Utilizing Social Media For Lead Generation

Adjitesh Prakash
School of Computing,
National College of Ireland, Dublin
adjitesh.prakash@student.ncirl.ie

Simon Caton
School of Computing,
National College of Ireland, Dublin
simon.caton@ncirl.ie

Christian Haas
University of Nebraska at Omaha
Omaha, NE, USA
christianhaas@unomaha.edu

Abstract

Social Media is the most prevalent platform for communication, forming and maintaining professional as well as social 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 their widespread acceptance. They pose many opportunities for businesses to exploit this facet of digitally mediated relationships, for example 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 in sourcing new employees, or collaborators. The paper utilizes data from social networking sites Twitter and LinkedIn and presents an automated approach for the discovery of leads. For the considered business cases, Twitter was found to be irrelevant for lead generation due to its emphasis on personal vs. professional user positioning. The presented final approach utilizes only four attributes from LinkedIn users' profiles to generate high quality leads, and is tested for robustness to variations in input data, different business contexts and vulnerability to noise in the input data. The results show the robustness and consistency of the presented approach to generate leads despite utilizing a small subset of features.

1. Introduction

The exponential growth in user interaction and relationships on social media has generated a huge amount of personal and professional data, which is often publicly available on various social media platforms. It has also been noted that digital relationships are becoming increasingly indistinguishable from their non-virtual counterparts [1]. Similarly, social media data exposes user preferences, habits and their personal as well as professional standings. This has opened the possibility of analyzing these data to understand and predict their behavior and preferences [2, 3]. Not only does social media allow for understanding individual behavior, but also community behavior and the identification of people who are similar in terms of their interests and intellect [4]. This knowledge opens 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, employees and collaborators by using users own shared information to identify interests and fit. Rather than reactively targeting users who search for certain products, businesses can proactively target relevant users even before they begin their search [5].

Lead generation traditionally is the initiation of interest or inquiry into products or services of a business. 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 a deep view of their personal and/or professional characteristics. The amount of information available on social media platforms concerning individuals' preferences, accomplishments as well as personal and professional aspirations, provides opportunities for businesses to identify these potential leads. LinkedIn and Twitter, for example, are two of the biggest platforms to analyze individuals' professional and personal representations. However, unlike platforms with restricted data access such as Facebook, information is more accessible. Thus allowing businesses to filter and identify the most relevant individuals more easily.

While social media data is readily available, in the absence of any automated methodology to generate quality leads the process of lead generation requires high manual efforts if social media data should be included. Typically it relies on manually scanning social media for specified attributes without any measure of relevancy of the leads generated, apart from individual discretion. As such, this is not scalable and is prone

URI: https://hdl.handle.net/10125/59669
ISBN: 978-0-9981331-2-6

to the rejection of relevant leads due to limitation posed by traditional filtering methods. Hence, this research was conceived on the need for a solution for a smarter and faster mechanism to tap into the information repository to generate quality leads for the teams which have traditionally relied on a semi-manual effort. The approach proposed in this paper is developed with and for an industry partner to address their clients' requests in lead generation, i.e. identify more individuals who qualify or meet a given set of characteristics.

Considering the identification of potential leads, the concept of *similarity* will be used. Recommendations based on similarity is used in various social media platforms (e.g., people you might know on Facebook or LinkedIn), and often filter / identify potentially relevant other users in the network. However, these recommendations are commonly implemented from an individual user's point of view, and the generalization to identify profiles similar to one or more exemplary desirable profiles for various purposes, e.g. head hunting, is largely missing. Hence, this paper investigates the following two research questions:

- RQ1: How can an automated approach for lead generation be implemented that makes use of available social media data to generate high quality leads?
- RQ2: Which types of social media data are relevant for producing "*good*" leads?

We present the findings of four case studies leveraging our approach to illustrate its capabilities and discuss design choices. The approach utilizes the techniques of natural language processing and text mining, information retrieval technique term frequency inverse document frequency (TF-IDF), and the distance measuring technique cosine similarity to calculate the similarity scores between profiles. Our approach is not a filtering technique in the sense that it doesn't discard profiles if they fail to meet a condition on a given attribute. In contrast, it matches the *similarity* of the profiles' attributes with the selected attributes from any set of ad-hoc requirements. So, if a profile has more common terms with the target profile, it is ranked higher than a profile with has less terms in common with the target profile. This not only ensures that the generated leads are relevant but also permits profiles to be ranked. Thus, our approach increases operational and performance efficiency by ordering leads with respect to their "relevancy".

This paper is structured as follows: Section 2 presents relevant background theory and related work in lead generation and the various principles of social media relationships. Section 3 presents the adopted methodology: the CRoss-Industry Standard Process for Data Mining (CRISP-DM) [6]. It also discusses the main features of the devised approach, key implementation details, as well as the employed evaluation methodology. Section 4 presents the evaluation of our approach using case studies from discovering leads in the multiple areas. It also discusses the results and rationale behind selection and rejection of different approaches to devise a data mining model for lead generation. Finally, in Section 5, the paper is concluded by highlighting the main findings and proposing future research avenues.

2. Background and Related Work

With the emergence of online social networks (OSN) and recent developments in data mining and machine learning, the process of recruitment and lead generation has seen a massive disruption. OSNs such as Twitter, Facebook, and LinkedIn are built around the principle that users willingly share information about themselves, their interests, skills, and connections to other users. Due to the sheer size of these networks with their millions of users, data mining techniques are necessary to utilize the available data and make it useful to the lead generation and search process. This section gives an overview of relevant concepts for lead generation as well as finding similarities in OSNs.

2.1. Social Media and Lead Generation

The goal of lead generation, and by extension this work, is to leverage the available information about known individuals (e.g. customers, potential employees or collaborators) and to use these examples to identify similar individuals (prospects) based on certain (preselected, or predefined) attributes which best define a prospect for the business. Typically, prospects will have a lot in common, and as such, notions of "similarity" are a convenient mechanism for their discovery, i.e. once a single "relevant" prospect is identified, other prospects are likely to be "similar". Here similar can be defined in terms of attributes like professional role, company, or speciality. However, similarity doesn't need to be limited only to professional descriptors, but prospects may also have a high degree of topic similarity, i.e., similar interests, post or tweet about similar topics, follow or like similar things or people. [7] have shown that topic similarities between users of an OSN can be used to predict the existence of links between users with a high accuracy. At the same time, it is known that user similarity can differ substantially based on the source of similarity, e.g., if other people or

activities are considered [8].

The availability of Social Media data presents an opportunity for companies to utilize the additional information to make better decisions. *Social Media Analytics* is an umbrella term that captures tools, methods, and approaches to leverage Social Media data for a given purpose [9]. From an enterprise perspective, companies have used social media analytics for topics such as understanding customer sentiment or improving their marketing strategies. For example, [10] discuss how to include social media data into customer relationship management to generate better or more promising sales leads. The notion of similarity in social media / OSNs is central concept of this type of social media-based recommendation. The relevance and need for this type of data mining-based use of available OSN data is evidenced by the emergence of dedicated platforms and service providers. Software-as-a-services (SaaS) solutions, such as Socedo[1], and InsideView[2], focus on providing companies with high quality information on relevant leads, mostly focusing on the B2B Marketing domain. The services provided by these SaaS providers do not substantially vary in terms of offered functionality, but the underlying algorithm that is used to identify leads are, for obvious reasons, closely guarded. Consequently, there is very little documentation of the algorithms and approach employed to identify leads in the sector.

While the term "lead generation" is often used in a business (in particular marketing and sales) setting, "recommendation" is more commonly used in consumer settings, albeit using the same principles. Common examples of social media analytics targeted at consumers include recommendations for people one might know, interests (e.g., movies, discussion topics, etc.) a user might like, or location-based activities. For example, [11] propose a recommender system that suggests items (people or tags) based on various types of information available in the OSN. [12] use homophily along with a second social principle, proximity, to suggest collaboration in academic networks. For ranking their recommendations, they also consider additional aspects such as diversity and novelty, which can be used in addition to the basic similarity metrics. While many studies have focused on one social media platform, [13] discuss the relevance and potential of cross-platform social media analytics, an aspect which is also central to our analysis. Typically, the focus of these recommender systems is the user of the OSN. In contrast, our approach aims to explicitly find recommendations based on a one or more sample desirable profiles.

[1] http://www.socedo.com – Last Visited: 06/11/2018

[2] https://www.insideview.com/ – Last Visited: 06/11/2018

2.2. Social Media and Recruiting

The advent of OSNs also fundamentally changed recruitment practices in general. Web-based recruiting and online applications are commonplace, and employers often use information available in OSNs in their search and hiring decisions [14]. However, from an academic point of view there is no definite answer whether using social media information is actually helpful in the recruitment and selection process for potential candidates [15]. In contrast, the self-representation of users in OSNs such as LinkedIn has significant effects on a recruiters hiring recommendation. [16] show that recruiters use available profile information in their assessment of the fit of a person to a job or company description, indicating that the self representation in OSNs can affect job recommendations. Finally, while personal OSNs (Facebook, Twitter, etc.) and professional OSNs (LinkedIn, etc.) have a different focus (self-representation vs self-promotion), both types of OSNs do not necessarily represent the actual identity of the user. Rather, they are often a mixture of identity (actual or represented) as well as restrictions in representation provided by the OSN provider, such as interface layout or available categories [17]. Overall, while social media has been used in the general recruitment process as described above, its use for generating high-quality leads in an automated fashion, as described in this paper, is a new approach.

3. Methodology

Our approach is an application of data mining and information retrieval from social media platforms. As such, we follow the CRoss-Industry Standard Process for Data Mining (CRISP-DM) [6]. As an extension to the Knowledge Discovery in Databases (or KDD) method [18], CRISP-DM permits us to encapsulate the business context and objectives into the research process. It comprises six activities, which we outline here to articulate our approach to utilize social media in lead generation.

3.1. Business Understanding

The objective of this research is to devise a methodology to generate leads for the sales team of an industry partner. This can be an exercise in actively sourcing employees (digital head-hunting), identifying new business or collaboration partners, an exercise in

competitive intelligence to scan competitor employee rosters, or their potential future hires etc. An initial list of prospects shared with the client is generated by a simple filtering based on the designation of an individual. As an example, if the request was to locate Marketing managers, a list of n prospects (number based on their subscription plan) is shared with the client generated by only considering the designation of the prospects. Our approach only commences after this activity, as it acts as a simple method of requirements engineering and preference elicitation. It is not the remit of this work to modify this part of the process. Typically clients' provide feedback on the list of n prospects shared with them, categorizing them as either good, mediocre, or bad leads and giving reasons for the same. Good leads (selected by the client from the list shared), then serve as the *seed profiles*.[3] The business objective then becomes: identify more profiles like these seeds. Clients' choice of leads exposes a lot of information about their intent, focus, and preferences. Attributes like Industry, Specialties, etc., that LinkedIn profiles should contain, also give valuable information about the relevant industry and skill sets that are attractive.

3.2. Data Understanding

At this stage, m seed candidates are provided, and their LinkedIn profiles retrieved. Continuing with our example of a client looking for marketing professionals, the target objective is to find relevant profiles in marketing. Here we commence identifying data from social media platforms (LinkedIn and Twitter) that can better articulate this domain. For example, profiles for all users associated with the marketing industry can be obtained by filtering for profiles with the word 'marketing' in their LinkedIn headline. Aside from the headline, four other attributes from LinkedIn can be utilized: Industry (the industry a user specifies in their profile), Current Employer, Company Industry (the industry which an individual's employer specifies on their LinkedIn page) and finally, Specialties (the areas of expertise of the employer as specified on their LinkedIn page). Traversing LinkedIn in an ad hoc manner to support this data curation is technically very challenging, therefore we employ crawlers that continuously populate an offline database.

We also harvest data from Twitter. Twitter is very different from LinkedIn with respect to user activity and expectations. LinkedIn is a highly professional network of individuals and the activities of users on this site focuses 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 representations. From Twitter, we populate another database on core areas of clients and key Twitter users in these areas. Twitter also provides content pertinent to contemporary issues in areas specific to our objective: it provides a view on vernacular, i.e. specific keywords or phrases provided by thought leaders as well as the masses. In the context of our example, a part of this database pertains to tweets and users tweeting about marketing. From Twitter, aspects such as a user Bio description is the personal representation of the user and typically reflects users' personal interests and preferences. We construct new user records every time a previously unseen user corresponds to a tweet harvested from Twitter. The Twitter database, at the time of writing, had approximately sixteen million entries, compared to seventy thousand entries for LinkedIn. Where possible we also link Twitter and LinkedIn profiles based on metadata within each of the two profiles pointing at each other.

Due to this disparity in the database sizes, many users with a Twitter profile (having been selected based on the presence of a keyword in their Twitter Bio) did not have a corresponding available LinkedIn profile (and vice versa). The challenges of a cross-platform design are exemplified in this disparity and are further discussed in [13]. Overall, the hit rate from Twitter to LinkedIn was significantly lower than that from LinkedIn to Twitter, meaning that it was much more likely that an existing LinkedIn profile could be linked to one of the Twitter accounts. Also, from the business point of view, self-representation [19] via the Twitter description is not of high relevance since clients are interested in the professional standings of an individual for their business aspirations. Similarly, the personal representation of two individuals is likely substantially different even if their LinkedIn profiles are highly related. That aside, their Twitter discourse may be highly aligned, or similar to their corresponding area as a whole. Thus, we cannot simply dismiss Twitter information just yet.

In summary, we have collected profile information of LinkedIn and Twitter users aligned to the normal areas of client requests and harvested a corpus of Tweets in and around the same areas.

3.3. Data Preparation

The text data collected from Twitter and LinkedIn is used to form a corpus. We have experimented

[3] Note that, depending on the size n, this could be a labor-intensive process. As we need the seed profiles for the subsequent steps, the client could also alternatively provide a set of good profiles themselves, potentially reducing their workload.

with several means to construct this corpus, which are illustrated in Figure 1, and are explained below in the context of our marketing example. We will also highlight the effect of the information bias by comparing each approach with and without Twitter Bio descriptions. These approaches build on the idea of "similarity" between leads and the example(s) as discussed earlier. The main focus of these approaches is to capture users' intent and propensity for engagement utilizing the attributes from their profiles. The attributes utilized from LinkedIn capture similarity of the professional networks of individuals. Specifically, the attributes used from LinkedIn: Headline, Current Employer, Company Speciality, and Company Industry. Attributes from Twitter capture user similarity of personal networks. For example, the bio description generally reflects user interests and opinions, and a high similarity between bio descriptions indicates shared interests and habits. Overall, the 5 considered approaches are summarized below:

Approach 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 compare with the seed profiles.

Approach 2: Twitter with LinkedIn User and Company (TLC): All profiles with keyword 'marketing' in their LinkedIn headlines and Twitter Bio Description. Extend to attributes for their companies from LinkedIn, namely, Company Specialities and Industry. Combine all these attributes for everyone identified, preprocess the corpus and then compare with the seed profiles.

Approach 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 available tweets from these identified individuals. Create and prepare a second text corpus of tweets and performing Topic Modeling (via Latent Dirichlet Allocation: LDA [20]) on the corpus of their tweets, and manually pick the most relevant topic(s). Filter the initial population so that only users the contributed to these topics are carried forward, then continue as in approach 2.

Approach 4: User Tweets with LinkedIn and Company (UTLC): Collect the 5000 most recent tweets about marketing. Pull the tweeting user Bio description and combine them with the LinkedIn headline and company attributes. Build the individual corpus, preprocess and then compare with the seed profiles.

Approach 5: Tweets with Synonyms, LinkedIn User and Company (SYN): Collect the 5000 randomly sampled tweets with keyword 'marketing' or synonyms of marketing. For all the individual Twitter handles, obtain their Twitter Bio description and combine them with the LinkedIn headline and the company attributes. Build the individual corpus, preprocess and then compare with the seed profiles.

3.4. Modeling

Once extracted from the LinkedIn and Twitter databases, the corpora need to be prepared for analysis. This involves cleaning of redundant characters (e.g. emoji, URLs etc.), determining the language (our approach is currently focused only on English content) and stop words, the punctuation, etc., before the subsequent analysis. We also word stem to reduce corpus dimensionality. After the corpora have been cleaned, they can be used for analysis and modeling. From each corpus, we construct a Document Term Matrix (or DTM), which consists of all the words from all selected users' corpora from Twitter and LinkedIn along the columns and individual users along the rows. 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.

Our objective is to identify leads similar to the seed profiles. One of the simplest ways of measuring similarity between two profiles is by using a concept of distance between them: profiles that have higher number of aspects in common are closer to each other than with less in common. The problem with distance, however, is that it can be skewed by non-normalised frequencies or occurrences within data. Similarly, profiles with more content can equally skew notions of similarity between profiles. For this reason, we utilize a measure of similarity that does not suffer from this problem: cosine similarity. The cosine similarity measure polarises frequencies, thus considering 1 occurrence equivalent to say 100 occurrences of a keyword. This is beneficial as it neutralises excessive use of specific terms.

A profile is 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. Taking the cosine of this angle projects the angle into 0 to 1, depending on whether two profiles (vectors) are dissimilar (at right angles to each other – 0) or similar

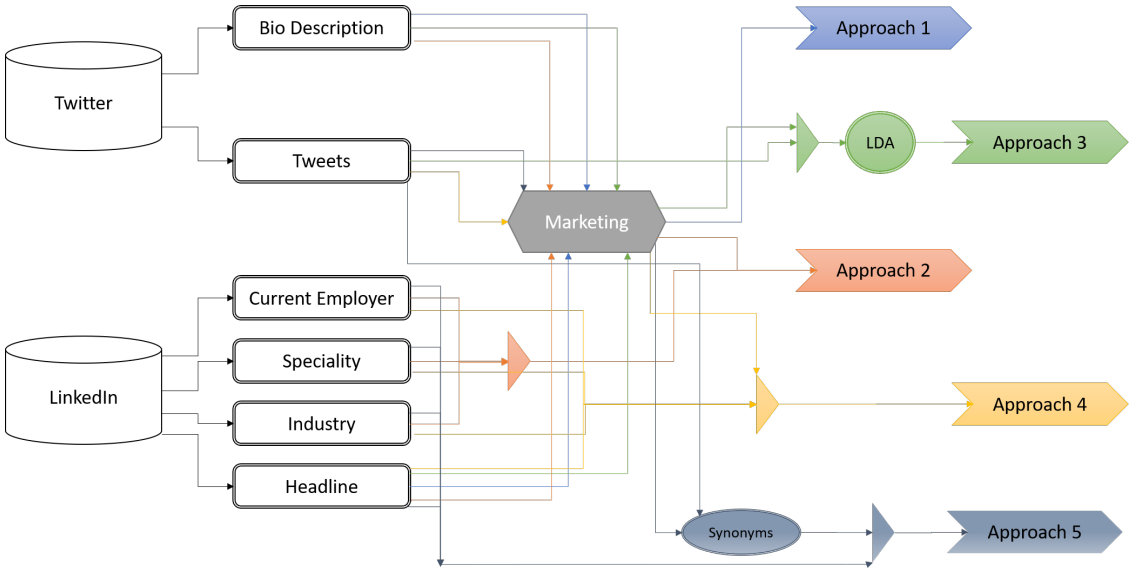

Figure 1. Flow Diagram of the 5 approaches employed

(coincide, or are parallel to each other – 1). If there are N profiles, leveraging the cosine similarity principle allows us to measure similarity between every pair of individual profiles with each other as well as the seed profiles. This would give us a set of $N * N$ numbers between 0 and 1, which corresponds to an Adjacency Matrix. Thus, we can identify the profiles which are most similar to each other, be them seed profiles or otherwise.

The output from this stage is an ordered list of leads corresponding to their similarity scores based on the input data from the five approaches noted above. We note here, that the data preparation and modeling stages can be undertaken multiple times. For example, top leads from a first cycle could equally be used in subsequent iterations to further traverse the profile space. Key here is to keep in mind that implementing too many consecutive cycles will result in the manifestation of an echo chamber. The idea of subsequent iterations is to identify potential leads that are similar to well-regarded leads, thus increasing the cardinality of the seed set.

3.5. Evaluation

The focus of the evaluation is a) whether the leads generated are relevant as established by domain experts (clients) as well as representatives of the industry partner's sales team; and b) how robust an approach is with respect to variations in input seed profile and business context. The latter aspect is important as we cannot assume that clients ex-ante precisely know what their requirements are. Similarly, some requirements may be under- or overemphasized in the requirements gathering process and others may emerge or change with time. Therefore, not only have we endeavored to identify a method capable of lead generation, but also one that is resilient to disparities in the seed quality.

3.6. Deployment

Our approached has been deployed in the company, and in the next section, we showcase a few case studies illustrating its merits and capabilities.

4. Evaluation of the Methodology

Prior to addressing the two core questions highlighted in section 3.5, we first need to briefly discuss the evaluation methodology applied in this paper, and discuss some key decisions. The lead generated from an approach depends on a selected cut-off level: the minimum similarity score. Such a threshold is reflective of the sensitivity and specificity of the business context. In cases where the business context aims to identify a lot of leads meeting a basic set of criterion, setting a lower threshold enables 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 threshold results in more stringently assessing profiles as potential leads. In this research, 0.25 is used to accommodate a thorough analysis of the quality of leads generated. Thus, we not only focus

on highly similar profiles, but also profiles which are deemed somewhat dissimilar. This facilitates a degree of blind testing by human evaluators (employees of our industry partner), as we can mix in leads predicted to be good with those predicted to be less good.

Another factor is the number of seed profiles used to perform lead generation. We have selected to initialise each run of the approach with 5 seed profiles. However, in section 4.2, we discuss the impact of smaller seed pools. Finally, we need to determine the lead generation context, i.e. what kind of leads we search for. For simplicity, we explain our results using our previous example of marketing. However, the discussed results stem from the general findings of the marketing use case, plus the following 3 other use cases: 1) seeking a new head of HR, 2) seeking a web developer and 3) a data scientist for a research project. The data in the subsequent analysis is coming from two social media platforms: Twitter and LinkedIn. The Twitter database corresponds to approximately sixteen million profiles, and the LinkedIn database around 70,000 profiles.

As described before, the quality of leads generated was manually verified by the domain experts at the company. Overall, the generated leads were found to be highly relevant and of high quality, and the client feedback on the leads generated at a given threshold was positive. This is a positive result regarding RQ1: with the proposed approach, we can leverage the available social media information for automated, high-quality lead generation.

4.1. Selecting Corpora

We begin with obtaining a list of prospective seeds. In this case, the client was looking for marketing professionals, 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 LinkedIn database. A random list of 20 profiles is shared with client from these 4,600 prospects. The client selects or qualifies leads from these 20 and sends back the list to the sales team. This ensures that the seeds are in the database. The client feedback also highlighted why a prospective seed was or was not selected. Next, we compare each of the five approaches discussed in Section 3.3.

Approach 1 only leverages the headline of the profiles for lead generation. This is essentially filtering where everybody with similar headlines and bio descriptions would be a good lead. Thus, it is not rich enough a corpus, and small changes in the seed set will have significant impacts. Approach 3 has three issues. First, it relies on user interaction to relevant output topics, making it vulnerable to observer, selection, and cognitive bias. Second, LDA is computationally expensive. Processing LDA on all the tweets of over 4,600 profiles leads to significant computation time. Third, the reliance on Twitter still exposes it to the problem highlighted earlier – it tends to be too aggressive in filtering out users; even more so if we do not have Twitter feeds for them. Approach 2 was found to be a much more consistent, feasible, and reliable method. It balances user as well as company specific data. It does not under-specify the domain like Approach 1, and does not aggressively cut out in cases where users lacked ample Twitter data.

Aside from the source components of the corpora, there are two additional choices concerning the seed corpus: combine all seeds together into one super-profile or consider them individually and look for individuals that are similar to more seeds. Both these paths were traversed with Approach 2. Path 2 is logically more relevant to the business context, since using the combination of attributes of the seed profile to filter for similar profiles allows us to consider all the relevant attributes that are of interest to the business. By qualifying leads from an initial list of prospects, a client indicates their most relevant attributes.

For example, consider the 5 seed profile (cleaned indicated via strikethrough) headlines: 'User Acquisition Manage~~r~~', 'Head ~~of~~ Market~~ing~~', 'Vice President ~~of PR~~ Market~~ing~~', 'Digital Market~~ing~~ Executive' and 'Market~~ing~~ Director'. This selection indicates the preference towards professional standing. As a *super-profile*, there are more keywords to match. So a user with a headline, say, 'Vice President Digital Market~~ing and~~ Acquisition' then overlaps with of 3 of the 5 seeds. If we consider the seeds individually, only one word is found in the 3 keywords from the first seed, which pushes the score down. If we consider the recall differences, this corresponds to information loss for *individual seeds*. Collectively as a *super-profile*, the profiles maintain a high recall by utilizing all relevant keywords to filter for similar leads. The same is reflected in the high relevancy of the leads generated. Approach 4 and Approach 5 suffer here from their reliance on Twitter to filter for the relevant profiles. Consequently, Twitter adds an element of smoothing to the ranking process, as highlighted in Table 1.

As we can see in Table 1, Lead D (Headline: 'Digital Marketing Assistant at XYZ' from Industry: 'Pharmaceuticals') is not relevant for the business context when we use the LinkedIn without Twitter Approach, as the score is below the threshold of 0.25. Including Twitter attributes puts them in Top 5 due to similarity in its Twitter profile with a seed profile.

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

However, this similarity is not relevant for the business context at hand. Lead A and C were high quality leads acknowledged by clients, and their scores with the LinkedIn only approach reflects this. Hence, the issues faced with leveraging Twitter is that even in cases where LinkedIn and Twitter profiles match, the additional data carries too high a precedence and distorts very fundamental aspects like the industry sector of a lead. In other words, the approach is experiencing the curse of dimensionality: as the number of dimensions increases, notions of distance become less meaningful, as all points become close to as well as far from each other. Aside from this as a dimensionality issue, similar effects have been noted in other scenarios [21].

More formally, the precision of the approach is affected due to the increase in the total number of terms in the corpus. And interestingly, the terms that increase in the corpus are mainly elements of personal representation of the individual, which are seemingly not relevant for the business context at hand. This same observation has been reflected in the quality of leads generated using the 5 approaches. Hence, we excluded Twitter from the analysis and all the 5 approaches were repeated considering only the LinkedIn attributes of the profiles and qualified leads. The best result was still Approach 2; its balance between individual and company attributes resulted in leads which were most sought after in multiple trials and client consultations, including areas other than marketing. Overall, with respect to RQ2, LinkedIn attributes provide the most relevant social media data for lead generation, whereas Twitter does not appear useful in this context.

4.2. Varying the Cardinality of Seed Profiles

While the previous approach for lead generation utilizes all the seeds as one entity, the *super-profile*, the next step is to test the impact of less seed profiles. We will focus on Approach 2 here, as it was the best performing approach as mentioned earlier. This consideration enables a discussion on the sensitivity of Approach 2 towards variations in the input data. This is done by randomly removing seeds from the client-selected starting pool (across multiple use cases). A significant drop in the score of the leads generated with less seed profiles, as compared to the baseline case with all 5 seeds, would indicate the level at which the approach becomes vulnerable to the scale of the entity corpus. Thus we explore the number of seeds needed to identify new leads based on the super-profile approach.

To find the minimum number of seeds, we consider the average score generated for the leads as well as the variance in leads. We use two hypotheses: First, considering the difference in average scores for Lead A and Lead B, respectively, we use the null hypothesis that there should be no difference in the score generated using all 5 seeds and the leads generated using 2, 3, and 4 seed profiles. Second, considering the variance in scores themselves, we test the hypotheses if the variation in generated scores increases with a decreasing number of seeds. For the first hypothesis, the results of the study show that the hypothesis seems to be true until we reach only 2 seeds, in which case we have to reject the null hypothesis. Specifically, the difference in scores generated for Lead B becomes significant at the 0.05 level, using a corresponding t test. The standard deviation and variance in the scores of the leads with the 3 sample and 2 sample seed approach is summarized in Table 2. Considering the second hypothesis, the variance in scores generated significantly increases when going from 3 to 2 seed profiles, as indicated by a chi-squared test for variance equality at the 0.05 level. These results highlight the volatility of the scores with 2 sample seeds compared to 3 sample seeds. Hence, based on the results we infer that the minimum number of leads that would work towards consistently generating relevant leads should be 3, using Approach 2. Also, when considering more than just the top 5 leads, we can observe that the deviations in rankings become more noticeable; often the top 5 leads no longer appear in the Top 5 when 2 or fewer seeds are used. This makes sense, as the effect of an individual seed on the suggested leads is much higher for a smaller number of seeds. While 3 seeds is the minimum according to the analysis, there is a practical consideration why we do not consider more than 5 seeds: clients tend to express mild frustration at choosing more than 5 seeds.

To further test the robustness and consistency of the approach, noise was introduced in the corpus by including non-qualified leads and irrelevant leads. Three different approaches were carried out to add noise in the input seed profile: 1) Adding two non-qualified seeds, 2) Adding two irrelevant seeds and 3) Adding two non-qualified and two irrelevant seeds. In doing so, the general following observations were noted. First, the scores of top profiles decrease in some cases and

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

increased in others. Second, certain low scoring profiles have high scores after introducing noise. Third, some new leads which are not from a relevant industry appear in the list. Finally, the relevant lead scores are generally reduced. This scenario highlights the problems of unclear client preferences, or when a client is unwilling to pick seeds. Both situations are understandable, as sometimes clients may not wish to specify a clear objective, may not have yet formulated one and wish to simply explore the digital space, or they may simply not wish to invest time in seed selection. However, the results clearly show that this can severely impact the quality and usefulness of the results.

In such cases, we can run the approach iteratively, i.e., take the initial seeds (while being aware of the limitations), pick the top n, and relay these back to the client along with a yes/no decision on the suggested leads. Positive responses are injected as seeds for another iteration. Iterating only two or three times in this fashion has revealed similar success rates to a well curated seed set, which mitigates the effects of bad starting seed sets. Similarly, in the event that no feedback is required, reasonable results were noted when company staff translated the client requirements into an initial seed populations.

4.3. Summary

Several approaches for lead generation utilizing social media platforms were proposed and evaluated. We found Twitter to be insignificant for the context of lead generation, hence the presented methodology only utilizes a set of few LinkedIn attributes to generate high quality leads for the business. Testing the approach for variations in the input seed profiles and their corresponding corpora revealed consistency in lead generation when a minimum of 3 relevant seed profiles (taken as a *super-profile*) is used. We also noted that when initial seed sets are substandard, we can iterate our approach to effectively discard poor starting seeds. However, iterating too many times will result in a lead echo chamber, where the same combinations of leads are produced every time. This is due to the over-representation of specific attributes that contribute towards profile similarity, which are key to our approach.

From the results and discussion presented above, it may appear that lead generation can be solved with high level of accuracy utilizing relatively few features from the individual LinkedIn profiles. However, in reality there are several challenges around social media data. Social media platforms are a stylised self representation of personal or professional projections of real life: information presented by users is likely exaggerated (we refer to [19] for a discussion on this in the context of Facebook users). It is immensely important to validate the information provided by users identified as leads. Ultimately, there is a substantial difference between finding a relevant collaboration partner, or actively source a potential new employee, and that person being fit for purpose. What we have highlighted here is a means to automate what many individuals do manually, thus enabling them to spend more time interfacing with potential leads than looking for them.

5. Conclusions and Future Work

The paper presents a semi-automatic approach to identify new leads for a business by leveraging the information on LinkedIn profiles about potential customers or leads, where leads can be new potential employees or collaboration partners. It describes a semi-automatic approach that enables the use of large amounts of social media data to generate leads (RQ1). To identify which types of social media data are most helpful in generating "*good*" leads (RQ2), we have experimented with different approaches to utilize Twitter and LinkedIn data, finding that adding Twitter data does not lead to better predictions (rather, it leads to smoothing, which makes finding high quality leads more difficult). The attributes picked from the user profile on LinkedIn for lead generation were: Headline, Current Employer, Company Speciality, and Company Industry. These attributes best captured the preferences of real clients for our industry partner. These attributes are also a good representation of the social capital of the individual and provide a stable measure of profile similarity. If a customer is from industry A, working in company X which has a given set of specialities, 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 profiles, the addition of bad or mediocre profiles

as seeds alongside good leads, and changing the nature of the seed profile by testing the approach to identify leads for 4 different business contexts. The approach consistently generated relevant leads across all business contexts when a minimum of 3 seed profiles is used.

It is worth noting that there are a number of potential limitations of this work, which could be addressed as future work. First, are the potential privacy concerns: the general ease of data availability does raise some concerns. Yet, a large portion of LinkedIn's business model is predicated around the use of the same data for the discovery of individuals. These concerns have been raised before in the use of social media data, see: [22], but research the mitigation of these concerns is needed. Second, there are elements of self-representation and social posturing at play in a study such as this, and thus some consideration could be necessary akin to [19]. Third, as a cross-platform study, it is painfully evident that using multiple platforms to better represent different perspectives of prospects is challenging due in part to sample size. This may also be related to the observation that iterating the method multiple times even with "sub-optimal" seed profiles, still resulted in "good" leads, and such further work would be required to unravel potential effects of echo chambers within the approach by increasing the sample size, and undertaking additional 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.

[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. Brandão, 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.