IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 3, January 2022
Impact Factor: 6.252

Using Social Media to Generate Leads

Rinku Pal[1] and Vivek Dubey[2]
Assistant Professor, BMS, Suman Education Society's LN College, Borivali East, Mumbai, India[1]
Student, BMS, Suman Education Society's LN College, Borivali East, Mumbai, India[2]

Abstract: *The most popular channel for communicating, establishing and sustaining both social and professional ties is social media. Their widespread acceptability is demonstrated by the expansion of platforms and the exponential development in the user base of social media websites like LinkedIn, Facebook, and Twitter. They provide several chances for organisations to take use of this aspect of digitally mediated interactions, such raising brand recognition and connecting with potential clients. This study focuses on the usage of social media to find appropriate profiles or "leads" for businesses looking to hire new people or collaborate with others. The study provides an automated method for lead finding using data from Twitter and LinkedIn, two social networking websites. Due to Twitter's emphasis on personal vs. professional user positioning, it was determined that it was not significant for lead generation in the business cases under consideration. The proposed final technique is evaluated for resilience to variations in input data, different business settings, and vulnerability to noise in the input data. It uses only four attributes from LinkedIn users' profiles to provide high quality leads. Despite only using a small portion of data, the findings demonstrate the resilience and consistency of the suggested technique to produce leads.*

Keywords: Social Media

I. INTRODUCTION

A significant amount of personal and professional data has been produced as a result of the social media platforms' exponential rise in user interactions and interaction. Additionally, it has been observed that virtual connections are increasingly similar to their physical counterparts. Similar to this, user interests, habits, and both personal and professional status are revealed by social media data. This has made it possible to examine these data in order to comprehend and anticipate their behaviour and preferences. Social media not only makes it possible to comprehend interpersonal conduct but also group behaviour and the identification of individuals who share similar intellectual and intellectual interests. Businesses may utilise this information to simplify their operations and be proactive rather than reactive to shifting customer preferences and interests by analysing trends and hot topics in social groups. Utilizing users' shared information to determine interests and fit, social media may assist organisations in finding new customers, workers, and collaborators. Businesses may proactively target relevant customers even before they start their search, as opposed to reactively targeting people who seek for certain items.

Traditionally, lead generation refers to the beginning of interest or enquiry about a company's goods or services. Here, the key term is initiation. We must be completely certain of a person's purpose and capacity to consume the good or service in order to pique their attention. We require an in-depth understanding of a person's personal and/or professional traits in order to have this degree of trust in their capability and intent. Businesses have the opportunity to find these prospective leads because to the wealth of information about people's preferences, achievements, and personal and professional goals that is available on social media platforms. For instance, two of the most popular venues for examining people's professional and personal portrayals are LinkedIn and Twitter. Information is more easily available than on networks like Facebook that have limited data access. Thus, making it simpler for organisations to sort through and pick out the most pertinent folks.

Even though social media data is easily accessible, lead production still needs significant manual labour in the absence of an automated system to produce quality leads. Typically, it depends only on an individual's judgement to manually search social media for certain traits without assessing the relevance of the results produced. Because of the limitations imposed by conventional filtering techniques, this is not scale able and is susceptible to the rejection of significant leads. This research was therefore inspired by the requirement for a system that can access the information repository more quickly and intelligently while also producing high-quality leads for teams that have previously depended on a

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 3, January 2022
Impact Factor: 6.252

semi-manual effort. The strategy outlined in this paper was created in collaboration with and for the benefit of an industry partner to satisfy the lead generation needs of their clients, i.e., find more people who qualify or fit a specific set of criteria.

The idea of resemblance will be utilised when identifying possible leads. Recommendations based on similarity are frequently used to filter and find possibly relevant other network members in various social media platforms (such as individuals you might know on Facebook or LinkedIn). The generalisation to identify profiles similar to one or more exemplary ideal profiles for various reasons, e.g. head hunting, is mostly absent, and these suggestions are often applied from the perspective of the particular user. Therefore, the following two research issues are examined in this paper:

- RQ1: How can social media data be used to build an automated lead generation strategy that produces high-quality leads?
- RQ2: Which social media data kinds are essential for generating "excellent" leads?

In order to demonstrate the possibilities of our technique and talk about design decisions, we share the results of four case studies that used it. The method calculates the similarity scores across profiles using the text mining and natural language processing techniques, the term frequency inverse document frequency (TF-IDF) information retrieval methodology, and the cosine similarity distance measuring technique. Our method is different from filtering in that it doesn't exclude profiles if they don't fulfil a requirement for a certain property. In contrast, it compares how well the characteristics of the profiles fit the characteristics that have been chosen from any set of ad hoc requirements. As a result, a profile that shares more terms with the target profile is given a better ranking than one that shares fewer words. This allows profiles to be prioritised in addition to guaranteeing that the leads provided are pertinent. By organising leads according to their "relevancy," our method improves operational and performance efficiency.

The format of this essay is as follows: The necessary underlying theory, associated research on lead generation, and numerous social media connection rules are all presented in Section 2. The Cross-Industry Standard Process for Data Mining is the approach that was chosen, and it is presented in Section 3. The primary aspects of the developed strategy, important implementation details, and the chosen assessment technique are also covered. Using case studies from finding leads across several fields, Section 4 evaluates our strategy. It also covers the outcomes and justifications for choosing certain methods over others when creating a data mining model for lead creation. The conclusion of the study summarises the key findings and suggests directions for further research in Section 5.

II. BACKGROUND AND RELATED WORK

The process of lead generation and recruiting has undergone a significant upheaval as a result of the rise of online social networks (OSN) and current advancements in data mining and machine learning. The foundation of OSNs like Twitter, Facebook, and LinkedIn is the idea that users voluntarily disclose information about themselves, their interests, abilities, and relationships to other users. The sheer magnitude of these networks, with their millions of members, necessitates the deployment of data mining tools to make the data accessible relevant for lead creation and search. This section provides an overview of pertinent ideas for lead generation and looks for OSNs that are related.

2.1. Social Media and Lead Generation

The purpose of lead generation, and consequently this work, is to use the information that is already available about known individuals (such as clients, potential partners, or employees) in order to identify comparable individuals (prospects) based on particular (preselected, or predefined) attributes that best define a prospect for the business. Prospects typically have a lot in common, making the concept of "similarity" a useful tool for finding them. In other words, if one "relevant" prospect is found, subsequent prospects are likely to be "similar." Similarity in this context might be described in terms of characteristics like professional function, business, or specialty. Prospects may share a great deal of topical similarity, such as comparable hobbies, post or tweet about similar themes, follow or like similar items or people. Similarity isn't only restricted to professional descriptions, though have demonstrated that topical similarity between OSN users may be utilised to accurately identify whether linkages exist between users. In addition, it is well known that user similarity can vary greatly depending on the source of similarity, such as when other persons or activities are taken into account.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 3, January 2022
Impact Factor: 6.252

Businesses have the chance to use the additional knowledge provided by the availability of Social Media data to make wiser decisions. The name "Social Media Analytics" serves as a catch-all for all the instruments, procedures, and strategies used to utilise Social Media data. From a business standpoint, organisations have embraced social media analytics for issues like comprehending client sentiment or enhancing their marketing plans. Consider how social media data may be incorporated into customer relationship management to provide better or more promising sales leads. The core idea behind this kind of social media-based recommendation is the idea of similarity between OSNs and social media. The advent of specialised platforms and service providers provides further proof of the relevance and necessity for this sort of data mining-based utilisation of the accessible OSN data. Software-as-a-services (SaaS) products, such Socedo1 and InsideView2, concentrate on giving businesses high-quality data about pertinent prospects, primarily in the B2B Marketing space. The functionality of the services given by these SaaS providers does not much change, but the underlying methodology that is utilised to identify leads is, for obvious reasons, highly guarded. As a result, there is relatively little information available on the techniques and methodology used in the sector to find leads.

While "recommendation" is more frequently used in consumer contexts, it nonetheless follows the same concepts as "lead generation," which is frequently used in corporate settings (particularly marketing and sales). Common examples of consumer-focused social media analytics include suggestions for individuals one may know, hobbies (such as movies, conversation topics, etc.), or location-based activities. Consider developing a recommender system that makes suggestions for things (people or tags) based on the many kinds of OSN information that is accessible. employ proximity, a second social concept, in conjunction with homophily to propose cooperation in academic networks. In addition to the standard similarity criteria, they also take into account other factors like variety and originality when rating their suggestions. Despite the fact that many studies have concentrated on a single social media platform, emphasise the importance and promise of cross-platform social media analytics, a factor that is equally crucial to our research. These recommender systems usually have the OSN user as their primary emphasis. Our method, in contrast, seeks to explicitly discover suggestions based on one or more acceptable sample profiles.

2.2. Social Media and Recruiting

The introduction of OSNs also significantly altered hiring procedures in general. Employers frequently use data from OSNs in their employment and search processes, and web-based recruiting and online applications are prevalent. From an academic perspective, it is unclear if using information from social media is genuinely beneficial in the recruitment and selection of suitable candidates. The self-representation of users on OSNs like LinkedIn, however, has a substantial impact on a recruiter's recommendation to hire. demonstrate how recruiters evaluate a candidate's fit with a job or business description using accessible profile information, demonstrating how the self-representation in OSNs might affect job suggestions. The introduction of OSNs also significantly altered hiring procedures in general. Employers frequently use data from OSNs in their employment and search processes, and web-based recruiting and online applications are prevalent. From an academic perspective, it is unclear if using information from social media is genuinely beneficial in the recruitment and selection of suitable candidates. The self-representation of users on OSNs like LinkedIn, however, has a substantial impact on a recruiter's recommendation to hire. demonstrate how recruiters evaluate a candidate's fit with a job or business description using accessible profile information, demonstrating how the self-representation in OSNs might effect job suggestions.

III. METHODOLOGY

Our strategy uses social media platforms for data mining and information retrieval. We adhere to the Cross-Industry Standard Process for Data Mining as a result (CRISP-DM). The Knowledge Discovery in Databases (or KDD) approach is extended by CRISP-DM, which enables us to incorporate the business environment and goals into the research process. It consists of six steps, which we list below to explain how we plan to use social media for lead generation.

3.1. Business Understanding

The goal of this study is to develop a mechanism for producing leads for an industry partner's sales team. This can be a competitive intelligence activity to look at rival employee rosters or their possible future recruits, an exercise in actively seeking personnel (digital head-hunting), locating new business or collaboration partners, etc. An initial list of

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 3, January 2022
Impact Factor: 6.252

possibilities communicated with the customer is produced by a straightforward filtering based on a person's designation. For instance, if the customer requested the location of Marketing managers, a list of n prospects (number depends on their subscription plan) is prepared by merely taking the prospects' designation into account and supplied with the client. This exercise serves as a straightforward technique for requirements engineering and preference elicitation, thus our approach doesn't start until after it. Modifying this step of the process is outside the scope of this study. Typically, customers give input on the list of n prospects that has been supplied with them, classifying them as good, middling, or terrible leads and explaining why. The seed profiles are then good leads. Identifying more profiles that resemble these seeds then becomes the commercial goal. The leads that clients choose reveal a lot of details about their intentions, priorities, and preferences. The attributes that LinkedIn profiles should include, such as Industry and Specialties, provide useful details about the pertinent industries and desirable skill sets.

3.2. Data Understanding

The retrieval of m seed candidates' LinkedIn profiles is done at this point. Keeping with our earlier example of a client seeking for marketing specialists, the aim purpose is to locate pertinent marketing profiles. Here, we start looking for information from social media sites like Twitter and LinkedIn that might further explain this domain. By filtering for profiles with the term "marketing" in their LinkedIn headline, for instance, one may see profiles for all people connected to the marketing industry. In addition to the headline, four other LinkedIn attributes can be used: Industry (the industry the user chooses in their profile), Current Employer, Company Industry, and lastly Specialties (the employer's areas of specialty as shown on their LinkedIn page). Since it is technically very difficult to traverse LinkedIn in an ad hoc manner to enable this data curation, we use crawlers to continually fill an offline database.

Also, we gather information from Twitter. Regarding user activities and expectations, Twitter and LinkedIn are significantly different from one another. The activities of users on LinkedIn are centred on users' professional representation as it is a very professional network of people. As a result of user activity on Twitter, individuals' personal and, to some extent, professional representations are exposed, Twitter is more intimate than LinkedIn. We utilise Twitter to fill out a different database with information on the main clientele areas and the important Twitter users in those areas. Twitter also offers material relevant to current events in areas that are particular to our goal: it offers a perspective on vernacular, or certain words or phrases that are used by both thought leaders and the general public. In the example, a portion of this database is concerned with tweets and people who are talking about marketing. A user's bio description on Twitter is a personal representation of the user and often represents their own hobbies and preferences. Every time a person who has never been seen before correlates to a tweet taken from Twitter, we create new user records. At the time of writing, there were about sixteen million entries in the Twitter database, compared to seventy thousand in the LinkedIn database. Based on metadata inside each of the two accounts pointing towards one another, we additionally link Twitter and LinkedIn profiles where possible.

Due to the difference in database sizes, many persons who had a Twitter profile (and had been chosen because their Twitter bio contained a keyword) did not also have an associated available LinkedIn page (and vice versa). This difference serves as an example of the difficulties of a cross-platform design, which are further described in. Overall, there was a substantially greater chance that an existing LinkedIn profile could be linked to one of the Twitter accounts than there was that a LinkedIn profile could link to Twitter. Additionally, from a professional standpoint, how one presents themselves on Twitter is not very important because clients are more concerned with a person's professional rankings in relation to their business goals. Similarly, even if two people's LinkedIn accounts are closely comparable, their personal representations are probably very different. Apart from that, their Twitter conversation may be quite comparable to or aligned with their respective area as a whole. Therefore, we cannot completely write off Twitter data just yet.

In short, we have extracted a corpus of Tweets in and around the typical locations of customer requests and profile information from persons on LinkedIn and Twitter linked to those areas.

3.3. Data Preparation

A corpus is created using the text information that was obtained from LinkedIn and Twitter. We tested a number of methods to build this corpus; these are shown in Figure 1 and discussed below in the context of our marketing case. By

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 3, January 2022
Impact Factor: 6.252

contrasting each technique with and without Twitter Bio descriptions, we will also illustrate the impact of information bias. These strategies expand upon the notion of "similarity" between leads and the example(s) that was previously described. These strategies' primary goal is to leverage attributes from users' profiles to capture users' intent and propensity for involvement. The features taken from LinkedIn to represent the similarity of people's professional networks. LinkedIn's Headline, Current Employer, Company Speciality, and Company Industry properties were specifically utilised. Twitter attributes record how similar users are on personal networks. For instance, the bio description often represents the interests and viewpoints of the user, and a high degree of similarity across bio descriptions suggests shared preferences. The five methods that were taken into consideration are outlined below:

First method: Twitter with LinkedIn User

All profiles that include the term "marketing" in their Twitter and LinkedIn profile descriptions. After pre-processing the corpus and combining these two characteristics for each person found, compare the results with the seed profiles.

Method 2: Twitter with LinkedIn User and Company (TLC): All profiles containing the phrase "marketing" in their Twitter bio description and LinkedIn headline. Include LinkedIn company qualities for their businesses, such as Company Specialties and Industry. Put all of these characteristics together for each person who has been identified, pre-process the corpus, and then compare to the seed profiles

Method 3: LinkedIn User and Company (LDA) and Twitter: Identify all Twitter and LinkedIn accounts that have the word "marketing" in their bios. Gather all of the tweets that these specified people have ever sent. Create and construct a second text corpus of tweets, then use Latent Dirichlet to do Topic Modelling. Allocation: Use LDA to analyse their twitter corpus and manually select the most pertinent topic.

Method 4 The 5000 most recent tweets concerning marketing are gathered in (User Tweets with LinkedIn and Company). Combine the LinkedIn headline and corporate data with the bio description of the tweeting user. Create the individual corpora, perform the pre-processing, and then contrast them with the seed profiles. The 5000 randomly selected tweets containing the term "marketing" or synonyms of marketing should be collected.

Method 5: Tweets with Synonyms, LinkedIn User and Company (SYN) The LinkedIn headline and the corporate qualities should be combined with the Twitter Bio descriptions for each individual Twitter account. Create the individual corpora, perform the pre-processing, and then contrast them with the seed profiles.

3.4. Modelling

The corpora need to be prepared for analysis after being taken from the LinkedIn and Twitter databases. Prior to the analysis, this comprises removing extraneous characters (such as emoji and URLs), identifying the language and stop words, punctuation, etc. Word stem is also used to lessen corpus dimensionality. The corpora can be utilised for analysis and modelling after they have been cleansed. We create a Document Term Matrix (DTM) from each corpus, which is composed of all the words from all selected users' corpora from LinkedIn and Twitter along the columns and specific individuals along the rows. The existence or absence of the appropriate phrase in each user profile is indicated by a 0 or 1 in each cell of the matrix.

Finding leads that resemble the seed profiles is our goal. The idea of distance between two profiles is one of the easiest methods to gauge how similar they are: profiles with more elements in common are closer to one another than those with less. However, the issue with distance is that non-normalized frequencies or occurrences in the data might distort it. Similar to the last example, profiles with greater information can also distort perceptions of profile similarity. For this reason, we use cosine similarity, a measure of similarity that does not have this issue. The cosine similarity metric polarises frequencies and treats one occurrence of a term as equal to, say, 100 occurrences. This is advantageous since it balances out overuse of specific terms.

In terms of its individual keywords, a profile is represented as a point in a coordinate plane whose dimensionality is equal to the number of unique keywords it contains. According to this theory, comparable profiles are the ones where the keywords are highly overlapping. The vector that represents them in coordinate space will either coincide, showing identical profiles, or will have a very tiny angle between them, showing a high degree of resemblance. On the other hand, two distinct profiles' vectors will be highly separated from one another. Right angles between two vectors represent entirely different profiles.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 3, January 2022
Impact Factor: 6.252

The angle can be projected into a value between 0 and 1, depending on how distinct or similar two profiles (vectors) are. Using the cosine similarity concept, we can determine how similar each pair of individual profiles is to both the seed profiles and each other if there are N profiles. This would result in a collection of N integers ranging from 0 to 1, which is equivalent to an adjacency matrix. As a result, whether they are seed profiles or not, we can determine which profiles are most similar to one another.

Based on the input data from the five ways mentioned above, this step produces an ordered list of leads that are ranked according to how similar they are to each other. We point out that the processes of data preparation and modelling can be repeated. Top leads from a first cycle, for instance, may be employed in additional rounds to further explore the profile space. The important thing to remember in this situation is that too many consecutive cycles will emerge as an echo chamber. The goal of following iterations is to find possible leads that are comparable to established leads in order to broaden the seed set's cardinality.

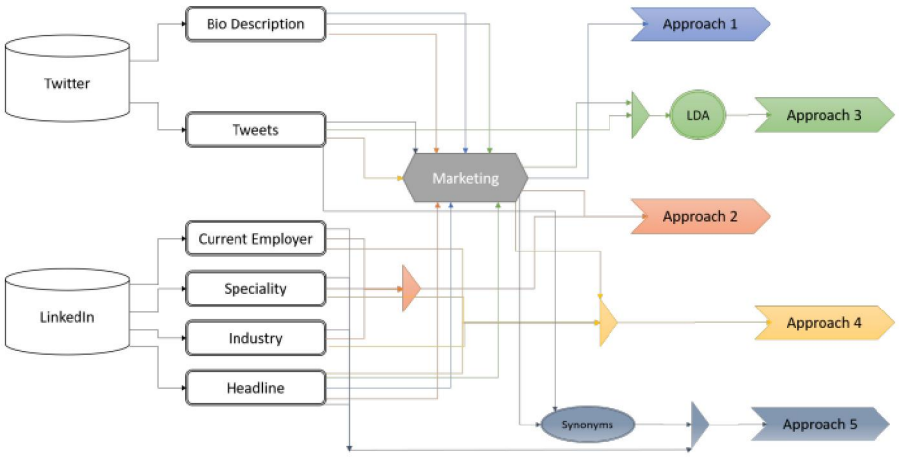

Figure 1. Flow Diagram of the 5 approaches employed

3.5. Evaluation

The evaluation is focused on two things: (a) if the leads produced are relevant as determined by industry partner sales team members and domain experts (clients); and (b) how resilient a strategy is with regard to fluctuations in input seed profile and business context. The latter is crucial since we cannot presume that customers know exactly what they need up front. Similar to this, while gathering requirements, certain requirements may be under- or overemphasised, while others may surface or alter over time. As a result, we have worked to find a strategy that can generate leads as well as one that is resistant to variations in seed quality.

3.6. Deployment

Our strategy has been used by the firm, and in the following area, we present a few case studies that demonstrate its benefits and potential.

IV. EVALUATION OF THE METHODOLOGY

Before responding to the two primary issues raised in section 3.5, it is important to briefly go over the assessment approach used in this study and several important choices. The lowest similarity score, which has been chosen as the cut-off level, determines the lead created by an approach. A threshold of this nature reflects the sensitivity and particularity of the business situation. Setting a lower threshold allows for a more flexible approach to identifying profiles as leads in situations when the business environment want to find many prospects who satisfy a fundamental set of criteria. Setting a greater threshold, on the other hand, results in a more rigorous evaluation of profiles as possible leads when the requirements are highly strict in terms of the classification and industry of the leads. In this study, 0.25 is chosen to allow for a complete examination of the calibre of leads produced. Therefore, we pay attention to profiles

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 3, January 2022
Impact Factor: 6.252

that are considered to be relatively different as well as those that are quite similar. As a result, we may combine leads that are expected to be good with those that are expected to be less good, which permits some degree of blind testing by human evaluators.

The quantity of seed profiles employed for lead creation is another aspect. Every time the method is performed, we've decided to start with five seed profiles. However, we cover the effects of smaller seed pools in section 4.2. The lead generation context, or the type of leads we are looking for, is the last step. We use the marketing example from our prior example to simply describe our findings. The mentioned outcomes, however, come from the basic conclusions of the marketing use case as well as the following 3 additional use cases: A data scientist is sought for a study project, a new HR director is needed, and a web developer is also needed. The information used in the study that follows originates from Twitter and LinkedIn. Around sixteen million profiles are included in the Twitter database and 70,000 in the LinkedIn database.

As previously said, the company's domain specialists personally checked the leads' quality. Overall, it was determined that the generated leads were very relevant and of a high calibre, and customer response on the leads created at a certain threshold was favourable. This is a good outcome for RQ1 since it enables us to use the existing social media data for automated, high-quality lead creation using the suggested technique.

4.1. Selecting Corpora:

Getting a list of potential seeds is where we start.

In this instance, the customer was seeking for marketing specialists, thus all pertinent profiles were gathered using the keyword "marketing" to filter them based on its appearance in their headlines. The LinkedIn database has about 4,600 prospects. Out of these 4,600 prospects, 20 profiles are chosen at random and shared with the customer. From these 20, the client chooses or qualifies prospects and returns the list to the sales team. The seeds will be present in the database as a result of this. The client response also explained the criteria used to choose or reject a potential seed. Then, we evaluate each of the five strategies covered in Section 3.3.

Approach 1 just uses the profiles' headlines to generate leads. Everybody who has comparable headlines and bio descriptions would make an excellent lead in this kind of filtering. Since of this, even little modifications to the seed set will have large effects because the corpus is not sufficiently rich. Three problems exist with Method 3. It is susceptible to selection, observer, and cognitive bias since it depends on user engagement with pertinent output issues. Second, LDA requires costly calculation. The amount of calculation time required to process LDA on all the tweets from over 4,600 accounts is substantial. Third, Twitter's continued reliance exposes it to the issue mentioned earlier — it has a tendency to be overly harsh when excluding users, which is made worse if we don't have Twitter accounts for them. It was discovered that Approach 2 was a far more trustworthy, practical, and consistent approach. It strikes a balance between user and business-specific data. It does not aggressively cut out when people lack sufficient Twitter data, unlike Approach 1, and it does not under-specify the domain.

In addition to the source corpora, there are two further options for the seed corpus: either aggregate all seeds into a single super-profile or seek for individuals who are similar to more seeds separately. Approach 2 was used to travel both of these routes. Path 2 is conceptually more applicable to the business context since it allows us to take into account all relevant features that are important to the company by using the seed profile's attributes to filter for comparable profiles. The most important characteristics of leads are revealed by a customer by qualifying them from a list of first prospects.

Consider the headlines for the following five seed profiles, which have been cleaned up and are highlighted by a strikethrough: "User Acquisition Manager," "Head of Marketing," "Vice President of PR Marketing," "Digital Marketing Executive," and "Marketing Director." This choice demonstrates a propensity for status in the profession. There are more keywords to match in a super-profile. A user who, for example, had the title "Vice President of Digital Marketing and Acquisition" would thus overlap with three of the five seeds. Only one word is included in the three keywords from the first seed when we look at the seeds separately, which lowers the score. When recall differences are taken into account, this amounts to a loss of knowledge for certain seeds. By using all pertinent keywords to filter for comparable leads, the profiles retain a high recall when taken as a whole as a super-profile. The high relevance of the

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 3, January 2022
Impact Factor: 6.252

leads generated reflects this. Here, approaches 4 and 5 struggle since they rely on Twitter to find the right profiles. Thus, as seen in Table 1, Twitter adds a smoothing component to the ranking process.

As we can see in Table 1, when we utilise the LinkedIn without Twitter Approach, Lead D (Headline: "Digital Marketing Assistant at XYZ" from Industry: "Pharmaceuticals") is not appropriate for the business context since the score is below the cut-off of 0.25. Due to the resemblance between their Twitter profile and a seed profile, using Twitter qualities places them in the Top 5.

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

Table 1. Score comparison of Leads with and without Twitter

This commonality, however, has no bearing on the current business situation. Leads A and C were seen as high-quality leads by clients, and this is reflected in their LinkedIn-only approach scores. Since of this, there are problems with exploiting Twitter because, even when LinkedIn and Twitter profiles match, the additional data has too much weight and distorts even basic factors like the lead's industrial sector. In other words, the technique is suffering from the curse of dimensionality, which states that as the number of dimensions rises, conceptions of distance lose their significance since all locations are both near and far from one another. Similar effects have been noticed in other settings; thus this is not only a dimensionality problem.

More technically, the growth in the total number of terms in the corpus has an impact on the approach's accuracy. Intriguingly, the phrases that become more prevalent in the corpus tend to be components of the individual's personal representation, which appear to be irrelevant for the current commercial setting. The calibre of leads produced utilising the 5 techniques reflects this similar fact. As a result, we did not include Twitter in our study and performed all five strategies using only the LinkedIn features of the profiles and qualified leads. The best outcome was still achieved by Approach 2; its harmony of human and corporate characteristics produced leads that were highly sought after in several trials and customer meetings, including those in fields other than marketing. Overall, in terms of RQ2, LinkedIn qualities offer the most pertinent social media data for lead creation, whereas Twitter doesn't seem to be helpful in this situation.

4.2. Varying the Cardinality of Seed Profiles

The next phase is to examine the impact of less seed profiles, whereas the prior strategy for lead creation uses all of the seeds as one single entity, the super-profile. Here, we'll concentrate on Approach 2, which, as was already indicated, performed the best. This factor makes it possible to talk about how sensitive Approach 2 is to changes in the input data. This is accomplished by deleting seeds at random from the beginning pool chosen by the client (across multiple use cases). The threshold at which the strategy becomes vulnerable to the size of the entity corpus would be shown by a considerable decrease in the score of the leads generated with fewer seed profiles when compared to the baseline instance with all 5 seeds. So, using the super-profile technique, we investigate the number of seeds required to find fresh leads.

We take into account both the variation in leads and the average score obtained for the leads to determine the bare minimum number of seeds. We employ two theories: First, we utilise the null hypothesis that there shouldn't be a difference between the score created using all 5 seeds and the leads generated using 2, 3, and 4 seed profiles. This is because Lead A and Lead B have different average scores. Second, we test the hypothesis that the variety in produced scores rises with a reduction in the number of seeds by taking into account the variance in scores themselves. According

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 3, January 2022
Impact Factor: 6.252

to the study's findings, the first hypothesis holds until there are just two seeds, at which point it must be rejected the null hypothesis. According to a matching t test, the difference in scores produced for Lead B is specifically significant at the 0.05 level. Table 2 provides a summary of the standard deviation and variation in the leaders' scores using the 3 sample and 2 sample seed approaches. When taking into account the second hypothesis, a chi-squared test for variance equality at the 0.05 level shows that the variation in scores obtained considerably rises when switching from 3 to 2 seed profiles. These outcomes show how much more erratic the scores were with two sample seeds as opposed to three. Consequently, based on the findings, we conclude that utilising Approach 2, 3 leads should be the minimal number needed to consistently provide meaningful leads. Additionally, when looking at more than just the top 5 leads, we can see that ranking variances become more obvious; frequently, the top 5 leads are no longer in the Top 5 when 2 or less seeds are employed. This makes sense because fewer seeds have a significantly greater impact on the proposed leads than more seeds do. Although the study indicates that 3 seeds are the absolute minimum, there is a practical reason why we do not consider more than 5 seeds: customers frequently show moderate annoyance when given the option of more than 5 seeds.

Non-qualified leads and irrelevant leads were added to the corpus in order to produce noise, further testing the approach's consistency and robustness. The following three methods were used to introduce noise into the input seed profile: Adding two non-qualified seeds, two irrelevant seeds, and two non-qualified and irrelevant seeds are the first two additions. In doing so, the broad observations listed below were made. First, the ratings of the top profiles fluctuate, sometimes falling and sometimes rising. Second, after adding noise, some low-scoring profiles start to score highly. Third, a few fresh leads that aren't from a relevant sector show up on the list. The applicable lead scores are then often decreased. This case illustrates the issues that arise when clients have ambiguous choices or refuse to choose seeds. Both scenarios are reasonable since clients occasionally may not want to designate a specific purpose, may not have yet created one and want to explore the digital environment instead, or they may just not want to spend time choosing seeds. However, the outcomes clearly demonstrate that this can have a significant influence on the accuracy and utility of the findings.

Lead	Score1	Score2	Score3	Standard Deviation	Variance
			3 Seeds		
A	0.571	0.572	0.56	0.00666	0.00004
B	0.572	0.571	0.571	0.00058	0
			2 Seeds		
A	0.488	0.583	0.522	0.04814	0.00232
B	0.45	0.535	0.496	0.04255	0.00181

Table 2. Standard Deviation and Variance of lead scores with 2 and 3 seed profiles

In these situations, we may run the strategy iteratively, which entails taking the initial seeds while being conscious of their limits, choosing the top n, and relaying them back to the client along with a yes/no choice regarding the recommended leads. Positive feedback is supplied as a seed for a subsequent iteration. Only two or three iterations in this way have shown success rates that are comparable to those of a carefully selected seed set, which lessens the impact of poor initial seed sets. Similar to this, reasonable outcomes were observed when firm personnel converted the customer needs into an initial seed populace in the case that no input is necessary.

4.3. Summary

Various strategies for lead generation via social media platforms were put forth and assessed.

Since Twitter was deemed to be negligible for the purpose of generating leads, the approach given here only makes use of a select few LinkedIn qualities to produce high-quality leads for the company. While a minimum of three pertinent seed profiles are employed, lead generation is consistent when testing the method for differences in the input seed

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 3, January 2022
Impact Factor: 6.252

profiles and their related corpora. We also highlighted that we can refine our method to successfully remove subpar starting seeds when initial seed sets are subpar. But if you iterate too frequently, you'll end up with a lead echo chamber, where the same lead combinations are generated repeatedly. This is because some characteristics, which are essential to our methodology and which contribute to profile similarity, are overrepresented.

It may appear from the results and discussion above that lead generation may be resolved with a high degree of accuracy by considering only a small number of features from the various LinkedIn accounts. However, social media data really faces a number of difficulties. Social media platforms are a stylized version of how individuals or organisations project their real-life experiences; as a result, the information offered by users is probably overstated. It is crucial to verify the data that people who have been designated as leads have submitted. In the end, there is a big difference between aggressively seeking out a possible new hire or discovering a relevant collaborative partner and that individual being the right match. What we've shown you here is a way to automate what many people do manually, allowing them to spend more time interacting with prospective leads than they would have to spend hunting for them.

V. CONCLUSION

The article outlines a semi-automatic method for finding new leads for a company by utilising the data on LinkedIn profiles of possible clients or leads, where leads can be new potential clients, employees, or partners in cooperation. It outlines a semi-automatic method that makes it possible to use a lot of social media data to produce leads (RQ1). We experimented with various methods to use LinkedIn and Twitter data in order to determine which types of social media data are most beneficial in generating "good" leads (RQ2). We discovered that adding Twitter data does not improve predictions; instead, it causes smoothing, which makes locating high quality leads more challenging. The following characteristics were chosen from the LinkedIn user profile for lead generation: Headline, Current Employer, Company Speciality, and Company Industry. For our industry partner, these qualities best reflected the tastes of actual clients. These characteristics offer a reliable indicator of profile similarity and serve as a decent reflection of the person's social capital. A prospect would be more likely to be from the same industry and working in a firm with a similar specialisation, holding a comparable designation as shown in the headline, if a customer is from industry A and working in company X which has a certain set of specialties. By examining the impact on the quality of the leads produced by varying the number of input seed profiles, adding poor or mediocre profiles as seeds alongside good leads, and changing the nature of the seed profile while testing the approach to identify leads for 4 different business contexts, the research also examines the robustness of the established methodology. When a minimum of 3 seed profiles are applied, the method reliably produces relevant leads across all business settings.

It is important to keep in mind that this work has a few possible flaws that might be fixed in further research. The possible privacy issues come first; the ease with which data is generally accessible does pose some questions. However, a significant element of LinkedIn's business strategy is based on the finding of others using the same data. These issues have been brought up previously in relation to the usage of social media data; for more information, read but research is required to address these issues. Second, research like this involves aspects of social posturing and self-representation, so similar considerations may be required. Third, given that this was a cross-platform research, it is painfully obvious that the sample size makes it difficult to effectively use several platforms to reflect the various viewpoints of prospects. This may also be connected to the finding that applying the method repeatedly even with "sub-optimal" seed profiles still produced "good" leads, and that additional research is needed to fully understand the effects of echo chambers within the approach by boosting sample sizes and running more scenarios.

REFERENCES

[1]. S. Caton, C. Dukat, T. Grenz, C. Haas, M. Pfadenhauer, and C. Weinhardt, “Foundations of trust: Contextualising trust in social clouds,” in Cloud and Green Computing (CGC), 2012 Second International Conference on, pp. 424–429, IEEE, 2012.

[2]. E. Gilbert and K. Karahalios, “Predicting tie strength with social media,” in Proceedings of the SIGCHI conference on human factors in computing systems, pp. 211–220, ACM, 2009.

[3]. E. Gilbert, “Predicting tie strength in a new medium,” in Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work, pp. 1047–1056, ACM, 2012.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 3, January 2022
Impact Factor: 6.252

[4]. L. A. Adamic and E. Adar, "Friends and neighbors on the web," Social networks, vol. 25, no. 3, pp. 211–230, 2003.

[5]. J. Hagel, "Net gain: Expanding markets through virtual communities," Journal of interactive marketing, vol. 13, no. 1, pp. 55–65, 1999.

[6]. P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer, and R. Wirth, "Crisp-dm 1.0 step-by-step data mining guide," 2000.

[7]. L. M. Aiello, A. Barrat, R. Schifanella, C. Cattuto, B. Markines, and F. Menczer, "Friendship prediction and homophily in social media," ACM Transactions on the Web (TWEB), vol. 6, no. 2, p. 9, 2012.

[8]. I. Guy, M. Jacovi, A. Perer, I. Ronen, and E. Uziel, "Same places, same things, same people?: mining user similarity on social media," in Proceedings of the 2010 ACM conference on Computer supported cooperative work, pp. 41–50, ACM, 2010.

[9]. D. Zeng, H. Chen, R. Lusch, and S.-H. Li, "Social media analytics and intelligence," IEEE Intelligent Systems, vol. 25, no. 6, pp. 13–16, 2010.

[10]. M. Rodriguez and R. M. Peterson, "The role of social crm and its potential impact on lead generation in business-to-business marketing," International Journal of Internet Marketing and Advertising, vol. 7, no. 2, pp. 180–193, 2012.

[11]. I. Guy, N. Zwerdling, I. Ronen, D. Carmel, and E. Uziel, "Social media recommendation based on people and tags," in Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval, pp. 194–201, ACM, 2010.

[12]. M. A. Brandao, M. M. Moro, G. R. Lopes, and ˜ J. P. Oliveira, "Using link semantics to recommend collaborations in academic social networks," in Proceedings of the 22nd International Conference on World Wide Web, pp. 833–840, ACM, 2013.

[13]. M. Hall, A. Mazarakis, M. Chorley, and S. Caton, "Editorial of the special issue on following user pathways: Key contributions and future directions in cross-platform social media research," International Journal of Human Computer Interaction, 2018.

[14]. J. Vicknair, D. Elkersh, K. Yancey, and M. C. Budden, "The use of social networking websites as a recruiting tool for employers," American Journal of Business Education, vol. 3, no. 11, p. 7, 2010.

[15]. D. Jeske and K. S. Shultz, "Using social media content for screening in recruitment and selection: pros and cons," Work, employment and society, vol. 30, no. 3, pp. 535–546, 2016.

[16]. J. K.-H. Chiang and H.-Y. Suen, "Self-presentation and hiring recommendations in online communities: Lessons from linkedin," Computers in Human Behavior, vol. 48, pp. 516–524, 2015.

[17]. J. Van Dijck, "'you have one identity': performing the self on facebook and linkedin," Media, Culture & Society, vol. 35, no. 2, pp. 199–215, 2013.

[18]. U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "The kdd process for extracting useful knowledge from volumes of data," Communications of the ACM, vol. 39, no. 11, pp. 27–34, 1996.

[19]. M. Hall and S. Caton, "Am I who I say I am? Unobtrusive self-representation and personality recognition on Facebook," PloS one, vol. 12, no. 9, p. e0184417, 2017.

[20]. D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," Journal of machine Learning research, vol. 3, no. Jan, pp. 993–1022, 2003.

[21]. M. Brauer, C. M. Judd, and M. D. Gliner, "The effects of repeated expressions on attitude polarization during group discussions.," Journal of Personality and Social psychology, vol. 68, no. 6, p. 1014, 1995.

[22]. M. Zimmer, ""but the data is already public": on the ethics of research in facebook," Ethics and information technology, vol. 12, no. 4, pp. 313–325, 2010.