# Utilizing Social Media for Lead Generation

MSc Reseach Project
Data Analytics

Ajitesh Prakash
15001113

School of Computing
National College of Ireland

Supervisor: Dr. Simon Caton

**National College of Ireland**
**Project Submission Sheet – 2015/2016**
**School of Computing**

| | |
|---|---|
| **Student Name:** | Ajitesh Prakash |
| **Student ID:** | 15001113 |
| **Programme:** | Data Analytics |
| **Year:** | 2016 |
| **Module:** | MSc Reseach Project |
| **Lecturer:** | Dr. Simon Caton |
| **Submission Due Date:** | 22/08/2016 |
| **Project Title:** | Utilizing Social Media for Lead Generation |
| **Word Count:** | 9.986 |

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

**ALL** internet material must be referenced in the bibliography section. Students are encouraged to use the Harvard Referencing Standard supplied by the Library. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action. Students may be required to undergo a viva (oral examination) if there is suspicion about the validity of their submitted work.

| | |
|---|---|
| **Signature:** | |
| **Date:** | 21st August 2016 |

**PLEASE READ THE FOLLOWING INSTRUCTIONS:**

1. Please attach a completed copy of this sheet to each project (including multiple copies).
2. **You must ensure that you retain a HARD COPY of ALL projects**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. Please do not bind projects or place in covers unless specifically requested.
3. Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

| **Office Use Only** | |
|---|---|
| Signature: | |
| Date: | |
| Penalty Applied (if applicable): | |

# Utilizing Social Media for Lead Generation

Ajitesh Prakash
15001113
MSc Reseach Project in Data Analytics

21st August 2016

**Abstract**

The Internet has not become the most widely used channel for communication but also the platform for forming and maintaining new relationships. The growth of platforms and the exponential rise in the user base of social media websites like LinkedIn, Facebook and Twitter, is evidence of the widespread acceptance of these networking platforms as the preferred means of communications and forging and maintaining relationships, posing an opportunity for business to exploit this facet of social media relationships for spreading awareness about the business and engage with prospective customers. The focus of this research is on the use of social media to identify relevant profiles or 'leads' for a business. The research utilizes data from social networking sites Twitter and LinkedIn in five different approaches and presents a methodology to generate leads. Twitter was found to be irrelevant for lead generation in the business context at hand. The presented methodology utilizes only four attributes from LinkedIn users' profile to generate high quality leads and is tested for robustness to variations in input data, different business context and vulnerability to noise in the input profiles. The results show the robustness and consistency of the presented methodology to generate leads utilizing a small subset of features.

## 1 Introduction

In marketing, lead generation is defined as the initiation of consumer interest or inquiry into products or services of a business [lea]. The word to note here is initiation. For initiating a persons' interest, we need to be fully confident about their intent and ability to consume the product or service. For achieving this level of confidence in a persons' ability and intent, we need to a 360 degree view of their personal and professional characteristics. In another approach to lead generation, a lead is a person who has indicated interest in the company's product or service in some way, shape, or form [hub]. Since, the interest was already triggered by the individual, we don't need to initiate their interest. They have already shown their interest in the product or services being offered. Their social media information can be utilized to understand them better and devise a personalized campaign to target them. The interesting part comes after a lead has been qualified. A qualified lead exposes the characteristics that the potential lead is more likely to have. The methodology described in this research utilizes the information of such qualified leads and leverages the information about their attributes to identify similar profiles.

The growth of social media in the past few years has exposed its immense possibilities. Exponential growth in user interaction and relationship on social media has generated a lot of personal and professional data which is publicly available on the various social media platforms. These data expose user preferences and habits and their personal and professional standings. This has opened the possibility of analyzing these data to understand and predict their behavior and preferences. Not only social media allows for understanding individual behavior but behavior of community of people and identify people who are similar in terms of their interests and intellect. This knowledge open doors for businesses to analyze trends and topics of discussion in communities of people and leverage this knowledge to streamline their practices to be proactive rather than reactive to changing user preferences and interests. Social media can help businesses find new clients by using users own shared information to identify interest. Rather than reactively targeting users who search a certain product or topic, businesses can proactively targets relevant users even before they begin their search for a product [Ganguly, 2015]. Information available on the social media provide the holistic view of personal and professional life of individuals giving the 360 degree view which we need to understand users better and make them an offer they can't refuse. Lead generation thrives on the abundance and diversity of user data available on the social media.

The amount of information available on these social media platforms about individuals' preferences, accomplishments - personal and professional and aspirations, enables business to identify opportunity to position their product or services in a prospect's mind. LinkedIn and Twitter, for example, are two of the biggest platforms for tapping in individuals' professional and personal representation, respectively. This available information allows for businesses to filter and identify the most relevant prospects for their product and services. Another approach, described in this research, to lead generation is leveraging the available information about an existing customer and scan the social media networks to filter for similar profiles.

The business context at hand was to expose more individuals who qualify or meet a given set of characteristics. In the absence of any automated methodology to generate quality leads, the process relies heavily on manually scanning the social media looking for specified attributes without any measure of relevancy of the leads generated apart from individual discretion and not to mention the long hours of work that is needed. This research was conceived on the need for a solution for a smarter and faster mechanism to tap in this information repository to generate quality leads for the inside sales team which have traditionally relied on the manual effort of its sales analysts to scan the database looking for specific keywords in the individuals profile and mark them as a lead. This approach is not only time consuming and exhausting but prone to rejection of relevant leads due to limitation posed by traditional filtering methods. This information loss means that sales team often loses out on significant portion of relevant leads and then cold calling the small number of identified leads, only to disqualify them later, at the cost of being branded as a 'spammer' and an upset person out there in the world who can in matter of minutes spoil the brand image of the company using the same social media platforms which lead the company to them at the first place.

The devised methodology is not a filtering technique in the sense that it doesn't discards a profile if it fails to meet a condition on a given attribute but it matches the 'similarity' of the profiles' attributes with the selected attributes from a existing customer or any ad-hoc requirement. So, if a profile have more common terms with the selected profile, its similarity score is higher than compared to a profile with has one or two terms in common with the selected profile. This not only insures that the generated leads are most relevant but also allows the sales team save time and increase operational and performance efficiency by ordering prospects with respect to their relevancy.

The data utilized in this research comes from two social media platforms: Twitter and LinkedIn. The Twitter database maintained had over 2 million line items however, the LinkedIn database at the time of documentation, had around 70,000 individual profiles. It was also noted in the course of research, of these 70,000 profiles not all profiles had an associated twitter profile. This was also a reason for not including twitter information in the final suggested methodology, as discussed in the 'Methodology' section. The research utilizes the techniques of natural language processing and text mining, information retrieval technique (TF-IDF) and distance measuring technique cosine similarity to calculate the similarity scores between profiles. The research question builds around identifying the most suitable attributes of social media representation of individuals, which best capture the desired characteristics of the leads in the business context at hand. During the evaluation, the research assumes the hypothesis, that that there should be no significant difference in the score of the profiles with removal of seed profiles at each iteration. Discussions and results of the hypothesis are presented in the evaluation section.

In the following sections of this research paper, related work in the area of lead generation is discussed. This section also discusses the various principles of social media relationship which the research utilizes to understand and select the attributes for lead generation. Following this section, the paper summarizes the adopted methodology and discusses the main features of the devised algorithm. Next, the paper presents an in depth discussion of the implementation of the methodology, the employed and tested approaches along with discussion on data collection and processing. Following the methodology, in the next section this paper presents the evaluation of the employed approaches, discussion on the results and rationale behind selection and rejection of approaches. In the final section, the paper concludes this research by highlighting the main findings, limitations and future work.

## 2 Related Work

Lead generation is the process of identifying new prospects based on individuals' relevancy from the point of view of the business for a business or based on the information available about any existing customers [hub]. Traditionally, leads were generated by reaching out to a huge chunk of population with relevant demography via email, direct phone calls or mass advertisements. However, these methods had been not very successful in generating the required number of customers for a business leading to undesired sales funnel and also exposing a brand to potential repercussion of unwanted intrusion into prospects' privacy.

With the growth of the Internet and rise of social media, the amount of information available about individuals' preferences and habits, has grown exponentially. This presents immense opportunity for businesses to better understand and target their customers and process these information to identify the attributes of a potential customer. This growth in data coupled with the advancement in computing power and machine learning, has led to a disruption in the sales industry seen in the form of smarter platforms for customer engagement and relationship management and smarter algorithms are constantly replacing humans for chore in the sales process of a business.

Lead generation in particular has seen a massive disruption in its approach caused by the advent of intelligent computers and algorithms. This has caused this process to gain a separate identity for itself in the corporate world, which has lead to it becoming an industry in itself. In the past few years, lead generation has attracted attention of both entrepreneurs and established firms, leading to a large number of platforms and service providers providing software as a services (SaaS), such as Socedo [1], Unomy[2], InsideView[3], especially focused on helping the inside sales team of a business with quality lead generation. The services provided by these SaaS providers is not very different from each other in terms of functionality. However, the main differentiating factor is the underlying algorithm that is used to identify leads which is essentially becomes a trade secret.

Consequently, there is very little documentation of the algorithms and approach employed to identify leads and no documented research was focused on solving the problem of lead generation. This is the reason why in the course of this research, very little information was available in the research arena regarding the concepts and techniques employed. Of what was available, a majority of them were patent applications not focused on sales lead generation and hence were not very relevant for the purpose of this research. In their paper *(a patent application)* [Wilkins and Zoken, 2005], Wilkins and Joken, discuss an approach of extracting information from the classified ads in the newspaper, magazines or web, to tap in intention, ability and willingness of people to buy a particular good or service. The paper uses example of automobile advertisement and asserts that someone selling a car is giving signal of possible purchase of a new car and the prospect identifier, number or email, uniquely identifies their likeliness to make the purchase. The paper comes in proximity of the scope of this research which leverages publicly available information to generate leads. In the sections below we discuss the attributes of social network dynamics which this research leverages to understand the intricacies of social network relationship and select attributes of individual profiles best suited for lead generation.

## 2.1 Social Network Relationships and their Dimensions

The main reasons behind social network relationships are trust and mutual benefit through exchange of information [Brown et al., 2007]. These interpersonal ties between the members of these social networks, motivates them to achieve mutual growth and goals. To understand the rationale behind these communities and communications, understanding the intrinsic characteristics the relationships between its member becomes very important. In this section we discuss the various dimensions of social media relationships.

### 2.1.1 Nature of Weak Ties

In 1973, Mark Granovetter published his paper titled "Strength of Weak ties". The paper argued that "the degree of overlap between two individuals' friendship network varies directly with the strength of their tie to one another and explored the impact of this principle on diffusion of influence and information" [Granovetter, 1973]. The underlying rationale to understand these interpersonal networks is defined in terms of the

[1] http://www.socedo.com, Last Visited: 08/19/2016

[2] https://www.unomy.com/, Last Visited: 08/19/2016

[3] https://www.insideview.com/, Last Visited: 08/19/2016

strengths of relationship, or tie strengths, between people in these interpersonal networks. "Tie strength is a multidimensional construct that represents the strength of dyadic interpersonal relationships in context of social networks" [R. Bruce Money and Graham, 1998].

Relationships on social networks are not created equally. Different people have different levels of relationship between them, Granovetter identified four critical dimensions of these interpersonal relationships: the amount of time the relationship has been existing, the emotional intensity level of the two people in a tie, the level of intimacy or mutual confiding/trust and the reciprocal services in a sense of social media it would mean the number of direct messages or tweets replied to, for example.

Based on these four criteria we can broadly identify interpersonal relationships to be either a strong tie or a weak tie. The focus of Granovetter was on studying the contribution of weak ties in the relationship. Weak tie or loose acquaintances can help a fiend generate creative ideas or get help them find a job. These weak ties also expedite the transfer of knowledge across different groups and this exposure to outside groups and consequently to the flowing information in those groups, is what gives the weak ties a leverage over strong ties. Strong ties are spatially limited to the wedged networks of similar people. Strong ties, generally refers to trusted friends and family, and there is a high level of intimacy and emotions involved in these ties.

Weak ties are critical in the process of information diffusion in a community and leading us to prospects. In context of lead generation, a prospect who has a few shared characteristic with the selected profile, is weakly tied to him or is a weak lead which will have a lower similarity score and is most likely to be disqualified from the sales funnel. Good quality leads tend to have a high similarity score with the selected profile, more similarity, more shared interests and aspirations, a tie which is strong in nature.

### 2.1.2 Nature of Strong Ties

Low similarity scores indicates weak ties. The main reason for this is the lack of intimacy and intensity in their relationship with other members i.e. fewer shared attributes. Good leads thrive on strong and credible social relationships and the strength of these relationships depends on number and type of information exchanged, frequency of all such exchanged information and the level of intimacy between the source and other community members.

Strong ties are characterized by (a) a sense of uniqueness and intimacy in relationship, (b) voluntary investment in the tie, (c) a desire of companionship with the partner, (d) interest in frequent interaction and (e) a sense of responsibility and commitment to grow the relationship [Walker et al., 1993]. Individuals who share such relationships interact more frequently and share more information compared to individuals in a weak tie. A relationship defined as strong would have high above average scores for similarity indicating more common attributes in user profiles.

Conclusion of various research carried out in the field suggest that information from a strong tie relation is perceived by members to have a positive effect on their decision making [Brown et al., 2007]. From a lead generation context, this conclusion is motivating since, if a profile is identified as a strong lead that means they are more likely to consider products and services offered, since someone whose opinion matters to them is already using that product or service.

The role of situation was not considered in the works of Granovetter in his theory of strength of weak ties. David Krackhardt in his paper [Krackhardt, 1992], leveraged the effect of situation to map the psychology of strong tie. Strong ties were found to be more useful to individuals in an insecure position, severe change or uncertainty. In the context of lead generation, this would relate to a situation of reaching a person who is seeking information about some product of which they are in a need, for whatever reason, or a person deciding to replace an existing product or process.

The beauty of corporate world is that if a competitor or a friend changes a habit, it can be assumed with certain degree of confidence that others in the same arena will change the habit in the near future. The change in habit is universally separated only by the time when the need for change is realized and materialized. Leveraging this aspect of corporate relationship, a strong tie between the an existing customer and the lead, lead generation may be seen as a proactive advertising of the product and services in anticipation of need for the services being offered in the near future.

In the next section the focus shifts to understand the rationale behind how humans form a relationship and ability of an individual to secure benefits by virtue of membership in social networks or other social structures.

## 2.2 Homophily and Social Relationships

The goal of lead generation is to leverage the available information about existing customers to identify similar individuals based on certain preselected attributes which best define a prospect for the business. “The similarity of individuals makes them susceptible to a greater level of interpersonal interaction, trust and understanding” [Martin Ruef, 2003] and hence makes the entire process of lead generation more fruitful.

Association of similar people is not a new concept and has been talked about by the likes of Aristotle and Plato. Aristotle in *Rhetoric* wrote the people *“love those who are like them”*. Plato in *Phaedrus* wrote “similarity begets friendship”. The homophily principle is that similarity breeds connection and the resulting connections are homogeneous with regards to many socio-demographic, behavioral and interpersonal characteristics. The result of this selective bonding is that homophily tailors the social networks of individuals significantly affecting “the information they receive, the attitudes they develop and the interactions they experience” [McPherson et al., 2001]. The principle of homophily also adverts that contact between similar people occurs at a much higher frequency than dissimilar people and that the information flow in such networks tends to be localized.

The principle of homophily best serves the purpose of lead generation. This research leverages LinkedIn and Twitter data of existing customers to find prospective customers. Attributes like customers’ professional role, company or speciality are selected to identify prospective customers. In a corporate world or otherwise, people make friends with people who are like them in terms of aspirations, intellectual or material. Birds of feather, flock together. This adage best represents corporate or business relationship. The needs of one company in a industry slowly or quickly becomes the desire of others and this desire is what the service providers thrive on. The entire rationale behind lead generation is to identify individuals who are swimming in the same waters as the businesses’ existing clients.

Homophily can broadly be caused by either similarity in demographic characteristics such as sex, race/ethnicity, age etc. or psychological characteristics like intelligence, attitudes and aspirations or geography characteristics such as location or organizations [McPherson et al., 2001]. Also, homophily tends to grow stronger as the diversity of relationship between two individuals increases. In a broad sense homophily can be categorized as being either a (a) status homophily, in which similarity is based on informal, formal, or society ascribed status and (b) value homophily, which is based on values, attitudes, and beliefs and decides future behaviors and any homophily would begin with the status, and slowly move to the latter as the social position changes [Lazarsfeld et al., 1954]. People who are more structurally similar to one another are more likely to "have issue-related interpersonal communication and to attend to each others issue positions, which, in turn, leads them to have more influence over one another" [McPherson et al., 2001] and this influence is not restricted to issue related communication but also seen in other types of advice, friendships and association.

The next section discusses the concept of social structure which basically is the recognition of the value people hold in a network. From lead generation point of view, valuable profiles exposes attributes that are of value to the business while looking for future customers.

## 2.3 Social Capital

"Social capital can be defined as the ability of an individual to secure benefits by virtue of membership in social networks or other social structures" [David and Jon, 2010]. Any individual has three kinds of capital that he can bring with himself in a relationship: (a) Financial Capital which is his monetary possessions, (b) Human capital which are attributes of their personality such as intellect and (c) Social capital, which is an attribute of his relationship with others, friends, families or acquaintances [Burt, 2004]. The social capital of an individual becomes the social capital of the community he is part of. Social capital is different from the other two in a sense that it is not an exclusive property of an individual but a joint ownership of him with people related to him in the community and it is through the use of this social capital than an individual creates opportunities for himself and transforms the other two into profit [Burt, 2004] From a lead generation point of view, the social capital of an individual is an essential factor deciding the relevance of the profile to act as the seed to generate leads. Higher the social capital of the seed, better the quality of leads generated. From the point of view of this research, the social capital of the seed is determined by the client which is provided in form of feedback on the list of seeds shared, for they understand their business requirements the best and know the exact quality they are looking for in their customers. Once the seeds have been selected, the characteristics of their different attributes are extracted and leveraged to find similar characteristics across respective attributes of prospective customers. In a given social network, not all individual have the power to influence others. Also, there are many people with financial and human capital at par with each other. It is the social capital of an individual which becomes critical in evaluation of the perceived value of the individual amongst the members and finally gives them the leverage that they enjoy to act as a seed.

In this research a core challenge was around selecting the attributes which best capture the preferences of the clients and relevancy of the profiles for the business context at hand. From weak ties and strong ties discussion, it becomes apparent that for someone to lead to someone similar, they must have a lot of commonality in their attributes. A strong tie connection is a reflection of shared interests and habits, thus looking for strong tie connections of a person, exposes the individuals who are most similar to him. But, strong tie relations are driven by homophily. People who are close, are close because they trust and understand each other, have similar social status, have common personality traits and enjoy identical professional accomplishments. The attributes selected for generating leads in this research are most reflective of the principles of homophily between two individuals and their current employers.

We have discussed the 4 main aspects of social relationship, how each of them plays its own part in forming of human relationship and in connecting individuals with people who are similar to them and how from the context of lead generation, homophily plays a pivotal role in determining the similarity in the attributes of individuals who are share a strong tie between them. In the next section we discuss related work in the field of lead generation and then go on to discuss the methodology, implementation and evaluation of the proposed methodology followed by a short discussion. We conclude with summarizing the findings in the 'Conclusion' section and proposing the extension of the research under 'Future Work' section.

# 3 Methodology

The focus of this research is to devise a methodology to generate leads for the sales team, leveraging the information available for the existing customers. The initial list of prospects shared with the client is generated by a simple filtering based on the designation of an individual. Suppose, the request was to locate Marketing managers, a list of 20 prospects *(number based on the subscription plan)* is shared with the client generated by only considering the designation of the prospects. The methodology developed in this research is adopted in the second and following iterations of the lead generation. Typically clients' provide feedback on the list of 20 prospects shared with them, categorizing them as either good, mediocre or bad leads and giving reasons for the same. The ones classified as good leads (selected by the client from the list shared), then serve as the 'seed' for the developed methodology, which leverages the attributes of qualified leads to identify more profiles like them. Client's choice of leads exposes a lot of information about their intent and focus, preference of leads. Attributes like *Industry*, *Specialties* etc, from prospect LinkedIn profile, gives valuable information about the relevant industry and the skill sets that are attractive or align most with the client's requirements and aspirations.

The research begins with obtaining a list of prospects which was shared with a client in response to their request. The client is a service provider and had some services for the companies in games and entertainment industry. The request was to share a list of managers from marketing department from the specified industry. A list of 20 prospects was shared with the client and the client had qualified certain prospects from the list as good leads and provided feedback on why a prospect was selected or why they were not. The research selects the qualified prospects, 5 in number(from the list shared, 5 were selected by client), obtains their relevant LinkedIn attributes and leverages the information to develop a methodology to utilize this information to generate a better list of leads.

The methodology begins with filtering for profiles with 'marketing' and 'manager' in LinkedIn Headline and filtering for relevant industry. The keywords for pulling the population from database comes from the client request. Here, in this case, client was looking for marketing professionals and hence all the relevant profiles were collected using 'marketing as the keyword. *Headline*, a LinkedIn attribute of a user profile, is a snippet of professional standings of a user, like their current designation and present company.

Profiles for all users associated with the marketing industry is obtained by filtering for profiles with word 'marketing' in their headline. The list of seed keyword can be extended to incorporate any specific request from the client. Apart from *Headline*, four other attributes from LinkedIn are utilized initially in building this methodology and these are: *Current Employer*, which essentially is the present company of the user. *Industry*, is the industry user specifies in his profile. *CompInd*, or Company Industry, is the industry which users' company specifies on the companies page and finally, *Specialties* which is the the area of expertise of the company as specified on the company's page on LinkedIn.

Another social media platform utilized in this methodology initially, was Twitter. Twitter is very different from LinkedIn in the sense of the user activity and expectations. LinkedIn is a highly professional network of individual and the activities of users on this site focused on the professional representation of users. Twitter on the other hand, is more personal than LinkedIn and user activities on twitter exposes users' personal and to some extent professional representation. *Bio Description*, a twitter attribute of user profile, is the personal representation of the user and typically reflects users' personal interests and preferences. This methodology utilizes users' tweets and bio description initially, in generating leads for clients.

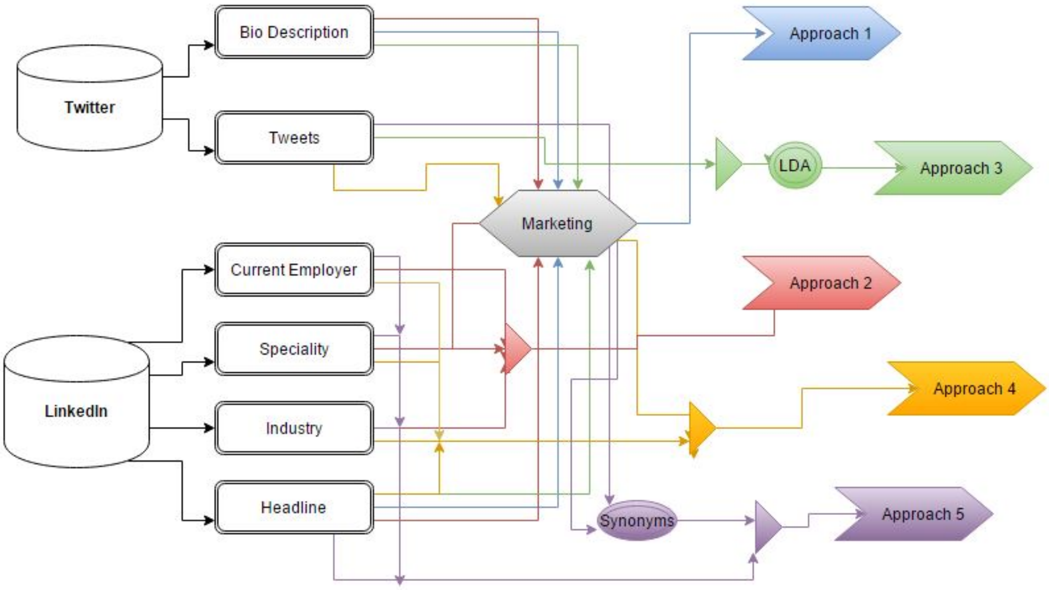


Figure 1: Flow Diagram of the 5 approaches employed

## 3.1 Cosine Similarity

One of the many ways of measuring similarity between profiles is by using the concept of distance between the two profiles.The rationale behind it is simple: profiles that have higher number of terms in common are closer to each other than those that share fewer terms between them. The problem with distance, however, is that it can be skewed by the term count in the corpus of the profile: profiles that have an high term count owing to more descriptive headline or specialities, or repetitive occurrence of the keyword, will show up as most similar even though they may be moderately similar to other profiles in the corpus. For this reason, the methodology utilizes a related measure of similarity that does not suffer from this problem.

A profile can be represented in terms of its constituent keywords as a point in a coordinate plane with its dimensionality equal to the number of distinct keywords it has in it. Using this concept, similar profiles are the ones which have a high degree of overlap between their keywords. On the coordinate space, the vector representing them will coincide indicating identical profiles or would have a very small angle between them reflecting high degree of similarity. On the other hand, two dissimilar profiles will have high degree of separation between their vectors. Two vectors at right angles indicate completely dissimilar profiles. Thus, what we need, therefore, is a quantity that varies from 0 to 1 depending on whether two profiles (vectors) are dissimilar(at right angles to each other) or similar (coincide, or are parallel to each other). The cosine trigonometric ratio has exactly this property. The cosine of the angle between the vector representation of two profiles is a reasonable measure of similarity between them and is referred to as cosine similarity.

## 3.2 Adjacency Matrix

If there are N profiles, leveraging the cosine similarity principle we can measure similarity between every pair of individual profile. This would give us a set of N*N numbers between 0 and 1, which can be represented as a matrix. This matrix is what is called the *DTM* or Document Term Matrix also referred to as a Adjacency Matrix. The diagonals of the adjacency matrix are set to 0 because the intersection represents similarity for each individual with himself and it doesn't require much thinking to agree that a profile would always be similar to itself.

## 3.3 Research Question

This research builds on the question of identifying the attributes of social media which best capture the intent and ability of individuals be relevant for a business to engage with. To reach at final methodology with best results of leads, five different approaches were tried utilizing the Twitter and LinkedIn data. These approaches are summarized in the implementation section with discussions on the results obtained and why a particular approach was discarded. Approach 2, without utilizing the twitter bio description, presents the best results and is what goes into devising this methodology. To validate the robustness of the methodology, it was tested with addition of noise in the corpus in form of bad or mediocre leads to the corpus and also by removing a seed profile in each iteration from the corpus to study the effect on the result. The results are presented in the evaluation section. During the evaluation, the research assumes the hypothesis, that that there should be no significant difference in the score of the profiles with removal of seed profiles at each iteration. Discussions and results of the hypothesis are presented in the evaluation section.

# 4 Implementation

The research begins with obtaining a list of prospects which was shared with a client and the client had qualified certain prospects from the list and provided feedback on why a prospect was selected or why they were not. The research selects the qualified prospects, 5 in number, obtains all their relevant LinkedIn and Twitter attributes and leverages the information to develop a methodology to utilize this information to generate a better list of leads.

The methodology begins with filtering for relevant profiles from the database based on the keywords specified by the client. In this case, client was looking for marketing professionals and hence all the relevant profiles were collected using marketing as the keyword to filter based on its presence in their headline. Around 4,600 prospects appear in the database. A random list of 20 profiles is shared with client from these 4,600 prospects. Client selects or qualifies leads from these 20 and sends back the list to the sales team. The methodology uses this information to select more relevant leads. The rationale is simple, identify all the relevant prospects from the database by leveraging the information about clients' preference of prospects.

During the course of the research, 5 different approaches were considered to identify all the relevant prospects from the database by utilizing Twitter and LinkedIn attributes of these profiles. These approaches build on the concepts of homophily and social capital discussed earlier. The main focus of the approaches implemented was to capture the user intent and ability for a business engagement utilizing the attributes from their social media. The attributes utilized from LinkedIn capture homophily around the professional networks of individual and their headline reflects their social capital. Attributes from Twitter capture user homophily around personal networks, for example, bio description generally reflects user interests and opinions and high similarity between bio description indicates shard interests and habits. These 5 approaches are summarized below.

## 4.1 Twitter with LinkedIn User (TLU)

All profiles with keyword 'marketing' in their LinkedIn headlines and Twitter Bio Description. Combine these two attributes for everyone identified, preprocess the corpus and then run similarity check of these individuals with the corpus of the 5 qualified leads headline and Bio Description.

## 4.2 Twitter with LinkedIn User and Company (TLC)

All profiles with keyword 'marketing' in their LinkedIn headlines and Twitter Bio Description. Tap in attributes for their companies from LinkedIn, namely, Company Speciality and Industry. Combine all these attributes for everyone identified, preprocess the corpus and then run similarity check of these individuals with the corpus of attributes for the 5 qualified leads.

## 4.3 Twitter with LDA and LinkedIn User and Company (LDA)

Collect all profiles with keyword 'marketing' in their LinkedIn headlines and Twitter Bio Description. Collect all tweets from these identified individuals. Create corpus of tweets for each one of them, run LDA on the corpus of their tweets, pick the most relevant bucket. For the users in these picked bucket, collect their company attributes. Combine the headline, description and company attributes, preprocess the corpus and then run similarity check of these individuals with the 5 qualified leads.

## 4.4 User Tweets with LinkedIn and Company (UTLC)

Collect the top(in terms of recency), 5000 tweets about marketing. Get their twitter bio description combine with LinkedIn headline, and company attributes and build individual corpus. Preprocess and run similarity check with the existing customer.

## 4.5 Tweets with Synonyms, LinkedIn and User Company (SYN)

Collect the top 5000(arbitrary) tweets about marketing and synonyms of marketing. For all the individual twitter handle, obtain their twitter bio description combine with LinkedIn headline, and company attributes and build individual corpus. Preprocess and run similarity check with the 5 qualified leads.

## 4.6 Data Collection

The data utilized in this research comes from the database of a company in lead generation industry. The current interest of the firm aligns with the outcome of this research and the data was provided on the condition of anonymity and utilizing the findings of the research to improve their understanding of client's requirements and meliorating their method of lead generation. The firm provided access to their Twitter and LinkedIn database for the purpose of this research. Twitter database had an approximate line item count of over sixteen million and for LinkedIn, it was over seventy thousand, at the time of the data extraction. The two data tables, Twitter and LinkedIn related by a key named 'UserId' which is system generated.

The rationale behind utilizing both Twitter and Linkedin is based on the concept of homophily discussed earlier. Homophily is substantially explained by Demographic and Psychological characteristics and is of two types, 'status' and 'value' [McPherson et al., 2001]. Demographic characteristics are generally ascribed such as education, race/ethnicity, age etc. and psychological characteristics includes characteristics like intellect or aspirations etc. Status homophily is based on socio-demographic factors such as education and occupation while Value homophily is based on psychological characteristic. The user details captured from Twitter are their tweets and bio-description and from LinkedIn, their skill set, current industry and headline, which is a brief professional summary. Through these fields users' status and value are captured.

## 4.7 Pre-processing and Data Preparation

The data collected from Twitter and LinkedIn is used to form a corpus. The corpus needs to be cleaned to remove redundant characters and English language stop words before analysis. The tool utilized to clean the corpus is R and package *TM* or Topic Modeling. TM package allows numerous transformations for cleaning the data such as *removeNumbers*, *removePunctuation*, *removeWords*, *stemDocument* and *stripWhitespace*. After the corpus has been cleaned it can be used for analysis. The TM package utilized to treat the corpus is used to generate a list of terms in form of matrix, called DTM or the Document Term Matrix. This matrix consists of all the terms in all of the users' corpus from Twitter and LinkedIn along the columns and individual users along the rows and each cell of the matrix is either a 0 or 1, indicating the presence or absence of the corresponding term in the respective user profile.

### 4.8 Data Description

The two social media platforms utilized in this research are Twiiter and Linkedin. The size of the twitter and the LinkedIn databases are very different, Twitter runs in millions but LinkedIn had just around 60K profiles at the time of research. Owing to this disparity in the database sizes, most of the profiles who were filtered owing to presence of keyword in their twitter bio description, for most of them, their LinkedIn profile was not available thereby inducing availability of information bias, i.e everybody who has a twitter profile doesn't necessarily have a LinkedIn profile and vice versa. However the number of incomplete rollback from Twitter to LinkedIn was significantly larger than that from LinkedIn to twitter. Also, from the business point of view, self-representation, which twitter description actually is, is not of high relevance since clients are interested in the professional standings of an individual for their business aspirations.

Generally Twitter Bio description is up to 160 characters which is larger than the corresponding Headline of individual profiles on LinkedIn, which can have a maximum of 100 characters. Also, personal representation of two individual is very different from each other however their could be a high level of similarity in their LinkedIn Headline. Suppose, two individual who are Chief Marketing Officers would typically have the almost similar thing mentioned in their LinkedIn headline. However, their twitter bio could be miles apart from each other and this induces a anomaly in their similarity score. The cosine similarity score for these two individuals should be high considering their identical headlines but owing to presence of large number of non matching keywords, pulled from twitter bio, their scores go down. Statistically speaking, the precision *(How many selected items are relevant from the whole box?)* of the algorithm in identifying similar leads decreases owing to the increase in the total number of terms which increases the size of the denominator. And interestingly, the terms that increase in the denominator are mainly elements of personal representation of the individual and are not relevant for the business context at hand. The same was reflected in the quality of leads generated using the 5 approaches. Hence, the research excluded twitter database from the analysis and all the 5 approaches were repeated considering only the LinkedIn attributes of the profiles and qualified leads.

## 5 Evaluation

The methodology presented for lead generation in this paper feeds on the LinkedIn and Twitter profile attributes of the users and checks for similarity of the user profiles and the seed profiles. There were 5 different approaches employed which utilized Twitter and LinkedIn attributes to generate leads. The evaluation of the approach and the quality of the leads generated from it is centred around the similarity scores of leads with the seed profiles and the affect on the scores with the variations in the seed profile. The main focus of the evaluation is a) whether the leads generated are relevant and b) how robust is an approach to variations in input seed profile and business context. Another aspect of evaluation is the quality assessment of the leads generated from the methodology, by domain experts available in the sales team of the industry partner of this research.

The lead generated from an approach depends on the selected cut off level. The cut off level is reflective of the sensitivity and specificity of the business context. In cases where the business context is to identify a lot of leads meeting a basic set of criterion, setting a lower cut off level allows the methodology to take a liberal approach towards qualifying profiles as leads. On the other hand, when the requirements are very specific in terms of the designation and industry of the leads, setting a higher cut off would make the methodology stringent in its approach to qualify profiles as leads. In this research, cut off level was selected to be 25% to accommodate for a through analysis of the methodology and the quality of leads generated.

The profiles qualified as 'good' leads for the client, generally have a similarity score with the 5 leads above a pre determined cut off in the similarity score. The cut-off can be changed at discretion to qualify more or less individuals as prospects. Approach 4 and Approach 5, suffer from their reliance on twitter to filter for the relevant profiles. The problem, mentioned earlier, of large number of incomplete rollbacks from twitter to LinkedIn, fails to leverage LinkedIn details of more than 90% of the profiles and hence essentially is not a good lead generation mechanism.

In the context of not considering Twitter attributes, Approach 1 only leverages the headline of the profiles for lead generation. This is essentially is filtering where everybody with identical headline would be a good lead. Approach 3, has three problems. First, it relies on the user to make a selection of the relevant bucket, exposing itself to observer, selection and cognitive bias. Second, LDA is computationally expensive. Processing LDA on all the tweets of over 5,000 profiles requires heavy computation power and a lot of time. The research selected a random sample of 100 profiles from all the relevant profiles to continue with the approach to generate leads. Third, the reliance on Twitter exposes it to the problem highlighted earlier. Approach 2, was found to be most consistent, feasible and reliable method of lead generation. The quality of leads generated was manually verified by the domain experts at the company and found to be highly relevant and of high quality and the client feedback on the leads generated at a given cutoff was positive. Hence, Approach 2 was selected to be the most efficient methodology of lead generation.

The central element of the process is deciding what goes into the corpus. The methodology essentially calculates scores based on the similarity or common terms between the corpus of the individual and that of the seed profiles. With the LinkedIn attributes corpus available, there are two paths to traverse:

a. To find the similarity of each of the filtered profile with each of the 5 prospects. i.e an individual list of leads for each of the 5 qualified leads. or,

b. To find the similarity of each of the filtered profile with these 5 prospects as 'one entity i.e. one lead list consisting details of individuals who are closest in terms of their attributes with the 5 prospects as a single entity i.e. most similar to everybody considered at once.

Both these paths were traversed with Approach 2. Path 2 is logically more relevant to the business context since by combining attributes of the seed profile and using the combination to filter for similar profiles, allows us to consider all the relevant attributes that are of interest to the business. By qualifying leads from initial list of prospects, a client basically indicates the attributes that are most relevant to them. For example, the 5 seed profiles utilized in this research have following headlines: 'User Acquisition Manager', 'Head of Marketing', 'Vice President of PR  Marketing', 'Digital Marketing Executive' and 'Marketing Director'. This selection indicates the preference of client for the prospect professional standing. By leveraging all these together we give the algorithm more keywords to match for. So somebody with a headline, say, 'Vice President Digital Marketing and Acquisition' is a hot lead since he is reflecting characteristics of 3 of the 5 qualified leads and consequently is assigned a higher score. Now consider his possible range of score with individual seed profiles. Of the 3 keywords from the first seed, he has only one keyword in common, which pushes his score down. Statistically speaking, the recall *(How many relevant items are selected?)* of the algorithm decreases, which is essentially loss of information. With the second path the algorithm maintains a high recall by utilizing all of the relevant keywords to filter for similar profiles. The same is reflected in the high relevancy of the leads generated with the second path of considering all the seeds as one entity to generate leads. Summarized below is the scores of the top 5 profiles with and without considering Twitter attributes

| Lead | With Twitter | Without Twitter |
|---|---|---|
| A | 0.479 | 0.669 |
| B | 0.451 | 0.368 |
| C | 0.431 | 0.647 |
| D | 0.418 | 0.216 |
| E | 0.410 | 0.467 |

Figure 2: Score comparison of Leads with and without Twitter

It is interesting to note, that Lead D with Headline: ‘Digital Marketing Assistant at XYZ’ from Industry: ‘Pharmaceuticals’ is not relevant for the business context and the LinkedIn without Twitter Approach, at the cut off of 25% doesn’t qualify him as a lead. While considering Twitter attributes puts him in Top 5 due to similarity in its Twitter profile with the seed profile, however, this similarity is not relevant for the business context at hand. Lead A and C were high quality leads acknowledged by the client, and their scores with the LinkedIn only approach reflects it.

With the approach established for lead generation utilizing all the seeds as one entity, the next step is to test the affect of decreasing the number of keywords made available to the algorithm for generating leads. The research proceeded further to test the sensitivity of Approach 2 towards variations in the input data. With the similarity scores available for each of the filtered profiles with the ‘one entity approach, the methodology starts removing one prospects from the ‘entity at a time and calculate the score after each removal. This is done by employing the method of random selection. We randomly select 2 profiles, 3 profiles and 4 profiles from the seed profiles and feed it to the algorithm to leverage these seed profiles to generate the leads. The process is iterated multiple times to remove each profile at least once from the seed.

This is where the research also test the hypothesis stated earlier. Any significant drop in the score of the leads generated with the all 5 seeds as an entity compared to their score in each of the sample approach stated earlier, would indicate the level at which the methodology becomes vulnerable to the scale of the entity corpus, indicating the minimum number of prospects to be considered for new leads identification bases on the one entity approach, i.e. the minimum number of seeds that the algorithm needs to maintain the quality of the leads generated. We move ahead with the hypothesis, that there should be no significant difference in the score of the leads generated using all 5 seeds and the leads generated using 2, 3, and 4 seed profiles. The similarity scores of Top 5 leads is with the 5 seed approach compared to their scores with 4/3/2 sample approach is summarized below.

Thus, we noted that the methodology is fairly robust and consistent when minimum 3 prospects are fed to create the entity. Hence we infer that the minimum number of leads that would work towards consistently generating relevant leads is 3 with leveraging the

| Lead | 5 seed | 4 seed | 3 seed | 2 seed |
|---|---|---|---|---|
| A | 0.557 | 0.584 | 0.571 | 0.520 |
| B | 0.487 | 0.485 | 0.488 | 0.450 |
| C | 0.435 | 0.356 | 0.325 | 0.342 |
| D | 0.378 | 0.334 | 0.334 | 0.313 |
| E | 0.374 | 0.391 | 0.409 | 0.367 |

Figure 3: Scores with variations in the seed Profiles

LinkedIn attributes of prospects. Also, with dropping the number of seeds some of the top 5 leads don't appear in the Top 5 with the other approaches. The change in the scores after variations in the number of seed profile fed to create the corpus, provides evidence to reject the hypothesis that the variations in the input seed would have no effect on the similarity scores.

## 5.1 DTM Normalization

The research moved ahead with testing the effect of normalizing the term frequency in the corpus, i.e. adjusting the relative importance of the terms in the corpus per their frequency of occurrence. Although, cosine similarity distance measure employed to calculate the similarity score between leads, automatically nullifies the effect of multiple occurrence of a term in driving the score, we still wanted to test it. For this 3 methods were deployed. Two R functions to normalize the document term matrix and TF-IDF to generate new document term matrix based on the term frequency in a document and number of documents having this term. The two R functions utilized to normalize the dtm were, a. weightTfIdf and b. weightSMART

The 3 approaches revealed following observations, with the normalization approach using the two functions, the scores decreased drastically for all the prospects. Also, the number of leads generated at 15% cut off is small, in fact if we remove the case of similarity between the seed profiles, we only get 3 profiles as leads across the 5 customers. With the TF-IDF approach the leads generated were exactly the same as obtained with the cosine similarity approach 2 leveraging only LinkedIn attributes which validates the statement that the cosine similarity automatically takes into consideration the frequency of the terms within the individual corpus.To further test the robustness and consistency of the methodology, noise was induced in the corpus by including non-qualified leads and irrelevant leads was carried out.

Three different approaches were carried out to add noise in the input seed profile: a. Adding two non-qualified leads, b. Adding two irrelevant leads and c. Adding two non-qualified and two irrelevant leads. With the 2 non-qualified prospects, following observations were recorded, a. Decrease in scores of top profiles for some increases for some, b. certain low scoring profiles now have high score, c. some new names appear but not from a relevant industry i.e. IT/computer games etc. and d. relevant lead scores pushed down. The observations with 2 not relevant prospects were as follows: a. Drop in scores of relevant profiles, b. relevant lead scores pushed down. and c. New names in the top 10 some of them consistent with the above approach. Summarized below is the score comparison of top two leads under the 3 different approaches of noise induction (in parenthesis is the rank of the lead under the approaches utilized):

Lead A :

| | |
|---|---|
| Original: | 0.557 (1) |
| Maybe: | 0.474 (2) |
| NotRelevant: | 0.461 (3) |
| Maybe-NotRelevant: | 0.461 (3) |

Figure 4: Scores of Lead A with Noise

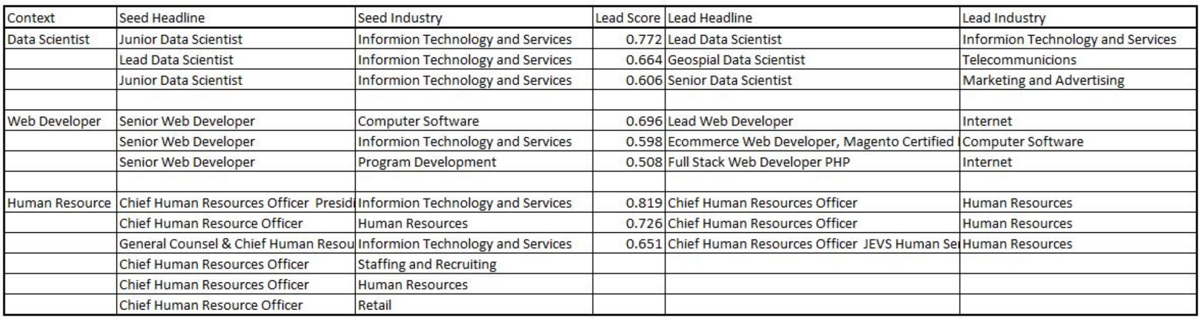
Lead B :

| | |
|---|---|
| Original: | 0.487 (2) |
| Maybe: | 0.495 (1) |
| NotRelevant: | 0.513 (1) |
| Maybe-NotRelevant: | 0.513 (1) |

Figure 5: Scores of Lead B with Noise

So after performing the above stated procedures and their various combination, it was established that the approach of utilizing attributes only from LinkedIn with a minimum of 3 leads produces the best list of leads leveraging the data about the existing customers using the cosine similarity distance measure approach to calculate the similarity between prospects and leads. Moving forward, the research also shifted the focus to profiles from other industries to validate the established methodology. Prospects from 3 different industries were obtained and used to generate leads. These seeds were acquired from the following industries:

a. Chief Human Resource Officer
b. Web Developer
c. Data Scientist



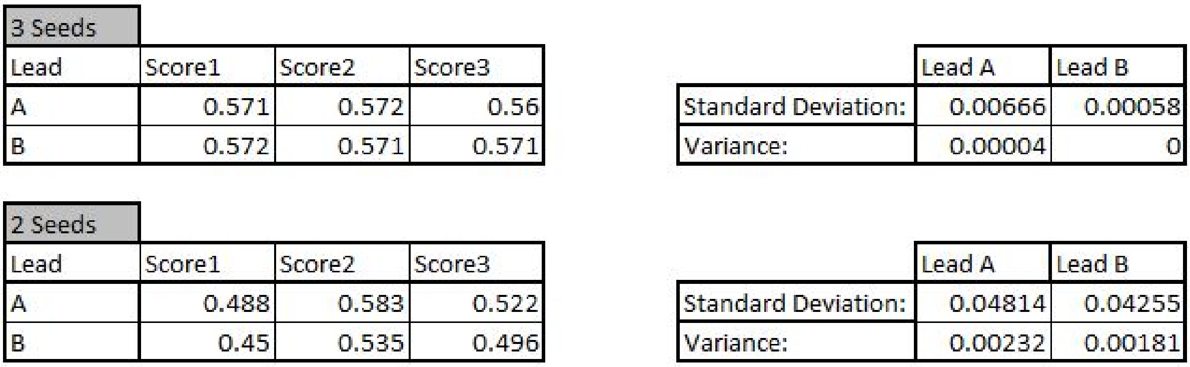
| Context | Seed Headline | Seed Industry | Lead Score | Lead Headline | Lead Industry |
|---|---|---|---|---|---|
| Data Scientist | Junior Data Scientist | Informion Technology and Services | 0.772 | Lead Data Scientist | Informion Technology and Services |
| | Lead Data Scientist | Informion Technology and Services | 0.664 | Geospial Data Scientist | Telecommunicions |
| | Junior Data Scientist | Informion Technology and Services | 0.606 | Senior Data Scientist | Marketing and Advertising |
| | | | | | |
| Web Developer | Senior Web Developer | Computer Software | 0.696 | Lead Web Developer | Internet |
| | Senior Web Developer | Informion Technology and Services | 0.598 | Ecommerce Web Developer, Magento Certified | Computer Software |
| | Senior Web Developer | Program Development | 0.508 | Full Stack Web Developer PHP | Internet |
| | | | | | |
| Human Resource | Chief Human Resources Officer Presidi | Informion Technology and Services | 0.819 | Chief Human Resources Officer | Human Resources |
| | Chief Human Resource Officer | Human Resources | 0.726 | Chief Human Resources Officer | Human Resources |
| | General Counsel & Chief Human Resou | Informion Technology and Services | 0.651 | Chief Human Resources Officer JEVS Human Se | Human Resources |
| | Chief Human Resources Officer | Staffing and Recruiting | | | |
| | Chief Human Resources Officer | Human Resources | | | |
| | Chief Human Resource Officer | Retail | | | |

Figure 6: Scores of Generated Leads

In Figure 6, the a sample of 3 leads is presented for each of the 3 business context. It is clear that the leads generated have farily high scores due to high similarity in the Headline of the leads and the seed profiles. It is interesting to note that lower scores such as a 60% compared to above 70% is due to the difference in the industry and some keywords in the headline but in general, the leads generated using the methodology for these business context have fairly high score and are very relevant as can be seen.

## 5.2 Discussion

In the course of this research, several approaches for lead generation utilizing the social media information of user were tried and evaluated. The research found Twitter to be insignificant from the context of lead generation and the presented methodology only utilizes a set of few LinkedIn attributes to generate high quality relevant leads for the business. Testing the methodology for variations in the input seed profiles and their corpus revealed that the methodology maintains its consistency of generating high quality relevant leads with a minimum of 3 seed profiles and all the using these seeds as one entity in the corpus. The standard deviation and variance in the scores of the leads with the 3 sample and 2 sample seed approach is summarized in figure 7 below. The presented statistics highlight the volatility of the scores with 2 sample approach compared to 3 sample. Based on the results, the methodology suggests a minimum of 3 seed profiles to be used for lead generation. The research also tests the methodology for different business cases and finds the methodology to be generating high quality leads. Approach to normalize the Document Term matrix decreased the scores significantly, however, with the TF-IDF approach, the results generated where exact copy of what was achieved using the cosine similarity approach of lead generation utilized in the methodology validating the methodology employed.



3 Seeds

| Lead | Score1 | Score2 | Score3 |
|---|---|---|---|
| A | 0.571 | 0.572 | 0.56 |
| B | 0.572 | 0.571 | 0.571 |

| | Lead A | Lead B |
|---|---|---|
| Standard Deviation: | 0.00666 | 0.00058 |
| Variance: | 0.00004 | 0 |

2 Seeds

| Lead | Score1 | Score2 | Score3 |
|---|---|---|---|
| A | 0.488 | 0.583 | 0.522 |
| B | 0.45 | 0.535 | 0.496 |

| | Lead A | Lead B |
|---|---|---|
| Standard Deviation: | 0.04814 | 0.04255 |
| Variance: | 0.00232 | 0.00181 |

Figure 7: Standard Deviation and Variance of lead scores with 2 and 3 seed profile

From the results and discussion presented above, readers may be convinced that lead generation problem can be solved with high level of accuracy utilizing few features from the individual LinkedIn profile. However, in real life there are several challenges around the social media data. Social media platforms are self representation, personal or professional. In that sense the information presented by users about their personal or professional standing could be exaggerated. It becomes immensely important to validate the information provided by users. The company uses schemes of entity resolution to tackle this problem of validating user data to some extent. However, more work is needed especially when collecting data from different social media platforms. Another area of concern is exposed in the scalability of language. English is the preferred means of communicating in social media. However, geographical focus of business context in areas where English is not the first language, languages with tens of millions of speakers are under-served. This would handicap the presented methodology as such, in taking into consideration all the relevant profiles. Translation tools could come in handy here before feeding the data to the methodology. Dealing with colloquial, misspelled content and sarcasm is another area of concern in dealing with social media data. Although, it makes sense to assume that professional representation should be devoid of any such exposure, but people are people and people do things they like. Word sense disambiguation is another area of vulnerability especially when dealing with specific regions, same words could mean different things in different languages. Although the focus of this research was concentrated in English speaking countries, taking this vulnerability into consideration presents an interesting future work.

The methodology presented in this research gives an approach to solving the lead generation problem, which has traditionally been a very time consuming and hectic process. Not only this research scores the leads based on their similarity to the desired traits and attributes, but also allows for businesses to rank their prospects based on their relevancy for a business context as indicated by their similarity scores. The presented methodology only utilizes a small set of attributes from the user LinkedIn profile and automates the lead generation problem, thus increasing the efficiency and productivity of the sales team.

# 6 Conclusion and Future Work

The research presents the methodology to identify new leads for a business by leveraging the LinkedIn information about any existing customer or qualified lead. In the course of research there were 5 different approaches employed, utilizing Twitter and LinkedIn data to identify new leads for business leveraging the concepts of Tie strengths, Homophily and Social Capital, to understand the nature of the relevant profiles and pick attributes which best align with these concepts of social network relationships, to identify individuals who are most relevant for the business context at hand. The attributes picked from the user profile on LinkedIn for lead generation were, Headline, Current Employer, Company Speciality and Company Industry. These attributes best capture the preference of a client for a prospective customer.

These attributes are also a good representation of the social capital of the individual and agree with the homophily principle. If a customer is from industry A, working in company X which has a given set of specialties, a prospect would more likely be from the same industry working in a company with similar speciality, holding a similar designation as reflected in the headline. The research also tests the robustness of the established methodology by studying the effect on the quality of the leads generated by variations in the number of input seed profile, addition of bad or mediocre profiles as seed alongside good leads and changing the nature of seed profile by testing the methodology to identify leads for 3 different business context. The methodology generated relevant leads consistently with a minimum of 3 seed profiles across all business context. The research validates the robustness and consistency of the established methodology across different business context to produce quality leads for clients leveraging only few attributes from LinkedIn profiles of prospects.

In future, this research can be extended to find the best way to reach out to these prospects on social media after identifying them. This can be done leveraging the same dynamics of social network relationships discussed in this research such as tie strengths and social capital. However, this would require data from social media beyond Twitter and LinkedIn, such as, Facebook, AngelList or community platforms like meetups. The idea here is to actively track user interests and habits and leverage the information to devise a channel to best engage with them. The research also builds a platform for identifying influencers on social media in communities of interests and connecting with them for brand promotion and content marketing.

## Acknowledgement

I would like to express my gratitude to my supervisor Dr. Simon Caton for the useful comments, remarks and engagement through the learning process of this master thesis. Furthermore I would like to thank him for introducing me to the topic as well for the support on the way. Also, I like to thank the the company for agreeing to share the database to make this research possible and employees who have willingly shared their precious time during the process of feedback and discussion . I would like to thank my loved ones, who have supported me throughout entire process, both by keeping me harmonious and helping me putting pieces together. I will be grateful forever for your love, guidance and support.

# References

Social exchange theory. *Encyclopedia Article.* URL https://en.wikipedia.org/wiki/Social_exchange_theory.

Lead generation: A beginner's guide to generating business leads the inbound way. URL http://blog.hubspot.com/marketing/beginner-inbound-lead-generation-guide-ht#sm.000001dcgf6jeadkdxzb877esfteh.

Lead generation. *Encyclopedia Article.* URL https://en.wikipedia.org/wiki/Lead_generation.

Sense of community. *Encyclopedia Article.* URL https://en.wikipedia.org/wiki/Sense_of_community.

L. A. Adamic and E. Adar. Friends and neighbors on the web. *SOCIAL NETWORKS*, 25:211–230, 2001.

T. O. AlbertM. Muniz, Jr. Brand community. *Journal of Consumer Research*, 27(4): 412–432, 2001. ISSN 00935301, 15375277. URL http://www.jstor.org/stable/10.1086/319618.

J. A. Barnes and F. Harary. Graph theory in network analysis. *Social Networks*, 5(2): 235–244, 1983.

D. J. Boorstin. The americans: The democratic experience. 1974.

G. M. Brauer M, Judd CM. The effects of repeated expressions on attitude polarization during group discussions. *Journal of Personality and Social Psychology*, 68(6):1014–1029, 1995.

J. Brown, A. J. Broderick, and N. Lee. Word of mouth communication within online communities: Conceptualizing the online social network. *Journal of Interactive Marketing*, 21(3):2–20, 2007. ISSN 1520-6653.

R. S. Burt. *Structural Holes: The Social Structure of Competition.* Cambridge, MA: Harvard UP, 1995.

R. S. Burt. Structural holes and good ideas1. *American journal of sociology*, 110(2): 349–399, 2004.

CMI. What is content marketing? 2015. URL http://contentmarketinginstitute.com/what-is-content-marketing/.

J. S. Coleman. The vision of foundations of social theory. *Analyse, Kritik*, 14(2):117–128, 1992.

J. W. Creswell. The selection of a research approach. *Sage Publications*, 2, 2003.

E. David and K. Jon. *Networks, Crowds, and Markets: Reasoning About a Highly Connected World.* Cambridge University Press, 2010. ISBN 0521195330, 9780521195331.

S. Dawson. A study of the relationship between student social networks and sense of community. *Educational Technology and Society*, 11(3):224–238, 2008.

D. Easley and J. Kleinberg. Strong and weak ties. *Networks, Crowds, and Markets: Reasoning about a Highly Connected World*, pages 47–84, 2010.

S. Fournier. Consumers and their brands: Developing relationship theory in consumer research. *Journal of consumer research*, 24(4):343–353, 1998.

N. Friedkin. A test of structural features of granovetter's strength of weak ties theory. *Social Networks*, pages 411–422, 1980.

S. Ganguly. Why social media advertising is set to explode in the next 3 years. 2015.

L. Garton, C. Haythornthwaite, and B. Wellman. Studying online soical networks. *Journal of Computer-Mediated Communication*, 3(1), 1997.

E. Gilbert. Predicting tie strength in a new medium. In *Proceedings of the SIGCHI conference on Computer Supported Cooperative Work*, pages 1047–1056. ACM, 2012.

E. Gilbert and K. Karahalios. Predicting tie strength with social media. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 211–220. ACM, 2009.

J. Golbeck. *Analyzing the Social Web*, volume 1. Morgan Kaufmann, 2013.

M. S. Granovetter. The strength of weak ties. *American journal of sociology*, pages 1360–1380, 1973.

J. Hagel. Net gain: Expanding markets through virtual communities. *Journal of Interactive Marketing*, 13(1):55 – 65, 1999. ISSN 1094-9968.

D. Krackhardt. The strength of strong ties: The importance of philos in organizations. *Networks and organizations: Structure, form, and action*, 216:239, 1992.

P. F. Lazarsfeld, R. K. Merton, et al. Friendship as a social process: A substantive and methodological analysis. *Freedom and control in modern society*, 18(1):18–66, 1954.

N. M. C. Martin Ruef, Howard E. Aldrich. The structure of founding teams: Homophily, strong ties, and isolation among u.s. entrepreneurs. *American Sociological Review*, 68 (2):195–222, 2003.

D. W. McMillan and D. M. Chavis. Sense of community. *Journal of community psychology*, 24(4):315–325, 1996.

M. McPherson, L. Smith-Lovin, and J. M. Cook. Birds of a feather: Homophily in social networks. *Annual review of sociology*, pages 415–444, 2001.

S. Milgram and J. Travers. An experimental study of the small world problem. *Sociometry*, 32(4):425–443, 1969.

G. A. Moore. Crossing the chasm: Marketing and selling high-tech products to mainstream customers. 2006.

N. Patel. 90-of-startups-will-fail. *Forbes*, 2015. URL `http://www.forbes.com/sites/neilpatel/2015/01/16/90-of-startups-will-fail-heres-what-you-need-to-know-about-the-10/#3b9c792e55e1`.

M. C. G. R. Bruce Money and J. L. Graham. Explorations of national culture and word-of mouth referral behaviour in the purchase of industrial services in the united states and japan. *Journal of Marketing*, page 79, 1998.

J. F. Sherry. *Servicescapes the concept of place in contemporary markets*. Lincolnwood, Ill. : NTC Business Books, 1998. ISBN 0844230057.

M. E. Walker, S. Wasserman, and B. Wellman. Statistical models for social support networks. *Sociological Methods and Research*, 22(1):71–98, 1993.

J. K. Wilkins and J. M. Zoken. Internet-enabled lead generation, Mar. 15 2005. US Patent 6,868,389.

R. L. Williams and J. Cothrel. Four smart ways to run online communities. *MITSloan Management Review*, 2000.

S. Wuchty. What is a social tie? *Proceedings of the National Academy of Sciences*, 106 (36):15099–15100, 2009. doi: 10.1073/pnas.0907905106. URL `http://www.pnas.org/content/106/36/15099.short`.

L. A. A. Xiaolin Shi and M. J. Strauss. Networks of strong ties. pages 33–47, 2007.