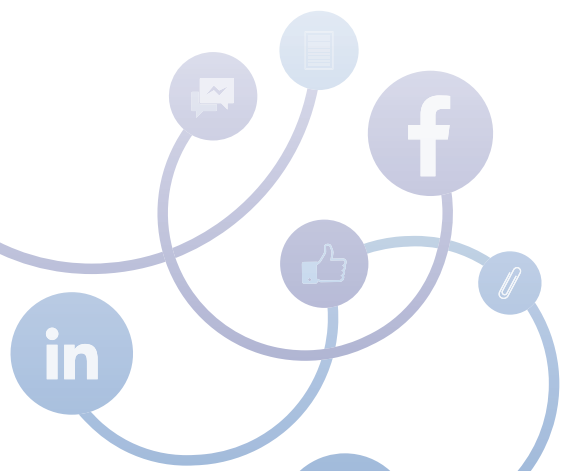


08

LIES AND EXAGGERATIONS

Many people chose to exaggerate or rely on half-truths to bolster their LinkedIn or Facebook profiles. At the very least, they may tend to be selective, including only positive or strong points, and omitting inconvenient aspects. Such gaps and wrong information can wreck havoc with the data miner's trends and predictions.

About 40% of people resort to outright lies on their resumes, one in two resumes have at least a small amount of misleading content, and about 78% of resumes are misleading.

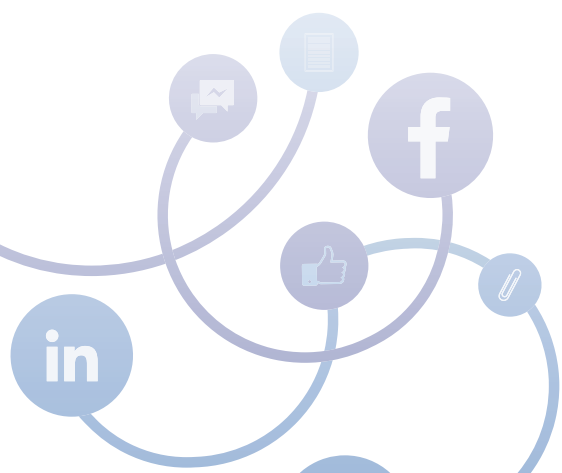


09

THE MENACE OF REPUTATION MANAGERS

The very popularity of social media as a platform where people can get genuine insights, different from the marketer's propaganda, has paradoxically led to marketing agents of all hues to converge on social media, in a bid to manufacture positive image for their brands. Reputation managers purchase likes, set up false profiles to communicate their marketing pitch through deception, and indulge in many practices such as black-hat SEO that range from unethical to outright fraud, all in a bid to present their brand in a positive light. Data mining tools are unable to cull the genuine from the fakes, and the widespread presence of fakes can even turn the results opposite to what the actual trend is.

A marketing research executive, on being flustered by the thousands of positive reviews and likes on a competitor offering, may decide to offer a similar service, only to find out the hard way that there is little or no demand for such a service in the first place, and all the likes and reviews were an attempt to generate interest and manufacture demand.



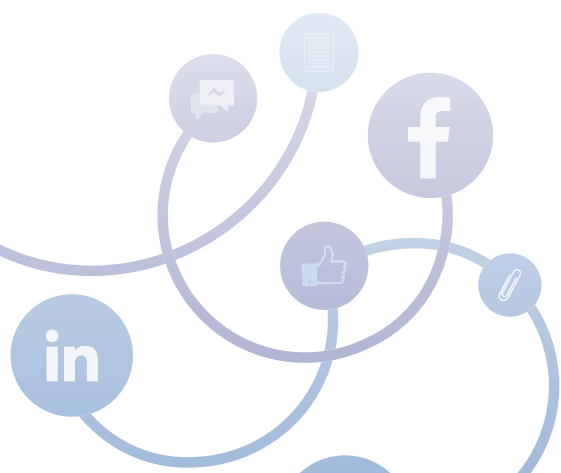
10 DISTORTIONS

“There are three kinds of lies: lies, damned lies, and statistics.”

– Mark Twain, Chapters from My Autobiography (North American Review, 1906).

Distortions need not be caused by lies, exaggerations and reputation managers alone. A case in point is a single high score boosting the average of a string of low scores. The risk of such distortion is more in social media channels such as Facebook, compared to anywhere else. A case in point is the celebrity page getting a million likes while all other pages in the target group getting only a few hundred likes. The million likes of the celebrity would skew the average to the higher side, creating an impression that more people like an average page.

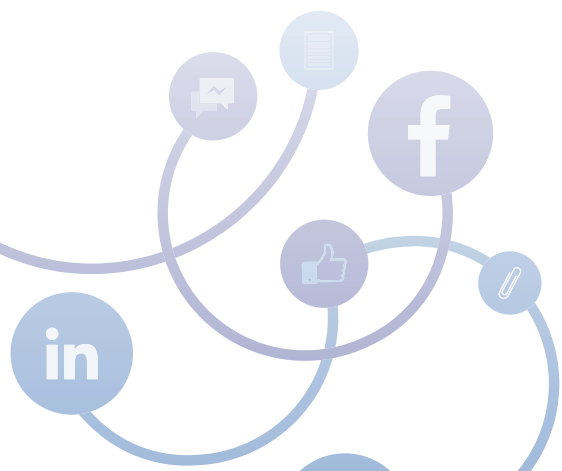
The big challenge before data miners is to remove outliers to get an accurate measure. However, in some cases, outliers do represent useful patterns. Making the system intelligent to distinguish between when to remove outliers and when to retain them works in theory, but is difficult, and invariably requires manual intervention.



11 ISSUE OF SERIOUSNESS

The accuracy of social media data depends to a large extent on the seriousness with which the members post their data. For instance, members may post serious information in LinkedIn, which is widely regarded as a channel for professionals. However, the same member may take his or her Facebook profile jocularly, indulging in bantering and other information that does not reflect their true characteristics.

A major problem is people liking pages and posts simply to humor their friends, or as a means of acknowledging them, even when they have no affinity to the contents of a page, or would not even have read a post.



12

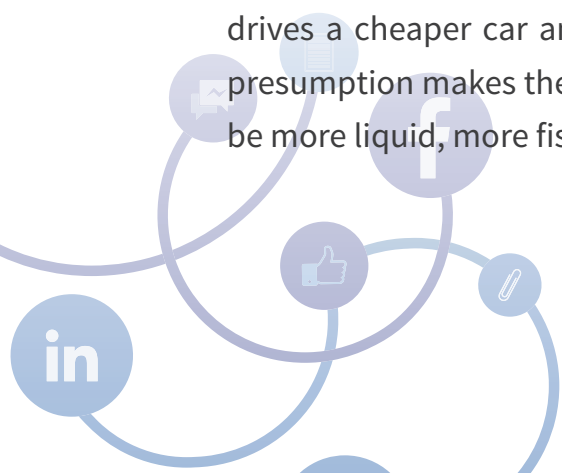
BIAS AND PRESUMPTION

At first glance, data mining seems to be a holy grail that would improve marketing effectiveness manifold. For instance, data mining helps to eliminate less desirable customers who are unlikely to buy, and direct resources towards customers most likely to buy. However, such an approach is rife with pitfalls.

The biggest pitfall is the evaluation dilemma. Traditional data mining is based on established ground-truths. Datasets are divided into training and test data, with training data used in learning, and the test data serving as ground truth for testing. However, such ground truth is often not available in the fast changing and highly fluid social media space, where trends change by the day and there is very little value for precedents.

In the absence of ground rules, subjective opinions take its place. The worth of such opinions depends wholly on the competence of the person making such opinions, and needless to say, bias, prejudices, and presumptions rule the roost in most cases.

For instance, a marketer may look to target a consumer who lives in the right neighborhood, has the right credit cards and receive the right catalogs, while eliminating a potential customer who lives in a less desirable neighborhood, drives a cheaper car and does no catalog shopping. However, their bias and presumption makes them oblivious to the fact that the second customer could be more liquid, more fiscally responsible and, in the end, a better prospect.



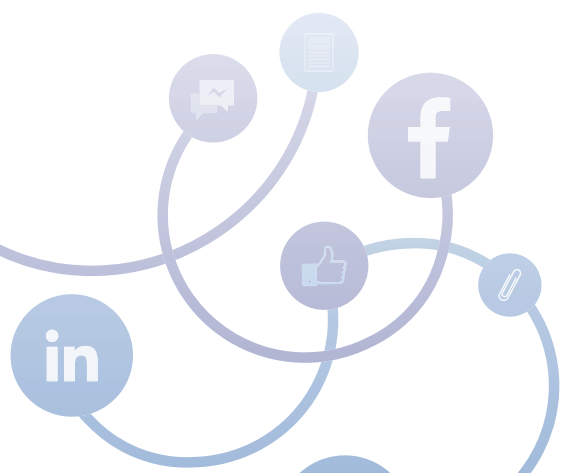
13

PRIVACY

Last but not the least is the issue of privacy. Data mining provides rich insights to marketers and other analysts that would allow them to package their services better for the customers whose data is mined. However, this is not always a justification for extracting and making use of data not intended for them in the first place. There is also the possibility of the mined data being abused, should it fall under wrong hands. The entire issue of data mining is clouded with ethical and legalistic roadblocks.

The debate on whether new opportunities can be allowed to subvert fundamental rights to privacy is very rife, and this has prompted many legislations in the US and elsewhere to put restrictions on unbridled data mining, and make it mandatory to let consumers know and get their consent before data mining can take place. Strictly legal data mining takes place only with the participant's consent.

Effective social media mining yields rich dividends, but only when such challenges are resolved.



SUYATI TECHNOLOGIES

Suyati is a young, upwardly mobile company focused on delivering niche IT services to support myriad Digital Engagement strategies. Our expertise also includes integration and delivery of CRM, CMS and Ecommerce solutions.

www.suyati.com

services@suyati.com

References:

1. http://books.google.co.in/books?hl=en&lr=&id=_VkrAQAAQBAJ&oi=fnd&pg=PR4&dq=facebook+mining+challenges&ots=JqjvqFSslH&sig=YZkH6MSQoA9rBdQo_YswB7P6WBo#v=onepage&q=facebook%20mining%20challenges&f=false
2. <http://www.sas.com/reg/web/corp/1845117>
3. <http://searchsoa.techtarget.com/tip/Data-mining-social-media-Twitters-untapped-potential>
4. <http://dmml.asu.edu/smm/SMM.pdf>
5. <http://www.insurancejournal.com/magazines/coverstory/2000/05/15/21117.htm>
6. <http://www.insurancejournal.com/magazines/coverstory/2000/05/15/21117.htm>
7. <http://www.techrepublic.com/article/can-data-mining-predict-the-future-of-your-enterprise/>
8. <http://www2.cs.siu.edu/~dche/publications/From%20Big%20Data%20To%20Big%20Data%20Mining.pdf>
9. https://www.cs.purdue.edu/homes/clifton/DistDM/Clifton_PPDM.ppt
10. <http://www.cnet.com/news/facebook-processes-more-than-500-tb-of-data-daily/>
11. <http://www.gradschoolhub.com/resume/>